

Homework 4

Ishita Dutta

6/02/2022

1

a)

```
N <- 828
n <- 85
M <- 215
number_errors <- c(0,1,2,3,4)
frequency <- c(57,22,4,1,1)
t_sum <- sum(number_errors * frequency)
t <- t_sum*(N/n)

s2 <- (1/(n-1))*sum(frequency*((number_errors - (t_sum/n))^2))

y_hat <- t/(N*M)

SE <- (sqrt((1-(n/N))*(s2/n)))/M
cat('error rate : ', y_hat, '\n', 'standard error : ', SE)
```

```
## error rate : 0.002024624
## standard error : 0.0003570679
```

b)

```
total_errors <- N*sqrt((1-(n/N))*(s2/n))
cat("Total number of errors in the 828 claims : ", total_errors)
```

```
## Total number of errors in the 828 claims : 63.56523
```

c)

```
n <- 18275
N <- 178020

se_p <- sqrt((1-(n/N))*((1-y_hat)*y_hat)/n)
cat(se_p)
```

```
## 0.0003149797
```

We can see from the above result that if we just took SRS sampling, the variance would have been smaller.

2

a)

```
N <- 29
n <- 4
total_students = c(1471, 890, 1021, 1587) # Mi
female_stu = c(792,447,511,800) #mi

female_students_interviewed = c(25,15,20,40)
smokers = c(10,3,6,27)

wij = (N * total_students) / (n * female_stu)

#sum(smokers)/sum(female_students_interviewed)

#sum(smokers) / sum(female_stu)

thatclu = sum(wij * (smokers / female_students_interviewed) * total_students)* (N/n)
thatclu / (sum(wij * total_students) * (N/n))
```

```
## [1] 0.4317106
```

b)

```
thatclu
```

```
## [1] 219937.2
```

c)

```
## This one I don't know ...
```

3

a)

```

measles <- read.csv("measles.csv")

school_avg <- c()
for (i in 1:10) {
  school_avg[i] <- mean(measles[measles$school==i & measles$returnf!=9,]$returnf)
}

school <- c(1:10)
data.frame(school, school_avg)

```

```

##      school school_avg
## 1         1  0.5000000
## 2         2  0.5277778
## 3         3  0.7647059
## 4         4  0.6000000
## 5         5  0.4615385
## 6         6  0.5416667
## 7         7  0.6818182
## 8         8  0.5833333
## 9         9  0.6571429
## 10        10  0.4117647

```

b)

```

N<-46
n<-10

W <-c()

M <- unique(measles$Mitotal)

for (i in 1:10) {
  #M <- nrow(measles[measles$school==i & measles$previmm==0 ,])
  #m <- nrow(measles[measles$school==i,])
  a <-length(measles[measles$school==i,]$mi)
  m <- sum(measles[measles$school==i,]$mi)/a
  Mi <- M[i]
  cat(Mi,' : ',m)
  cat('\n')

  w <- (Mi*N)/(m*n)
  W[i] <- w
}

```

```

## 78 : 40
## 238 : 38
## 261 : 19
## 174 : 30
## 236 : 30

```

```
## 188 : 25
## 113 : 23
## 170 : 43
## 296 : 38
## 207 : 21
```

```
W
```

```
## [1] 8.97000 28.81053 63.18947 26.68000 36.18667 34.59200 22.60000 18.18605
## [9] 35.83158 45.34286
```

```
measles[measles$school==i,]
```

```
##      school form returnf consent hadmeas previmm sideeff gp noshot notser gpadv
## 287      10    1        1        0        1        1        0  0 0        0        0  0
## 288      10    1        1        0        0        0        1  0 0        0        0  0
## 289      10    1        1        0        1        0        0  0 0        0        0  0
## 290      10    1        0        0        1        0        0  0 0        0        0  1
## 291      10    1        0        0        1        0        0  0 0        0        0  0
## 292      10    1        0        0        0        1        0  0 0        0        0  1
## 293      10    0        9        9        0        0        0  0 0        0        0  0
## 294      10    1        0        0        0        0        0  1 0        0        0  0
## 295      10    1        0        0        1        0        0  0 0        0        0  0
## 296      10    1        1        0        1        0        0  0 0        0        0  0
## 297      10    1        0        0        1        0        0  0 0        0        0  0
## 298      10    0        9        9        0        0        0  0 0        0        0  0
## 299      10    1        0        0        1        0        1  0 0        0        0  0
## 300      10    1        1        1        0        0        0  0 0        0        0  0
## 301      10    0        9        9        0        0        0  0 0        0        0  0
## 302      10    1        0        0        1        0        1  1 0        0        0  0
## 303      10    1        0        0        1        1        1  0 0        0        0  0
## 304      10    1        1        1        0        0        0  0 0        0        0  0
## 305      10    9        9        9        0        0        0  0 0        0        0  0
## 306      10    1        0        0        1        0        0  1 0        0        0  0
## 307      10    1        1        1        0        0        0  0 0        0        0  0
##      Mitotal mi
## 287      207 21
## 288      207 21
## 289      207 21
## 290      207 21
## 291      207 21
## 292      207 21
## 293      207 21
## 294      207 21
## 295      207 21
## 296      207 21
## 297      207 21
## 298      207 21
## 299      207 21
## 300      207 21
## 301      207 21
## 302      207 21
## 303      207 21
```

```
## 304      207 21
## 305      207 21
## 306      207 21
## 307      207 21
```

```
#measles[measles$school==i & measles$previmm==0,]
```

c)

```
# I'm not sure about this but the concept follows (2b)
```

d)

4

a)

```
states <- read.csv("statepps.csv")

landarea <- states$landarea
states$psi_landarea <- states$landarea / sum(states$landarea)
n <- 10
states_n <- 51

cumulative <- cumsum(landarea)

samples <- sample(1:max(cumsum(landarea)), n, replace = TRUE)

group <-c()
index <-c()
for (i in 1:10) {
  temp <- states_n - sum(samples[i]<cumulative) + 1
  index[i] <-temp
  group[i] <- states$state[temp]
}
group
```

```
## [1] "North Dakota" "North Carolina" "Maine" "Virginia"
## [5] "California" "Arizona" "Texas" "Texas"
## [9] "New York" "Virginia"
```

b)

```

pop <- states$pop2019
psi_pop <- states$pop2019 / sum(states$pop2019)
states["psi_pop"] <- psi_pop
n <- 10
states_n <- 51

cumulative <- cumsum(pop)

samples <- sample(1:max(cumsum(pop)), n, replace = TRUE)

group <- c()
index <- c()
for (i in 1:10) {
  temp <- states_n - sum(samples[i]<cumulative) + 1
  index[i] <- temp
  group[i] <- states$state[temp]
}
group

## [1] "New Mexico"      "Utah"              "California"        "Nevada"
## [5] "Tennessee"         "Maine"             "South Carolina"    "Michigan"
## [9] "South Carolina"    "New York"

```

c)

- We can see that there are repetitions in some samples but also can find out that the states with large states result in large populations.

5

a)

```

dat <- states[index,]

t_hat <- (1/n)*(sum(dat$counties/dat$psi_landarea))
sum(states$counties)

## [1] 3143

t_hat

## [1] 3467.937

se <- sqrt((1/n)*(1/(n-1))*sum(((dat$counties/dat$psi_landarea) - t_hat)^2))
se

## [1] 821.2637

```

b)