Math 255 - Homework 7

Colin Pi

Due in class, Friday May 10

Problem 1

Lohr textbook ch. 5 exercise 1.
In one-stage cluster sampling, $\hat{p}_{cluster,\ unbiased} = \hat{p}_{srs} = 112/134$ only if clusters have equal sizes; however, cluster sizes won't obviously be the same in this case (you have to use ratio estimate because you don't know M_0). Also, SE$[\hat{p}]$ must be $\sqrt{(1-n/N)\frac{S^2e}{n\bar{M}^2}}$, where $S^2e = 1/(n-1)\sum_{i=1}^{n}(t_i - \hat{p}M_i)^2$, t_i = total number of people disagree with having a incinerator, M_i = total number of people inside the block; therefore, $\hat{V}(\hat{p}) = \hat{p}(1-\hat{p})$ is an inaccurate approximation as well.

Problem 2

Lohr textbook ch. 5 exercise 3.

(a)

The sampling unit of the research is not the same as the element. It is a one-stage cluster sampling with wetland as psus and the sites within each of the wetland as ssus (element). We first estimate the total pH level of suburban wetlands by taking the average total of pH level per site and multiplying it by the number of sites selected (2-4). Then we estimate the total pH level of the two wetlands by getting the average of total pH level per wetland and multiplying it by the number of wetlands investigaged (which is 2). Finally, we get the average pH level in suburban area by dividing the estimated total pH level of the two wetlands with total number of sites in the two suburban wetlands.

(b)

We can't treat that each site as independent because the sites within the same cluster (wetland in this case) are more likely to be correlated to each other. Therefore, student t-test is not appropriate for this case.

Problem 3

(a)

Instead of randomly pooling out the articles from 1285 scholarly journals, the researchers randomly picked 26 journals from 1285 journals and worked on the articles published in 1988 in each journal. This is one-stage cluster sampling whose psus is a scholarly journal and ssus (element) is an article published in each journal.

(b)

```
> journal$N <- 1285
> journal$n <- 26
> journal$wts <- journal$N/journal$n
>
> design3.cluster <- svydesign(id = ~1, fpc = ~N, weights = ~wts,
+     data = journal)
>
> svyratio(~nonprob, ~numemp, design3.cluster)
Ratio estimator: svyratio.survey.design2(~nonprob, ~numemp, design3.cluster)
Ratios=
```

```
           numemp
nonprob 0.9256757
SEs=
            numemp
nonprob 0.03398672
```

$\hat{p}_{nonprob} = 0.9256757$
$SE[\hat{p}_{nonprob}] = 0.03398672$

(c)

Our estimate suggests that more than 92.56% of the articles are using nonprobability sampling with standard error of 3%. So, it supports the statement that an overwhelming proportion of articles rely on nonprobability sampling. But high reliance on nonprobabiliy sampling does not necessarily imply that court should give legitimacy to nonprobability sampling because nonprobability samples (voluntary, convenience, etc.), if biased, rather give an inaccurate picture of population.

Problem 4

```
> spanish <- read.csv("http://math.carleton.edu/kstclair/data/spanish.csv")
>
> spanish$N <- 72
> spanish$n <- 10
> spanish$wts <- spanish$N/spanish$n
>
> design4.cluster <- svydesign(id = ~class, fpc = ~N, weights = ~wts,
+     data = spanish)
```

(a)

```
> svytotal(~trip, design4.cluster)
     total     SE
trip 453.6 111.82
> confint(svytotal(~trip, design4.cluster), df = degf(design4.cluster))
        2.5 %   97.5 %
trip 200.6411 706.5589
>
> 72/10 * sum(spanish$trip)
[1] 453.6
> s.t.1 <- var((spanish %>% group_by(class) %>% summarise(t = sum(trip)) %>%
+     ungroup())$t)
> 72/10 * sum(spanish$trip) - qt(c(0.975, 0.025), df = 9) * 72 *
+     sqrt((1 - 10/72) * s.t.1/10)
[1] 200.6411 706.5589
```

$\hat{t}_{unb} = 453.6$
95% CI: (200.6411, 706.5589)

(b)

```
> svymean(~score, design4.cluster)
        mean     SE
score 66.796 2.7091
> confint(svymean(~score, design4.cluster), df = degf(design4.cluster))
         2.5 %   97.5 %
score 60.66752 72.92432
```

Since we don't have an information about M_0, we should use ratio estimate.

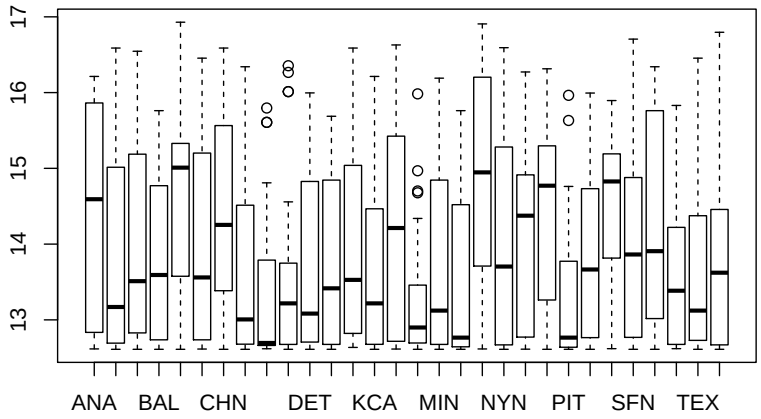
$\hat{\bar{y}}_r = 66.796$
95% CI: (60.66752, 72.92432)

Problem 5

```
> pop <- read.csv("http://math.carleton.edu/kstclair/data/baseball.csv",
+     header = FALSE, na.strings = c("NA", " ", "."))
> names(pop) <- c("team", "league", "player", "salary", "POS",
+     "G", "GS", "InnOuts", "PO", "A", "E", "DP", "PB", "GB", "AB",
+     "R", "H", "SecB", "ThiB", "HR", "RBI", "SB", "CS", "BB",
+     "SO", "IBB", "HPB", "SH", "SF", "GIDP")
> pop$logsal <- log(pop$salary)
```

(a)

```
> boxplot(logsal ~ team, data = pop)
```



Within variance is big, which enchance the precision of cluster sampling (cluster is more representative of the population), but variance between teams are big as well, which offsets the positive impact of higher within variance on the precision. The precision of cluster sampling may not be significantly different from the preficions of SRS.

(b)

```
> max((pop %>% group_by(team) %>% count() %>% ungroup)$n)
[1] 29
> min((pop %>% group_by(team) %>% count() %>% ungroup)$n)
[1] 24
```

The smallest cluster has a size of 24, and the biggest one has a size of 29. The size does not vary that much by cluster, so we may assume that cluster sizes are equal.

(c)

```
> summary(aov(logsal ~ team, data = pop))
             Df Sum Sq Mean Sq F value   Pr(>F)
team         29  135.2   4.661   3.291 2.13e-08 ***
Residuals   767 1086.2   1.416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> var(pop$logsal)
[1] 1.534428
```

MSB $= 4.661$, $S^2 = 1.534428$, $\therefore \frac{MSB}{S^2} = 3.0376134$.
The result from SRS will be way more precise than the result from cluster sampling because the variance of the result from cluster sampling is expected to be three times that of the result from SRS.

(d)

```
> set.seed(80)
> library(dplyr)
> mean((pop %>% group_by(team) %>% count() %>% ungroup)$n)
[1] 26.56667
> n <- 6
> team.names <- levels(pop$team)
> samp.teams <- sample(team.names, size = n, replace = FALSE)
> baseball.1cluster <- filter(pop, team %in% samp.teams) %>% droplevels()
```

I calculated the average size of teams (26.57) and divide the average to approximate n (≈ 5.6). So I set n $= 6$ and got a total of 156 observations.

(e)

```
> baseball.1cluster$N <- nrow(pop)
> baseball.1cluster$n <- 6
> baseball.1cluster$wts <- baseball.1cluster$N/baseball.1cluster$n
>
> design5.cluster <- svydesign(id = ~team, fpc = ~N, weights = ~wts,
+     data = baseball.1cluster)
> svymean(~logsal, design5.cluster)
         mean     SE
logsal 13.895 0.1785
```

$\hat{\bar{y}} = 13.895$
$SE[\hat{\bar{y}}] = 0.1785$

(f)

```
> svymean(~logsal, design5.cluster, deff = T)
           mean       SE   DEff
logsal 13.89502  0.17846 3.1225
```

If the variance of `logal` within the clusters I sampled happens to be smaller/larger than the average within cluster variance, Deff estimate from the sample can be different from the one calculated based on population level. Also, assuming equal size clusters also affected the calculation as well because the clusters actually don't have equal size.