Math 255 - Homework 2

Colin Pi

Due in class, Wednesday April 10

Problem 1

Lohr textbook ch. 2 exercise 1

(a) 142

(b)

```
> pop.1 <- c(98, 102, 154, 133, 190, 175)
>
> mean.func1 <- function(a) {
+     return(mean(pop.1[a]))
+ }
```

- Plan 1

```
> sample.plan1 <- list(c(1, 3, 5), c(1, 3, 6), c(1, 4, 5), c(1,
+     4, 6), c(2, 3, 5), c(2, 3, 6), c(2, 4, 5), c(2, 4, 6))
>
> sample.means.plan1 <- unlist(lapply(sample.plan1, mean.func1))
> mean.plan1 <- mean(sample.means.plan1)
> var.plan1 <- mean(sample.means.plan1^2) - mean.plan1^2
```

i. 142
ii. 18.94
iii. 142 - 142 = 0
iv. 18.94 + 0 = 18.94

- Plan 2

```
> sample.plan2 <- list(c(1, 4, 6), c(2, 3, 6), c(1, 3, 5))
> sample.means.plan2 <- unlist(lapply(sample.plan2, mean.func1))
> mean.plan2 <- weighted.mean(sample.means.plan2, c(1/4, 1/2, 1/4))
> var.plan2 <- weighted.mean(sample.means.plan2^2, c(1/4, 1/2,
+     1/4)) - mean.plan2^2
> bias.plan2 <- mean.plan2 - 142
> mse.plan2 <- var.plan2 + bias.plan2^2
```

i. 142.5
ii. 19.36
iii. 0.5
iv. $19.36 + 0.5^2 = 19.61$

(c) The esimator of Plan 1 is more precise (smaller variance) and accurate (unbiased). Also, the MSE of Plan 1 is smaller than that of Plan 2. So, Plan 1 is preferred over Plan 2.

Problem 2

Lohr textbook ch. 2 exercise 2

(a)

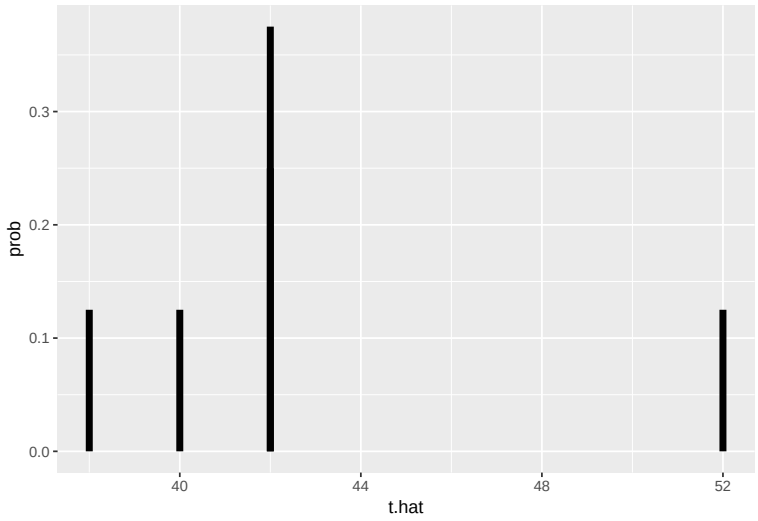
$\pi_1 = P(1\ in\ S) = P((1,3,5,6)\ or\ (1,4,6,8)) = 1/8 + 1/8 = 1/4$
$\pi_2 = P(2\ in\ S) = P((2,3,7,8)\ or\ (2,4,6,8)) = 1/4 + 3/8 = 5/8$
$\pi_3 = P(3\ in\ S) = P((1,3,5,6)\ or\ (2,3,7,8)) = 1/8 + 1/4 = 3/8$
$\pi_4 = P(4\ in\ S) = P((1,4,6,8)\ or\ (2,4,6,8)\ or\ (4,5,7,8)) = 1/8 + 3/8 + 1/8 = 5/8$
$\pi_5 = P(5\ in\ S) = P((1,3,5,6)\ or\ (4,5,7,8)) = 1/8 + 1/8 = 1/4$
$\pi_6 = P(6\ in\ S) = P((1,3,5,6)\ or\ (1,4,6,8)\ or\ (2,4,6,8)) = 1/8 + 1/8 + 3/8 = 5/8$
$\pi_7 = P(7\ in\ S) = P((2,3,7,8)\ or\ (4,5,7,8)) = 1/4 + 1/8 = 3/8$
$\pi_8 = P(8\ in\ S) = P((2,3,7,8)\ or\ (1,4,6,8)\ or\ (2,4,6,8)\ or\ (4,5,7,8)) = 1/4 + 1/8 + 3/8 + 1/8 = 7/8$

(b)

```
> pop.2 <- c(1, 2, 4, 4, 7, 7, 7, 8)
>
> mean.func2 <- function(a) {
+     return(mean(pop.2[a]))
+ }
>
> sample.2 <- list(c(1, 3, 5, 6), c(2, 3, 7, 8), c(1, 4, 6, 8),
+     c(2, 4, 6, 8), c(4, 5, 7, 8))
> sample.means.2 <- unlist(lapply(sample.2, mean.func2))
> t.hat.2 <- 8 * sample.means.2
> mydists.2 <- data.frame(t.hat = t.hat.2, prob = c(1, 2, 1, 3,
+     1)/8)
> knitr::kable(mydists.2 %>% group_by(t.hat) %>% summarize(prob = sum(prob)))
```

t.hat	prob
38	0.125
40	0.125
42	0.625
52	0.125

```
> ggplot(mydists.2, aes(x = t.hat, y = prob)) + geom_segment(aes(yend = 0,
+     xend = t.hat), size = 2)
```



Problem 3

Lohr textbook ch. 2 exercise 30 part (a) only (purposive means don't use random sampling!)

```
> set.seed(70)
> starting.3 <- sample(10, 1)
> starting.3
[1] 1
> sample.3 <- c(1, 11, 21, 31, 41, 51, 61, 71, 81, 91)
> area.sample.3 <- c(12, 2, 40, 54, 45, 49, 48, 9, 16, 40)
```

I used systematic sampling to pick the 10 rectangles. I first picked the starting rectangle for my sample (from rectangle 1 to 10) by random and picked every 10th rectangles from the starting point.

$\bar{y} = 31.5$
$\hat{t} = 100 \cdot \bar{y} = 3150$

Problem 4

Lohr textbook ch. 2 exercise 8

(a) SRS is an appropriate way of sampling out the students from the e-mail list. We do not have any other criteria that allows us to classify them into different clusters or strata.
(b) This case we won't have a list of patients but only board certified allergist. So using cluster sampling is more appropriate than SRS.
(c) SRS is a good way to sample the topics and estimate the probability of error.
(d) SRS is appropriate in this example as well, but we can use either cluster or stratified sampling based on which city the ballots are from.

Problem 5

Lohr textbook ch. 2 exercise 10
$\hat{V}(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n}$, where S^2 = Population variance.

(a) n = 400, N = 4,000: $\hat{V}(\bar{y}) = (1 - \frac{400}{4000})\frac{S^2}{400} = 0.00225 \cdot S^2$
(b) n = 30, N = 300: $\hat{V}(\bar{y}) = (1 - \frac{30}{300})\frac{S^2}{30} = 0.03 \cdot S^2$
(c) n = 3,000, N = 300,000,000: $\hat{V}(\bar{y}) = (1 - \frac{3000}{300000000})\frac{S^2}{3000} = 0.00033333 \cdot S^2$

As (c) has the lowest $\hat{V}(\bar{y})$, it will give the most precise estimation of the population mean.

Problem 6

Lohr textbook ch. 2 exercise 26
Let's consider systematic sampling of choosing $n = \frac{N}{k}$ samples from population with size N by first choosing the starting point a list of population members (using random number) and picking every k^{th} unit thereafter. The starting number can be any number from 1 upto k because we cannot pick n number of samples if the starting number is bigger than k (we will end up choosing less than n samples if the starting point is bigger than k). The total number of samples we can get is k, so $P(S_1) = P(S_1) = P(S_2) = ... = P(S_k) = \frac{1}{k}$. One characteristic of systematic sampling is that the same element does not appear more than one sample. For instance, if the starting point is 1, and we get to choose every 5th element from the population, the sample

will be composed of $1^{st}, 6^{th}, 11^{th}$... elements. These elements won't appear in another sample whose starting point is 2 because there is no way to choose $1^{st}, 6^{th}, 11^{th}$... elements if the starting point is 2 instead of 1. This suggests $\pi_1 = \pi_2 = ... = \pi_i = \frac{1}{k} = \frac{n}{N}$ (because $n = \frac{N}{k}$), one of the properties of SRS. But as noted earlier, $P(S_1) = P(S_2) = ... = P(S_k) = \frac{1}{k} \neq \frac{1}{\binom{N}{n}}$. This shows that $\pi_i = \frac{n}{N}$ does not necessarily indicates that the sampling scheme is SRS.

Problem 7

Lohr textbook ch. 2 exercise 13

(a)

n = 745, N = 2700, $\hat{p} = 0.2$

$SE(\hat{p}) = \sqrt{(1 - \frac{n}{N})(\frac{\hat{p}(1-\hat{p})}{n-1})} = \sqrt{(1 - \frac{745}{2700})(\frac{0.2 \cdot 0.8}{744})} = 0.0124786$

$\hat{t} = N\hat{p} = 540$

$SE(\hat{t}) = SE(N\hat{p}) = N \cdot SE(\hat{p}) = 33.6921547$

95% CI: $\hat{t} \pm q \cdot SE(\hat{t})$, where q = $Z_{0.975}$ from N(0,1) = 1.96

We are 95% confident that the true total numbers of nurses reported bullying in the county lies in between 473.9646 to 606.0354.

(b)

We should assume that that there is no nonresponse or other measurement errors for this estimate to work. So, we may consider that we took an SRS of size = 745 instead of 935.

Problem 8

```
> pubs <- c(rep(0, 28), rep(1, 4), rep(2, 3), rep(3, 4), rep(4,
+     4), rep(5, 2), rep(6, 1), rep(7, 0), rep(8, 2), rep(9, 1),
+     rep(10, 1))
>
> table(pubs)
pubs
 0  1  2  3  4  5  6  8  9 10
28  4  3  4  4  2  1  2  1  1
> summary(pubs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00    1.78    3.00   10.00
> sd(pubs)
[1] 2.682445
```

(a)

$\hat{\bar{y}}_U = \bar{y}_S = 1.78$

$SE(\bar{y}) = \sqrt{(1 - \frac{n}{N})}\frac{s}{\sqrt{n}}$, where s^2 = sample variance

$= \sqrt{1 - \frac{50}{807}}\frac{2.682445}{\sqrt{50}} = 0.3674151$

(b)

$\hat{p} = 28/50 = 0.56$

$SE(\hat{p}) = \sqrt{1 - \frac{n}{N}\frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{1 - \frac{50}{807}\frac{0.56 \cdot 0.44}{49}} = 0.0686805$