IS 3001

SAMPLING TECHNIQUES

GROUP 05

Group Members

s15593 - Kawya Alagoda
s15594 – Maheshika Amarasingha
s15596 - Sithumini Arsakularatne
s15599 - Dushan Chamika

Contents

Introduction

This dataset provides birth rates and related data across the 50 states and DC from 2016 to 2021. The data was sourced from the Centers for Disease Control and Prevention (CDC) and includes detailed information such as number of births, gender, birth weight, state, region and year of the delivery. A particular emphasis is given to detailed information on the mother's educational level.

Each row in the dataset is considered a category defined by the state, Region birth year, baby's gender, Low Birth Weight and mother's educational level

Three quantities are given for each category: number of births, mother's average age and average baby weight.

The dataset was analyzed using Simple Random Sampling, Stratified Sampling, and Cluster Sampling separately. "rsampcal" function in R was used for determining samples.

In simple random sampling, the "rsampcalc" function was used to determine the sample size, and the obtained sample size was Stratified Sampling is based on dividing the population into various strata, and individuals are selected randomly from these strata to suit the sample size. (In these cases, the strata must be homogenous, collectively exhaustive, and mutually exclusive.) Here, "mother's educational level" was used as the stratifying variable, and individuals were randomly selected from the groups proportionally to the sizes of strata, to suit the sampling size determined in the random sampling method.

In the two-stage sampling design the population is devide into groups, like cluster sampling and then randomly select some clusters from all clusters. In this design new samples are taken from each cluster sampled. And here, initially the population is divided into N clusters based on the variable "Region", a sample of n clusters are selected from N(First stage) and then individual elements are selected from these clusters randomly(Second stage).

Methodology

Sample Size calculation

Normally we use below equations for sample size calculations.

$$n = \frac{n0}{1 + \frac{n0}{N}} \qquad n0 = \left(\frac{Z_{\alpha/2}S}{e}\right)^2$$

n = Sample Size

N = Population Size

$Z\alpha/2$ = Z value of the significance level

S = Population Variation

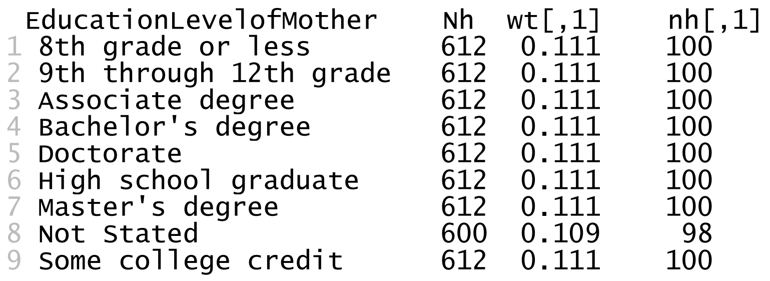
e = Margin of error

Simple Random Sampling

For this project rsampcalc function included in R software is used for calculate the sample size. We keep a margin of error of 3 and 5% type 1 error. Here we get sample size as 894 units.

Stratified Random Sampling

In the Stratified sample technique the population should be divided into strata. Then observations are selected from each stratum proportionally to the size of each stratum to obtain a sample size of n. here we devided population into 9 stratums. Then we select individuals from each and every stratum by using allocation.Using ssampcalc function that is included in sampler package we obtain sample size as 898.



```
  EducationLevelofMother     Nh  wt[,1]    nh[,1]
1 8th grade or less         612  0.111     100
2 9th through 12th grade    612  0.111     100
3 Associate degree          612  0.111     100
4 Bachelor's degree         612  0.111     100
5 Doctorate                 612  0.111     100
6 High school graduate      612  0.111     100
7 Master's degree           612  0.111     100
8 Not Stated                600  0.109      98
9 Some college credit       612  0.111     100
```

Two-Stage Cluster Sampling

Cluster sampling involves dividing the specific population of interest into geographically distinct groups. Summation of selected clusters populations is taken as the population size of the cluster sampling. Here we use our clustering variable as “Region”. First we divide our population into five clusters. Among them we randomly select 3 clusters and then we select samples from each selected clusters using SRS. Here we calculate sample sizes using above mentions equations in SRS.R function give our first selected clusters as Midwest, Northeast and Southeast.R function give our second selected clusters as Midwest, Southeast and West.

	Cluster	M_i	m_i
Sample one	Northeast	1293	585
	Midwest	1293	585
	Southeast	972	509
Sample two	West	1290	584
	Southeast	972	509
	Midwest	1293	585

Results of the study

Simple Random Sampling (SRS)

Mean

	Sample 1		Sample 2		Population
	Estimated Population Mean	SE	Estimated Population Mean	SE	Mean
Number of Births	4072.7	218.01	4355.1	232.03	4115.444
Birth weight of children	3248.8	3.7488	3252.8	3.7756	3250.888
Mother's age	29.554	0.0934	29.644	0.0925	29.55227

- Consider the number of births Estimated population mean of sample one less than actual mean and Estimated population mean of sample two greater than actual mean.
- Consider the birth weight of children Estimated population mean of sample one less than actual mean and Estimated population mean of sample two greater than actual mean.
- Consider the mother's age Estimated population mean of sample one approximately same to actual mean and Estimated population mean of sample two greater than actual mean.
- In conclusion approximately all three variables are nearly same with actual population parameters.

Total

		Sample 1		Sample 2		Population
		Estimated Population Total	SE	Estimated Population Total	SE	Total
Gender	Male	453	14.957	471	14.937	2749
	Female	441	14.957	423	14.937	2747
Low Birth Weight	Yes	84	8.7288	82	8.6349	508
	No	810	8.7288	812	8.6349	4988
Number of Births		3640960	194905	3893436	207436	22618480
Birth weight of children		2904453	3351.4	2908033	3375.4	17866878

- Estimated population total can be calculated by using the equation N multiply by sample mean. It can be shown below the table.

	Total using Sample 1	Total using Sample 2
Number of Births	22383559.2	23935629.6
Birth weight of children	17855404.8	17877388.8

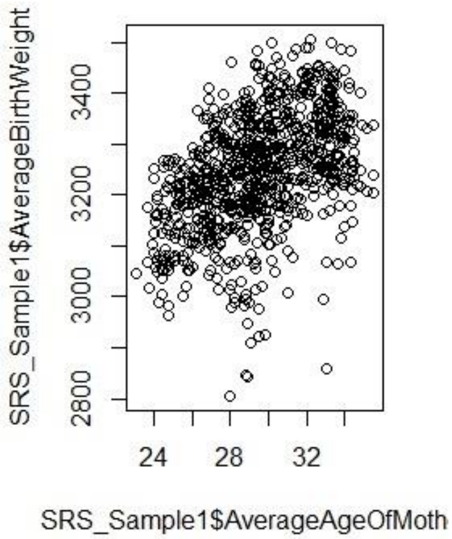
- According to the above table it's clear that our Estimated population totals nearly equal to the actual population total.

Proportion

		Sample 1		Sample 2		Population
		Estimated Proportion	SE	Estimated Proportion	SE	Proportion
Region	Northeast	0.24944	0.0145	0.25391	0.0146	0.2352620
	Southeast	0.12138	0.0127	0.16219	0.0123	0.1768559
	Midwest	0.21700	0.0138	0.22371	0.0139	0.2352620
	Southwest	0.11635	0.0107	0.11298	0.0106	0.1179039
	West	0.24385	0.0144	0.24720	0.0144	0.2347162
Gender	Male	0.50671	0.0167	0.47315	0.0167	0.500182
	Female	0.49329	0.0167	0.52685	0.0167	0.499818
Low Birth Weight	Yes	0.09396	0.0098	0.091723	0.0097	0.09243086
	No	0.90604	0.0098	0.908277	0.0097	0.90756914

Regression Estimation

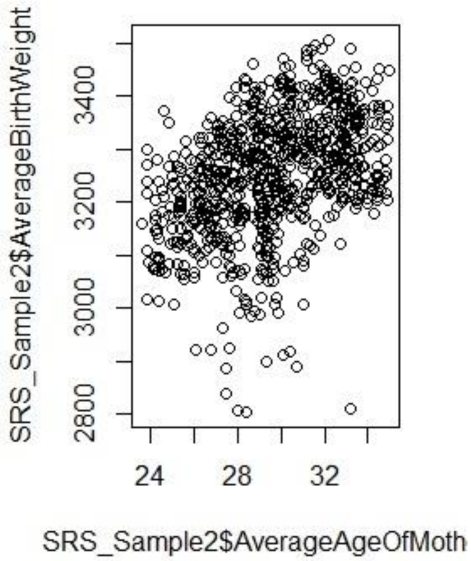
Sample One



In here we can see there is strong correlation between Mother's age and child birth weight. So we can represent relationship as regression equation.

Average Birth Weight = 2690.46 + 18.89*Average age of mother

Sample Two



In here we can see there is strong correlation between Mother’s age and child birth weight. So we can represent relationship as regression equation.
Average Birth Weight = 2745.02 + 17.13*Average age of mother

Stratified Sampling

Mean

	Sample 1		Sample 2		Population
	Estimated Population Mean	SE	Estimated Population Mean	SE	Mean
Number of Births	4089.5	205.3	4095.2	197.62	4115.444
Birth weight of children	3258	2.9471	3244.9	3.1283	3250.888
Mother’s age	29.582	0.0296	29.559	0.0306	29.55227

- Consider the mean value of number of births Estimated population mean of sample one and Estimated population mean of sample two less than actual mean.
- Consider the mean value of birth weight of children Estimated population mean of sample one greater than actual mean and Estimated population mean of sample two less than actual mean.
- Consider the mean value of mother’s age Estimated population mean of sample two approximately same to actual mean and Estimated population mean of sample one greater than actual mean.
- In conclusion approximately all three variables are nearly same with actual population parameters.

Total

		Sample 1		Sample 2		Population
		Estimated Population Total	SE	Estimated Population Total	SE	Total
Gender	Male	446	14.993	440	15.008	2749
	Female	452	14.933	458	15.008	2747
	Yes	75	7.3132	79	7.384	508

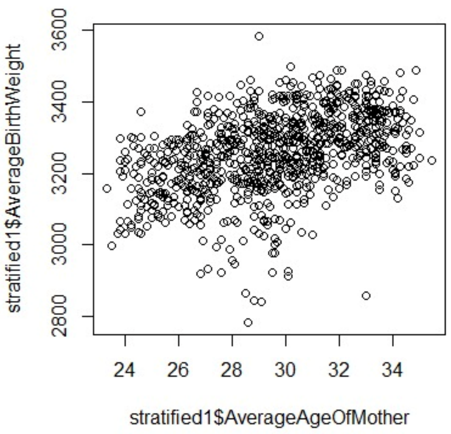
Low Birth Weight	No	823	7.3132	819	7.384	4988
Number of Births		3672345	184355	3677453	177459	22618480
Birth weight of children		2925664	2646.5	2913934	2809.3	17866878

Proportion

		Sample 1		Sample 2		Population
		Estimated Proportion	SE	Estimated Proportion	SE	Proportion
Region	Northeast	0.255011	0.0146	0.22383	0.0139	0.2352620
	Southeast	0.157016	0.0121	0.17706	0.0128	0.1768559
	Midwest	0.250557	0.0145	0.22717	0.0140	0.2352620
	Southwest	0.094655	0.0098	0.11359	0.0106	0.1179039
	West	0.242762	0.0143	0.25835	0.0146	0.2347162
Gender	Male	0.50334	0.0167	0.48998	0.0167	0.500182
	Female	0.49666	0.0167	0.51002	0.0167	0.499818
Low Birth Weight	Yes	0.083519	0.0081	0.087973	0.0082	0.09243086
	No	0.916481	0.0081	0.912027	0.0082	0.90756914

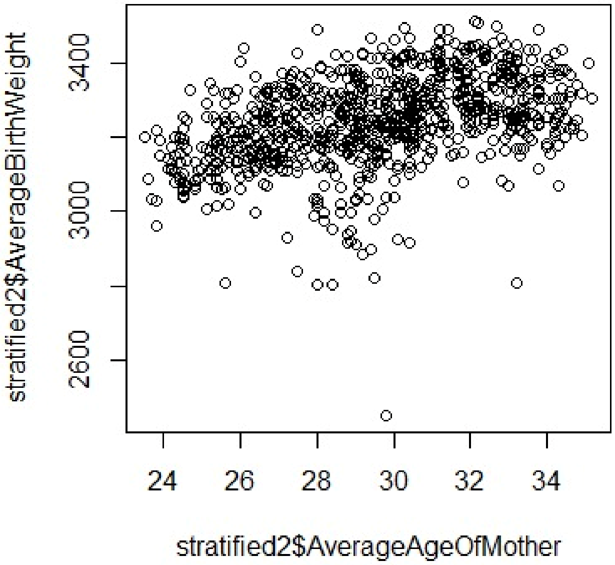
Regression Estimation

Sample one



In here we can see there is strong correlation between Mother's age and child birth weight. So we can represent relationship as regression equation.
Average Birth Weight = 2733.01 + 17.75*Average age of mother

Sample two



In here we can see there is strong correlation between Mother's age and child birth weight. So we can represent relationship as regression equation.
Average Birth Weight = 2713 + 18*Average age of mother

Cluster Sample

Mean

	Sample 1		Sample 2		Population
	Estimated Population Mean	SE	Estimated Population Mean	SE	Mean
Number of Births	4024.8	122.39	4034	141.7	4115.444
Birth weight of children	3258.1	2.8231	3253.8	2.7092	3250.888
Mother's age	29.6	0.0688	29.369	0.0669	29.55227

- Consider the mean value of number of births Estimated population mean of sample one and Estimated population mean of sample two less than actual mean.
- Consider the mean value of birth weight of children Estimated population mean of sample one and Estimated population mean of sample two greater than actual.
- Consider the mean value of mother’s age Estimated population mean of sample two less than actual mean and Estimated population mean of sample one greater than actual mean.
- In conclusion approximately all three variables are nearly same with actual population parameters.

Total

		Sample 1		Sample 2		Population
		Estimated Population Total	SE	Estimated Population Total	SE	Total
Gender	Male	833	20.493	829	20.486	2749
	Female	846	20.493	849	20.486	2747
Low Birth Weight	Yes	149	11.656	143	11.441	508
	No	1530	11.656	1535	11.441	4988
Number of Births		6757640	205495	6769032	237778	22618480
Birth weight of children		5470415	4740	5459948	4546.1	17866878

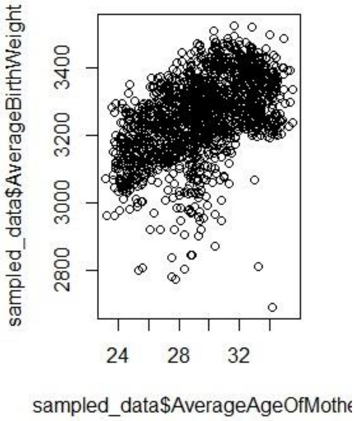
Proportion

		Sample 1		Sample 2		Population
		Estimated Proportion	SE	Estimated Proportion	SE	Proportion
Region	Northeast	0.34842	0.0116	----------	--------	0.2352620
	Southeast	0.30316	0.0112	0.30334	0.0112	0.1768559
	Midwest	0.34842	0.0116	0.34863	0.0116	0.2352620
	Southwest	----------	--------	----------	--------	0.1179039
	West	----------	--------	0.34803	0.0116	0.2347162
Gender	Male	0.49613	0.0122	0.49404	0.0122	0.500182

	Female	0.50387	0.0122	0.50596	0.0122	0.499818
Low Birth Weight	Yes	0.088743	0.0069	0.085221	0.0068	0.09243086
	No	0.911257	0.0069	0.914779	0.0068	0.90756914

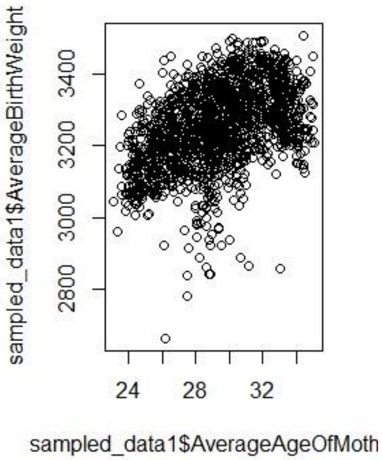
Regression Estimation

Sample One



In here we can see there is strong correlation between Mother's age and child birth weight. So we can represent relationship as regression equation.
Average Birth Weight = 2645.20 + 20.71*Average age of mother

Sample Two

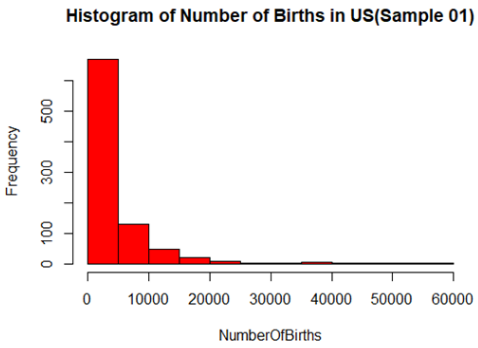


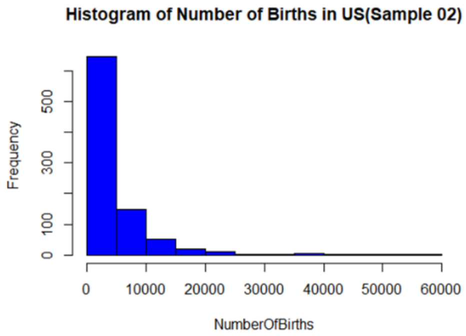
In here we can see there is strong correlation between Mother's age and child birth weight. So we can represent relationship as regression equation.
Average Birth Weight = 2705.64 + 18.67*Average age of mother.

Graphical Analysis

Simple random sampling Graphical Analysis

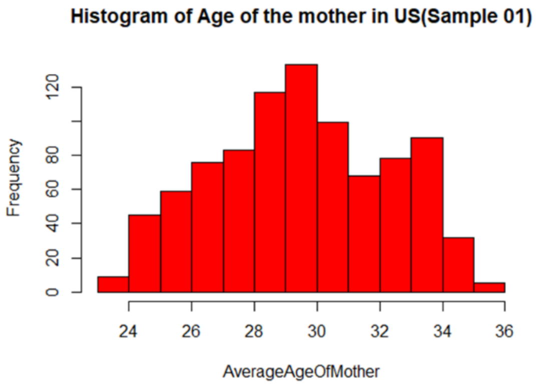
Number of births

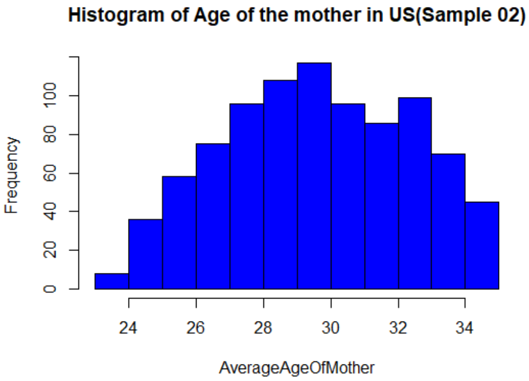




According to the above histograms it's clear that the number of births in US is skewed to right (positively skewed) in both samples. That is most common number of births in the states USA is in between 0 to 3000

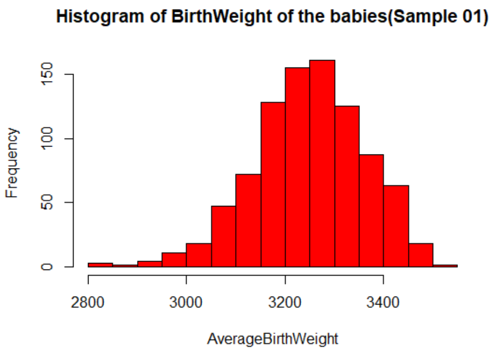
Average Age of the mother

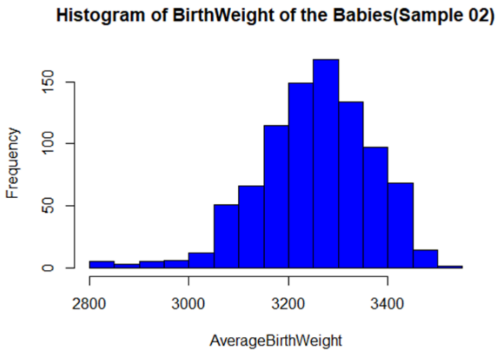




According to the above histograms nearly the mothers' age is vary in normally distributed manner. That's mean the majority of the mothers' age is is middle age range in the birth in the US 2016 to 2021

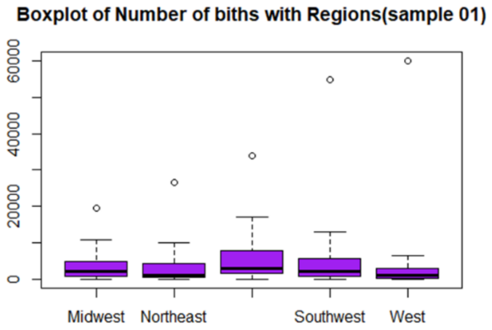
Births Weight of children

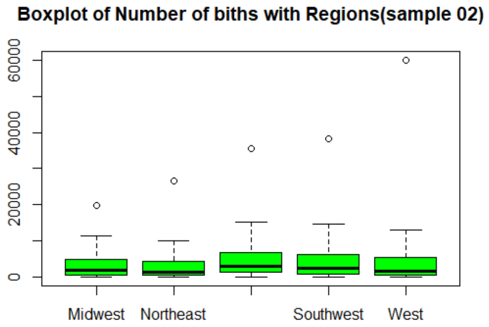




According to the above histograms nearly average age of birthweight vary in left skewed manner. That’s mean the majority of the birthweights are in above the 3000 g s.

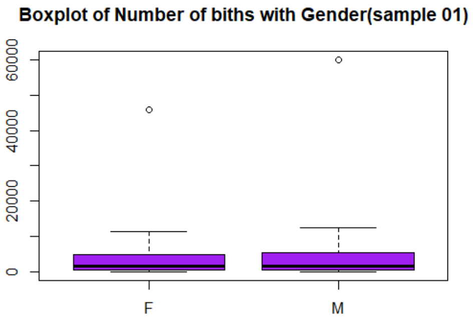
Number of births with region

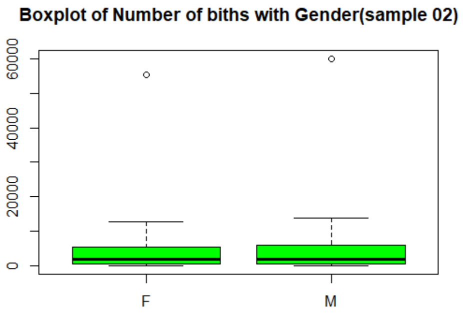




According to the above boxplots nearly average no of births in Midwest, northeast, southwest and west are more similar to each other. But average no of births in southeast is greater than other four. That’s mean more babies are born in southeast than other regions.

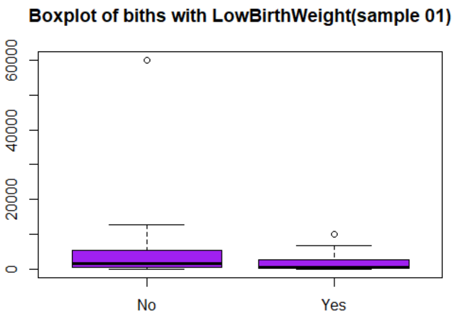
Number of births with gender

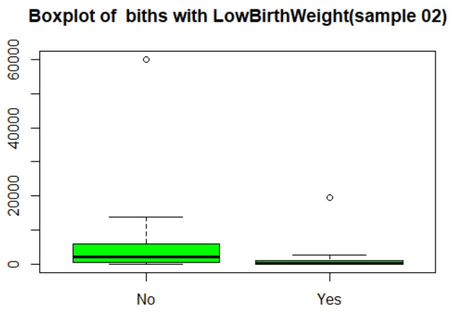




According to the above boxplots we can see all boxes have same spreads. Which means no of male baby births and female baby births are approximately equal.

Number of births with low birth weight





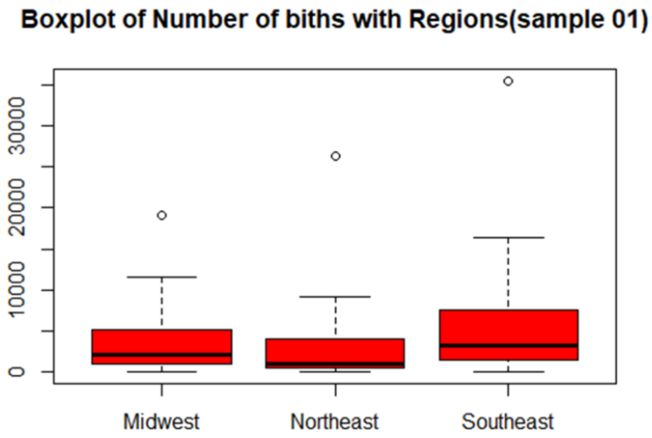
According to the above boxplots we can see no low birth weight have large spreads than have low birth weight. Which means few no of births have low birth weight.

Cluster sampling Graphical Analysis

Number of births with region

In here there is a brief explanation about the number of births in US with our Clustering variable Region.

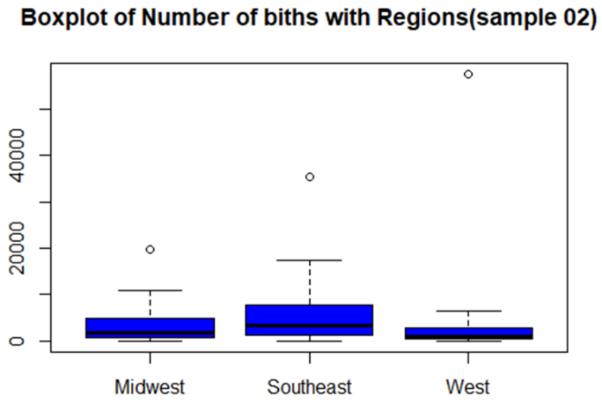
Sample 01



In here above boxplot shows how the Region vary with births in two stage cluster sampling. In sample 01 our selected clusters was Midwest Northeast and Southeast. So that above plot shows how those clusters vary with the number of births.

It is visible that median of births of the Southeast region is high than to other regions. There is clear outlier in three boxplots.

Sample 02



In here above boxplot shows how the Region vary with births in two stage cluster sampling. In sample 02 our selected clusters was Midwest Southeast and west. So that above plot shows how those clusters vary with the number of births.

It is visible that median of births of the Southeast region is high than to other regions. There is clear outlier in three boxplots.

Conclusion of the Analysis

The results of this study which regards to the sampling designs; simple random sampling, stratified random sampling and the two-stage cluster sampling for the dataset 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. Regression estimation or the ratio estimation also give the similar findings. Therefore, we can conclude that, it is possible to draw any of these probabilistic sampling designs, under the other practical limitations such as time, effort & cost for a proper analysis of the data set.

Results of the analysis does not differ significantly with the method of sampling but the standard error of the estimations in the two-staged cluster sampling is lower when compared to the other two which should be considered when conducting the analysis.

R Codes

R code for Population parameters estimates

```
install.packages("survey")
library("survey")
data = read.csv("us_births_2016_2021.csv")
data




#------------------------------------------------calculate the population parameters----------------------------
--------------------------------------------#
##mean
N=nrow(data)
mean(data$`NumberOfBirths`,na.rm=TRUE)
mean(data$`AverageAgeOfMother`,na.rm=TRUE)
mean(data$`AverageBirthWeight`,na.rm=TRUE)


##Proportion
data.frame((table(data$Region))/length(data$Region))
data.frame((table(data$Year))/length(data$Year))
data.frame((table(data$Gender))/length(data$Gender))
data.frame((table(data$LowBirthWeight))/length(data$LowBirthWeight))

##Total

table(data$Gender)
table(data$LowBirthWeight)
table(data$Year)

sum(data$NumberOfBirths)
sum(data$AverageBirthWeight)
```

R Code for simple random sampling

```
# -----------sample 01------------------------------------------
install.packages("sampler")
library(sampler)


n=rsampcalc(nrow(data),e=3,ci=95)
set.seed(15596)
SRS_Sample1 = rsamp(df = data,n, rep = FALSE)
SRS_Sample1


SRS1=svydesign(id=~1,data=SRS_Sample1)


### calculate the mean of the variables (sample 01)
svymean(~`NumberOfBirths`,design=SRS1,na.rm=TRUE)
svymean(~`AverageAgeOfMother`,design=SRS1,na.rm=TRUE)
svymean(~`AverageBirthWeight`,design=SRS1,na.rm=TRUE)


### calculate sample Proportion of the variable (sample 01)
svymean(~`Region`,design=SRS1,na.rm=TRUE)
svymean(~`Year`,design=SRS1,na.rm=TRUE)
svymean(~`Gender`,design=SRS1,na.rm=TRUE)
svymean(~`LowBirthWeight`,design=SRS1,na.rm=TRUE)


## calculate the sample total of the variable (sample 01)
svytotal(~`Gender`, design=SRS1,na.rm=TRUE)
svytotal(~`LowBirthWeight`, design=SRS1,na.rm=TRUE)
svytotal(~`Year`, design=SRS1,na.rm=TRUE)
svytotal(~`NumberOfBirths`, design=SRS1,na.rm=TRUE)

svytotal(~`AverageBirthWeight`, design=SRS1,na.rm=TRUE)


# -----------sample 02------------------------------------------

n=rsampcalc(nrow(data),e=3,ci=95)

set.seed(15595)

SRS_Sample2 = rsamp(df = data,n, rep = TRUE)

SRS_Sample2

SRS2=svydesign(id=~1,data=SRS_Sample2)


### calculate the mean of the variables (sample 02)

svymean(~`NumberOfBirths`,design=SRS2,na.rm=TRUE)

svymean(~`AverageAgeOfMother`,design=SRS2,na.rm=TRUE)

svymean(~`AverageBirthWeight`,design=SRS2,na.rm=TRUE)


### calculate sample Proportion of the variable (sample 02)

svymean(~`Region`,design=SRS2,na.rm=TRUE)

# svymean(~`EducationLevel`,design=SRS2,na.rm=TRUE)

svymean(~`Year`,design=SRS2,na.rm=TRUE)

svymean(~`Gender`,design=SRS2,na.rm=TRUE)

svymean(~`LowBirthWeight`,design=SRS2,na.rm=TRUE)


## calculate the sample total of the variable (sample 02)

svytotal(~`Gender`, design=SRS2,na.rm=TRUE)

svytotal(~`LowBirthWeight`, design=SRS2,na.rm=TRUE)

svytotal(~`Year`, design=SRS2,na.rm=TRUE)

svytotal(~`NumberOfBirths`, design=SRS2,na.rm=TRUE)
svytotal(~`AverageBirthWeight`, design=SRS2,na.rm=TRUE)
#--------------------regression estimate for sample 01--------------------#
install.packages(sampler)
library(sampler)
SRS_Sample1
plot(SRS_Sample1$`AverageAgeOfMother`,SRS_Sample1$`AverageBirthWeight`)
#plot(SRS_Sample2$`NumberOfBirths`,SRS_Sample2$`AverageBirthWeight`)##No correlation
SLR1 = lm(`AverageBirthWeight`~`AverageAgeOfMother`,SRS_Sample1)
SLR1
# mean_BirthWeight= 2690.46  +  18.89 *mean(data$`AverageAgeOfMother`)
# mean_BirthWeight



#--------------------regression estimate for sample 02--------------------#

SRS_Sample2
plot(SRS_Sample2$`AverageAgeOfMother`,SRS_Sample2$`AverageBirthWeight`)
#plot(SRS_Sample2$`NumberOfBirths`,SRS_Sample2$`AverageBirthWeight`)##No correlation
SLR2 = lm(`AverageBirthWeight`~`AverageAgeOfMother`,SRS_Sample2)
SLR2
# mean_BirthWeight=  2745.02    +   17.13  *mean(data$`AverageAgeOfMother`)
# mean_BirthWeight
```

R Code for stratified sampling

```
N=nrow(data)
mean(data$NumberOfBirths,na.rm=TRUE)
mean(data$AverageAgeOfMother,na.rm=TRUE)
mean(data$AverageBirthWeight,na.rm=TRUE)


data.frame((table(data$Region))/length(data$Region))
data.frame((table(data$Year))/length(data$Year))
data.frame(table(data$Gender)/length(data$Gender))
data.frame(table(data$EducationLevel)/length(data$EducationLevel))
data.frame((table(data$Region))/length(data$Region))
data.frame((table(data$LowBirthWeight))/length(data$LowBirthWeight))


table(Year)
table(Gender)
colnames(data)[colnames(data) == "LowBirthWeight - Yes if it is a low birth weight, No otherwise"] = "LowBirthWeight"
table(data$LowBirthWeight)
sum(data$NumberOfBirths)
sum(data$AverageBirthWeight)

###-----------------------------------sample 1--------------------------------------###
library(sampler)
size=rsampcalc(nrow(data),3,95,0.5)
sample1_size = ssampcalc(df=data,n=size,strata = EducationLevelofMother)
sample1_size

#================================================================================
========
# #---- Check for missing values in Cluster1-------------
# missing_values <- sum(is.na(sample1_size$NumberOfBirths))
# if (missing_values > 0) {
#   # Handle missing values (remove or impute)
#   # Example: Remove rows with missing values
#   sample1_size <- Cluster1[!is.na(sample1_size$NumberOfBirths), ]
# }
# variance_sample1_size <- var(sample1_size$NumberOfBirths)
# cat("Variance for Cluster1:", variance_cluster1, "\n")
```

```
#
#
#
#==============================================================================
============

set.seed(013001)
stratified1 = ssamp(df=data,n=894, strata =EducationLevelofMother)
stratified1

for(i in 1:898){
  if(stratified1$EducationLevelofMother[i]=="8th grade or less"){
    pw1 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="9th through 12th grade "){
    pw1 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Associate degree "){
    pw1 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Bachelor's degree "){
    pw1 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Doctorate"){
    pw1 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="High school graduate"){
    pw1 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Master's degree"){
    pw1 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Not Stated"){
    pw1 = round((600/98),2)
  }else if(stratified1$EducationLevelofMother[i]=="Some college credit"){
    pw1 = round((612/100),2)
  }
  #pw=c(print(pw))
}
strat1 = cbind(stratified1,pw1)
strat1


###---Calculating estimated population means for Sample 1---###
library(survey)
sample1=svydesign(id=~1,strata =~EducationLevelofMother ,data=strat1)
svymean(~NumberOfBirths,design=sample1)
svymean(~AverageAgeOfMother,design=sample1)
svymean(~AverageBirthWeight,design=sample1)


svymean(~Region,design=sample1)
svymean(~Gender,design=sample1)
svymean(~LowBirthWeight,design=sample1)

#-----------Totals
svytotal(~Gender, design=sample1)
svytotal(~LowBirthWeight, design=sample1)
svytotal(~NumberOfBirths,design =sample1 )
svytotal(~AverageBirthWeight , design =sample1)

###-------------------------------sample 2----------------------------------###
library(sampler)
size=rsampcalc(nrow(data),3,95,0.5)
sample1_size = ssampcalc(df=data,n=size,strata = EducationLevelofMother)
sample1_size

set.seed(023001)
stratified2 = ssamp(df=data,n=898, strata =EducationLevelofMother)
stratified2
pw2=0
for(i in 1:898){
  if(stratified1$EducationLevelofMother[i]=="8th grade or less"){
    pw2 = round((612/100),2)
}else if(stratified1$EducationLevelofMother[i]=="9th through 12th grade "){
    pw2 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Associate degree "){
    pw2 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Bachelor's degree "){
    pw2 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Doctorate"){
    pw2 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="High school graduate"){
    pw2 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Master's degree"){
    pw2 = round((612/100),2)
  }else if(stratified1$EducationLevelofMother[i]=="Not Stated"){
    pw2 = round((600/98),2)
  }else if(stratified1$EducationLevelofMother[i]=="Some college credit"){
   pw2 = round((612/100),2)
  }
  #pw=c(print(pw))
}
strat2 = cbind(stratified2,pw2)
strat2

###---Calculating estimated population means for Sample 2---###
library(survey)
sample2=svydesign(id=~1,strata =~EducationLevelofMother ,data=strat2)
##----------means
svymean(~NumberOfBirths,design=sample2)
svymean(~AverageAgeOfMother,design=sample2)
svymean(~AverageBirthWeight,design=sample2)

##---prportions
svymean(~Region,design=sample2)
svymean(~Gender,design=sample2)
svymean(~LowBirthWeight,design=sample2)


#-----------Totals
svytotal(~Gender, design=sample2)
svytotal(~NumberOfBirths,design =sample2 )
svytotal(~AverageBirthWeight , design =sample2)
svytotal(~LowBirthWeight, design=sample2)
#Regression Estimation
#sample 01
sample1
plot(stratified1$AverageAgeOfMother,stratified1$AverageBirthWeight)
SLR1 = lm(AverageBirthWeight~AverageAgeOfMother,stratified1)
SLR1
mean_BirthWeight=2733.01+17.75*mean(data$AverageAgeOfMother)
mean_BirthWeight



#sample 02
sample2
plot(stratified1$AverageAgeOfMother,stratified1$AverageBirthWeight)
SLR2 = lm(AverageBirthWeight~AverageAgeOfMother,stratified2)
```

```
SLR2
mean_BirthWeight=2713.01+18*mean(data$AverageAgeOfMother)
mean_BirthWeight
```

R Code for cluster sampling

```
# install.packages("survey")
# install.packages("dplyr")
# install.packages("sampler")
library(survey)
library(dplyr)
library(sampler)

N=nrow(data)
Regions <- c("Southeast", "West", "Southwest", "Northeast", "Midwest")

#Assume Regions  as Clusters
#select randomly 3 clusters from 5 clusters

N=5
n=3

#========================================================sample
01======================================================================
=======#
set.seed(15599)
Clus_1=sample(Regions,size=3,replace=F)
Cluster1<-data[data$Region=="Northeast",]
Cluster2<-data[data$Region=="Midwest",]
Cluster3<-data[data$Region=="Southeast",]

#---- Check for missing values in Cluster1-------------
missing_values <- sum(is.na(Cluster1$NumberOfBirths))
if (missing_values > 0) {
  # Handle missing values (remove or impute)
  # Example: Remove rows with missing values
  Cluster1 <- Cluster1[!is.na(Cluster1$NumberOfBirths), ]
}
variance_cluster1 <- var(Cluster1$NumberOfBirths)
cat("Variance for Cluster1:", variance_cluster1, "\n")
```

```
# ------Check for missing values in Cluster2-------------
missing_values <- sum(is.na(Cluster2$NumberOfBirths))
if (missing_values > 0) {
  # Handle missing values (remove or impute)
  # Example: Remove rows with missing values
  Cluster2 <- Cluster2[!is.na(Cluster2$NumberOfBirths), ]
}

variance_cluster2 <- var(Cluster2$NumberOfBirths)
cat("Variance for Cluster2:", variance_cluster2, "\n")

#-------------- Check for missing values in Cluster3---------------
missing_values <- sum(is.na(Cluster3$NumberOfBirths))
if (missing_values > 0) {
  # Handle missing values (remove or impute)
  # Example: Remove rows with missing values
  Cluster3 <- Cluster3[!is.na(Cluster3$NumberOfBirths), ]
}
variance_cluster3 <- var(Cluster3$NumberOfBirths)
cat("Variance for Cluster3:", variance_cluster3, "\n")

#---- Check for missing values in Clusters-------------

#cluster sizes
M1 <- nrow(Cluster1)
M2 <- nrow(Cluster2)
M3 <- nrow(Cluster3)


#calculating sample sizes each cluster
#Take Margin of error = 0.03
#Take CL=95%
n0=(1.96^2*0.5*0.5)/0.03^2


m1=ceiling(n0/1+(n0/M1))#sample size of the cluster 01
m2=ceiling(n0/1+(n0/M2))#sample size of the cluster 02
m3=ceiling(n0/1+(n0/M3))#sample size of the cluster 03
# m1=585#sample size of the cluster 01
# m2=585#sample size of the cluster 02
# m3=509#sample size of the cluster 03




#
========================================================================
==========

# Assuming each cluster is selected, let's now take m element sample within clusters

# Sample within Cluster 1
set.seed(15690) # Set seed for reproducibility
sampled_indices_c1 <- sample(1:nrow(Cluster1), m1, replace = FALSE)
sampled_data_c1 <- Cluster1[sampled_indices_c1, ]

# Sample within Cluster 2
set.seed(15691) # Set seed for reproducibility
sampled_indices_c2 <- sample(1:nrow(Cluster2), m2, replace = FALSE)
sampled_data_c2 <- Cluster2[sampled_indices_c2, ]

# Sample within Cluster 3
set.seed(15619) # Set seed for reproducibility
sampled_indices_c3 <- sample(1:nrow(Cluster3), m3, replace = FALSE)
sampled_data_c3 <- Cluster3[sampled_indices_c3, ]

# Combine sampled data from all clusters
sampled_data <- rbind(sampled_data_c1, sampled_data_c2, sampled_data_c3)



# Create the survey design objects
svy_design1 <- svydesign(id = ~1,  data = sampled_data)


# Calculate the mean for the sample 01 from cluster sampling
svymean(~NumberOfBirths, design = svy_design1)
svymean(~AverageAgeOfMother, design = svy_design1)
svymean(~AverageBirthWeight, design = svy_design1)
```

```
# Calculate the proportion for the sample 01 from cluster sampling
svymean(~Region, design = svy_design1)
svymean(~Year, design = svy_design1)
svymean(~Gender, design = svy_design1)
svymean(~`LowBirthWeight`,design=svy_design1)

# Calculate the total for the sample 01 from cluster sampling
svytotal(~`Gender`, design=svy_design1)
svytotal(~`LowBirthWeight`, design=svy_design1)
svytotal(~`Year`, design=svy_design1)
svytotal(~`NumberOfBirths`, design=svy_design1)
svytotal(~`AverageBirthWeight`, design=svy_design1)



#=====================================================sample
02=========================================================================
#
set.seed(15596)
Clus_2=sample(Regions,size=3,replace=F)
Cluster1<-data[data$Region=="West",]
Cluster2<-data[data$Region=="Southeast",]
Cluster3<-data[data$Region=="Midwest",]

#---- Check for missing values in Cluster1-------------
missing_values <- sum(is.na(Cluster1$NumberOfBirths))
if (missing_values > 0) {
  # Handle missing values (remove or impute)
  # Example: Remove rows with missing values
  Cluster1 <- Cluster1[!is.na(Cluster1$NumberOfBirths), ]
}
variance_cluster1 <- var(Cluster1$NumberOfBirths)
cat("Variance for Cluster1:", variance_cluster1, "\n")

# Check for missing values in Cluster2
missing_values <- sum(is.na(Cluster2$NumberOfBirths))
if (missing_values > 0) {
  # Handle missing values (remove or impute)
  # Example: Remove rows with missing values
  Cluster2 <- Cluster2[!is.na(Cluster2$NumberOfBirths), ]
}
```

```
variance_cluster2 <- var(Cluster2$NumberOfBirths)
cat("Variance for Cluster2:", variance_cluster2, "\n")

# Check for missing values in Cluster3
missing_values <- sum(is.na(Cluster3$NumberOfBirths))
if (missing_values > 0) {
  # Handle missing values (remove or impute)
  # Example: Remove rows with missing values
  Cluster3 <- Cluster3[!is.na(Cluster3$NumberOfBirths), ]
}
variance_cluster3 <- var(Cluster3$NumberOfBirths)
cat("Variance for Cluster3:", variance_cluster3, "\n")

#---- Check for missing values in Clusters-------------

#cluster sizes
M1 <- nrow(Cluster1)
M2 <- nrow(Cluster2)
M3 <- nrow(Cluster3)
M1
M2
M3


#calculating sample sizes each cluster
#Take Margin of error = 0.03
#Take CL=95%
n0=(1.96^2*0.5*0.5)/0.03^2


m1=ceiling(n0/1+(n0/M1))#sample size of the cluster 01
m2=ceiling(n0/1+(n0/M2))#sample size of the cluster 02
m3=ceiling(n0/1+(n0/M3))#sample size of the cluster 03
m1=584#sample size of the cluster 01
m2=509#sample size of the cluster 02
m3=585#sample size of the cluster 03
```

```
#
========================================================================
==========

# Assuming each cluster is selected, let's now take m element sample within clusters

# Sample within Cluster 1
set.seed(15663) # Set seed for reproducibility
sampled_indices_c1 <- sample(1:nrow(Cluster1), m1, replace = FALSE)
sampled_data_c1 <- Cluster1[sampled_indices_c1, ]

# Sample within Cluster 2
set.seed(15618) # Set seed for reproducibility
sampled_indices_c2 <- sample(1:nrow(Cluster2), m2, replace = FALSE)
sampled_data_c2 <- Cluster2[sampled_indices_c2, ]

# Sample within Cluster 3
set.seed(15619) # Set seed for reproducibility
sampled_indices_c3 <- sample(1:nrow(Cluster3), m3, replace = FALSE)
sampled_data_c3 <- Cluster3[sampled_indices_c3, ]

# Combine sampled data from all clusters
sampled_data1 <- rbind(sampled_data_c1, sampled_data_c2, sampled_data_c3)



# Create the survey design objects
svy_design2 <- svydesign(id = ~1,  data = sampled_data1)


# Calculate the mean for the sample 02 from cluster sampling
svymean(~NumberOfBirths, design = svy_design2)
svymean(~AverageAgeOfMother, design = svy_design2)
svymean(~AverageBirthWeight, design = svy_design2)

# Calculate the Proportion for the sample 02 from cluster sampling
svymean(~Region, design = svy_design2)
svymean(~Year, design = svy_design2)
svymean(~Gender, design = svy_design2)
svymean(~`LowBirthWeight`,design=svy_design2)
```

```
# Calculate the total for the sample 02 from cluster sampling
svytotal(~`Gender`, design=svy_design2)
svytotal(~`LowBirthWeight`, design=svy_design2)
svytotal(~`Year`, design=svy_design2)

svytotal(~`NumberOfBirths`, design=svy_design2)
svytotal(~`AverageBirthWeight`, design=svy_design2)

#Regression Estimation for sample 01

plot(sampled_data$`AverageAgeOfMother`,sampled_data$`AverageBirthWeight`)
#plot(sampled_data$`NumberOfBirths`,sampled_data$`AverageBirthWeight`)#no
correlation
SLR1 = lm(`AverageBirthWeight`~`AverageAgeOfMother`,sampled_data)
SLR1

#mean_BirthWeight= 2645.20+ 20.71 *mean(data$`AverageAgeOfMother`)
#mean_BirthWeight


#Regression Estimation for sample 02

plot(sampled_data1$`AverageAgeOfMother`,sampled_data1$`AverageBirthWeight`)
#plot(sampled_data$`NumberOfBirths`,sampled_data$`AverageBirthWeight`)#no
correlation
SLR2 = lm(`AverageBirthWeight`~`AverageAgeOfMother`,sampled_data1)
SLR2

#mean_BirthWeight=2705.64  +  18.67*mean(data$`AverageAgeOfMother`)
#mean_BirthWeight
```

Graphical analysis code

SRS

```
# Create  bar plot
svyhist(~NumberOfBirths,SRS1,main="Histogram of Number of Births in US(Sample 01)",col="Red",probability=FALSE)
svyhist(~NumberOfBirths,SRS2,main="Histogram of Number of Births in US(Sample 02)",col="Blue",probability=FALSE)
svyhist(~AverageAgeOfMother ,SRS1,main="Histogram of Age of the mother in US(Sample 01)",col="Red",probability=FALSE)
svyhist(~AverageAgeOfMother ,SRS2,main="Histogram of Age of the mother in US(Sample 02)",col="Blue",probability=FALSE)
svyhist(~AverageBirthWeight  ,SRS1,main="Histogram of BirthWeight of the babies(Sample 01)",col="Red",probability=FALSE)
svyhist(~AverageBirthWeight  ,SRS2,main="Histogram of BirthWeight of the Babies(Sample 02)",col="Blue",probability=FALSE)

#create box plot

svyboxplot(~NumberOfBirths~Region,SRS1,main="Boxplot of Number of biths with Regions(sample 01)",col="Purple")
svyboxplot(~NumberOfBirths~Region,SRS2,main="Boxplot of Number of biths with Regions(sample 02)",col="Green")


svyboxplot(~NumberOfBirths~EducationLevel,SRS1,main="Boxplot of biths with EducationLevel(sample 01)",col="Purple")
svyboxplot(~NumberOfBirths~EducationLevel,SRS2,main="Boxplot of  biths with EducationLevel(sample 02)",col="Green")

svyboxplot(~NumberOfBirths~LowBirthWeight,SRS1,main="Boxplot of biths with LowBirthWeight(sample 01)",col="Purple")
svyboxplot(~NumberOfBirths~LowBirthWeight,SRS2,main="Boxplot of  biths with LowBirthWeight(sample 02)",col="Green")
```

Stratified

```
svyboxplot(~NumberOfBirths~EducationLevel,sample1,main="Boxplot of Number of biths with Education level(sample 01)",col="Red")
svyboxplot(~NumberOfBirths~EducationLevel,sample2,main="Boxplot of Number of biths with Education level(sample 02)",col="Blue")
```

Cluster

```
svyboxplot(~NumberOfBirths~Region,svy_design1,main="Boxplot of Number of biths with Regions(sample 01)",col="Red")
svyboxplot(~NumberOfBirths~Region,svy_design2,main="Boxplot of Number of biths with Regions(sample 02)",col="Blue")
```