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,]

#Remove unnecessary columns
#We keep Sex,diet,diabetes,family history heart problems,obesity, BMI
heartattack <-
  heartattack[,(colnames(heartattack)
                %in% c("Sex","Diabetes","Family.History","Obesity","Diet","BMI"))]
head(heartattack)
```

```
##       Sex Diabetes Family.History Obesity      Diet      BMI
## 6  Female        1              1       0 Unhealthy 20.14684
## 7    Male        0              0       0   Healthy 28.88581
## 8    Male        0              0       1   Average 22.22186
## 13   Male        1              1       1 Unhealthy 35.10224
## 14   Male        1              1       0   Healthy 25.56490
## 16   Male        1              1       0   Average 36.52440
```

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

```r
#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
```

```r
#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
```

```r
#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
```

```r
#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
```

```r
#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
```

```r
#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)

#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,])
```

```r
#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
```

```r
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

```r
# set seed
set.seed(2)

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

```
##         Sex Diabetes Family.History Obesity      Diet      BMI
## 2772   Male        1              0       1   Healthy 29.65312
## 2043 Female        1              0       1   Average 36.52504
## 7828   Male        0              0       0 Unhealthy 21.60942
## 1224   Male        1              1       0   Healthy 22.68139
## 1152 Female        1              1       1   Healthy 24.21819
## 831    Male        1              0       1 Unhealthy 26.88142
```

```
#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)
```

```
##       Sex Diabetes Family.History Obesity      Diet      BMI
## 2621 Male        1              1       0 Unhealthy 36.16253
## 1338 Male        1              1       1   Average 21.84712
## 3776 Male        1              1       0 Unhealthy 28.15095
## 6685 Male        1              0       1   Average 37.04400
## 4694 Male        0              1       1 Unhealthy 36.39712
## 3791 Male        0              0       0   Healthy 39.47205
```

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

```
##         Sex Diabetes Family.History Obesity      Diet      BMI
## 462  Female        1              1       0   Healthy 36.98066
## 3659 Female        1              1       1 Unhealthy 25.21583
## 3933 Female        1              0       0 Unhealthy 23.69793
## 2407 Female        1              1       0   Average 24.88832
## 316  Female        1              0       0   Healthy 22.86218
## 3004 Female        1              0       1   Healthy 21.31734
```

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 29.09488
## 2 Stratified Estimate 29.13750
```

```
#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

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     0.3240470
## 2  Stratification     0.3136828
```

```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       28.45975       29.73001
## 2  Stratification       28.52268       29.75232
```

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,])

#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                                  0.4577259
## 2  Stratification                                  0.4566627
```

```
#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

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                          0.02690080
## 2  Stratification                          0.02533275
```

```
# 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)
```

```r
# 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      0.4050004      0.5104515
## 2  Stratification      0.4070105      0.5063149
```