STA304 A2

Q1

a) Choose one of the survey questions and identify one parameter of interest.

Survey Question: Do you support or oppose defunding the police and redirecting substantial portions of the police budget to other government services?

Parameter of Interest: Proportion of the voting population that is not sure about the defunding and redirecting proposal.

b) Based on the relevant cross tabulation table below your selected survey, choose one stratification variable and show the following:

Stratification variable: Gender

i.Weighted frequency:

```
## Estimate of the population parameter:
## (498×0.078 + 516×0.099)/1014 = 0.089
## (0.0886 rounded to 3 decimal place, N_Male = 498, N_Female = 516)
##
## variance_hat:
## ((498/1014)(0.078(1-0.078)/(577-1)) + (516/1014)(0.099(1-0.099)/(437-1)))/1014 +
## ((516/1014)(0.078(1-0.078)/(577-1)) + (498/1014)(0.099(1-0.099)/(436)))/1014^2
## = 1.63302693×10^-7
##
## A bound on the error of estimation:
## 2×sqrt(variance_hat) = 2×sqrt(1.63302693×10^-7) = 8.08×10^-4
##
## We ignored the FPC since the calculated value is close to 1
```

ii.Unweighted frequency:

```
## Estimate of the population parameter:
## (577×0.078 + 437×0.099)/1014 = 0.087
## (0.0871 rounded to 3 decimal place, N_Male = 577, N_Female = 437)
##
## variance_hat:
## (498/1014)^2(0.078(1-0.078)/(577-1)) + (516/1014)^2(0.099(1-0.099)/437-1)
## = 8.3093379×10^-5
##
## A bound on the error of estimation:
## 2×sqrt(variance_hat) = 2×sqrt(8.3093379×10^-5) = 0.018
##
## We ignored the FPC since the calculated value is close to 1
```

c) Compare the two estimates in part (b) above. Explain which is a post-stratified estimate.

The estimate from the weighted frequency (i.) is post-stratified since the weights are adjusted so that they match the known ratio between male and female in the Ontario population. The estimate from the unweighted frequency (ii.) is not balanced according to the Ontario population since it only represents the ratios of the observed sample, hence it is not a post-stratified estimate.

Q2

a) Take a stratified random sample of 150 players, using proportional allocation with the different teams as strata (teams are in column 1 of the data file). Describe how you selectedthe sample. Show your R codes used to obtain your stratified sample.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(sampling)

baseball_data=read.csv("baseball.csv")
N=dim(baseball_data)[1]; N
```

```
## [1] 797
```

```r
n=150
set.seed(2440)

round(n*(table(baseball_data$team)/N))
```

```
##
## ANA ARI ATL BAL BOS CHA CHN CIN CLE COL DET FLO HOU KCA LAN MIL MIN MON NYA NYN
##   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5
## OAK PHI PIT SDN SEA SFN SLN TBA TEX TOR
##   5   5   5   5   5   5   5   5   5   5
```

```r
st_sample=strata(baseball_data,stratanames="team",size=rep(5,30), method="srswor")
stratified_sample<-getdata(baseball_data,st_sample)
stratified_sample<-stratified_sample %>%
  group_by(team) %>%
  mutate(s2_i=var(log(salary))) %>%
  mutate(pitcher_proportion_per_team=sum(position=="P")/5)
```

We first use the table function to find how many observations there are in each team, then we use those numbers to get their proportions to the N (797), We multiply those proportions with n (150) and round the results to get the stratified sample population of each team by proportional allocation (all are 5). Lastly, we use strata() and getdata() to draw our stratified sample.

b) Find the mean of the variable logsal= ln(salary), using your stratified sample, and give a 95% CI.

```
#sample means
mu_hat=mean(log(stratified_sample$salary)); mu_hat
```

```
## [1] 13.85341
```

```
#construct a table with N_i,n_i and sample variance for each team
table_b<-data.frame(table(baseball_data$team))
table_b<-table_b %>%
  mutate(n_i=5) %>%
  rename(N_i=Freq) %>%
  rename(team=Var1)

table_b<-table_b %>%
  mutate(s2_i=unique(stratified_sample$s2_i)) %>%
  mutate(var_mu_hat_hat_team=((N_i^2/N^2)*(1-(n_i/N_i))*(s2_i/n_i)))

glimpse(table_b)
```

```
## Rows: 30
## Columns: 5
## $ team                <fct> ANA, ARI, ATL, BAL, BOS, CHA, CHN, CIN, CLE, CO...
## $ N_i                 <int> 26, 28, 28, 25, 27, 26, 29, 27, 28, 27, 26, 26,...
## $ n_i                 <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,...
## $ s2_i                <dbl> 2.9084565, 0.8763391, 2.0213903, 0.8464057, 0.5...
## $ var_mu_hat_hat_team <dbl> 4.999984e-04, 1.776934e-04, 4.098731e-04, 1.332...
```

```
#computing a 95% CI using the sample
var_mu_hat_hat<-sum(table_b$var_mu_hat_hat_team)
stderror=sqrt(var_mu_hat_hat); stderror
```

```
## [1] 0.08876673
```

```
myCI=c(mu_hat-2*stderror,mu_hat+2*stderror)
myCI
```

```
## [1] 13.67588 14.03094
```

For simplicity, we round all the reported outputs (estimates and CIs) in this assignment to two decimal places. The mean of the variable logsal from the stratified sample is 13.85 and (13.68, 14.03) is the 95% CI.

c) Estimate the proportion of players in the data set who are pitchers, using your stratified sample, and give a 95% CI.

```
#sample proportion
sample_pitcher_count=sum(stratified_sample$position=="P")
sample_pitcher_proportion=sample_pitcher_count/n;
sample_pitcher_proportion
```

```
## [1] 0.4866667
```

```
#construct a table with p_i_hat,q_i_hat and sample variance for each team
table_c<-table_b %>%
  mutate(p_i_hat=stratified_sample[seq(1, n, 5), 35]) %>%
  mutate(q_i_hat=1-stratified_sample[seq(1, n, 5), 35]) %>%
  mutate(var_p_hat_hat_team=((N_i^2/N^2)*(1-(n_i/N_i))*(p_i_hat*q_i_hat)/(n_i-1)))

#computing a 95% CI using the sample
var_p_hat_hat_c<-sum(table_c$var_p_hat_hat_team)
stderror_c=sqrt(var_p_hat_hat_c); stderror_c
```

```
## [1] 0.03908482
```

```
myCI_c=c(sample_pitcher_proportion-2*stderror_c,
         sample_pitcher_proportion+2*stderror_c)
myCI_c
```

```
## [1] 0.4084970 0.5648363
```

The estimated proportion is 0.49 and (0.41, 0.56) is the 95% CI.

d) Take a simple random sample of 150 players and repeat part (c). How does your estimate compare with that of part (c).

```
set.seed(2440)
srs<-baseball_data[sample(1:nrow(baseball_data), n),]
glimpse(srs)
```

```
## Rows: 150
## Columns: 30
## $ team         <chr> "LAN", "FLO", "MON", "MIL", "PHI", "NYN", "MIL", "SDN"...
## $ leagueID     <chr> "NL", "NL", "NL", "NL", "NL", "NL", "NL", "NL", "NL", ...
## $ player       <chr> "roberda0", "castrra0", "ohkato01", "grievbe0", "worre...
## $ salary       <int> 975000, 400000, 2337500, 700000, 2750000, 800000, 3760...
## $ position     <chr> "LF", "C", "P", "RF", "P", "RF", "RF", "1B", "P", "P",...
## $ g_played     <int> 68, 32, 14, 108, 73, 62, 138, 147, 32, 5, 23, 49, 18, ...
## $ g_started    <int> 45, 27, 15, 64, 0, 44, 77, 142, 0, 0, 0, 1, 2, 105, 42...
## $ InnOuts      <int> 1136, 729, 254, 1411, 235, 1159, 2354, 3622, 104, 204,...
## $ put_out      <int> 77, 192, 4, 106, 10, 91, 219, 1132, 3, 1, 1, 7, 2, 264...
## $ assists      <int> 0, 12, 11, 0, 15, 0, 4, 91, 4, 10, 2, 4, 2, 304, 4, 9,...
## $ errors       <int> 2, 2, 1, 4, 1, 3, 4, 13, 0, 0, 0, 0, 0, 3, 1, 0, 0, 4,...
## $ double_plays <int> 0, 2, 3, 0, 2, 0, 2, 108, 0, 0, 0, 0, 2, 76, 1, 0, 0, ...
## $ pass         <chr> ".", "0", ".", ".", ".", ".", ".", ".", ".", ".", ".",...
## $ GB           <int> 68, 32, 14, 108, 73, 62, 138, 147, 32, 5, 23, 49, 18, ...
## $ AB           <int> 233, 96, 25, 234, 0, 192, 353, 547, 1, 0, 0, 3, 38, 50...
## $ R            <int> 45, 9, 0, 28, 0, 24, 41, 78, 0, 0, 0, 0, 4, 74, 37, 0,...
## $ H            <int> 59, 13, 2, 61, 0, 45, 99, 158, 0, 0, 0, 0, 4, 150, 80,...
## $ SecB         <int> 4, 3, 0, 15, 0, 7, 18, 31, 0, 0, 0, 0, 0, 21, 18, 0, 0...
## $ ThiB         <int> 7, 0, 0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...
## $ HR           <int> 2, 3, 0, 7, 0, 7, 7, 26, 0, 0, 0, 0, 1, 17, 10, 0, 0, ...
## $ RBI          <int> 21, 8, 0, 29, 0, 22, 46, 105, 0, 0, 0, 0, 4, 55, 40, 0...
## $ SB           <int> 33, 0, 1, 0, 0, 3, 15, 0, 0, 0, 0, 0, 0, 7, 4, 0, 0, 0...
## $ CS           <int> 1, 0, 0, 0, 0, 0, 8, 0, 0, 0, 0, 0, 0, 4, 1, 0, 0, 0, ...
## $ BB           <int> 28, 11, 1, 39, 0, 10, 53, 66, 0, 0, 0, 0, 4, 27, 23, 0...
## $ SO           <int> 31, 30, 8, 65, 0, 35, 48, 121, 0, 0, 0, 2, 9, 39, 50, ...
## $ IBB          <chr> "0", "2", "0", "5", "0", "0", "2", "5", "0", "0", "0",...
## $ HBP          <chr> "4", "1", "0", "0", "0", "0", "9", "5", "0", "0", "0",...
## $ SH           <int> 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 7, 2, 3, 0, 0, ...
## $ SF           <chr> "3", "0", "0", "2", "0", "0", "3", "5", "0", "0", "0",...
## $ GIDP         <int> 2, 1, 0, 4, 0, 6, 9, 16, 0, 0, 0, 0, 3, 13, 4, 0, 0, 7...
```

```
#population and sample proportion
srs_sample_pitcher_count=sum(srs$position=="P")
srs_pitcher_proportion=srs_sample_pitcher_count/n;
srs_pitcher_proportion
```

```
## [1] 0.4666667
```

```
population_pitcher_count=sum(baseball_data$position=="P")
population_pitcher_proportion=population_pitcher_count/N;
population_pitcher_proportion
```

```
## [1] 0.4717691
```

```
#computing a 95% CI using the sample
var_mu_hat_hat_d<-(1-(n/N))*(srs_pitcher_proportion*(1-srs_pitcher_proportion))/(n-1)
stderror_d=sqrt(var_mu_hat_hat_d); stderror_d
```

```
## [1] 0.03682414
```

```
myCI_d=c(srs_pitcher_proportion-2*stderror_d,
         srs_pitcher_proportion+2*stderror_d)
myCI_d
```

```
## [1] 0.3930184 0.5403149
```

The estimated proportion is 0.47 and (0.39, 0.54) is the 95% CI. When compare with the estimate from part (c), we see that the estimate produced from the simple random sample is smaller than that of the stratified sample.

e) Examine the sample variances of logsal in each stratum. Do you think optimal allocation would be worthwhile for this problem?

```
table_b
```

```
##    team N_i n_i      s2_i var_mu_hat_hat_team
## 1   ANA  26   5 2.9084565        4.999984e-04
## 2   ARI  28   5 0.8763391        1.776934e-04
## 3   ATL  28   5 2.0213903        4.098731e-04
## 4   BAL  25   5 0.8464057        1.332484e-04
## 5   BOS  27   5 0.5518074        1.032018e-04
## 6   CHA  26   5 1.2580753        2.162781e-04
## 7   CHN  29   5 1.2391743        2.715532e-04
## 8   CIN  27   5 1.8426826        3.446278e-04
## 9   CLE  28   5 1.8236471        3.697771e-04
## 10  COL  27   5 3.5700417        6.676873e-04
## 11  DET  26   5 0.1678319        2.885231e-05
## 12  FLO  26   5 1.4927463        2.566209e-04
## 13  HOU  25   5 0.8075251        1.271275e-04
## 14  KCA  27   5 2.1302675        3.984134e-04
## 15  LAN  24   5 3.1315177        4.496070e-04
## 16  MIL  25   5 0.7315926        1.151735e-04
## 17  MIN  25   5 1.6782195        2.641996e-04
## 18  MON  28   5 1.4347242        2.909160e-04
## 19  NYA  29   5 1.0501829        2.301376e-04
## 20  NYN  26   5 1.7279872        2.970616e-04
## 21  OAK  27   5 1.6174551        3.025046e-04
## 22  PHI  25   5 1.8005139        2.834522e-04
## 23  PIT  27   5 0.8050032        1.505558e-04
## 24  SDN  26   5 1.6777214        2.884203e-04
## 25  SEA  27   5 1.8712320        3.499673e-04
## 26  SFN  28   5 0.1673112        3.392533e-05
## 27  SLN  26   5 2.0247698        3.480821e-04
## 28  TBA  26   5 0.8978644        1.543536e-04
## 29  TEX  27   5 0.5828015        1.089985e-04
## 30  TOR  26   5 1.2054095        2.072243e-04
```

The s2_i column from table_b gives the sample variances of logsal in each stratum, we observe that the variances are unequal across strata (s2_i's are not all the same) and we can see from the N_i column that the stratum sizes are roughly equal, hence we can assume that the cost is about the same in each stratum. Therefore, a Neyman allocation would be worthwhile for this problem.

f) Using the sample variances from (e) to estimate the population stratum variances,determine the optimal allocation for a sample in which the cost is the same in each stratum and the total sample size is 150. How much does the optimal allocation differ from proportional allocation for this scenario?

```
table_b<-table_b %>%
  mutate(sigma_i=sqrt(s2_i)) %>%
  mutate(N_i_by_sigma_i=N_i*sigma_i) %>%
  mutate(proportional_allocation=N_i*(150/797)) %>%
  mutate(optimal_allocation=n*N_i_by_sigma_i/sum(N_i_by_sigma_i)) %>%
  mutate(rounded_proportional_allocation=round(proportional_allocation)) %>%
  mutate(rounded_optimal_allocation=round(optimal_allocation)) %>%
  mutate(population_stratum_variances=(N_i^2/N^2)*((N_i-n_i)/N_i)*(s2_i/n_i))

table_b_compare<-table_b %>%
  select(team,rounded_proportional_allocation,rounded_optimal_allocation,
         population_stratum_variances)
glimpse(table_b_compare)
```

```
## Rows: 30
## Columns: 4
## $ team                            <fct> ANA, ARI, ATL, BAL, BOS, CHA, CHN, ...
## $ rounded_proportional_allocation <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,...
## $ rounded_optimal_allocation      <dbl> 7, 4, 6, 4, 3, 5, 5, 6, 6, 8, 2, 5,...
## $ population_stratum_variances    <dbl> 4.999984e-04, 1.776934e-04, 4.09873...
```

```
mean(table_b$rounded_proportional_allocation)
```

```
## [1] 5
```

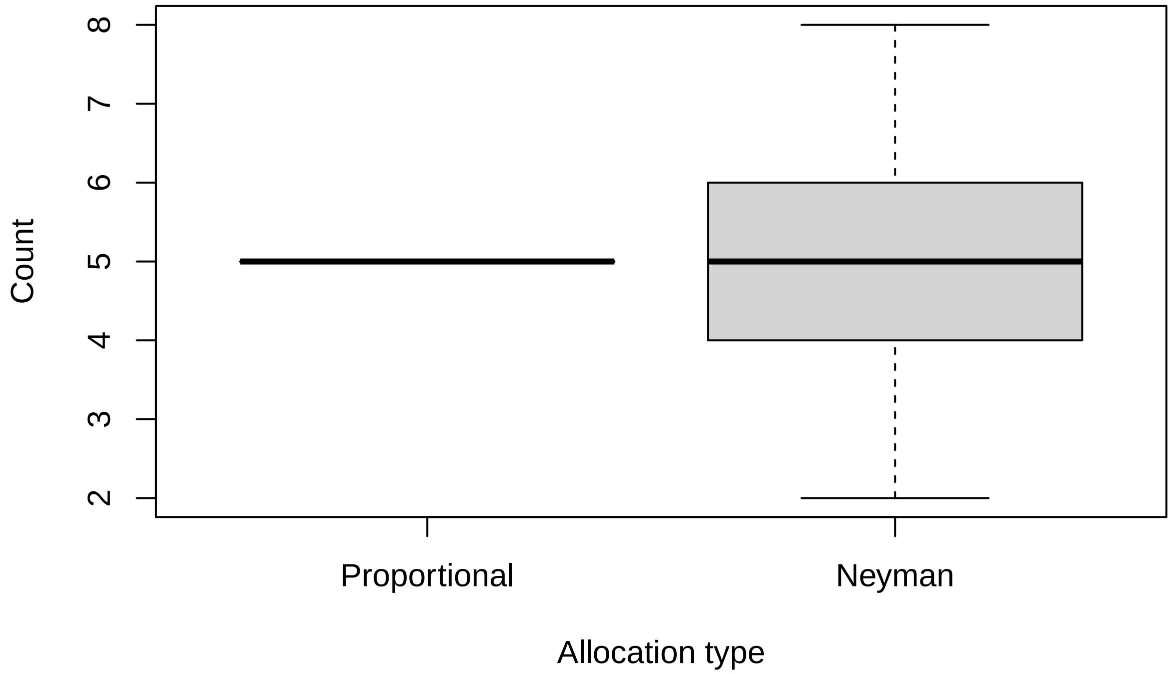
```
mean(table_b$rounded_optimal_allocation)
```

```
## [1] 4.933333
```

```
sqrt(var(table_b$rounded_optimal_allocation)) #standard deviation for the Neyman allocation
```

```
## [1] 1.436791
```

```
boxplot(table_b_compare$rounded_proportional_allocation,
        table_b_compare$rounded_optimal_allocation, ylab="Count",
        xlab="Allocation type", names=c("Proportional","Neyman"))
```



The estimates of the population stratum variances are shown by the table above. The optimal allocation for a sample in which the cost is the same in each stratum and the total sample size is 150 is the Neyman allocation. For the optimal allocation: n_i= = N_i*(150sigma_i/sum(N_i_by_sigma_i)) For proportional allocation: n_i= n(N_i/N) = N_i*(150/797)

From the boxplot and the table above, we can see that their median is about the same and the average of the proportional allocations is 5 while the Neyman allocations has an average 4.93 and a standard deviation of 1.44 Hence, we conclude that the optimal allocation does not differ by a lot from proportional allocation for this scenario.