IS3001

IS3001

Sampling Techniques

D. C. Rupasinghe

Contents

Introduction

The dataset includes information on pregnancy women. It has several variables that can be used to describe their health.

The variables in the dataset are, “Patient ID” which is a single variable created for identify each patient and Hospital state, Hospital division, Hospital ownership, Ethnicity, no of pregnancies, Glucose level, Blood pressure, Age and Outcome which is a Boolean variable that says if the patient has diabetics.

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, “Ethnicity” 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 cluster sampling, unlike in stratified sampling homogeneity is external however heterogeneity is internal within the clusters. In the two-stage sampling design the population is partitioned into groups, like cluster sampling, but 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 “Hospital state”, a sample of n clusters are selected from N and then individual elements are selected from these clusters randomly. All these methods are explained in detail, in the next parts of the report.

Methodology

Simple Random Sample

- First, we must determine the sample size we have to get. We can obtain that from the below mentioned formula.

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

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

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

n = Sample Size

N = Population Size

$Z_{\frac{\alpha}{2}}$ = Z value of the significance level

S = Data Variation

e = Margin of error

- 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 445 as the sample size.

Stratified Random Sample

- To get a lower variability it is highly recommended to divide the population into strata's and do separate analysis on each one.
- So, we divide the entire population by ethnicity, then analyze the no of pregnancies in each stratum separately.
- Analyzing the population variability of no of pregnancies, we observed that they are similar in size.
- Therefore, we used the proportional allocation to find the sample size for each stratum.

Ethnicity	Population size	Sample size
White	383	223
Black	228	133
Mixed	76	44
Asian	76	44
Total	763	444

Two stage cluster sample

- In two stage cluster sampling an SRS of a cluster is selected, then another SRS in each cluster is taken.
- As the clustering variable we used the variable “Hospital state” which has 10 clusters.
- Then we selected 6 clusters out of 10.
- Then we took an SRS of each cluster

```
> ClusterDetails
  Hospital state Sample Size Population Size
1             FL          41              42
2             CO          73              78
3             CA         258             340
4             AR          71              75
5             AL          83              89
6             DE           7               7
>
```

```
> ClusterDetails
  Hospital state Sample Size Population Size
1             AZ          74              79
2             FL          41              42
3             AR          71              75
4             DC           8               8
5             AK          14              14
6             CO          73              78
>
```

Results of the study

Population Data

Mean	pop_mean_BloodPressure	72.32634
	pop_mean_Glucose	121.6868
	pop_mean_NoOfPregnancies	3.676404
	pop_Age	33.2713
Total	pop_total_NoOfPregnancies	2939
	NoOfPatientsHavingDiabetics	266
	NoOfPatientsNotHavingDiabetics	497
	Hospital_Ownership	
	Government – Federal	11
	Government - Hospital District or Authority	108
	Government – Local	37
	Government – State	13
	Physician	6
	Proprietary	158
	Tribal	5
	Voluntary non-profit – Church	60
	Voluntary non-profit – Other	95
	Voluntary non-profit – Private	270
Proportion	pop_proportion (1: diabetic patient, 0: not)	
	0 1	
	0.6513761 0.3486239	
	Hospital_Ownership	
	Government – Federal	0.014416
	Government - Hospital District or Authority	0.141546
	Government – Local	0.048492
	Government – State	0.017038
	Physician	0.007863
	Proprietary	0.207077
	Tribal	0.006553
	Voluntary non-profit – Church	0.078636
	Voluntary non-profit – Other	0.124508
	Voluntary non-profit – Private	0.353866

Simple Random Sample (SRS)

a. Sample size : 445
b. Estimations:

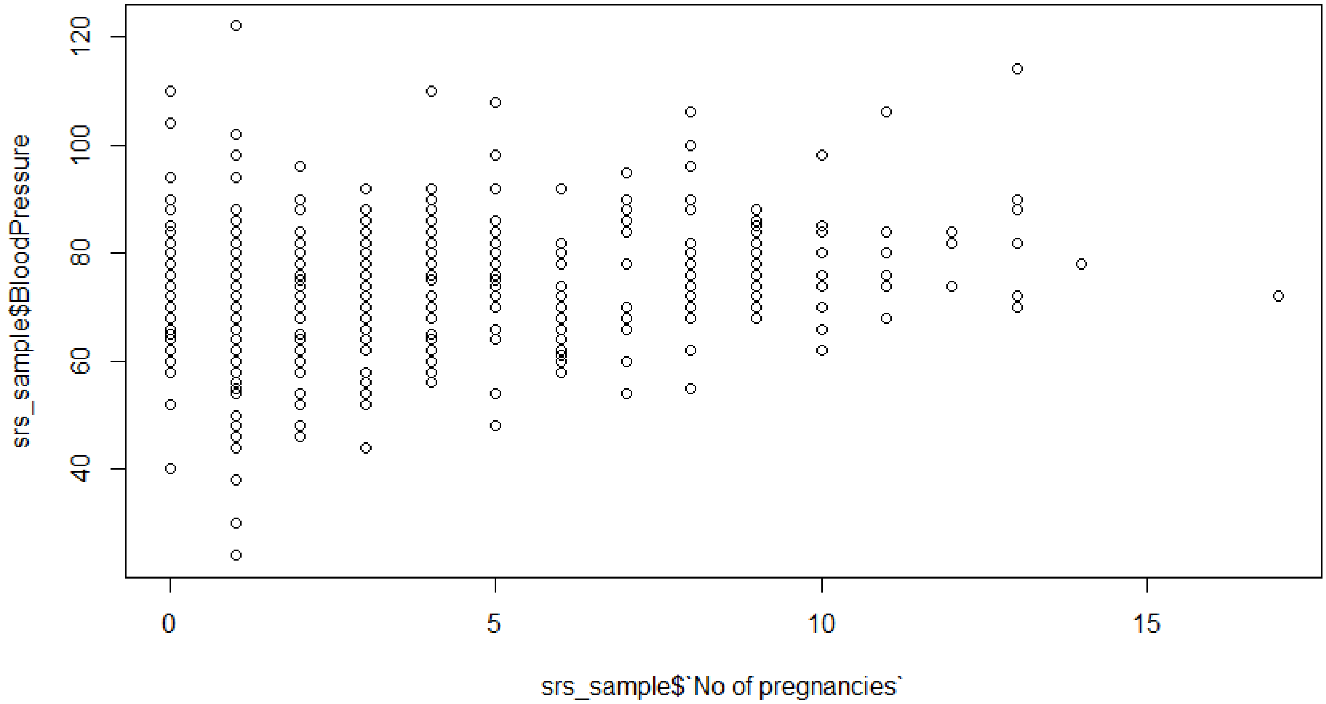
Mean	mean SE BloodPressure 72.299 0.6011 mean SE Glucose 123.01 1.4758 mean SE NoOfPregnancies 3.676 0.1582 Age 32.598 0.5472
Total	total SE NoOfPregnancies 1636 70.403 NoOfPatientsHavingDiabetics 158 NoOfPatientsNotHavingDiabetics 287 Hospital_Ownership Government – Federal 7 Government - Hospital District or Authority 57 Government – Local 21 Government – State 9 Physician 6 Proprietary 93 Tribal 2 Voluntary non-profit – Church 35 Voluntary non-profit – Other 56 Voluntary non-profit – Private 159
Proportion	mean SE NoOfPregnancies 3.6764 0.1582 NoOfPatientsHavingDiabetics 0.3550 NoOfPatientsNotHavingDiabetics 0.6449 Hospital_Ownership Government – Federal 0.01573033 Government - Hospital District or Authority 0.12808988 Government – Local 0.04719101 Government – State 0.02022471 Physician 0.01348314 Proprietary 0.20898876 Tribal 0.00449438 Voluntary non-profit – Church 0.07865168 Voluntary non-profit – Other 0.12584268 Voluntary non-profit – Private 0.3573037

- When considering the estimated values of this simple random sampling design to actual population values,
- The variable BloodPressure has lower standard errors than the variable Glucose.

- The estimated proportions for hospital ownership variable has lower standard error, when the population proportion is 0.5.

c. Regression Estimation

For Sample 01



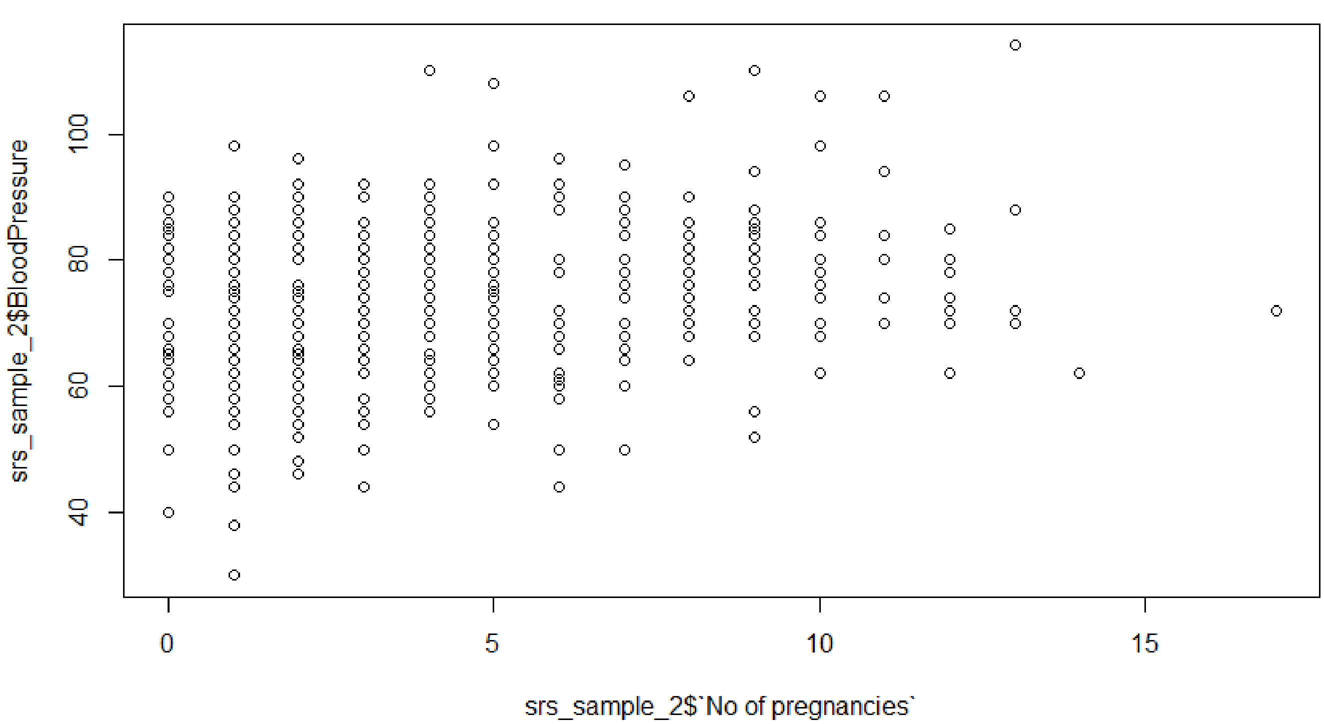

```
Call:
lm(formula = BloodPressure ~ `No of pregnancies`, data = srs_sample)

Coefficients:
        (Intercept)  `No of pregnancies`
            68.9729               0.9047
```

```
#mean no of pregnancies in population=3.676404
#then calculate expected mean no BloodPressure using regression model
mean_BloodPressure_1= 68.9729 +0.9047*3.676404
mean_BloodPressure_1
```

Calculate expected mean blood Pressure using regression model = 72.29894

For sample 02




```
Call:
lm(formula = BloodPressure ~ `No of pregnancies`, data = srs_sample_2)

Coefficients:
        (Intercept)  `No of pregnancies`
             69.610                0.801
```

```
#mean no of pregnancies in population=3.676404
#then calculate expected mean no BloodPressure using regression model
mean_BloodPressure_2=  69.610+0.801*3.676404
mean_BloodPressure_2
```

Calculate expected mean blood Pressure using regression model = 72.5548

The expected mean marks obtained from the regression model for sample 1 and sample 2 are approximately equivalent

d. Comparison

	Sample 01_SRS	Sample 02_SRS
Mean	Mean SE BloodPressure 72.299 0.6011 Mean SE Glucose 123.01 1.4758 Mean SE NoOfPregnancies 3.6764 0.1582	mean SE BloodPressure 72.636 0.5919 mean SE Glucose 121.53 1.4006 mean SE NoOfPregnancies 3.6742 0.1534
Total	total SE NoOfPregnancies 1636 70.403	total SE NoOfPregnancies 1635 68.251
Total	NoOfPatientsHavingDiabetics 158 NoOfPatientsNotHavingDiabetics 287 Hospital Ownership Government – Federal 7 Government - Hospital District or Authority57 Government – Local 21 Government – State 9 Physician 6 Proprietary 93 Tribal 2 Voluntary non-profit – Church 35 Voluntary non-profit – Other 56 Voluntary non-profit – Private 159	NoOfPatientsHavingDiabetics 152 NoOfPatientsNotHavingDiabetics 293 Hospital Ownership Government – Federal 6 Government - Hospital District or Authority55 Government – Local 17 Government – State 2 Physician 100 Proprietary 2 Tribal 34 Voluntary non-profit – Church 35 Voluntary non-profit – Other 58 Voluntary non-profit – Private 165

Here, SRS 01 and SRS 02 are given nearly equivalent estimated values for both variable, BloodPressure and Glucose.

The estimated total for no of pregnancies variable has nearly same estimated total as population total.

Stratified Random Sample

a. Stratification Variable - Ethnicity
b. Sample size - 444

```
> strata_size
  Ethnicity  Nh         wt  nh
1     Asian  76 0.09960682  44
2     Black 228 0.29882045 133
3     Mixed  76 0.09960682  44
4     White 383 0.50196592 223
>
```

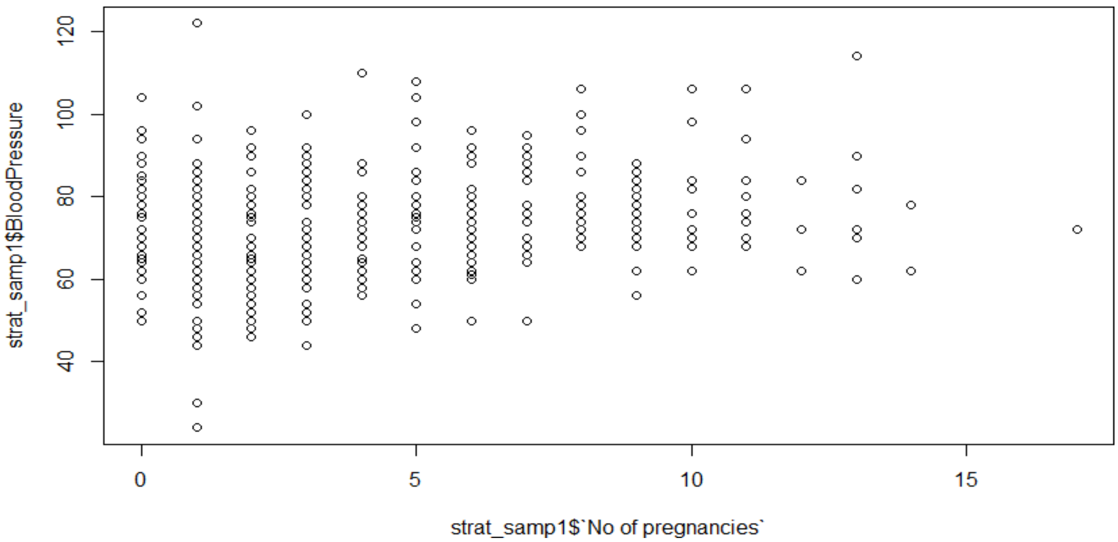
c. Estimations

Mean BloodPressure Glucose No of pregnancies Diabetic patients Age	mean SE 72.723 0.5973 121.94 1.4931 3.9662 0.1634 0.35811 0.0227 33.302 0.5437
Total No of pregnancies	Total SE 3028.9 124.75
Proportion	mean SE NoOfPregnancies 3.6764 0.1582 NoOfPatientsHavingDiabetics 0.3550 NoOfPatientsNotHavingDiabetics 0.6449 Hospital_Ownership Government – Federal 0.01573043 Government - Hospital District or Authority 0.12808956 Government – Local 0.04719201 Government – State 0.02022471 Physician 0.01348414 Proprietary 0.20898876 Tribal 0.00449438 Voluntary non-profit – Church 0.07865168 Voluntary non-profit – Other 0.12584268 Voluntary non-profit – Private 0.35730345

- When considering the estimated values of this stratified random sampling design to actual population values,The variable BloodPressure & Glucose has nearly similar mean and lower standard error. The estimated total for no of pregnancies variable has deviated from the population.

d. Regression Estimation:

For sample 01

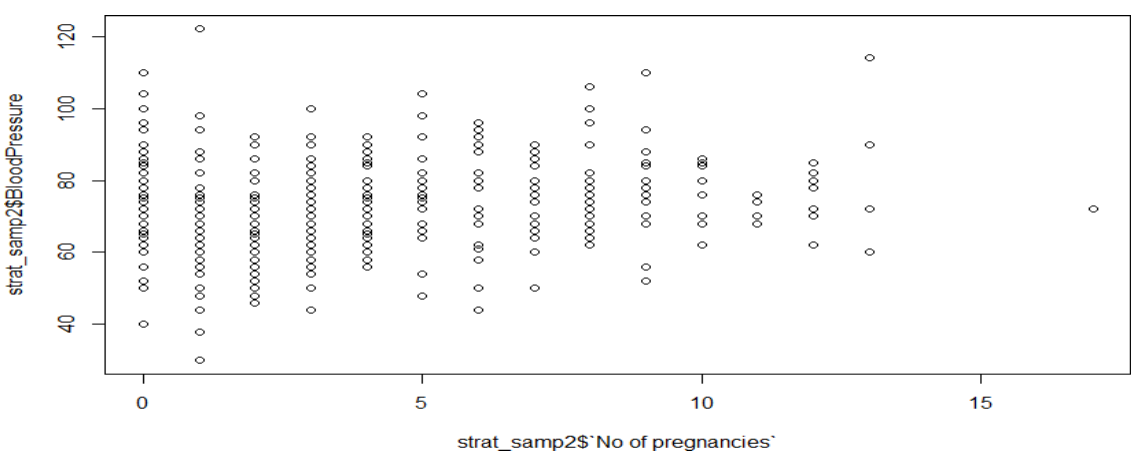



```
Call:
lm(formula = BloodPressure ~ `No of pregnancies`, data = strat_samp1)

Coefficients:
        (Intercept)  `No of pregnancies`
            69.7095               0.7598
```

Calculate expected mean blood Pressure using regression model = 72.50283

For sample 02




```
Call:
lm(formula = BloodPressure ~ `No of pregnancies`, data = strat_samp2)

Coefficients:
        (Intercept)  `No of pregnancies`
            70.6117               0.5275
```

Calculate expected mean blood Pressure using regression model = 72.551

The expected mean marks obtained from the regression model for sample 1 and sample 2 are approximately equivalent.

e. Comparison

	Sample 01	Sample 02
Mean	mean SE BloodPressure 72.723 0.5973 Glucose 121.94 1.493 Age 33.302 0.543 NoOfPreg 3.966 0.1634	mean SE BloodPressure 72.552 0.5826 Glucose 121.23 1.4685 Age 32.87 0.5478 NoOfPreg 3.677 0.1575
Total	No of Pregnancies Total SE 3028.9 124.75	No of Pregnancies total SE 2400.5 102.78
Proportion	mean SE Diabetics 0.35811 0.0227 Ethnicity Asian 0.099099 Black 0.299550 Mixed 0.099099 White 0.502252	mean SE Diabetics 0.33784 0.0225 Ethnicity Asian 0.098099 Black 0.299650 Mixed 0.098099 White 0.512252

- Here, Sample 01 and Sample 02 are given nearly equivalent estimated values for the variables BloodPressure, Glucose, Age & No Of Pregnancies.
- The noticeable fact is, the standard errors of proportions Bloodpressure, Glucose, Age, No of pregnancies variables are almost similar. In the other hand the variable diabetics, Ethnicity also shows similar standard errors.

When comparing the sample 1 and sample 2 estimations with the population values all three estimators mean, total and proportion are approximately equivalent with lower standard errors.

Two – Stage Cluster Sample

a. Clustering Variable: Hospital State

b. Sample Size - Number clusters in the population: 10

```
> ClusterDetails
  Hospital state Sample Size Population Size
1             FL          41              42
2             CO          73              78
3             CA         258             340
4             AR          71              75
5             AL          83              89
6             DE           7               7
>
```

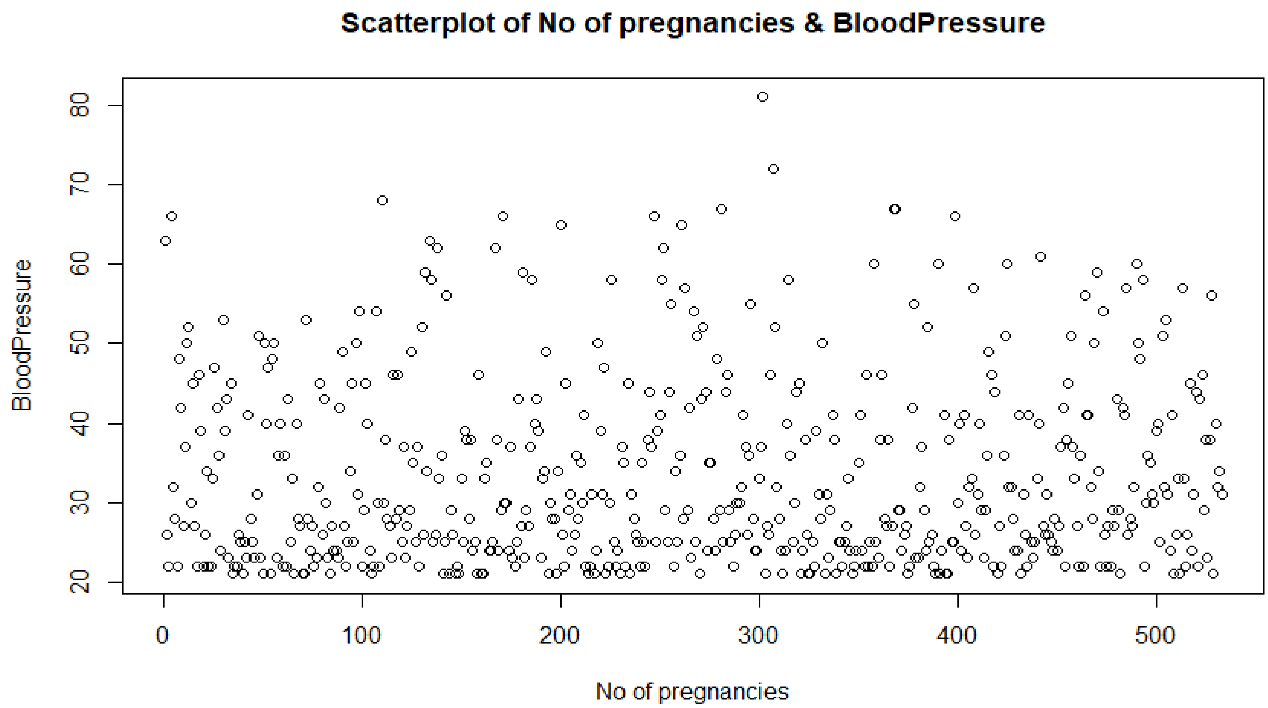
c. Estimations

Mean	> Cluster_BP_1
	                mean SE
	BloodPressure 72.143 0.3072
	> Cluster_GC_1
	          mean SE
	Glucose 122.74 1.4603
	> Cluster_mean_NoOfPregnancies_1
	                       mean SE
	`No of pregnancies` 3.8493 0.2578
	>
	> Cluster_mean_Age_1
	      mean SE
	Age 32.66 0.3083
	>

Total	> Cluster_total_NoOfpegnancies_1 total SE `No of pregnancies` 4048.2 1645.2 >
Proportions	> ClusterDetails Hospital state Sample Size Population Size 1 AZ 74 79 2 FL 41 42 3 AR 71 75 4 DC 8 8 5 AK 14 14 6 CO 73 78 >

d. Regression Estimation :

For sample 1

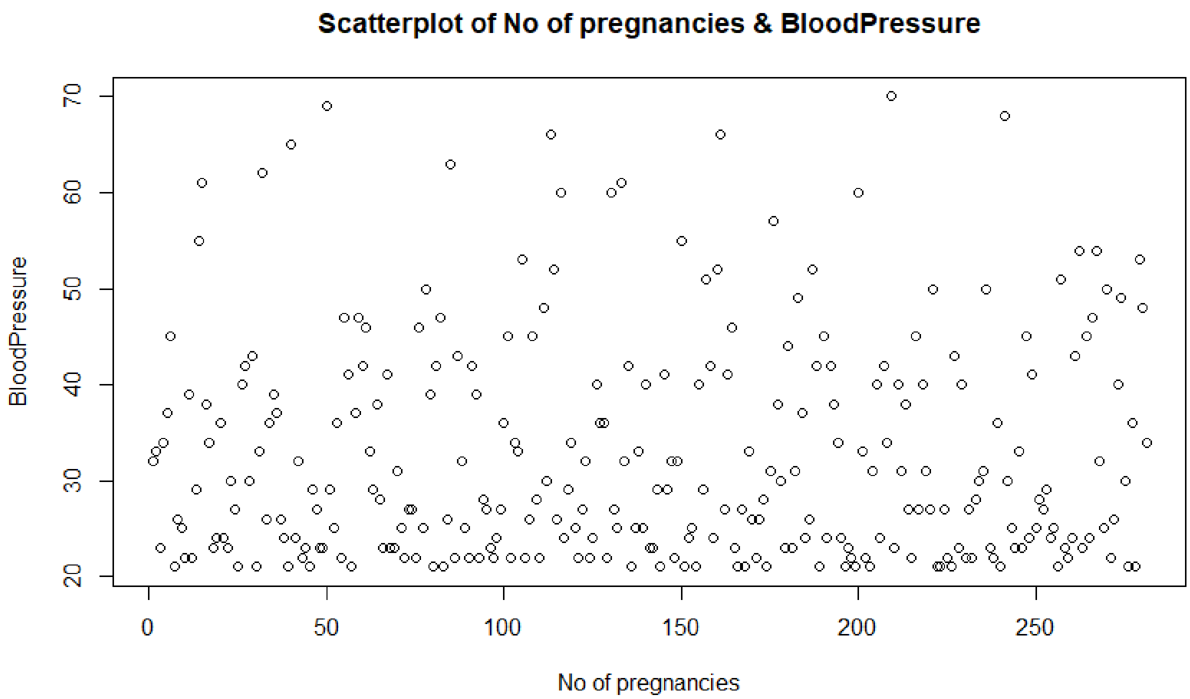



```
Call:
lm(formula = BloodPressure ~ `No of pregnancies`, data = Cluster1)

Coefficients:
        (Intercept)  `No of pregnancies`
             69.182                0.772

> 
> #mean NO of pregnancies in population=3.676404
> #then calculate expected mean BloodPressure using regression model
> mean_NoOfPregnancies_1= 69.182+ 0.772  *3.676404
> mean_NoOfPregnancies_1
[1] 72.02018
> 
```

For sample 2




```
Call:
lm(formula = BloodPressure ~ `No of pregnancies`, data = Cluster2)

Coefficients:
        (Intercept)   `No of pregnancies`
            70.6529                0.5437

> #mean NO of pregnancies in population=3.676404
> #then calculate expected mean BloodPressure using regression model
> mean_NoOfPregnancies_2=  70.6529+ 0.5437  *3.676404
> mean_NoOfPregnancies_2
[1] 72.65176
>
```

- The expected mean marks obtained from the regression model for sample 1 and sample 2 are approximately equivalent.

e. Comparison

Mean

Sample 01	Sample 02

`> Cluster_BP_1`
`mean SE`
`BloodPressure 72.709 0.3664`
	`> Cluster_BP_1`
	`mean SE`
	`BloodPressure 72.143 0.3072`

`> Cluster_GC_1`
`mean SE`
`Glucose 122.99 2.3382`
	`> Cluster_GC_1`
	`mean SE`
	`Glucose 122.74 1.4603`

`> Cluster_mean_NoOfPregnancies_1`
`mean SE`
`` `No of pregnancies` 3.7967 0.1681 ``
	`> Cluster_mean_NoOfPregnancies_1`
	`mean SE`
	`` `No of pregnancies` 3.8493 0.2578 ``
	`>`

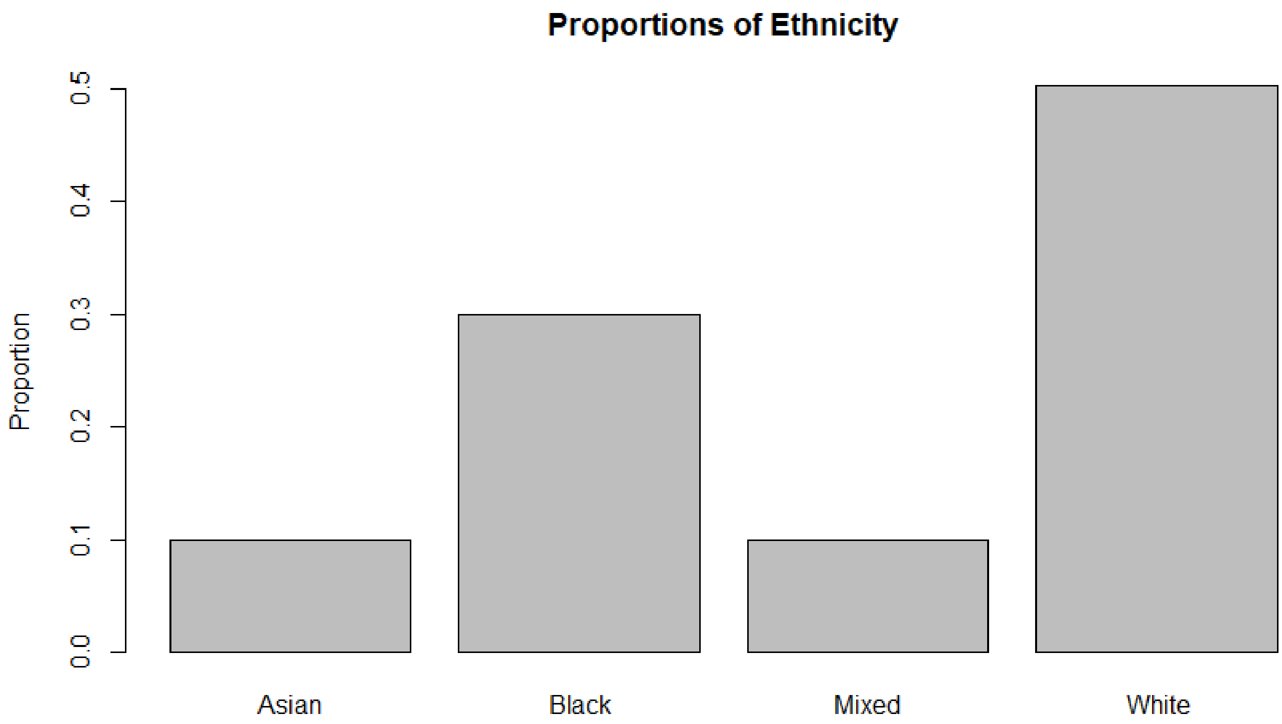
`> Cluster_mean_Age_1`
`mean SE`
`Age 32.66 0.3083`
`>`
	`> Cluster_mean_Age_1`
	`mean SE`
	`Age 32.66 0.3083`
	`>`

Sample 01	**Sample 02**

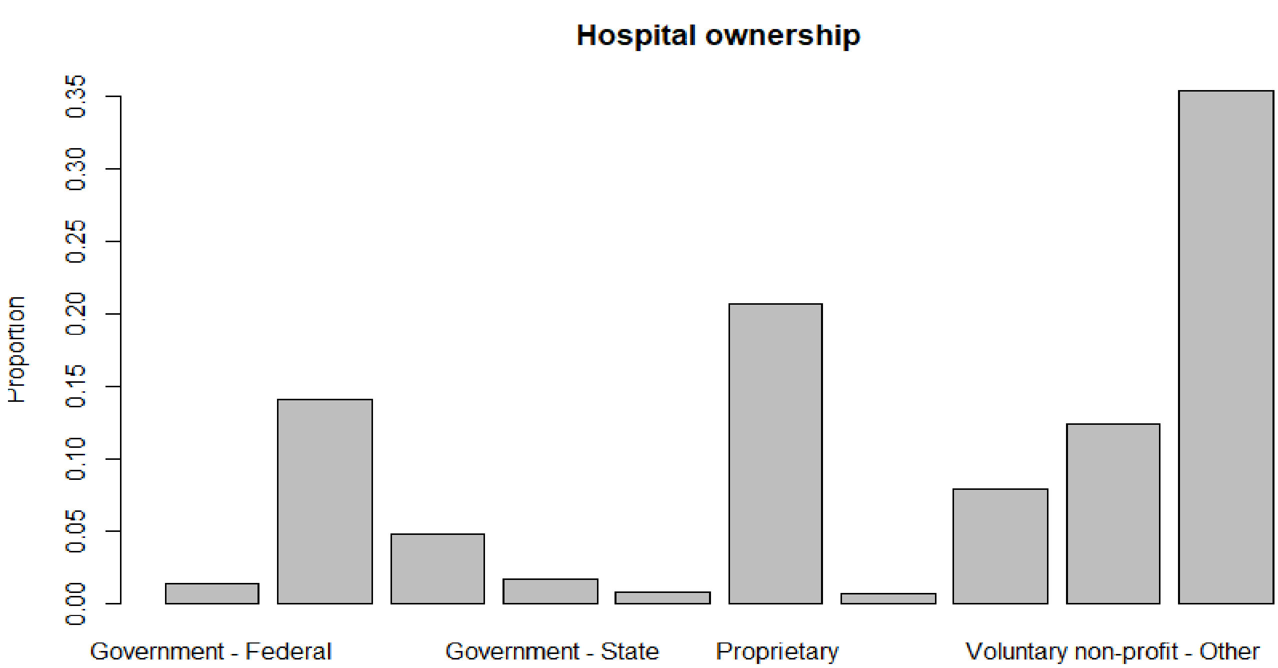
`> Cluster_total_NoOfpegnancies_1`
`total SE`
`` `No of pregnancies` 4048.2 1645.2 ``
	`> Cluster_total_NoOfpegnancies_2`
	`total SE`
	`` `No of pregnancies` 1873 541.86 ``

- The estimated means of both clusters for both variables (BloodPressure and Glucose) are almost equivalent. Estimated totals are not similar. Mainly the estimated proportions of two clusters for the variable hospital state is nearly similar as the mean and total. But there is a small difference between the estimated proportions of the variable ethnicity in the above two clusters.

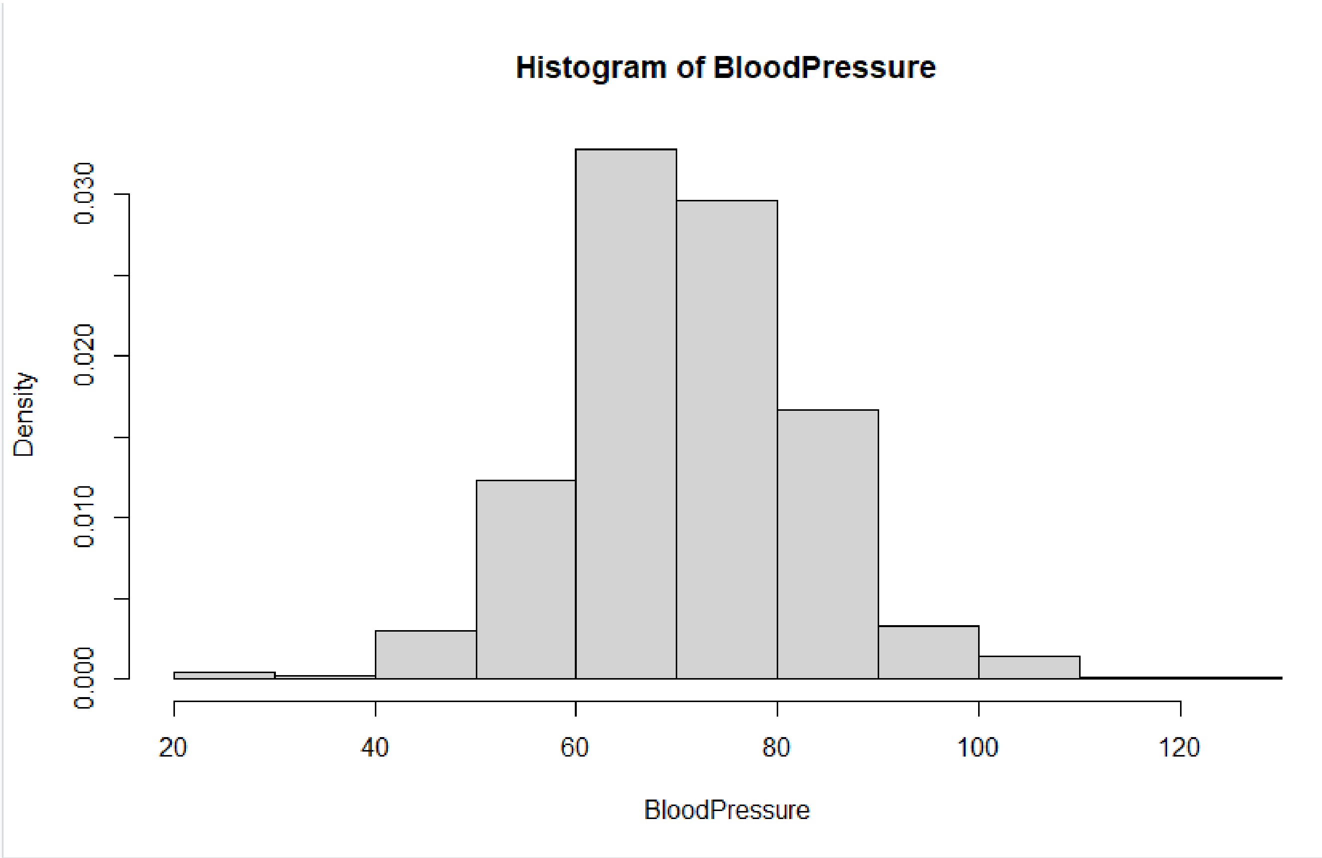
Graphical Analysis



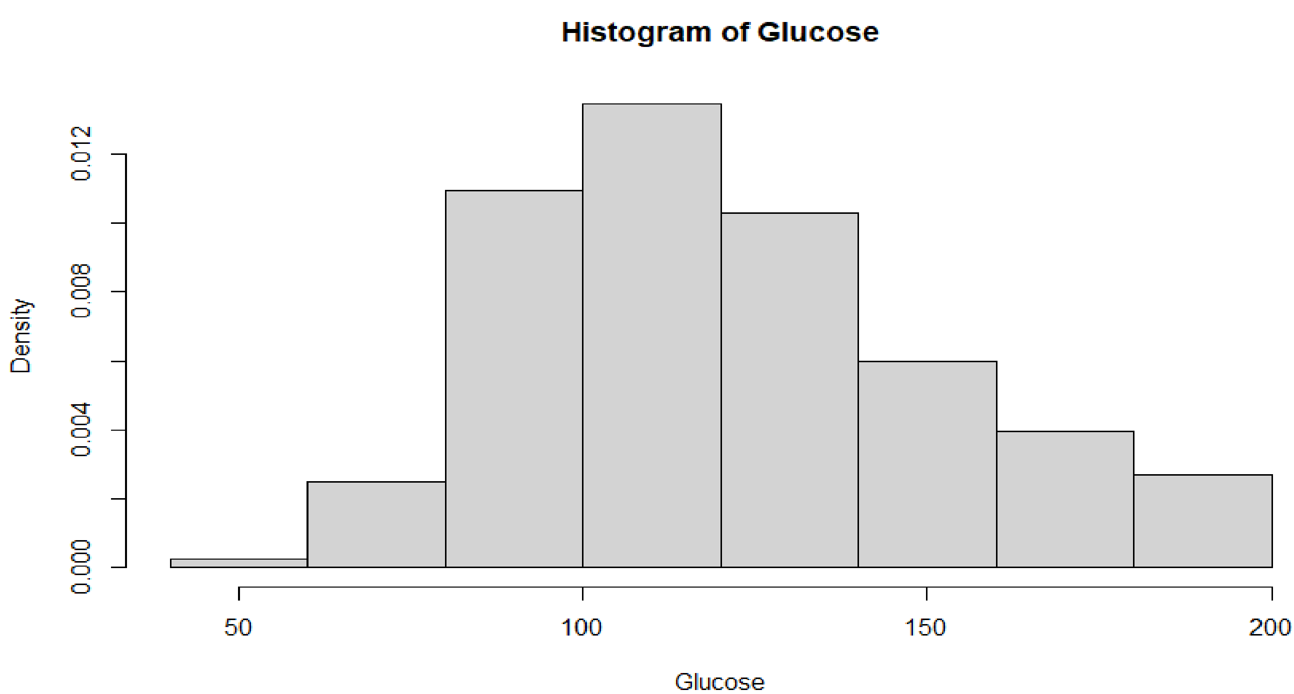
The above graph interprets the proportions of each “Asian”, “Black”, “Mixed” & “white” in the variable Ethnicity.



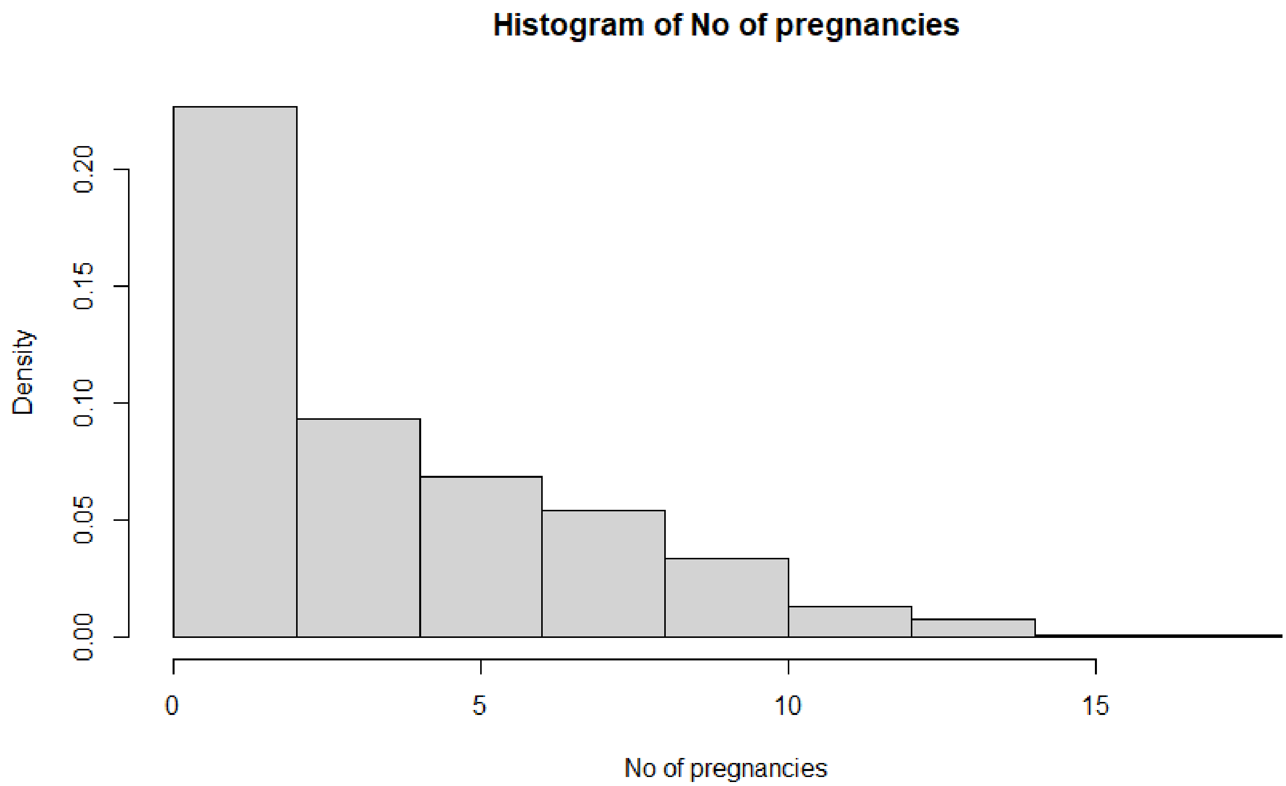
The above graph compares the proportion values in the variable of hospital ownership.



The above histogram interprets the distribution of variable blood pressure.



The above histogram interprets the distribution of variable blood pressure.



The above histogram interprets the distribution of variable No of pregnancies.

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 05 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 05.
- 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

Simple Random Sample Code

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

library(sampler)

Dataset_05 <- read_excel("C:/Users/ACER/Downloads/IS 3001 Group 11/Dataset_05.xlsx")

set.seed(123)

#sample size for SRS

srs_size=rsampcalc(nrow(Dataset_05),e=3,ci=95)

srs_size


#drawing a SRS

srs_sample=rsamp(Dataset_05,n=srs_size,rep =FALSE)

srs_sample

#Estimations

#install.packages("survey")

library(survey)

# sample mean for BloodPressure

attach(srs_sample)

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

srs_sample_mean_for_BloodPressure=svymean(~BloodPressure,srs_sample_design)

srs_sample_mean_for_BloodPressure


# sample mean for Glucose

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

srs_sample_mean_for_Glucose=svymean(~Glucose,srs_sample_design)

srs_sample_mean_for_Glucose
```

```
# sample mean for Age

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

srs_sample_mean_for_Age=svymean(~Age,srs_sample_design)

srs_sample_mean_for_Age


# sample mean for No_of_pregnancies

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

srs_sample_mean_for_NoOfPregnancies = svymean(~`No of pregnancies`,srs_sample_design)

srs_sample_mean_for_NoOfPregnancies


# sample total for No of pregnancies

srs_total_for_NoOfPregnancies=svytotal(~`No of pregnancies`,srs_sample_design)

srs_total_for_NoOfPregnancies


# No of Diabetic patients for srs

table(srs_sample$`Outcome (1: diabetic patient, 0: not)`)

# Hospital Ownership Distribution for srs

table(srs_sample$`Hospital Ownership`)

# Hospital State Distribution for srs

table(srs_sample$`Hospital state`)

# Hospital Division Distribution for srs

table(srs_sample$`Hospital division`)

# Ethnicity distribution for srs

table(srs_sample$Ethnicity)

detach(srs_sample)
```

```
# Actual values
attach(Dataset_05)

pop_mean_BloodPressure=mean(BloodPressure)
pop_mean_BloodPressure

pop_mean_Glucose=mean(Glucose)
pop_mean_Glucose

pop_mean_Age=mean(Age)
pop_mean_Age

pop_mean_NoOfPregnancies=mean(`No of pregnancies`)
pop_mean_NoOfPregnancies

# Total No of pregnancies
pop_total_NoOfPregnancies=sum(`No of pregnancies`)
pop_total_NoOfPregnancies

# No of Diabetic patients
table(Dataset_05$`Outcome (1: diabetic patient, 0: not)`)
pop_proportion=table(`Outcome (1: diabetic patient, 0: not)`)/length(`Outcome (1: diabetic patient, 0: not)`)
pop_proportion

# Hospital Ownership Distribution for population
```

```
table(Dataset_05$`Hospital Ownership`)

pop_proportion=table(`Hospital Ownership`)/length(`Hospital Ownership`)

pop_proportion



# Hospital State Distribution for population

table(Dataset_05$`Hospital state`)

pop_proportion=table(`Hospital state`)/length(`Hospital state`)

pop_proportion



# Hospital Division Distribution for population

table(Dataset_05$`Hospital division`)

pop_proportion=table(`Hospital division`)/length(`Hospital division`)

pop_proportion



# Ehnicity Distribution for population

table(Dataset_05$'Ethnicity')

pop_proportion=table(Ethnicity)/length(Ethnicity)

pop_proportion


detach(Dataset_05)


#Regression estimation
```

```
plot(srs_sample$`No of pregnancies`,srs_sample$BloodPressure)


#fitting linear regression model

lm(BloodPressure~`No of pregnancies`,srs_sample)


#mean no of pregnancies in population=3.676404

#then calculate expected mean no BloodPressure using regression model

mean_BloodPressure_1= 68.9729 +0.9047*3.676404

mean_BloodPressure_1



 #Sample 02 _SRS

set.seed(321)

#SRS sample

srs_sample_2=rsamp(Dataset_05,n=srs_size,rep =FALSE)

srs_sample_2


attach(srs_sample_2)

srs_sample_design_2=svydesign(id=~1,strata=NULL,data =srs_sample_2)


srs_sample_2_mean_Bloodpressure=svymean(~BloodPressure,srs_sample_design_2)

srs_sample_2_mean_Bloodpressure



srs_sample_2_mean_Gluecose=svymean(~Glucose,srs_sample_design_2)
```

```
srs_sample_2_mean_Gluecose

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

srs_sample_2_mean_for_NoOfPregnancies = svymean(~`No of pregnancies`,srs_sample_design_2)
srs_sample_2_mean_for_NoOfPregnancies

# sample total for No of pregnancies
srs_total_for_NoOfPregnancies=svytotal(~`No of pregnancies`,srs_sample_design_2)
srs_total_for_NoOfPregnancies

# No of Diabetic patients for srs_2
table(srs_sample_2$`Outcome (1: diabetic patient, 0: not)`)

# Hospital Ownership Distribution for srs_2
table(srs_sample_2$`Hospital Ownership`)

# Hospital State Distribution for srs_2
table(srs_sample_2$`Hospital state`)

# Hospital Division Distribution for srs_2
table(srs_sample_2$`Hospital division`)

# Ethnicity distribution for srs_2
table(srs_sample_2$Ethnicity)
```

```
#Regression estimation

plot(srs_sample_2$`No of pregnancies`,srs_sample_2$BloodPressure)

#fitting linear regression model
lm(BloodPressure~`No of pregnancies`,srs_sample_2)

#mean no of pregnancies in population=3.676404
#then calculate expected mean no BloodPressure using regression model
mean_BloodPressure_2= 69.610+0.801*3.676404
mean_BloodPressure_2

detach(srs_sample_2)
```

Stratified Sampling Code

```
set.seed(1234)
attach(Dataset_05)
library(sampler)
#Q1
#Stratified variable is school setting
#sample size for stratified sampling
strata_size=ssampcalc(Dataset_05,445,`Ethnicity`)
strata_size
```

```
#getting stratified samples

strat_samp1=ssamp(Dataset_05,445,`Ethnicity`)

strat_samp1


#Q2

#stratified

attach(strat_samp1)


# sample weight for Asian = 76/44  =1.72

# sample weight for Black = 228/133 =1.72

# sample weight for Mixed = 76/44  =1.72

# sample weight for White = 383/223 =1.72


strat_samp1$w=1.72

library(survey)


#define survey design object

strat_design=svydesign(id=~1,strata =`Ethnicity`,data = strat_samp1,weights=~w)


#Estimate sample mean of Bloodpressure

str_mean_Bloodpressure=svymean(~BloodPressure,strat_design)

str_mean_Bloodpressure


#Estimate sample mean of Gluecose
```

```
str_mean_Gluecose=svymean(~Glucose,strat_design)

str_mean_Gluecose


#Estimate sample mean of Age

str_mean_Age=svymean(~Age,strat_design)

str_mean_Age


#Estimate mean No of pregnancies

str_mean_NoOfPregnancies=svymean(~`No of pregnancies`,strat_design)

str_mean_NoOfPregnancies


#Estimate sample proportion of Diabetic patients

str_prop_Diabetics=svymean(~`Outcome (1: diabetic patient, 0: not)`,strat_design)

str_prop_Diabetics


#Estimate sample proportion of Ethnicity

str_prop_Ethnicity=svymean(~Ethnicity,strat_design)

str_prop_Ethnicity


#Estimate total No of pregnancies

population_total=svytotal(~`No of pregnancies`,strat_design)

population_total


detach(strat_samp1)
```

```
#Q3

#actual values from the population

attach(Dataset_05)


population_mean=mean(BloodPressure)

population_mean


population_mean=mean(Glucose)

population_mean



population_proportion_Hospital_Ownership_wise=table(`Hospital Ownership`)/length(`Hospital Ownership`)

population_proportion_Hospital_Ownership_wise


population_proportion_Hospital_state_wise=table(`Hospital state`)/length(`Hospital state`)

population_proportion_Hospital_state_wise


population_proportion_Hospital_Division_wise=table(`Hospital division`)/length(`Hospital division`)

population_proportion_Hospital_Division_wise


detach(Dataset_05)


#Q4

# regression estimation

plot(strat_samp1$`No of pregnancies`,strat_samp1$BloodPressure)
```

```
#fitting linear regression model

lm(BloodPressure~`No of pregnancies`,strat_samp1)


#mean NO of pregnancies in population=3.676404

#then calculate expected mean bloodpressure using regression model

mean_BloodPressure_1= 69.7095+ 0.7598 *3.676404

mean_BloodPressure_1


# getting on other stratified sample

strat_samp2=ssamp(Dataset_05,445,`Ethnicity`)

strat_samp2


strat_samp2$w=1.47

attach(strat_samp2)


strat_design2=svydesign(id=~1,strata =`Ethnicity`,data = strat_samp2,weights=~w)


#Estimate sample mean of Bloodpressure

str_mean2_Bloodpressure=svymean(~BloodPressure,strat_design2)

str_mean2_Bloodpressure


#Estimate sample mean of Gluecose

str_mean2_Gluecose=svymean(~Glucose,strat_design2)

str_mean2_Gluecose
```

```
#Estimate sample mean of Age

str_mean2_Age=svymean(~Age,strat_design2)

str_mean2_Age


#Estimate mean No of pregnancies

str_mean2_NoOfPregnancies=svymean(~`No of pregnancies`,strat_design2)

str_mean2_NoOfPregnancies


#Estimate sample proportion of Diabetic patients

str_prop_Diabetics2=svymean(~`Outcome (1: diabetic patient, 0: not)`,strat_design2)

str_prop_Diabetics2


#Estimate total No of pregnancies

population_total2=svytotal(~`No of pregnancies`,strat_design2)

population_total2


#Estimate sample proportion of Ethnicity

str_prop_Ethnicity2=svymean(~Ethnicity,strat_design2)

str_prop_Ethnicity2
```

```
# regression estimation

plot(strat_samp2$`No of pregnancies`,strat_samp2$BloodPressure)



#fitting linear regression model

lm(BloodPressure~`No of pregnancies`,strat_samp2)



#mean NO of pregnancies in population=3.676404

#then calculate expected mean BloodPressure using regression model

mean_NoOfPregnancies_2= 70.6117+0.5275 *3.676404

mean_NoOfPregnancies_2

detach(strat_samp2)
```

Two Stage Cluster Sampling Code

```
#Cluster sampling
set.seed(1234)
e=3
ci=95

#1) Obtaining a sample from two stage Cluster Sampling
#Selecting the number of clusters
#Cluserting variable = Hospital state

n = 6 #No of clusters selected

#Number of clusters in the population
N=length(unique(Dataset_05$`Hospital state`))
N


#Selecting the First Cluster Sample
#Selecting the clusters using SRS
clusters1 = sample(x = unique(Dataset_05$`Hospital state`),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 = Dataset_05[Dataset_05$`Hospital state`==clusters1[i],]
ClusterSize[i] = nrow(dat)
#Selecting sample sizes for each cluster
library(sampler)
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("Hospital state","Sample Size","Population Size")
ClusterDetails

#View(Cluster1)
#Calculating sample weights
pw = numeric(0)
for (i in 1:nrow(Cluster1)){
pw[i] = (N*ClusterDetails[ClusterDetails$`Hospital state`==Cluster1[i,]$`Hospital state`,]$`Population Size`)/
(n*ClusterDetails[ClusterDetails$`Hospital state`==Cluster1[i,]$`Hospital state`,]$`Sample Size`)
}
#Adding weights column to the main data frame
Cluster1=cbind(Cluster1,pw)

#View(Cluster1)
#Survey  Design
library(survey)
#Clustering variables are Hospital state & division
Cluster_Design = svydesign(ids = ~`Hospital state`+~`Hospital division`, weights = ~pw, data = Cluster1)
#Calculating mean, proportion, total and their Standard errors
#Proportions
Cluster_BP_1 = svymean(~BloodPressure,design = Cluster_Design)
Cluster_BP_1

Cluster_GC_1 = svymean(~Glucose,design = Cluster_Design)
Cluster_GC_1

#Means
Cluster_mean_NoOfPregnancies_1 = svymean(~`No of pregnancies`,design = Cluster_Design)
Cluster_mean_NoOfPregnancies_1
```

```
Cluster_mean_Age_1 = svymean(~Age,design = Cluster_Design)
Cluster_mean_Age_1

#totals
Cluster_total_NoOfpegnancies_1 = svytotal(~`No of pregnancies`,design = Cluster_Design)
Cluster_total_NoOfpegnancies_1

#4) Perform ratio or regression estimations
#Regression estimation
#plot(Cluster1$Age,Cluster1$BloodPressure,main="Scatterplot of Age &
Bloodpressure",xlab="Age",ylab="BloodPressure")
#plot(Cluster1$Age,Cluster1$Glucose,main="Scatterplot of Age & Glucose",xlab="Age",ylab="Glucose")
plot(Cluster1$Age,Cluster1$NoOfPregnancies,main="Scatterplot of No of pregnancies & BloodPressure",xlab="No of
pregnancies",ylab="BloodPressure")

RegressionLm = lm(`BloodPressure`~`No of pregnancies`,data = Cluster1)
RegressionLm

#mean NO of pregnancies in population=3.676404
#then calculate expected mean BloodPressure using regression model
mean_NoOfPregnancies_1= 69.182+ 0.772 *3.676404
mean_NoOfPregnancies_1

#Ratio Estimation
r=svyratio(~`BloodPressure`,~`Glucose`,Cluster_Design)
r

predict(r,mean(Dataset_05$BloodPressure))

#Selecting the second Cluster sample
set.seed(4567)
#Selecting the clusters using SRS
clusters2 = sample(x = unique(Dataset_05$`Hospital state`),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 = Dataset_05[Dataset_05$`Hospital state`==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("Hospital state","Sample Size","Population Size")
ClusterDetails

#View(Cluster2)
#Calculating sample weights
pw = numeric(0)
for (i in 1:nrow(Cluster2)){
  pw[i] = (N*ClusterDetails[ClusterDetails$`Hospital state`==Cluster2[i,]$`Hospital state`,]$`Population Size`)/
    (n*ClusterDetails[ClusterDetails$`Hospital state`==Cluster2[i,]$`Hospital state`,]$`Sample Size`)
}
attach(Cluster2)
#Adding weights column to the main data frame
Cluster2=cbind(Cluster2,pw)
#View(Cluster2)
#Survey   Design
library(survey)
#Clustering variables are Hospital state & division
Cluster_Design = svydesign(ids = ~`Hospital state`+~`Hospital division`, weights = ~pw, data = Cluster2)

#Calculating mean, proportion, total and their Standard errors
#Proportions
Cluster_BP_2 = svymean(~BloodPressure,design = Cluster2_Design)
Cluster_BP_2


Cluster_GC_2 = svymean(~Glucose,design = Cluster_Design)
Cluster_GC_2

#Means
Cluster_mean_NoOfPregnancies_2 = svymean(~`No of pregnancies`,design = Cluster_Design)
Cluster_mean_NoOfPregnancies_2

Cluster_mean_Age_2 = svymean(~Age,design = Cluster_Design)
Cluster_mean_Age_2

#totals
Cluster_total_NoOfpegnancies_2 = svytotal(~`No of pregnancies`,design = Cluster_Design)
Cluster_total_NoOfpegnancies_2

# 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_BPS = table(Dataset_05$BloodPressure)/length(Dataset_05$BloodPressure)
Pop_BPS

Pop_Glucose = table(Dataset_05$Glucose)/length(Dataset_05$Glucose)
Pop_Glucose


#Means
pop_mean_BloodPressure = mean(Dataset_05$BloodPressure)
pop_mean_BloodPressure


pop_mean_Glucose = mean(Dataset_05$Glucose)
pop_mean_Glucose

pop_mean_NoOfPregnancies = mean(Dataset_05$`No of pregnancies`)
pop_mean_NoOfPregnancies

#Totals
pop_total_NoOfPregnancies = sum(Dataset_05$`No of pregnancies`)
pop_total_NoOfPregnancies

#Regression estimation
#plot(Cluster1$Age,Cluster1$BloodPressure,main="Scatterplot of Age & Bloodpressure",xlab="Age",ylab="BloodPressure")
#plot(Cluster1$Age,Cluster1$Glucose,main="Scatterplot of Age & Glucose",xlab="Age",ylab="Glucose")
plot(Cluster2$Age,Cluster2$NoOfPregnancies,main="Scatterplot of No of pregnancies & BloodPressure",xlab="No of pregnancies",ylab="BloodPressure")

RegressionLm = lm(`BloodPressure`~`No of pregnancies`,data = Cluster2)
RegressionLm

#mean NO of pregnancies in population=3.676404
#then calculate expected mean BloodPressure using regression model
mean_NoOfPregnancies_2= 70.6529+ 0.5437 *3.676404
mean_NoOfPregnancies_2
```

Graphical Analysis Code

```
View(Dataset_05)
#Ethnicity
library(sampler)
library(survey)
Ethnicity_table = table(Dataset_05$Ethnicity)/length(Dataset_05$Ethnicity)
Ethnicity_table
##    Asian      Black      Mixed      White
## 0.09960682 0.29882045 0.09960682 0.50196592



barplot(Ethnicity_table,main="Proportions of Ethnicity",names.arg =c("Asian"," Black"," Mixed "," White"), ylab =
"Proportion" )

Hospital_Ownership_table = table(Dataset_05$`Hospital Ownership`)/length(Dataset_05$`Hospital Ownership`)
Hospital_Ownership_table
## Government - Federal Government - Hospital District or Authority
##  0.014416776                         0.141546527
#Government - Local                    Government - State
#0.048492792                         0.017038008
#Physician                          Proprietary
#0.007863696                         0.207077326
#Tribal             Voluntary non-profit - Church
#0.006553080                         0.078636959
#Voluntary non-profit - Other            Voluntary non-profit - Private
#0.124508519                         0.353866317

barplot(Hospital_Ownership_table,main="Hospital ownership",names.arg =c("Government - Federal","Government -
Hospital District or Authority"," Government - Local "," Government - State"," Physician"," Proprietary","
Tribal","Voluntary non-profit - Church ", "Voluntary non-profit - Other","Voluntary non-profit - Private"), ylab =
"Proportion" )

#BloodPressure
#par(mfrow=c(1,2))
#svyhist(~BloodPressure,design = Cluster_Design)
hist(x = Dataset_05$BloodPressure,prob=T,main="Histogram of BloodPressure",xlab="BloodPressure")


#Gluecose
#par(mfrow=c(1,2))
#svyhist(~Glucose,design = Cluster_Design)
hist(x = Dataset_05$Glucose,prob=T,main="Histogram of Glucose",xlab="Glucose")

#Gluecose
#par(mfrow=c(1,2))
#svyhist(~`No of pregnanices`,design = Cluster_Design)
hist(x = Dataset_05$`No of pregnancies`,prob=T,main="Histogram of No of pregnancies",xlab="No of pregnancies")
```