

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 Z007941 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( ((2*z)^2*(0.5)*(0.5)) / (MOE^2)) # 1537
# since we know N = 8763, using FPC
n = ceiling( n_0 / (1 + (n_0/pop_size)) ) # 1032

#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
```

```
##      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 <- 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

average_sample_size <- round(average_size_proportion*n)
healthy_sample_size <- round(healthy_size_proportion*n)
unhealthy_sample_size <- round(unhealthy_size_proportion*n)
```

Assuming the worst case proportions 0.5, the sample size used if we ignored FPC is 1537. Whereas including FPC the sample size used in SRS will be 1032. In the three strata: Diet =: (Average,Healthy,Unhealthy), sample size for average is 341, sample size for Healthy is 355, sample size for Unhealthy is 336

Selecting Samples through SRS and Stratification from diet

```
# set seed
set.seed(2023)

# take SRS of n = 1308
SRS.index <- sample.int(pop_size, n, replace=F)
heartsample <- heartattack[SRS.index, ]
head(heartsample)
```

##	Patient.ID	Age	Sex	Cholesterol	Blood.Pressure	Heart.Rate	Diabetes
##	5342	RQF3517	66 Female	169	134/107	66	1
##	4153	PDP7568	36 Male	362	168/103	106	1
##	6867	IGX5007	47 Male	204	179/102	49	1
##	3892	WH04445	32 Male	329	171/88	91	1
##	5579	LQJ4049	76 Female	289	103/86	93	0
##	2448	MXU7515	72 Male	197	178/60	50	1
##	Family.History	Smoking	Obesity	Alcohol.Consumption	Exercise.Hours.Per.Week		
##	5342	0	1	0	1		4.1293715
##	4153	0	1	0	1		15.8852288
##	6867	1	1	0	1		12.3257250
##	3892	1	1	1	1		15.8284110
##	5579	1	1	1	0		5.1937069
##	2448	0	1	0	0		0.2085372
##	Diet	Previous.Heart.Problems	Medication.Use	Stress.Level			
##	5342	Unhealthy	1	1		1	
##	4153	Unhealthy	0	0		4	
##	6867	Unhealthy	0	1		5	
##	3892	Healthy	1	0		1	
##	5579	Average	0	1		9	
##	2448	Average	0	0		1	
##	Sedentary.Hours.Per.Day	Income	BMI	Triglycerides			
##	5342	7.243322	238240	21.07242		568	
##	4153	10.701283	79281	19.72057		281	
##	6867	11.100653	24184	30.13575		540	
##	3892	7.533750	143838	36.47466		366	
##	5579	1.919237	222725	38.46187		506	
##	2448	2.174866	210200	28.04375		607	
##	Physical.Activity.Days.Per.Week	Sleep.Hours.Per.Day					Country
##	5342	3		10			Argentina
##	4153	5		10			Germany
##	6867	3		10			Argentina
##	3892	2		7			Argentina
##	5579	7		5			United Kingdom
##	2448	4		7			Spain
##	Continent	Hemisphere	Heart.Attack.Risk				
##	5342	South America	Southern Hemisphere			1	
##	4153	Europe	Northern Hemisphere			1	
##	6867	South America	Southern Hemisphere			1	
##	3892	South America	Southern Hemisphere			1	
##	5579	Europe	Northern Hemisphere			1	
##	2448	Europe	Southern Hemisphere			1	

The

Calculating Estimates

```
#Calculate mean BMI from SRS
```