R Notebook

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

# Our population of interest are people at risk of heart attack
heartattack <- heartattack[heartattack$Heart.Attack.Risk == 1,]
head(heartattack)
```

```
##    Patient.ID Age    Sex Cholesterol Blood.Pressure Heart.Rate Diabetes
## 6     ZOO7941  54 Female         297         172/86         48        1
## 7     WYV0966  90   Male         358         102/73         84        0
## 8     XXM0972  84   Male         220         131/68        107        0
## 13    FPS0415  77   Male         228         101/72         68        1
## 14    YYU9565  60   Male         259         169/72         85        1
## 16    DCY3282  73   Male         122         114/88         97        1
##    Family.History Smoking Obesity Alcohol.Consumption Exercise.Hours.Per.Week
## 6               1       1       0                   1                0.625008
## 7               0       1       0                   1                4.098177
## 8               0       1       1                   1                3.427929
## 13              1       1       1                   1               19.633268
## 14              1       1       0                   1               17.037374
## 16              1       1       0                   1               14.559664
##         Diet Previous.Heart.Problems Medication.Use Stress.Level
## 6  Unhealthy                      1              1            2
## 7    Healthy                      0              0            7
## 8    Average                      0              1            4
## 13 Unhealthy                      0              0            9
## 14   Healthy                      1              1            1
## 16   Average                      0              0            5
##    Sedentary.Hours.Per.Day Income      BMI Triglycerides
## 6                 7.798752 241339 20.14684           795
## 7                 0.627356 190450 28.88581           284
## 8                10.543780 122093 22.22186           370
## 13               10.917524  29886 35.10224           590
## 14                8.727417 292173 25.56490           506
## 16               10.086479 265839 36.52440           773
##    Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day Country     Continent
## 6                                5                  10 Germany        Europe
## 7                                4                  10  Canada North America
## 8                                6                   7   Japan          Asia
## 13                               7                   6 Vietnam          Asia
## 14                               1                   4   China          Asia
## 16                               5                   8   Italy        Europe
##             Hemisphere Heart.Attack.Risk
## 6  Northern Hemisphere                 1
## 7  Northern Hemisphere                 1
## 8  Northern Hemisphere                 1
```

```
## 13 Northern Hemisphere                 1
## 14 Northern Hemisphere                 1
## 16 Southern Hemisphere                 1
```

Find recommended sample size for this study

```
# calculate min sample size needed
pop_size <- nrow(heartattack) # 3139

# using 95% CI, find n for worst case scenario: p = 0.5
MOE <- 0.05
z <- 1.96
p_guess <- 0.5

# if N is large enough to ignore FPC
n_0 = ceiling( (z^2*(0.5)*(0.5)) / (MOE^2)) # 385
# since we know N = 3139, using FPC
n = ceiling( n_0 / (1 + (n_0/pop_size)) ) # 343
```

Assuming the worst case proportions 0.5, the sample size used if we ignored FPC is 385. Whereas including FPC the sample size used in SRS will be 343.

Compare study design for stratification

```


#Calculate within variance of each sex: Male, Female
variance_within_strata <- aggregate(BMI ~ Sex, heartattack, var)
colnames(variance_within_strata) <- c("Sex","Within Variance Sex")
print(variance_within_strata)
```

Method 1: stratify by sex

```
##      Sex Within Variance Sex
## 1 Female            38.33507
## 2   Male            40.77213
```

```
#Get stratum sizes
male_stratum_size <- nrow(heartattack[heartattack$Sex == "Male",])
female_stratum_size <- nrow(heartattack[heartattack$Sex == "Female",])


#Sample size n_h proportional to N_h*S_pw^2/sqrt(cost)
#Ignore costs

#total is used to normalize N_h*S_pw^2/sqrt(cost) to equal 1
total <- sum(male_stratum_size*variance_within_strata$`Within Variance Sex`[1],
             female_stratum_size*variance_within_strata$`Within Variance Sex`[2])

male_size_proportion <-
```

```
  male_stratum_size*variance_within_strata$`Within Variance Sex`[1]/total

female_size_proportion <-
  female_stratum_size*variance_within_strata$`Within Variance Sex`[2]/total

#total sample size * strata proportion = strata sample size
male_sample_size <- round(male_size_proportion*n)
female_sample_size <- round(female_size_proportion*n)

#Overall stratified variance
var.strata <- c(variance_within_strata$`Within Variance Sex`[1],
                variance_within_strata$`Within Variance Sex`[2])
wt.strata <- c(male_size_proportion, female_size_proportion)

overall.sex.var <- sum(wt.strata*var.strata)
data.frame(`Overall Sex Variation` = c(overall.sex.var))
```

```
##   Overall.Sex.Variation
## 1              39.09994
```

```
#Calculate within variance of each diet stratum: Average, Unhealthy, Healthy
variance_within_strata <- aggregate(BMI ~ Diet, heartattack, var)
colnames(variance_within_strata) <- c("Diet","Within Variance BMI")
variance_within_strata
```

Method 2: stratify by diet

```
##        Diet Within Variance BMI
## 1   Average            40.50160
## 2   Healthy            40.07035
## 3 Unhealthy            39.64113
```

```
#Get stratum sizes
average_stratum_size <- nrow(heartattack[heartattack$Diet == "Average",])
healthy_stratum_size <- nrow(heartattack[heartattack$Diet == "Healthy",])
unhealthy_stratum_size <- nrow(heartattack[heartattack$Diet == "Unhealthy",])


#Sample size n_h proportional to N_h*S_pw^2/sqrt(cost)
#Ignore costs
#total is used to normalize N_h*S_pw^2/sqrt(cost) to equal 1
total <- sum(average_stratum_size*variance_within_strata$`Within Variance BMI`[1],
             healthy_stratum_size*variance_within_strata$`Within Variance BMI`[2],
             unhealthy_stratum_size*variance_within_strata$`Within Variance BMI`[3])

average_size_proportion <-
  average_stratum_size*variance_within_strata$`Within Variance BMI`[1]/total
healthy_size_proportion <-
  healthy_stratum_size*variance_within_strata$`Within Variance BMI`[2]/total
unhealthy_size_proportion <-
```

```
  unhealthy_stratum_size*variance_within_strata$`Within Variance BMI`[3]/total

#multiply total sample size with proportions to get the sample size for each
#strata
average_sample_size <- round(average_size_proportion*n)
healthy_sample_size <- round(healthy_size_proportion*n)
unhealthy_sample_size <- round(unhealthy_size_proportion*n)

#Overall stratified variance
var.strata <- c(variance_within_strata$`Within Variance BMI`[1],
                variance_within_strata$`Within Variance BMI`[2],
                variance_within_strata$`Within Variance BMI`[3])
wt.strata <-
  c(average_size_proportion, healthy_size_proportion, unhealthy_size_proportion)

overall.diet.var <- sum(wt.strata*var.strata)
print(overall.diet.var)
```

```
## [1] 40.07295
```

```
#Calculate within variance of whether patient has diabetes: 1: Yes, 0: No
variance_within_strata <- aggregate(BMI ~ Diabetes, heartattack, var)
colnames(variance_within_strata) <- c("Diabetes","Within Variance Diabetes")
print(variance_within_strata)
```

Method 3: stratify by whether patient has diabetes

```
##   Diabetes Within Variance Diabetes
## 1        0                 39.23851
## 2        1                 40.46166
```

```
#Get stratum sizes
diabetes_stratum_size <- nrow(heartattack[heartattack$Diabetes == 1,])
no_diabetes_stratum_size <- nrow(heartattack[heartattack$Diabetes == 0,])


#Sample size n_h proportional to N_h*S_pw^2/sqrt(cost)
#Ignore costs
total <-
  sum(diabetes_stratum_size*variance_within_strata$`Within Variance Diabetes`[1],
      no_diabetes_stratum_size*variance_within_strata$`Within Variance Diabetes`[2])

diabetes_size_proportion <-
  diabetes_stratum_size*variance_within_strata$`Within Variance Diabetes`[1]/total
no_diabetes_size_proportion <-
  no_diabetes_stratum_size*variance_within_strata$`Within Variance Diabetes`[2]/total

diabetes_sample_size <- round(diabetes_size_proportion*n)
no_diabetes_sample_size <- round(no_diabetes_size_proportion*n)
```

```r
#Overall stratified variance
var.strata <- c(variance_within_strata$`Within Variance Diabetes`[1],
                variance_within_strata$`Within Variance Diabetes`[2])
wt.strata <- c(diabetes_size_proportion, no_diabetes_size_proportion)

overall.diabetes.var <- sum(wt.strata*var.strata)
print(overall.diabetes.var)
```

```
## [1] 39.65881
```

```r
#Calculate within variance of whether patient has
#family history of heart-related problems:#1: Yes, 0: No

variance_within_strata <- aggregate(BMI ~ Family.History, heartattack, var)
colnames(variance_within_strata) <- c("Family History","Within Variance Family History")
print(variance_within_strata)
```

Method 4: stratify by whether patient has family history of heart-related problems

```
##   Family History Within Variance Family History
## 1              0                       40.39519
## 2              1                       39.71046
```

```r
#Get stratum sizes
history_stratum_size <- nrow(heartattack[heartattack$Family.History == 1,])
no_history_stratum_size <- nrow(heartattack[heartattack$Family.History == 0,])


#Sample size n_h proportional to N_h*S_pw^2/sqrt(cost)
#Ignore costs
total <-
  sum(history_stratum_size*variance_within_strata$`Within Variance Family History`[1],
   no_history_stratum_size*variance_within_strata$`Within Variance Family History`[2])

history_size_proportion <-
  history_stratum_size*variance_within_strata$`Within Variance Family History`[1]/total
no_history_size_proportion <-
  no_history_stratum_size*variance_within_strata$`Within Variance Diabetes`[2]/total

history_sample_size <- round(history_size_proportion*n)
no_history_sample_size <- round(no_history_size_proportion*n)

#Overall stratified variance
var.strata <- c(variance_within_strata$`Within Variance Family History`[1],
                variance_within_strata$`Within Variance Family History`[2])
wt.strata <- c(history_size_proportion, no_history_size_proportion)

overall.history.var <- sum(wt.strata*var.strata)
print(overall.history.var)
```

```
## [1] 39.7444
```

```r
#Calculate within variance of obesity status: 1: Obese, 0: Not obese
variance_within_strata <- aggregate(BMI ~ Obesity, heartattack, var)
colnames(variance_within_strata) <- c("Obesity","Within Variance Obesity")
print(variance_within_strata)
```

Method 5: stratify by obesity status

```
##   Obesity Within Variance Obesity
## 1       0                39.83100
## 2       1                40.29621
```

```r
#Get stratum sizes
obesity_stratum_size <- nrow(heartattack[heartattack$Obesity == 1,])
not_obese_stratum_size <- nrow(heartattack[heartattack$Obesity == 0,])


#Sample size n_h proportional to N_h*S_pw^2/sqrt(cost)
#Ignore costs
total <- sum(obesity_stratum_size*variance_within_strata$`Within Variance Obesity`[1],
             not_obese_stratum_size*variance_within_strata$`Within Variance Obesity`[2])

obesity_size_proportion <-
  obesity_stratum_size*variance_within_strata$`Within Variance Obesity`[1]/total
not_obese_size_proportion <-
  not_obese_stratum_size*variance_within_strata$`Within Variance Obesity`[2]/total

history_sample_size <- round(obesity_size_proportion*n)
no_history_sample_size <- round(not_obese_size_proportion*n)

#Overall stratified variance
var.strata <- c(variance_within_strata$`Within Variance Obesity`[1],
                variance_within_strata$`Within Variance Obesity`[2])
wt.strata <- c(obesity_size_proportion, not_obese_size_proportion)

overall.obesity.var <- sum(wt.strata*var.strata)
print(overall.obesity.var)
```

```
## [1] 40.06844
```

```r
overall_var <-
  data.frame(overall.sex.var,
             overall.diet.var,
             overall.diabetes.var,
             overall.history.var,
             overall.obesity.var)

colnames(overall_var) <-
  c("Overall Sex Var.",
    "Overall Diet Var.",
    "Overall Diabetes Var.",
    "Overall History Var.",
```

```
    "Overall Obesity Var.")

print(overall_var)
```

```
##   Overall Sex Var. Overall Diet Var. Overall Diabetes Var. Overall History Var.
## 1         39.09994          40.07295              39.65881              39.7444
##   Overall Obesity Var.
## 1             40.06844
```

By computing and comparing the within variances based on different stratas, stratifying by sex gave the lowest overall within variance of 39.09994. Since the stratification study design performs the best for the largest between-strata variance, implying the lowest within-strata variance, we will stratify by sex.

In the two stratums: Sex = (Male, Female), sample size for Male is 235 and sample size for Female is 108

Selecting Samples through SRS and Stratification by sex

```
# set seed
set.seed(1)

# take SRS of n = 1032
SRS.index <- sample.int(pop_size, n, replace=F)
SRS_sample <- heartattack[SRS.index, ]
head(SRS_sample)
```

```
##      Patient.ID Age    Sex Cholesterol Blood.Pressure Heart.Rate Diabetes
## 2898    YMC7841  86 Female         361         150/67         45        0
## 1965    YDS4023  77   Male         160        103/106         82        1
## 6079    EDZ2722  30 Female         348        104/102         54        1
## 2625    YXX0164  61   Male         205        112/110         99        1
## 4262    DQQ3866  21   Male         140        180/103         48        0
## 1379    IDW3149  32 Female         262         179/80         81        0
##      Family.History Smoking Obesity Alcohol.Consumption Exercise.Hours.Per.Week
## 2898              1       1       1                   0               19.407365
## 1965              0       1       0                   0               14.888193
## 6079              1       0       1                   1               11.607732
## 2625              1       1       0                   0               17.874208
## 4262              1       1       1                   1                3.849926
## 1379              1       0       0                   0               17.839845
##           Diet Previous.Heart.Problems Medication.Use Stress.Level
## 2898 Unhealthy                      1              1            6
## 1965   Healthy                      1              0           10
## 6079 Unhealthy                      0              1            4
## 2625   Healthy                      0              0            9
## 4262   Average                      1              1            5
## 1379 Unhealthy                      1              1            5
##      Sedentary.Hours.Per.Day Income      BMI Triglycerides
## 2898               3.7473314 147131 19.50969           259
## 1965               5.7870381 258654 23.72228           182
## 6079               2.3421202  39298 23.03643           333
## 2625               9.5188653 171259 30.56734           753
```

```
## 4262               0.8926316 179903 37.96709           409
## 1379              11.7472568 252602 37.04031           158
##      Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day      Country
## 2898                               6                  10     Colombia
## 1965                               3                   5      Nigeria
## 6079                               4                   9  New Zealand
## 2625                               3                   8 South Africa
## 4262                               3                   5      Nigeria
## 1379                               0                   7    Australia
##          Continent          Hemisphere Heart.Attack.Risk
## 2898 South America Northern Hemisphere                 1
## 1965        Africa Northern Hemisphere                 1
## 6079     Australia Southern Hemisphere                 1
## 2625        Africa Southern Hemisphere                 1
## 4262        Africa Northern Hemisphere                 1
## 1379     Australia Southern Hemisphere                 1
```

```
#Stratify male and female stratums to take samples from
male_stratum <- heartattack[heartattack$Sex == "Male",]
female_stratum <- heartattack[heartattack$Sex == "Female",]

#Take Stratified samples of males (n = 708) and females (n = 324)
stratified_male.index <- sample.int(male_stratum_size, male_sample_size, replace = F)
male_sample <- male_stratum[stratified_male.index,]
head(male_sample)
```

```
##      Patient.ID Age  Sex Cholesterol Blood.Pressure Heart.Rate Diabetes
## 6171    KRG4242  60 Male         398         103/65         68        0
## 4634    OZR9308  87 Male         301        177/102         81        1
## 6600    FFB3370  19 Male         267         105/76         99        0
## 5839    ETF7967  41 Male         212        172/105         87        1
## 1258    ROS3937  70 Male         194         136/83        109        0
## 2803    NSO4719  42 Male         396          97/74         48        1
##      Family.History Smoking Obesity Alcohol.Consumption Exercise.Hours.Per.Week
## 6171              0       1       0                   1               5.4849440
## 4634              0       1       0                   1              14.9717581
## 6600              0       1       0                   1               0.1295951
## 5839              1       1       1                   0              11.8835229
## 1258              0       1       0                   0               0.8825470
## 2803              0       1       1                   1               4.0045040
##           Diet Previous.Heart.Problems Medication.Use Stress.Level
## 6171 Unhealthy                      0              1            1
## 4634 Unhealthy                      1              0            2
## 6600   Healthy                      0              0            3
## 5839   Average                      1              0            9
## 1258   Average                      0              0            8
## 2803 Unhealthy                      1              1            3
##      Sedentary.Hours.Per.Day Income      BMI Triglycerides
## 6171               9.5839070 281319 38.85762           638
## 4634               8.6223292 292806 26.21581           663
## 6600               1.9024796  24796 18.05761           700
## 5839               0.7238682 234710 36.28044           433
## 1258               9.0172259  79839 22.49204            72
## 2803               6.6440823 192132 21.37693           429
```

```
##      Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day     Country
## 6171                               4                   7       China
## 4634                               1                   5 South Korea
## 6600                               1                   6     Vietnam
## 5839                               7                   9   Australia
## 1258                               1                   4       Japan
## 2803                               2                   8      Brazil
##          Continent          Hemisphere Heart.Attack.Risk
## 6171          Asia Northern Hemisphere                 1
## 4634          Asia Northern Hemisphere                 1
## 6600          Asia Northern Hemisphere                 1
## 5839     Australia Southern Hemisphere                 1
## 1258          Asia Northern Hemisphere                 1
## 2803 South America Southern Hemisphere                 1
```

```
stratified_female.index <- sample.int(female_stratum_size, female_sample_size, replace = F)
female_sample <- female_stratum[stratified_female.index,]
head(female_sample)
```

```
##      Patient.ID Age    Sex Cholesterol Blood.Pressure Heart.Rate Diabetes
## 5184    ZLG7622  63 Female         382         143/63         98        1
## 4257    NGM7550  43 Female         348         105/80        102        1
## 6112    NZI0371  89 Female         185         115/86         85        0
## 2749    SCD0081  66 Female         330         98/110         84        1
## 736     NDO6933  19 Female         378        119/110         51        0
## 4518    SCR8032  36 Female         387         135/79         96        1
##      Family.History Smoking Obesity Alcohol.Consumption Exercise.Hours.Per.Week
## 5184              1       1       1                   1               0.5608616
## 4257              0       1       1                   1               8.5463448
## 6112              1       1       0                   0              12.7571257
## 2749              0       1       0                   0               6.8450335
## 736               1       0       0                   1               7.4047067
## 4518              1       0       0                   1              19.5108322
##           Diet Previous.Heart.Problems Medication.Use Stress.Level
## 5184 Unhealthy                      0              0            6
## 4257   Healthy                      0              0            3
## 6112 Unhealthy                      0              1           10
## 2749 Unhealthy                      0              0            6
## 736    Healthy                      0              0            5
## 4518 Unhealthy                      1              1            5
##      Sedentary.Hours.Per.Day Income      BMI Triglycerides
## 5184               0.2410653 144163 26.88929           490
## 4257              10.1811439 237268 38.60199           176
## 6112               7.2216589 287876 26.19829           485
## 2749               0.2959193 237453 27.96217           347
## 736                8.0047399 146158 33.07572            99
## 4518              10.7903673 195479 32.62334           243
##      Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day       Country
## 5184                               1                   5         India
## 4257                               3                   6      Colombia
## 6112                               6                   9        Brazil
## 2749                               0                   6 United States
## 736                                6                  10   South Korea
## 4518                               2                   7       Germany
```

```
##          Continent          Hemisphere Heart.Attack.Risk
## 5184          Asia Northern Hemisphere                 1
## 4257 South America Northern Hemisphere                 1
## 6112 South America Southern Hemisphere                 1
## 2749 North America Northern Hemisphere                 1
## 736           Asia Northern Hemisphere                 1
## 4518        Europe Northern Hemisphere                 1
```

Continuous Population

```
#Calculate mean BMI from SRS

SRS_BMI_mean <- mean(SRS_sample$BMI)

#Calculate mean BMI from male sample and female sample

male_BMI_mean <- mean(male_sample$BMI)
female_BMI_mean <- mean(female_sample$BMI)
#Calculate stratified estimator for BMI mean (sum of weighted BMI means)

strata_estimator_BMI_mean <- (male_stratum_size/pop_size)*male_BMI_mean +
                                (female_stratum_size/pop_size)*female_BMI_mean


data.frame(`Sampling Method` = c("SRS","Stratified Estimate"),
           `BMI Mean` = c(SRS_BMI_mean,strata_estimator_BMI_mean))
```

Estimate Mean

```
##       Sampling.Method BMI.Mean
## 1                 SRS 28.29644
## 2 Stratified Estimate 28.63382
```

```
#Calculate SE for SRS and Stratified

#SRS SE calculation
SRS_variance <- sum((SRS_sample$BMI - SRS_BMI_mean)^2)/(n-1)
SRS_FPC <- (1- n/pop_size)
SRS_SE <- sqrt(SRS_FPC * SRS_variance/n)

#Stratified SE calculation

#First calculate male and female strata variances
#and the strata FPC and proportions relative to population size squared
male_strata_variance <- sum((male_sample$BMI - male_BMI_mean)^2)/(male_sample_size-1)
male_strata_FPC <- (1 - male_sample_size/male_stratum_size)
male_proportion_squared <- (male_stratum_size/pop_size)^2
```

```r
female_strata_variance <-
  sum((female_sample$BMI - female_BMI_mean)^2)/(female_sample_size-1)
female_strata_FPC <- (1 - female_sample_size/female_stratum_size)
female_proportion_squared <- (female_stratum_size/pop_size)^2

# SE = sqrt(sum ((N_h/N)^2 * Strata_H_FPC * Strata Variance / strata sample size))
stratified_SE <- sqrt(
  (male_proportion_squared*male_strata_FPC*male_strata_variance/male_sample_size)+
(female_proportion_squared*female_strata_FPC*female_strata_variance/female_sample_size))

data.frame(`Sampling Method` = c("SRS","Stratification"),
           `Continuous SE` = c(SRS_SE,stratified_SE))
```

Calculate Standard Error

```
##   Sampling.Method Continuous.SE
## 1             SRS     0.3238444
## 2  Stratification     0.3397826
```

```r
# Construct 95% CI for mean BMI for SRS
SRS_cont_moe <- 1.96*SRS_SE
SRS_cont_ci <- c(SRS_BMI_mean - SRS_cont_moe,
                 SRS_BMI_mean + SRS_cont_moe)


# Construct 95% CI for mean BMI for Stratified
stratified_cont_moe <- 1.96*stratified_SE
stratified_cont_ci <- c(strata_estimator_BMI_mean - stratified_cont_moe,
                        strata_estimator_BMI_mean + stratified_cont_moe)

data.frame(`Sampling Method` = c("SRS","Stratification"),
           `CI Lower Bound` = c(SRS_cont_ci[1], stratified_cont_ci[1]),
           `CI Upper Bound` = c(SRS_cont_ci[2], stratified_cont_ci[2]))
```

Construct 95% Confidence Interval

```
##   Sampling.Method CI.Lower.Bound CI.Upper.Bound
## 1             SRS       27.66170       28.93117
## 2  Stratification       27.96784       29.29979
```

Binary Population

```r
#We use the previous samples

#SRS
#Find number of observations where BMI > 30 from SRS sample
num_obs_BMI_over_30 <- nrow(SRS_sample[SRS_sample$BMI > 30,])
```

```r
#Find estimated proportion of BMI over 30 by dividing observed BMI > 30 by sample size
SRS_proportion_obs_BMI_over_30 <- num_obs_BMI_over_30/n

#STRATIFIED
#male estimated proportion of BMI over 30
male_num_obs_BMI_over_30 <- nrow(male_sample[male_sample$BMI > 30,])
male_proportion_BMI_over_30 <- male_num_obs_BMI_over_30/male_sample_size

#female estimated proportion of BMI over 30
female_num_obs_BMI_over_30 <- nrow(female_sample[female_sample$BMI > 30,])
female_proportion_BMI_over_30 <- female_num_obs_BMI_over_30/female_sample_size

#Sum weighted stratified proportions to get overall stratified proportion estimate
stratified_overall_proportion <-
  (male_stratum_size/pop_size)*male_proportion_BMI_over_30 +
  (female_stratum_size/pop_size)*female_proportion_BMI_over_30

data.frame(`Sampling Method` = c("SRS","Stratification"),
           `Proportion of BMI Greater Than 30 Estimate` =
             c(SRS_proportion_obs_BMI_over_30,stratified_overall_proportion))
```

Estimate Proportion

```
##   Sampling.Method Proportion.of.BMI.Greater.Than.30.Estimate
## 1             SRS                                 0.3906706
## 2  Stratification                                 0.4496463
```

```r
#SRS

#variance = sqrt[p(1-p)/n]
SRS_proportion_SE <-
  sqrt(SRS_proportion_obs_BMI_over_30*(1-SRS_proportion_obs_BMI_over_30)/n)


# square root(sum(StratumProportion^2 * stratumFPC * variance/stratum_sample_size))

#Male proportions Variance
male_proportion_BMI_over_30_variance <-
  male_proportion_BMI_over_30 * (1 - male_proportion_BMI_over_30)
#Female proportions Variance
female_proportion_BMI_over_30_variance <-
  female_proportion_BMI_over_30 * (1 - female_proportion_BMI_over_30)

#   FPC used is same as the one used from calculated continuous SE:
#   male_strata_FPC, female_strata_FPC

#   Male and Female stratum proportions squared
#   is same as one used to calculate continuous SE:
#   male_proportion_squared, female_proportion_squared
```

```r
stratified_proportion_SE <-
  sqrt( (male_proportion_squared * male_strata_FPC *
  male_proportion_BMI_over_30_variance/male_sample_size) +

  (female_proportion_squared * female_strata_FPC *
  female_proportion_BMI_over_30_variance/female_sample_size) )

data.frame(`Sampling Method` = c("SRS","Stratification"),
           `Proportion of BMI greater than 30 SE` =
             c(SRS_proportion_SE,stratified_proportion_SE))
```

Calculate Standard Error

```
##   Sampling.Method Proportion.of.BMI.greater.than.30.SE
## 1             SRS                           0.02634416
## 2  Stratification                           0.02534392
```

```r
# Construct 95% CI for proportion of observations with BMI > 30 for SRS
SRS_binary_moe <- 1.96*SRS_proportion_SE
SRS_binary_ci <- c(SRS_proportion_obs_BMI_over_30 - SRS_binary_moe,
                   SRS_proportion_obs_BMI_over_30 + SRS_binary_moe)


# Construct 95% CI for proportion of observations with BMI > 30 for Stratified
stratified_binary_moe <- 1.96*stratified_proportion_SE
stratified_binary_ci <- c(stratified_overall_proportion - stratified_binary_moe,
                          stratified_overall_proportion + stratified_binary_moe)

data.frame(`Sampling Method` = c("SRS","Stratification"),
           `CI Lower Bound` = c(SRS_binary_ci[1], stratified_binary_ci[1]),
           `CI Upper Bound` = c(SRS_binary_ci[2], stratified_binary_ci[2]))
```

Construct 95% confidence interval

```
##   Sampling.Method CI.Lower.Bound CI.Upper.Bound
## 1             SRS      0.3390360      0.4423051
## 2  Stratification      0.3999722      0.4993204
```