Math 255 - Homework 6

Colin Pi

Due in class, Friday May 3

Problem 1

Lohr textbook ch. 4 exercise 4. Data set is missing from SDaA so use the file:

```
> ssc <- read.csv("http://math.carleton.edu/kstclair/data/ssc.csv")
>
> ssc$n <- nrow(ssc)
> ssc$N <- 864
> ssc$wts <- ssc$N/ssc$n
> design1.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = ssc)
```

(a)

```
> prop.sex <- svyby(~occupation, ~sex, design1.srs, svymean)
> prop.sex
  sex occupationa occupationi occupationn se.occupationa se.occupationi
f   f   0.5217391   0.2173913   0.2608696     0.06717767     0.05547014
m   m   0.6730769   0.1442308   0.1826923     0.04195495     0.03142211
  se.occupationn
f     0.05905247
m     0.03456058
> confint(prop.sex, df = degf(design1.srs))
                     2.5 %    97.5 %
f:occupationa 0.38899517 0.6544831
m:occupationa 0.59017339 0.7559805
f:occupationi 0.10778158 0.3270010
m:occupationi 0.08214026 0.2063213
f:occupationn 0.14418111 0.3775580
m:occupationn 0.11440015 0.2509845
```

$\hat{p}_d = 0.5217391$, 95% CI: (0.38899517, 0.6544831)

(b)

```
> total.sex <- svyby(~occupation, ~sex, design1.srs, svytotal)
> total.sex
  sex occupationa occupationi occupationn se.occupationa se.occupationi
f   f      138.24        57.6       69.12       23.58916       16.05039
m   m      403.20        86.4      109.44       32.10078       19.30341
  se.occupationn
f       17.45628
m       21.40100
> confint(total.sex, df = degf(design1.srs))
                   2.5 %    97.5 %
f:occupationa  91.62751 184.85249
m:occupationa 339.76843 466.63157
f:occupationi  25.88422  89.31578
m:occupationi  48.25620 124.54380
f:occupationn  34.62616 103.61384
```

```
m:occupationn  67.15135 151.72865
```

$\hat{t}_d = 138.24$, 95% CI: (91.62751, 184.85249)

Problem 2

Lohr textbook ch. 4 exercise 5

```
> golfsrs$n <- nrow(golfsrs)
> golfsrs$N <- 14938
> golfsrs$wts <- golfsrs$N/golfsrs$n
> design2.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = golfsrs)
> mean.18 <- svyby(~wkend18, ~holes, design2.srs, svymean)
> mean.18
   holes  wkend18      se
9      9       NA     NaN
11    11       NA     NaN
18    18 34.82882 2.14466
>
> nrow(golfsrs %>% filter(holes == 18))
[1] 85
```

$\bar{y}_d = 34.82882$, $SE[\bar{y}_d] = 2.14466$

Problem 3

Lohr textbook ch. 4 exercise 43.

```
> 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)
> n <- 150
>
> set.seed(30)  # put your favorite large integer here
> samp <- sample(1:nrow(pop), size = n, replace = FALSE)
> baseball.srs <- pop[samp, ]
>
> baseball.srs$N <- nrow(pop)
> baseball.srs$n <- 150
> baseball.srs$wts <- baseball.srs$N/baseball.srs$n
> design3.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = baseball.srs)
```

(a)

```
> mean.pos <- svyby(~logsal, ~POS, design3.srs, svymean)
> rownames(mean.pos) <- c()
> knitr::kable(mean.pos)
```

POS	logsal	se
1B	14.06918	0.3535321
2B	14.16321	0.2069902
3B	14.45517	0.3757621
C	14.35334	0.3041476
CF	13.59887	0.3160283
LF	14.09283	0.4809258
P	13.77970	0.1247193
RF	14.55502	0.4468839
SS	13.47226	0.3643839

(b)

```
> mean.hr <- svyratio(~HR, ~R, design3.srs)
> mean.hr
Ratio estimator: svyratio.survey.design2(~HR, ~R, design3.srs)
Ratios=
           R
HR 0.2416375
SEs=
            R
HR 0.01251667
> confint(mean.hr, df = degf(design3.srs))
          2.5 %    97.5 %
HR/R 0.2169044 0.2663706
```

$\hat{B} = 0.2416375$, 95% CI: (0.2169044, 0.2663706)

Problem 4

Read Lohr textbook ch. 4 exercise 15, then answer the questions below.

(a) Estimate the mean concentration of lead in the area using the systematic sample and give a SE. (Assume that this sample behaves like a SRS). Repeat for the mean concentration of copper.
As N is big enough, we can assume fpc ≈ 1

$$\bar{y}_{lead} = 127,\ SE[\bar{y}_{lead}] = \sqrt{\frac{s^2_{lead}}{n}} = \sqrt{\frac{146^2}{121}} = 13.27273$$

$$\bar{y}_{copper} = 35,\ SE[\bar{y}_{copper}] = \sqrt{\frac{s^2_{copper}}{n}} = \sqrt{\frac{16^2}{35}} = 2.704494$$

(b) Redo (a) but this time use the poststratified sample. Since points are sampled on a grid you can assume proportional allocation where Nh/N = nh/n.

$$\bar{y}_{lead,\ post} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n}\bar{y}_h = \frac{82}{121}\cdot 71 + \frac{31}{121}\cdot 259 + \frac{8}{121}\cdot 189 = 126.9669$$

$$SE[\bar{y}_{lead,\ post}] \approx \sqrt{\sum(\frac{n_h}{n})(\frac{n}{n-1})(\frac{n_h-1}{n_h})\frac{s_h^2}{n}} = \sqrt{\frac{82}{121}\frac{28^2}{121} + \frac{31}{121}\frac{232^2}{121} + \frac{8}{121}\frac{79^2}{121}} = 11.03471$$

$$\bar{y}_{copper,\ post} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n}\bar{y}_h = \frac{82}{121}\cdot 28 + \frac{31}{121}\cdot 50 + \frac{8}{121}\cdot 45 = 34.76033$$

$$SE[\bar{y}_{copper,\ post}] \approx \sqrt{\sum(\frac{n_h}{n})(\frac{s_h^2}{n})} = \sqrt{\frac{82}{121}\frac{9^2}{121} + \frac{31}{121}\frac{18^2}{121} + \frac{8}{121}\frac{15^2}{121}} = 1.123663$$

(c) Compare your estimates/SEs in (b) to those from part (a) and use your results to a recommendation about use of stratification in future surveys to increase precision.

Poststratification gives more precise results than systematic sampling. For the future research, we can include more data from Stratum B when researching lead because the variance of the area is relatively bigger than the rest. Oversampling such regions may increase the precision of the result.

(d) The poststratified variance in (eq. 4.22) requires some assumptions to be met. Discuss whether or not these assumptions are met for this problem and comment on violation of these assumptions might effect your conclusion to part (c).
It may not meet the assumptions as the sample size of stratum C is really small (8). SE for poststratified estimates are only approximate and are best used when sample sizes within the strata is large. Also, we assumed that $\frac{N_h}{N} \approx \frac{n_h}{n}$ due to lack of information about the population. If this assumption turns out to be not true, eq. 4.22 may give a wrong approximation of SE. In that we cannot accurately measure postratification SE, we cannot conclude whether poststratification produces more precise result than SRS.

Problem 5

Consider the golf example you first worked on in ch. 2 exercise 16 (golfsrs). Omit the entry (row) that claims to have 11 holes from the golfsrs data.

(a) Estimate the proportion of all courses that have a golf pro and give a SE.

```
> golfsrs2 <- golfsrs %>% filter(holes != 11)
>
> golfsrs2$n <- nrow(golfsrs2)
> golfsrs2$N <- 14938
> golfsrs2$wts <- golfsrs2$N/golfsrs2$n
>
> design4.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = golfsrs2)
> prop.pro <- svymean(~pro, design4.srs)
> prop.pro
        mean     SE
pron 0.26891 0.0407
proy 0.73109 0.0407
```

$\hat{p}_{pro} = 0.73109$, $SE[\hat{p}_{pro}] = 0.0407$

(b) Repeat part (a) using a poststratified estimate and SE using the fact that within the entire population, 3735 courses are 9 holes and 11203 are 18 holes.

```
> pop.ps <- data.frame(holes = c(9, 18), N.str = c(3735, 11203))
>
> design.post <- postStratify(design4.srs, strata = ~holes, population = pop.ps)
>
> svymean(~pro, design.post)
        mean     SE
pron 0.25296 0.0353
proy 0.74704 0.0353
```

$\hat{p}_{pro,\ post} = 0.74704$, $SE[\hat{p}_{pro,\ post}] = 0.0353$

(c) Compare estimate values and carefully explain why the postratified estimate is (higher/lower) than the SRS estimate.

```
> svyby(~pro, ~holes, design4.srs, svymean)
   holes      pron      proy    se.pron    se.proy
9      9 0.5882353 0.4117647 0.08442214 0.08442214
18    18 0.1411765 0.8588235 0.03777630 0.03777630
```

```
> 11203/14938
[1] 0.7499665
```

We can observe that the proportion of having golf pro in 9 holes courses is almost the half of that of 18 holes, but the fraction of 9 holes courses in population is way lower than that of 18 holes courses ($\frac{N_9}{N} = 0.2500335$ v. $\frac{N_{18}}{N} = 0.7499665$). In that SRS does not count the distribution of 9 and 18 holes courses in the population but simply gives equal weights to each unit in the sample, it overrepresents 9 holes courses to 18 holes courses. Therefore, the SRS estimate is relativel lower than postratified estimate, which accounts for the difference in the distribution of 9 and 18 holes courses.