Homework 3

yi Chen (yc3356)

9/28/2018

Assignment 3

problem 1

(a)

In this probem, we know that: the standard deviation of three households are different. Given that there is no extra information about the sample cost. I choose to use the method of **Neyman Allocation**

Thus, the strategy is $n_h \propto N_h \times S_h$, where $N_1 : N_2 : N_3 = 35000 : 45000 : 10000 = 7 : 9 : 2$ and $S_1 : S_2 : S3 = 2 : 1 : 1.$

Thus, we have: $n_1 : n_2 : n_3 = 14 : 9 : 2.$

$$n_1 = \frac{14}{14+9+2} \times 900 = 504$$

$$n_2 = \frac{9}{14+9+2} \times 900 = 324$$

$$n_3 = \frac{2}{14+9+2} \times 900 = 72$$

(b)

$$V_{prop}(p_{str}) = \sum_{h=1}^{3} (\frac{N_h}{N})^2 V(p_h)$$

$$= \sum_{h=1}^{3} (\frac{N_h - n_h}{N_h - 1})(\frac{N_h}{N})^2 \frac{p_h(1-p_h)}{n_h}$$

```
N <- 90000
N_h <- c(35000,45000,10000)
n_h <- c(350,450,100)
p <- c(0.45,0.25,0.12)
v_prop_p_str_hat <- sum(((N_h-n_h)/(N_h-1)) * (N_h/N)^2 * (p*( 1 - p))/(n_h))
v_prop_p_str_hat
```

```
## [1] 0.0002219133
```

$$V_{SRS}(p_{SRS}) = (\frac{N-n}{N-1})\frac{p(1-p)}{n}$$

Where, we know the true p value:

$$p = 45\% \times \frac{7}{7+9+2} + 25\% \times \frac{9}{7+9+2} + 12\% \times \frac{2}{7+9+2} \approx 31.33\%$$

```
p <- 7/18 * 0.45 + 9/18 * 0.25 + 2/18 * 0.12
n <- sum(n_h)
v_srs_p_srs_hat <- ((N-n)/(N-1)) * ((p*(1-p)) /(n))
v_srs_p_srs_hat
```

```
## [1] 0.0002366737
```

```
v_srs_p_srs_hat / v_prop_p_str_hat
```

```
## [1] 1.066515
```

The coefficient of efficiency is 1.06515 which indecate that stratified sampling is about 6.6% more efficient than SRS

problem 2

```
referred_publications <- c(0,1,2,3,4,5,6,7,8,9,10)
faculty_members <- c(28,4,3,4,4,2,1,0,2,1,1)
data <- rep(referred_publications,faculty_members)
N <- 807
n <- 50
y_bar_s <- mean(data)
t_hat <- N * y_bar_s
s <- sd(data)
sd_t_hat <- N * (s / sqrt(n)) * sqrt(1-n/N)
uppder_bond <- t_hat + qnorm(0.975) * sd_t_hat
lower_bond <- t_hat + qnorm(0.025) * sd_t_hat
lower_bond;uppder_bond
```

```
## [1] 855.3229
```

```
## [1] 2017.597
```

Thus, the 95% confidence interval for total number of refereed publication is [855.3229,2017.579]

problem 3

```
## acres87 denote number of acres devoted to farms in 1987
library(SDaA)
data <- agsrs$acres87
N <- 3078
n <- 300

y_bar_s <- mean(data)
t_hat <- N * y_bar_s

s <- sd(data)
sd_t_hat <- N * (s / sqrt(n)) * sqrt(1-n/N)
sd_srs <- sd_t_hat

uppder_bond <- t_hat + qnorm(0.975) * sd_t_hat
lower_bond <- t_hat + qnorm(0.025) * sd_t_hat
lower_bond;uppder_bond
```

```
## [1] 815311779
```

```
## [1] 1043515342
```

Thus, the 95% confidence interval for total number of acres to farms in the U.S. in 1987 is [815311779,1043515342]

problem 4

```
NC_data <- agstrat$acres87[which(agstrat$region=='NC')]
NE_data <- agstrat$acres87[which(agstrat$region=='NE')]
S_data <- agstrat$acres87[which(agstrat$region=='S')]
W_data <- agstrat$acres87[which(agstrat$region=='W')]

N_h <- c(1054,220,1382,422)
n_h <- c(103,21,135,41)

NC_mean <- mean(NC_data)
NE_mean <- mean(NE_data)
S_mean <- mean(S_data)
W_mean <- mean(W_data)

NC_var <- var(NC_data)
NE_var <- var(NE_data)
S_var <- var(S_data)
W_var <- var(W_data)
var_h <- c(NC_var,NE_var,S_var,W_var)

NC_t_hat <- 1054 * NC_mean
NE_t_hat <- 220 * NE_mean
S_t_hat <- 1382 * S_mean
W_t_hat <- 422 * W_mean

t_hat <- NC_t_hat + NE_t_hat + S_t_hat + W_t_hat
sd_t_hat <- sqrt(sum((1 - n_h/N_h) * (N_h ^2) * (var_h/n_h)))
sd_strtified <- sd_t_hat

uppder_bond <- t_hat + qnorm(0.975) * sd_t_hat
lower_bond <- t_hat + qnorm(0.025) * sd_t_hat
lower_bond;uppder_bond
```

```
## [1] 820636226
```

```
## [1] 1017219719
```

Thus, the 95% confidence interval for total number of acres to farms in the U.S. in 1987 is [820636226,1017219719]

We can see that the stratified sampling gives a much smaller confidence interval than Sample radomg sampling.

```
(sd_srs / sd_strtified)^2
```

```
## [1] 1.347568
```

As we can see from the efficient coefficient, we can say that the stratified sampling is 34.75% more efficient than sample random sampling.

problem 5

Since the cost is assumed to be the same in all of the stratums, optimal allocation actually is the same as the neyman allocation. We need to allocation based on the principle: $n_h \propto N_h \times S_h$

```
allocation <- sqrt(var_h) * N_h
allocation
```

```
## [1] 180801275  22055638 329903316 259270507
```

Here we have $n_{NC} : n_{NE} : n_S : n_W = 180801275 : 22055638 : 329903316 : 271558209$

Thus:

```
for (i in 1:4){
  print(round(allocation[i]/sum(allocation)*300))
}
```

```
## [1] 68
## [1] 8
## [1] 125
## [1] 98
```

According to the Neyman (Optimal) allocation principle, we can allocate 68 sample to northce central, 9 sample (the sum of the result above is 299, the extra 1 can given to the stratum with smallest sample size and this will not make a big difference), 125 sample to south and 98 sample to west.

problem 6

```
set.seed(5234)

NC_data <- sample(agpop$acres87[which(agpop$region=='NC')], 68)
NE_data <- sample(agpop$acres87[which(agpop$region=='NE')],9)
S_data <- sample(agpop$acres87[which(agpop$region=='S')],125)
W_data <- sample(agpop$acres87[which(agpop$region=='W')],98)

NC_data <- NC_data[which(NC_data>0)]
NE_data <- NE_data[which(NE_data>0)]
S_data <- S_data[which(S_data>0)]
W_data <- W_data[which(W_data>0)]

N_h <- c(1054,220,1382,422)
n_h <- c(length(NC_data),length(NE_data),length(S_data),length(W_data))

NC_mean <- mean(NC_data)
NE_mean <- mean(NE_data)
S_mean <- mean(S_data)
W_mean <- mean(W_data)

NC_var <- var(NC_data)
NE_var <- var(NE_data)
S_var <- var(S_data)
W_var <- var(W_data)
var_h <- c(NC_var,NE_var,S_var,W_var)

NC_t_hat <- 1054 * NC_mean
NE_t_hat <- 220 * NE_mean
S_t_hat <- 1382 * S_mean
W_t_hat <- 422 * W_mean

t_hat <- NC_t_hat + NE_t_hat + S_t_hat + W_t_hat
sd_t_hat <- sqrt(sum((1 - n_h/N_h) * (N_h ^2) * (var_h/n_h)))
sd_strtified_new <- sd_t_hat

uppder_bond <- t_hat + qnorm(0.975) * sd_t_hat
lower_bond <- t_hat + qnorm(0.025) * sd_t_hat
lower_bond;uppder_bond
```

```
## [1] 873564995
```

```
## [1] 1066611737
```

Thus, the 95% confidence interval for total number of acres to farms in the U.S. in 1987 is [873564995,1066611737]

We can see that the stratified sampling gives a smaller confidence interval than Sample radomg sampling.

```
(sd_strtified / sd_strtified_new) ^2
```

```
## [1] 1.036977
```

As we can see from the efficient coefficient, we can say that the stratified sampling based on the neyman allocation is 3.69% more efficient than stratified sampling based on the proportation allocation.

And for the standard deviation itself, the difference between these two allocation methods is samlle.

problem 7

```
data_1 <- otters$holts[which(otters$habitat == 1 )]
data_2 <- otters$holts[which(otters$habitat == 2 )]
data_3 <- otters$holts[which(otters$habitat == 3 )]
data_4 <- otters$holts[which(otters$habitat == 4 )]

N_h <- c(89,61,40,47)
n_h <- c(19,20,22,21)

mean_1 <- mean(data_1)
mean_2 <- mean(data_2)
mean_3 <- mean(data_3)
mean_4 <- mean(data_4)

var_1 <- var(data_1)
var_2 <- var(data_2)
var_3 <- var(data_3)
var_4 <- var(data_4)
var_h <- c(var_1,var_2,var_3,var_4)

t_hat_1 <- N_h[1] * mean_1
t_hat_2 <- N_h[2] * mean_2
t_hat_3 <- N_h[3] * mean_3
t_hat_4 <- N_h[4] * mean_4

t_hat <- t_hat_1 + t_hat_2 + t_hat_3 + t_hat_4
sd_t_hat <- sqrt(sum((1 - n_h/N_h) * (N_h ^2) * (var_h/n_h)))

uppder_bond <- t_hat + qnorm(0.975) * sd_t_hat
lower_bond <- t_hat + qnorm(0.025) * sd_t_hat
lower_bond;uppder_bond
```

```
## [1] 839.8317
```

```
## [1] 1129.597
```

Thus, the 95% confidence interval for total number of acres to farms in the U.S. in 1987 is [839.8317,5971129.597]