Midterm

Xinyi Lin

11/4/2019

```
library(tidyverse)
library(survey)
```

```
load("./nhanes1999.RData")
```

Question 1

```
# calculate mean
nona_data = data.2[!is.na(data.2$bmxbmi),]
est_mean = sum(nona_data$wtmec2yr*nona_data$bmxbmi)/sum(nona_data$wtmec2yr)
# calculate variance
jk1_est = rep(NA, 52)
for (i in 5:56) {
  jk1_est[i-4] = sum(nona_data[,i]*nona_data$bmxbmi)/sum(nona_data[,i])
}
k = 52
delta_q = jk1_est-est_mean
var = (k-1)*sum(delta_q^2)/k
# calculate CI
t = qt(0.975, k-1)
lower_CI = est_mean-t*sqrt(var)
higher_CI = est_mean+t*sqrt(var)
```

The variable name for the sampling weights is `wtmec2yr` which means full sample 2 year MEC exam weight. The estimated population mean for BMI is 25.733 and the 95% confidence interval is (25.371, 26.094) if variances are computed based on sum of squares around the point estimate.

Question 2

```
mis_per = length(data.2$bmxbmi[is.na(data.2$bmxbmi)])/length(data.2$bmxbmi)
```

There are 8.834 percentage of cases in this data set has missing data on BMI. Assume in reality, proportions of male and female are the same, if a higher proportion of females reported BMIs than males and females had higher BMIs than males, the above estimate of mean BMI would be over-estimated.

Question 3

```
JK1design = svrepdesign(variables = data.2[,c(2,3,57)], repweights = data.2[,5:56],
                        weights = data.2$wtmec2yr, combined.weights = TRUE, type = "JK1", scale=1, rsca
```

For male, 1st, 5th, 10th, 25th, median, 75th, 90th, 95th and 99th percentiles of BMI are as following:

```
sub1 = subset(JK1design, riagendr == 1)
svyquantile(~bmxbmi, sub1, c(0.01, 0.05, 0.1, 0.25, 0.75, 0.90, 0.95, 0.99),ci=F, na.rm=T, se=F)[,1]
```

```
##    q0.01    q0.05     q0.1    q0.25    q0.75     q0.9    q0.95    q0.99
## 14.17432 15.51000 16.78143 20.97098 29.07000 33.71808 36.29579 44.20526
```

For female, 1st, 5th, 10th, 25th, median, 75th, 90th, 95th and 99th percentiles of BMI are as following:

```
sub2 = subset(JK1design, riagendr == 2)
svyquantile(~bmxbmi, sub2, c(0.01, 0.05, 0.1, 0.25, 0.75, 0.90, 0.95, 0.99),ci=F, na.rm=T, se=F)[,1]
```

```
##    q0.01    q0.05     q0.1    q0.25    q0.75     q0.9    q0.95    q0.99
## 13.93000 15.34909 16.92469 20.77000 30.27000 36.43227 39.67664 48.34889
```

Question 4

```
rak_data = data.2 %>%
  mutate(rak_age = ifelse(ridageyr<5, 1, 6),
         rak_age = ifelse(ridageyr>=5&ridageyr<=17, 2, rak_age),
         rak_age = ifelse(ridageyr>=18&ridageyr<=24, 3, rak_age),
         rak_age = ifelse(ridageyr>=25&ridageyr<=44, 4, rak_age),
         rak_age = ifelse(ridageyr>=45&ridageyr<=64, 5, rak_age)) %>%
  select(-ridageyr) %>%
  select(seqn, riagendr, rak_age, wtmec2yr, everything())
#design = svydesign(id = ~1, weights = ~wtmec2yr, data = rak_data)
JK1design2 = svrepdesign(variables = rak_data[,c(2,3,57)], repweights = rak_data[,5:56],
                         weights = rak_data$wtmec2yr, combined.weights = TRUE, type = "JK1")

pop.gender = data.frame(riagendr = c(1,2), Freq = c(138053563, 143368343))
pop.age = data.frame(rak_age = 1:6, Freq = c(19175798, 53118014, 27143454, 85040251, 61952636, 34991753
rdesign = rake(JK1design2, list(~riagendr, ~rak_age), list(pop.gender, pop.age))

svytable(~riagendr, rdesign, round=TRUE)
```

```
## riagendr
##         1         2
## 138053563 143368343
```

```
svytable(~rak_age, rdesign, round=TRUE)
```

```
## rak_age
##        1        2        3        4        5        6
## 19175798 53118014 27143454 85040251 61952636 34991753
```

Question 5

```
BMI_rak = svymean(~bmxbmi, rdesign, na.rm = TRUE)
BMI_rak
```

```
##          mean     SE
## bmxbmi 25.84 0.1813
```

```
df=degf(rdesign)
t = qt(0.975, df)
lower_CI = est_mean-t*0.1813^2
higher_CI = est_mean+t*0.1813^2
CI = c(lower_CI, higher_CI)
CI
```

```
## [1] 25.66662 25.79859
```

The estimated population mean for BMI using new weights is 25.84 and 95% confidence interval is (25.667, 25.799).

Comparing to the estimated mean and 95% confidence interval I get before raking, we can find that after raking, the estimated mean of BMI is slightly larger and the 95% confidence interval is slightly narrower.

This means compared to population, this sample contains higher proportions of observations with lower BMI and led to slightly lower estimated mean before raking. Raking helps to adjust this unbalance and also increase accuracy, led to a narrower confidence interval.

The difference between estimated means before and after raking is small, which means the age group and gender proportions in this sample mimic proportions in reality.