R Notebook

```
# load data set
diabetes <- read.csv("diabetes.csv", header=T)

#Calculating necessary Sample Size
Population_Size <- nrow(diabetes)
MOE <- 0.05
Z <- 1.96
Proportion <- 0.5
Sample_Size <-
  Proportion*(1-Proportion)/(MOE^2/Z^2+Proportion*(1-Proportion)/Population_Size)
```

Using $n = \frac{p(1-p)}{\frac{\sigma^2}{z^2} + \frac{p(1-p)}{N}}$ results to the necessary sample size of 256.0711012 rounding to 257.

```
#Selecting the samples

#set seed
set.seed(1738)
#Random sample selection
SRS_Sample <- diabetes[sample(nrow(diabetes), Sample_Size), ]
head(SRS_Sample)
```

```
##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 437          12     140            85            33       0 37.4
## 161           4     151            90            38       0 29.7
## 586           1      93            56            11       0 22.5
## 115           7     160            54            32     175 30.5
## 38            9     102            76            37       0 32.9
## 436           0     141             0             0       0 42.4
##     DiabetesPedigreeFunction Age Outcome
## 437                    0.244  41       0
## 161                    0.294  36       0
## 586                    0.417  22       0
## 115                    0.588  39       1
## 38                     0.665  46       1
## 436                    0.205  29       1
```

```
#Stratified sample selection !!!To be decided how to stratify!!!
```