Math 255 - Homework 3

Colin Pi

Due in class, Wednesday April 15

Problem 1

Lohr textbook ch. 2 exercise 14. Data set is missing from SDaA so use the file:

```
> ssc <- read.csv("http://math.carleton.edu/kstclair/data/ssc.csv")
```

(a) The members of SSC not in the online directory is not considered. Also, it only covers workers in academics, government, and industry. These possibly lead to undercoverage issue.

(b)

```
> ssc$N <- 864
> ssc$wts <- ssc$N/nrow(ssc)
> design1.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = ssc)
> svymean(~sex, design1.srs)
        mean     SE
sexf 0.30667 0.0343
sexm 0.69333 0.0343
> confint(svymean(~sex, design1.srs), df = degf(design1.srs))  ## can I just use z distribution for prop
          2.5 %    97.5 %
sexf 0.2388099 0.3745234
sexm 0.6254766 0.7611901
>
> sex <- ifelse(ssc$sex == "f", 1, 0)
> se <- sqrt((1 - length(sex)/864) * (mean(sex) * (1 - mean(sex)))/(length(sex) -
+     1))
> mean(sex) + se * qt(c(0.025, 0.975), df = length(sex) - 1)
[1] 0.2388099 0.3745234
```

$\hat{p}_{female} = 0.30667$
95% CI: (0.2388099 0.3745234)

(c)

```
> svytotal(~sex, design1.srs)
      total    SE
sexf 264.96 29.67
sexm 599.04 29.67
> confint(svytotal(~sex, design1.srs), df = degf(design1.srs))
        2.5 %   97.5 %
sexf 206.3317 323.5883
sexm 540.4117 657.6683
>
> se <- 864 * sqrt((1 - nrow(ssc)/864) * (mean(sex) * (1 - mean(sex)))/(nrow(ssc) -
+     1))
> mean(sex) * 864 + se * qt(c(0.025, 0.975), df = nrow(ssc) - 1)
[1] 206.3317 323.5883
```

$\hat{t}_{female} = 264.96$
95% CI: (206.3317, 323.5883)

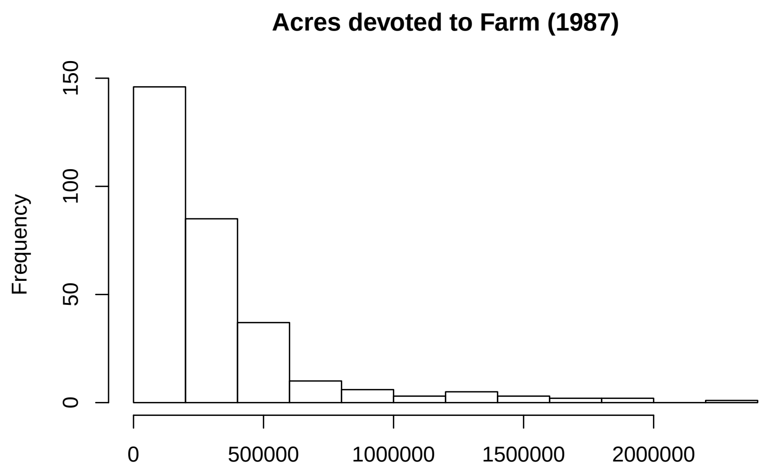
Problem 2

Lohr textbook ch. 2 exercise 15. Data set `agsrs`

```
> agsrs$N <- 3078
> agsrs$wts <- agsrs$N/nrow(agsrs)
> design2.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = agsrs)
```

(a)

```
> hist(agsrs$acres87, main = "Acres devoted to Farm (1987)", xlab = "")
```

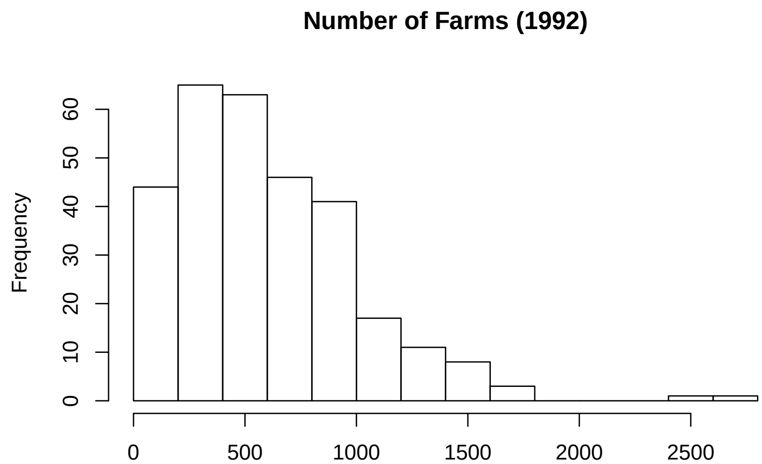



```
> svymean(~acres87, design2.srs)
          mean    SE
acres87 301954 18914
> confint(svymean(~acres87, design2.srs), df = degf(design2.srs))
         2.5 %   97.5 %
acres87 264733 339174.5
>
> se <- sqrt(1 - nrow(agsrs)/3078) * sd(agsrs$acres87)/sqrt(nrow(agsrs))
> mean(agsrs$acres87) + qt(c(0.025, 0.975), df = nrow(agsrs) -
+     1) * se
[1] 264733.0 339174.5
```

$\hat{\bar{y}}_{U,1987} = 301954$ Acres
95% CI: (264733, 339174.5)

(b)

```
> hist(agsrs$farms92, main = "Number of Farms (1992)", xlab = "")
```

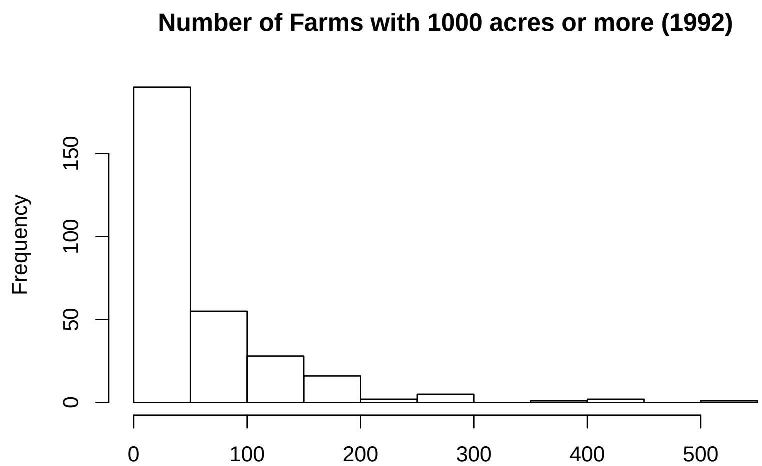



```
> svymean(~farms92, design2.srs)
          mean     SE
farms92 599.06 22.062
> confint(svymean(~farms92, design2.srs), df = degf(design2.srs))
           2.5 %   97.5 %
farms92 555.6426 642.4774
>
> se <- sqrt(1 - nrow(agsrs)/3078) * sd(agsrs$farms92)/sqrt(nrow(agsrs))
> mean(agsrs$farms92) + qt(c(0.025, 0.975), df = nrow(agsrs) -
+     1) * se
[1] 555.6426 642.4774
```

$\hat{\bar{y}}_{U,1987} = 599.06$
95% CI: (555.6426, 642.4774)

(c)

```
> hist(agsrs$largef92, main = "Number of Farms with 1000 acres or more (1992)",
+     xlab = "")
```



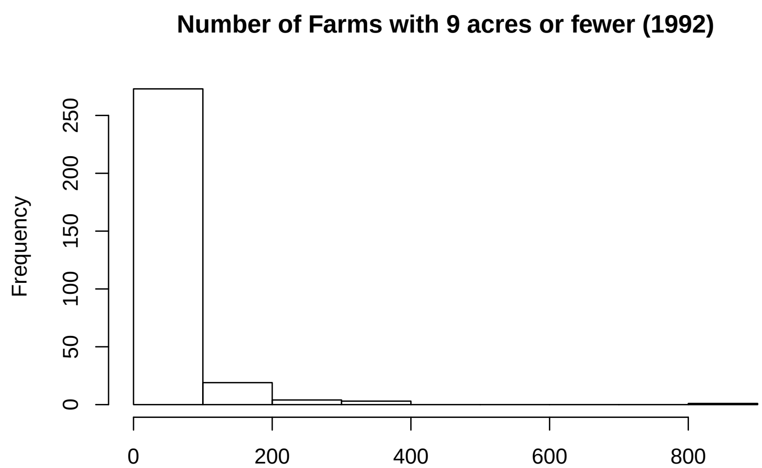

```
> svymean(~largef92, design2.srs)
           mean     SE
largef92 56.593 3.9904
> confint(svymean(~largef92, design2.srs), df = degf(design2.srs))
           2.5 %   97.5 %
largef92 48.7406 64.44606
```

```
>
> se <- sqrt(1 - nrow(agsrs)/3078) * sd(agsrs$largef92)/sqrt(nrow(agsrs))
> mean(agsrs$largef92) + qt(c(0.025, 0.975), df = nrow(agsrs) -
+     1) * se
[1] 48.74060 64.44606
```

$\hat{\bar{y}}_{U,1992} = 56.593$
95% CI: (48.7406, 64.44606)

(d)

```
> hist(agsrs$smallf92, main = "Number of Farms with 9 acres or fewer (1992)",
+     xlab = "")
```




```
> svymean(~smallf92, design2.srs)
           mean     SE
smallf92 46.823 3.6375
> confint(svymean(~smallf92, design2.srs), df = degf(design2.srs))
           2.5 %   97.5 %
smallf92 39.6649 53.98177
>
> se <- sqrt(1 - nrow(agsrs)/3078) * sd(agsrs$smallf92)/sqrt(nrow(agsrs))
> mean(agsrs$smallf92) + qt(c(0.025, 0.975), df = nrow(agsrs) -
+     1) * se
[1] 39.66490 53.98177
```

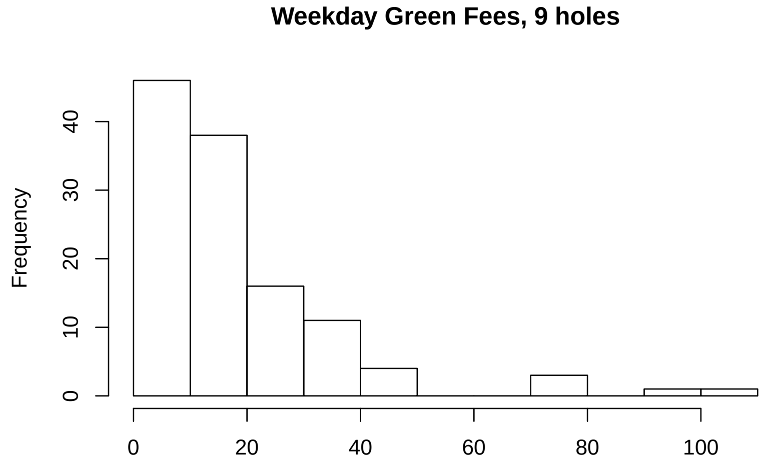
$\hat{\bar{y}}_{U,1992} = 46.823$
95% CI: (39.6649, 53.98177)

Problem 3

Lohr textbook ch. 2 exercise 16. Data set `golfsrs`

(a)

```
> hist(golfsrs$wkday9, main = "Weekday Green Fees, 9 holes", xlab = "")
```



The distribution is heavility skewed to the right.

(b)

```
> golfsrs$N <- 14938
> golfsrs$wts <- golfsrs$N/nrow(golfsrs)
> design3.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = golfsrs)
> svymean(~wkday9, design3.srs)
         mean     SE
wkday9 20.153 1.6299
>
> se <- sqrt(1 - nrow(golfsrs)/14938) * sd(golfsrs$wkday9)/sqrt(nrow(golfsrs))
> se
[1] 1.629866
```

Average weekday greens fee to play 9 holes of golf = \$20.153
SE: \$1.6299

Problem 4

Lohr textbook ch. 2 exercise 18

```
> holes18 <- ifelse(golfsrs$holes == 18, 1, 0)
> design3.srs <- update(design3.srs, holes18 = holes18)
> svymean(~holes18, design3.srs)
           mean     SE
holes18 0.70833 0.0415
> confint(svymean(~holes18, design3.srs), df = degf(design3.srs))
            2.5 %    97.5 %
holes18 0.6261612 0.7905054
>
> mean(holes18)
[1] 0.7083333
>
> se <- sqrt((1 - length(holes18)/14938) * (mean(holes18) * (1 -
+     mean(holes18)))/(length(holes18) - 1))
> mean(holes18) + se * qt(c(0.025, 0.975), df = length(holes18) -
+     1)
[1] 0.6261612 0.7905054
```

$\hat{p}_{18\ holes} = 0.70833$
95% CI: (0.6261612, 0.7905054)

Problem 5

Lohr textbook ch. 2 exercise 19.
n_0, the required sample size we would use for SRS without replacement, is defined as

$$n_0 = (\frac{z_{\alpha/2}S}{e})^2.$$

For a large populations, $S^2 \approx p(1-p)$, and the maximum of S^2 is at which p = 1/2. If we plug-in p = 1/2 to get the maximum S^2, $z_{\alpha/2} = 1.96$, and e = 0.04, $n_0 = (\frac{1.96}{0.04})^2 \cdot 1/4 = 600.25$. If we plug-in n_0 to the equation below,

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{600.25}{1 + \frac{600.25}{N}}$$

```
> city <- c("Buckeye", "Gillbert", "Gila Bend", "Phoenix", "Tempe")
> N <- c(4857, 59338, 1724, 1149417, 153821)
> n <- unlist(lapply(N, function(x) {
+     return(600.25/(1 + 600.25/x))
+ }))
> sample.size <- data.frame(City = city, Population = N, Sample.Size = n)
> knitr::kable(sample.size)
```

City	Population	Sample.Size
Buckeye	4857	534.2277
Gillbert	59338	594.2388
Gila Bend	1724	445.2322
Phoenix	1149417	599.9367
Tempe	153821	597.9168

For the cities with population above 50k, the required sample size to have margin of error of 4% is close to 600 (n_0, which does not put fpc into consideration). But the required sample size for small cities like Buckeye and Gila Bend shows a huge deviation from n_0, suggesting that fpc makes a huge difference for the populations with small size.

Problem 6

Lohr textbook ch. 2 exercise 32.

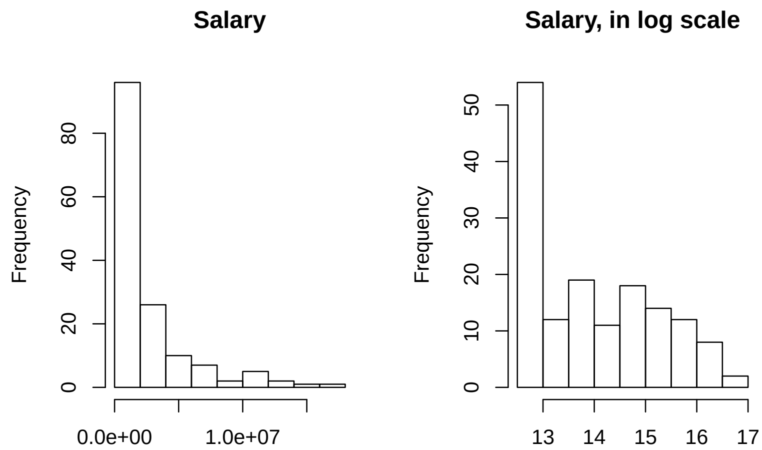
(a)

```
> 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, ]
> str(baseball.srs)
'data.frame':   150 obs. of  31 variables:
 $ team    : Factor w/ 30 levels "ANA","ARI","ATL",..: 3 15 11 13 9 5 27 7 29 5 ...
 $ league  : Factor w/ 2 levels "AL","NL": 2 2 1 2 1 1 2 2 1 1 ...
 $ player  : Factor w/ 791 levels "aardsda0","abbotpa0",..: 663 444 638 203 635 356 626 468 560 153 ...
 $ salary  : int  11666667 1400000 385000 370000 2700000 750000 8625000 1200000 1700000 500000 ...
 $ POS     : Factor w/ 9 levels "1B","2B","3B",..: 7 7 5 9 7 8 3 7 1 6 ...
 $ G       : int  69 44 79 104 1 136 142 66 49 30 ...
 $ GS      : int  0 0 77 97 30 59 139 0 11 5 ...
 $ InnOuts: int  245 85 1983 2526 564 1772 3684 159 321 132 ...
 $ PO      : int  9 1 177 137 1 133 93 1 104 7 ...
 $ A       : int  9 5 2 279 17 5 325 4 3 1 ...
 $ E       : int  0 0 9 10 0 2 10 0 0 0 ...
 $ DP      : int  0 1 1 56 2 0 23 0 12 0 ...
 $ PB      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ GB      : int  69 44 79 104 1 136 142 66 49 30 ...
 $ AB      : int  2 0 332 384 4 290 500 2 134 75 ...
 $ R       : int  0 0 41 66 0 51 109 0 13 9 ...
 $ H       : int  0 0 107 105 1 79 157 0 30 17 ...
 $ SecB    : int  0 0 9 15 0 14 32 0 2 8 ...
 $ ThiB    : int  0 0 3 2 0 1 4 0 1 0 ...
 $ HR      : int  0 0 2 8 0 6 34 0 5 2 ...
 $ RBI     : int  0 0 26 31 0 33 124 0 17 8 ...
 $ SB      : int  0 0 19 13 0 5 4 0 0 0 ...
 $ CS      : int  0 0 13 2 0 4 3 0 0 0 ...
 $ BB      : int  0 0 7 17 0 15 72 0 14 10 ...
 $ SO      : int  2 0 50 56 2 49 92 0 19 21 ...
 $ IBB     : int  0 0 0 0 0 0 5 0 0 0 ...
 $ HPB     : int  0 0 0 9 0 2 13 0 3 1 ...
 $ SH      : int  0 0 12 22 0 1 1 0 0 0 ...
 $ SF      : int  0 0 1 3 0 2 7 0 2 0 ...
 $ GIDP    : int  0 0 5 4 0 5 8 0 3 1 ...
 $ logsal  : num  16.3 14.2 12.9 12.8 14.8 ...
```

(b)

```
> par(mfrow = c(1, 2))
> hist(baseball.srs$salary, main = "Salary", xlab = "")
> hist(baseball.srs$logsal, main = "Salary, in log scale", xlab = "")
```




```
> par(mfrow = c(1, 1))
```

The distributions of both salary and logsal are heavily skewed to the right, but the distribution of logsal is less skewed than that of Salary.

(c)

```
> baseball.srs$N <- 797
> baseball.srs$wts <- 797/nrow(baseball.srs)
> design6.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = baseball.srs)
> svymean(~logsal, design6.srs)
         mean     SE
logsal 13.948 0.0903
> confint(svymean(~logsal, design6.srs), df = degf(design6.srs))
          2.5 %   97.5 %
logsal 13.76949 14.12649
>
> se <- sqrt(1 - nrow(baseball.srs)/797) * sd(baseball.srs$logsal)/sqrt(nrow(baseball.srs))
> mean(baseball.srs$logsal) + qt(c(0.025, 0.975), df = nrow(baseball.srs) -
+     1) * se
[1] 13.76949 14.12649
```

$\hat{logsal} = 13.948$
95% CI: (13.76949, 14.12649)

(d)

```
> pitcher <- ifelse(baseball.srs$POS == "P", 1, 0)
> design6.srs <- update(design6.srs, pitcher = pitcher)
> svymean(~pitcher, design6.srs)
           mean     SE
pitcher 0.48667 0.0369
> confint(svymean(~pitcher, design6.srs), df = degf(design6.srs))
            2.5 %   97.5 %
pitcher 0.4137654 0.559568
>
> se <- sqrt((1 - length(pitcher)/797) * (mean(pitcher) * (1 -
+     mean(pitcher)))/(length(pitcher) - 1))
> mean(pitcher) + se * qt(c(0.025, 0.975), df = length(pitcher) -
+     1)
[1] 0.4137654 0.5595680
```

$\hat{p}_{pitcher} = 0.48667$
95% CI: (0.4137654, 0.559568)

(e)

```
> mean(pop$logsal)
[1] 13.92706
> pitcher.pop <- ifelse(pop$POS == "P", 1, 0)
> mean(pitcher.pop)
[1] 0.4717691
```

$\bar{logsal}_U = 13.92706$
$p_{pitcher} = 0.4717691$
Both parameters are included in the CIs calculated above.

Problem 7

$$n_{min} = 28 + 25(\frac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^3}{NS^3})^2$$

```
> mean.diff.salary <- pop$salary - mean(pop$salary)
> N <- 797
> S.salary <- sd(pop$salary)
> n_min.salary <- 28 + 25 * ((sum(mean.diff.salary^3))/(N * S.salary^3))^2
>
> mean.diff.logsal <- pop$logsal - mean(pop$logsal)
> S.logsal <- sd(pop$logsal)
> n_min.logsal <- 28 + 25 * ((sum(mean.diff.logsal^3))/(N * S.logsal^3))^2
```

$n_{min,\ salary} = 166.7455208$
$n_{min,\ logsal} = 35.4787922$

Problem 8

$\frac{1}{\pi_i} = \frac{N}{n},\ E[Z_i] = \frac{n}{N}$
$E[\hat{t}] = E[\sum_{i \in S} \frac{y_i}{\pi_i}] = E[\sum_{i=1}^{N} \frac{y_i}{\pi_i} Z_i] = \frac{N}{n} E[\sum_{i=1}^{N} y_i Z_i] = \frac{N}{n} \frac{n}{N} E[\sum_{i=1}^{N} y_i] = E[t]$