Math 255 - Homework 4

Colin Pi

Due in class, Friday April 19

Problem 1

Lohr textbook ch. 3 exercise 5.

(a) Selected scholars in American Council of Learned Societies in seven disciplines who answered to the survey.

(b)

```
> Nh <- c(9100, 1950, 5500, 10850, 2100, 5500, 9000)
> N <- sum(Nh)
> p.hats <- c(0.37, 0.23, 0.23, 0.29, 0.19, 0.43, 0.41)
> phat.str.1 <- sum(Nh/N * p.hats)
> phat.str.1
[1] 0.3336591
>
> nh <- c(636, 451, 481, 611, 493, 575, 588)
> var.str <- sum((Nh/N)^2 * (1 - nh/Nh) * p.hats * (1 - p.hats)/(nh -
+     1))
> se.str.1 <- sqrt(var.str)
> se.str.1
[1] 0.007903364
```

$\hat{p}_{str} = \sum_{h=1}^{7} \frac{N_h}{N}\hat{p}_h = \frac{9100}{44000} \cdot 0.37 + ... + \frac{9000}{44000} \cdot 0.41 = 0.3336591$

$SE[\hat{p}_{str}] = \sqrt{\sum_{h=1}^{7} (\frac{N_h}{N})^2(1 - \frac{n_h}{N_h})\frac{\hat{p}(1-\hat{p})}{n_h - 1}}$

$= \sqrt{(\frac{9100}{44000})^2(1 - \frac{636}{9100})\frac{0.37(1-0.37)}{635} + ... + (\frac{9000}{44000})^2(1 - \frac{588}{9000})\frac{0.41(1-0.41)}{587}} = 0.0079034$

Problem 2

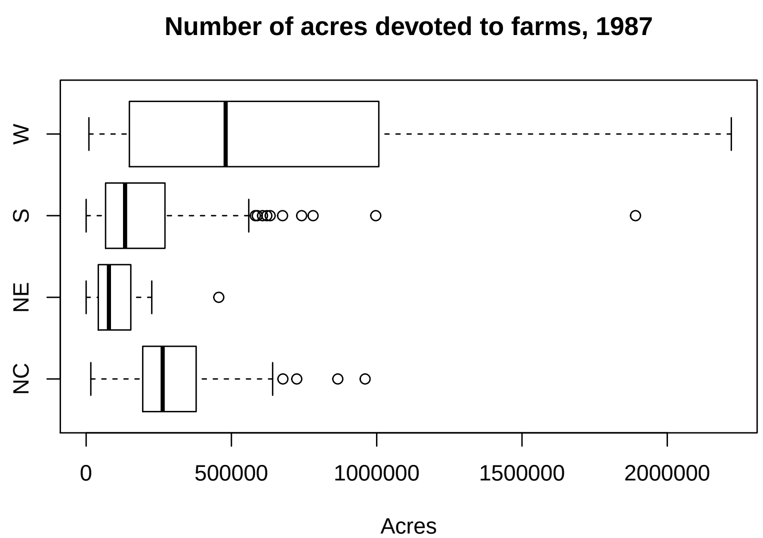
Lohr textbook ch. 3 exercise 9.

```
> agstrat$N <- recode(agstrat$region, NC = 1054, NE = 220, S = 1382,
+     W = 422)
> agstrat %>% group_by(region) %>% summarize(min(N), max(N))
# A tibble: 4 x 3
  region `min(N)` `max(N)`
  <fct>     <dbl>    <dbl>
1 NC         1054     1054
2 NE          220      220
3 S          1382     1382
4 W           422      422
>
> agstrat <- agstrat %>% group_by(region) %>% mutate(n = n())
> agstrat %>% group_by(region) %>% summarize(min(n), max(n))  # check
# A tibble: 4 x 3
```

```
  region `min(n)` `max(n)`
  <fct>      <dbl>    <dbl>
1 NC           103      103
2 NE            21       21
3 S            135      135
4 W             41       41
>
> agstrat$wts <- agstrat$N/agstrat$n
> agstrat %>% group_by(region) %>% summarize(min(wts), max(wts))  #check
# A tibble: 4 x 3
  region `min(wts)` `max(wts)`
  <fct>        <dbl>      <dbl>
1 NC            10.2       10.2
2 NE            10.5       10.5
3 S             10.2       10.2
4 W             10.3       10.3
>
> design.strat <- svydesign(id = ~1, fpc = ~N, weights = ~wts,
+     strata = ~region, data = agstrat)
```

(a)

```
> boxplot(acres87 ~ region, data = agstrat, horizontal = TRUE,
+     main = "Number of acres devoted to farms, 1987", xlab = "Acres")
```

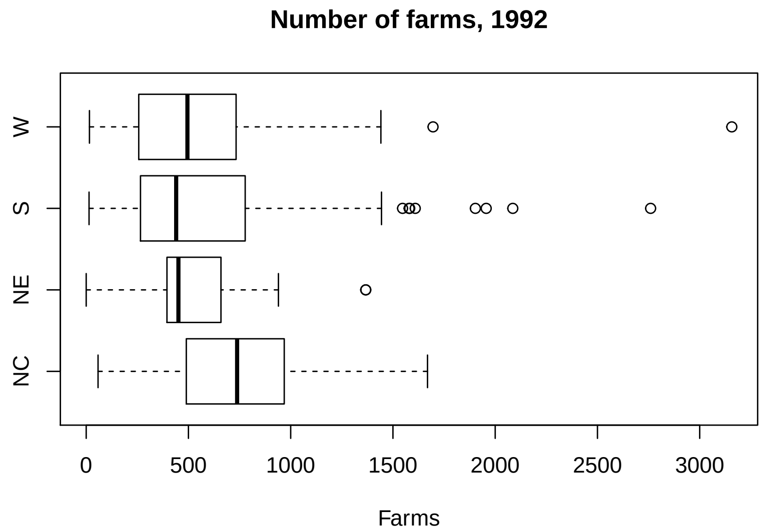



```
> svymean(~acres87, design.strat)
          mean    SE
acres87 298547 16293
> confint(svymean(~acres87, design.strat), df = degf(design.strat))
           2.5 %   97.5 %
acres87 266482.4 330611.8
```

$\bar{y}_{ssr}$ (301954) $>$ $\bar{y}_{str}$ (298547)
$SE[\bar{y}_{ssr}]$ (18914) $>$ $SE[\bar{y}_{str}]$ (16293)
$CI_{ssr} = (264733, 339174.5)$ is wider than $CI_{str} = (266482.4, 330611.8)$

(b)

```
> boxplot(farms92 ~ region, data = agstrat, horizontal = TRUE,
+     main = "Number of farms, 1992", xlab = "Farms")
```

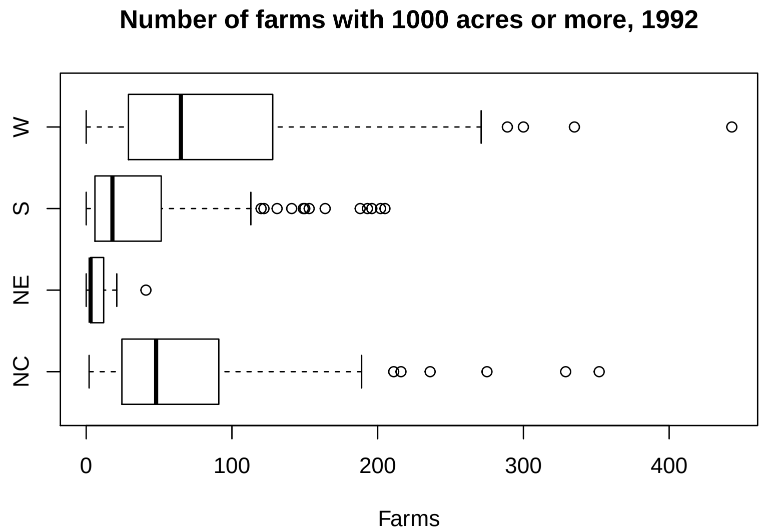



```
> svymean(~farms92, design.strat)
          mean     SE
farms92 637.16 24.277
> confint(svymean(~farms92, design.strat), df = degf(design.strat))
           2.5 %   97.5 %
farms92 589.3853 684.9422
```

$\bar{y}_{ssr}$ (599.06) $<$ $\bar{y}_{str}$ (637.16)
$SE[\bar{y}_{ssr}]$ (22.062) $<$ $SE[\bar{y}_{str}]$ (24.277)
$CI_{ssr} = (555.6426, 642.4774)$ is narrower than $CI_{str} = (589.3853, 684.9422)$.

(c)

```
> boxplot(largef92 ~ region, data = agstrat, horizontal = TRUE,
+     main = "Number of farms with 1000 acres or more, 1992", xlab = "Farms")
```



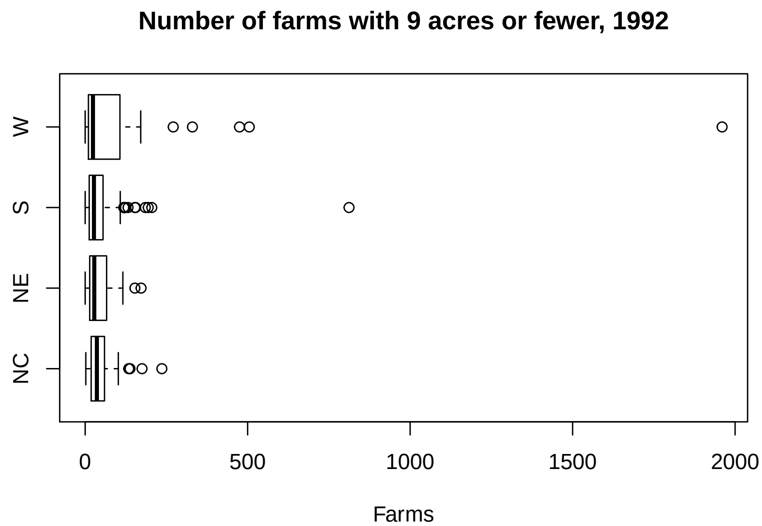

```
> svymean(~largef92, design.strat)
           mean     SE
largef92 56.698 3.5577
```

```
> confint(svymean(~largef92, design.strat), df = degf(design.strat))
            2.5 %   97.5 %
largef92 49.69636 63.69954
```

$\bar{y}_{ssr} = 56.593 < \bar{y}_{str} = 56.698$ (approximately the same)
$SE[\bar{y}_{ssr}] = 3.9904 > SE[\bar{y}_{str}] = 3.5577$
$CI_{ssr} = (48.77239, 64.41428)$ is wider than $CI_{str} = (49.69636, 63.69954)$.

(d)

```
> boxplot(smallf92 ~ region, data = agstrat, horizontal = TRUE,
+     main = "Number of farms with 9 acres or fewer, 1992", xlab = "Farms")
```




```
> svymean(~largef92, design.strat)
           mean     SE
largef92 56.698 3.5577
> confint(svymean(~largef92, design.strat), df = degf(design.strat))
            2.5 %   97.5 %
largef92 49.69636 63.69954
> svymean(~smallf92, design.strat)
           mean     SE
smallf92 56.863 7.2014
> confint(svymean(~smallf92, design.strat), df = degf(design.strat))
            2.5 %   97.5 %
smallf92 42.69033 71.03526
```

$\bar{y}_{ssr} = 46.823 < \bar{y}_{str} = 56.863$
$SE[\bar{y}_{ssr}] = 3.6375 < SE[\bar{y}_{str}] = 7.2014$
$CI_{ssr} = (39.69387, 53.95279)$ is narrower than $CI_{str} = (42.69033, 71.03526)$.

Problem 3

Lohr textbook ch. 3 exercise 15.

(a)

Advantage: proportional allocation provides the most precise result when within variance of all strata are similar.

Disadvantage: If the within variance varies on strata, proportional allocation is not the best way to produce the most precise result (optimal allocation is better than proportional allocation in this case).

(b)

```
> Nh <- c(190, 407, 811)
> ybarh <- c(3.925, 3.938, 3.942)
> N <- sum(Nh)
> ybar.str.3 <- sum(Nh/N * ybarh)
>
> sh <- c(0.037, 0.052, 0.07)
> nh <- c(21, 14, 22)
> se.str.3 <- sqrt(sum((Nh/N)^2 * (1 - nh/Nh) * sh^2/nh))
> se.str.3
[1] 0.009408975
> ci.3 <- ybar.str.3 - se.str.3 * qt(c(0.975, 0.025), 21 + 14 +
+     22 - 3)
> ci.3
[1] 3.919686 3.957414
```

$\bar{y}_U = \sum_{h=low}^{upper} \frac{N_h}{N}\bar{y}_h = \frac{190}{1408} \cdot 3.925 + \frac{407}{1408} \cdot 3.938 + \frac{811}{1408} \cdot 3.942 = 3.9385497$

$SE[\bar{y}_U] = \sqrt{\sum_{h=low}^{upper} (\frac{N_h}{N})^2(1 - \frac{n_h}{N_h})\frac{s_h^2}{n_h}}$

$= \sqrt{(\frac{190}{1408})^2 \cdot (1 - \frac{21}{190})\frac{0.037^2}{21} + (\frac{407}{1408})^2 \cdot (1 - \frac{14}{407})\frac{0.053^2}{14} + (\frac{811}{1408})^2 \cdot (1 - \frac{22}{811})\frac{0.070^2}{22}} = 0.009409$

95% CI: $\bar{y}_U \pm qt_{0.975,df=57-3} \cdot SE[\bar{y}_U] = (3.9196859, 3.9574136)$

(c)

We can answer this answer by estimating the difference in means of log prices among the three strata.

- Low v. Middle

```
> diff.1 <- 3.925 - 3.938
>
> se.1 <- sqrt(sum((1 - nh[1:2]/Nh[1:2]) * sh[1:2]^2/nh[1:2]))
> se.1
[1] 0.01563599
>
> ci.3.1 <- diff.1 - se.1 * qnorm(c(0.975, 0.025))   # what df should I use?
```

$$\bar{y}_{low} - \bar{y}_{middle} \pm qt_{0.975,\ df=} \cdot \sqrt{Var[\bar{y}_{low}] + Var[\bar{y}_{middle}]} = -0.013 \pm 1.96 \cdot \sqrt{0.000057985 + 0.000185}$$

95% CI: (-0.043646, 0.017646)
We are 95% confident that the log price of low income stratum is -0.043646 to 0.017646 higher than middle income stratum in average. We can't conclude that the log price of low income stratum is different from that of middle income stratum.

- Middle v. Upper

```
> diff.2 <- 3.938 - 3.942
>
> se.2 <- sqrt(sum((1 - nh[2:3]/Nh[2:3]) * sh[2:3]^2/nh[2:3]))
> se.2
[1] 0.02007945
```

```
>
> ci.3.2 <- diff.2 - se.2 * qnorm(c(0.975, 0.025))  # what df should I use?
> ci.3.2
[1] -0.04335501  0.03535501
```

$$\bar{y}_{middle} - \bar{y}_{upper} \pm qt_{0.975,\ df=} \cdot \sqrt{Var[\bar{y}_{middle}] + Var[\bar{y}_{upper}]} = -0.004 \pm 1.96 \cdot \sqrt{0.0001864991 + 0.0002166853}$$

95% CI: (-0.043355, 0.035355)
We are 95% confident that the log price of middle income stratum is -0.043355 to 0.035355 higher than upper income stratum in average. We can't conclude that the log price of middle income stratum is different from that of upper income stratum.

- High v. low

```
> diff.3 <- 3.925 - 3.942
>
> se.3 <- sqrt(sum((1 - nh[c(1, 3)]/Nh[c(1, 3)]) * sh[c(1, 3)]^2/nh[c(1,
+     3)]))
> se.3
[1] 0.01657319
>
> ci.3.3 <- diff.3 - se.3 * qnorm(c(0.975, 0.025))  # what df should I use?
> ci.3.3
[1] -0.04948285  0.01548285
```

$$\bar{y}_{low} - \bar{y}_{upper} \pm qt_{0.975,\ df=} \cdot \sqrt{Var[\bar{y}_{low}] + Var[\bar{y}_{upper}]} = -0.017 \pm 1.96 \cdot \sqrt{0.000057985 + 0.0002166853}$$

95% CI: (-0.0494829, 0.0154829)
We are 95% confident that the log price of low income stratum is -0.0494829 to 0.0154829 higher than upper income stratum in average. We can't conclude that the log price of low income stratum is different from that of upper income stratum.

Problem 4

Lohr textbook ch. 3 exercise 16. (Data in SDaA.)

```
> otters$N <- recode(otters$habitat, `1` = 89, `2` = 61, `3` = 40,
+     `4` = 47)
> otters %>% group_by(habitat) %>% summarize(min(N), max(N))
# A tibble: 4 x 3
  habitat `min(N)` `max(N)`
    <int>    <dbl>    <dbl>
1       1       89       89
2       2       61       61
3       3       40       40
4       4       47       47
>
> otters <- otters %>% group_by(habitat) %>% mutate(n = n())
> otters %>% group_by(habitat) %>% summarize(min(n), max(n))  # check
# A tibble: 4 x 3
```

```
  habitat `min(n)` `max(n)`
    <int>    <dbl>    <dbl>
1       1       19       19
2       2       20       20
3       3       22       22
4       4       21       21
>
> otters$wts <- otters$N/otters$n
> otters %>% group_by(habitat) %>% summarize(min(wts), max(wts))  #check
# A tibble: 4 x 3
  habitat `min(wts)` `max(wts)`
    <int>      <dbl>      <dbl>
1       1       4.68       4.68
2       2       3.05       3.05
3       3       1.82       1.82
4       4       2.24       2.24
>
> design4.strat <- svydesign(id = ~1, fpc = ~N, weights = ~wts,
+     strata = ~habitat, data = otters)
```

(a)

```
> svytotal(~holts, design4.strat)
       total     SE
holts 984.71 73.921
```

$\hat{t}_{str} = 984.71$
$SE[\hat{t}_{str}] = 73.921$

(b)

The study area is divided into stratum based on the predominant terrain type. In other words, some of the sections may exhibit characteristics of more than one classification (for example we can see Cliffs and Agriculture in certain section). So classifying such sections heaviliy relies on the researchers' judgement, possibly resulting in selection bias. Also, some of the dens can be either abandoned or belonged to other animals, implying that the study is not also free from measurement error.

Problem 5

(a) Households that are not listed in the county's telephone number have no chance to be sampled, so it is not free from selection bias. Also, the survey is not free from nonresponse issue as well.

(b)

```
> radon <- read.csv("http://math.carleton.edu/kstclair/data/radon.csv")
> options(survey.lonely.psu = "remove")
>
> radon$wts <- radon$popsize/radon$sampsize
> head(radon %>% group_by(countyname) %>% summarize(min(wts), max(wts)))  #check
# A tibble: 6 x 3
  countyname `min(wts)` `max(wts)`
  <fct>           <dbl>      <dbl>
1 Aitkin          1350       1350
2 Anoka           1261.      1261.
3 Becker          2750       2750
4 Beltrami        1643.      1643.
```

```
5 Benton            2375       2375
6 Big Stone          967.       967.
>
> design5.strat <- svydesign(id = ~1, fpc = ~popsize, weights = ~wts,
+     strata = ~countyname, data = radon)
> svymean(~radon, design5.strat)
        mean     SE
radon 4.8986 0.1543
```

$\hat{\bar{y}}_{str} = 4.8986$ pCi/L
$SE[\hat{\bar{y}}_{str}] = 0.1543$ pCi/L

(c)

```
> htf4radon <- ifelse(radon$radon >= 4, 1, 0)
> update(design5.strat, htf4radon = htf4radon)
Stratified Independent Sampling design
update(design5.strat, htf4radon = htf4radon)
> svytotal(~htf4radon, design5.strat)
           total    SE
htf4radon 722781 28101
> confint(svytotal(~htf4radon, design5.strat), df = degf(design5.strat))
             2.5 %   97.5 %
htf4radon 667632.1 777930.4
```

$\hat{t}_{radon>4pCi/L} = 722781$ households
95% CI: (667632.1, 777930.4) households

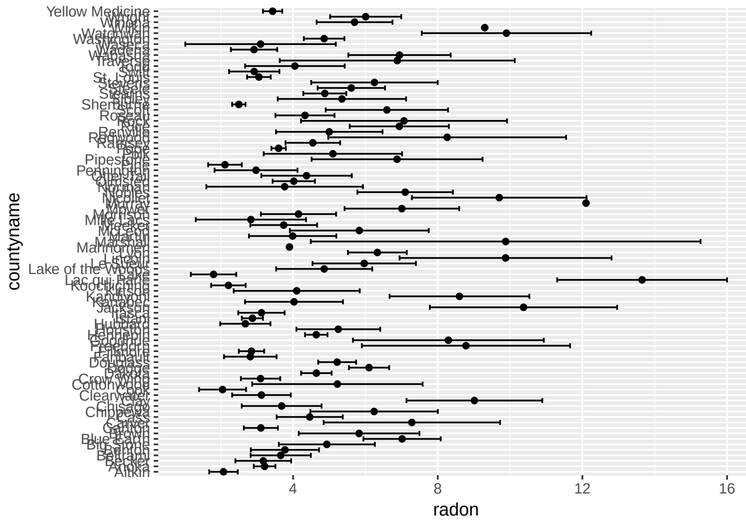
(d)

```
> head(radon %>% group_by(countyname) %>% arrange(sampsize))
# A tibble: 6 x 6
# Groups:   countyname [5]
  countynum countyname sampsize popsize radon   wts
      <int> <fct>         <int>   <int> <dbl> <dbl>
1        43 Mahnomen          1    1600   3.9  1600
2        51 Murray            1    3900  12.1  3900
3        84 Wilkin            1    2800   9.3  2800
4        16 Cook              2    1800   2.7   900
5        16 Cook              2    1800   1.4   900
6        23 Fillmore          2    7900   3.2  3950
```

Mahomen, Murray, and Wilkin county only have 1 home sampled.

(e)

```
> svyby.out <- svyby(~radon, ~countyname, design5.strat, svymean)
> ggplot(svyby.out, aes(y = countyname, x = radon)) + geom_point() +
+     geom_errorbarh(aes(xmax = radon + se, xmin = radon - se))
```



(f)

```
> head(svyby.out %>% arrange(se))
       countyname     radon        se
1        Mahnomen  3.900000 0.0000000
2          Murray 12.100000 0.0000000
3          Wilkin  9.300000 0.0000000
4       Sherburne  2.500000 0.1855169
5            Pope  3.600000 0.1999565
6 Yellow Medicine  3.433333 0.2665797
> head(svyby.out %>% arrange(desc(se)))
  countyname    radon       se
1   Marshall 9.877778 5.389193
2    Redwood 8.260000 3.288240
3   Traverse 6.880000 3.249807
4    Lincoln 9.875000 2.930659
5   Freeborn 8.780000 2.879234
6       Rock 7.066667 2.849216
```

The county with the largest SE for estimating the mean randon level of homes is Marshall county. The county with the smallest SE for estimating the mean randon level of homes is Sherburne county.

```
> radon %>% filter(countyname %in% c("Marshall", "Sherburne")) %>%
+     group_by(countyname) %>% summarize(n = n(), sd = sd(radon))
# A tibble: 2 x 3
  countyname     n     sd
  <fct>      <int>  <dbl>
1 Marshall       9 16.2
2 Sherburne      9  0.557
>
> 16.2/0.557
[1] 29.08438
>
> 5.389193/0.1855169
[1] 29.04961
```

$$SE[\hat{\bar{y}}] = \sqrt{(1 - \frac{n_h}{N_h})}\frac{s_h}{\sqrt{n_h}}$$

SE is proportional to sample standard deviation and inversely proportional to square root of sample size. 9 homes are sampled from both counties, but the sample standard deviation of Marshall county is almost 30 times that of Sherburne county, which is close to the ratio of standard errors between the two counties (let's ignore fpc as population size is big). In sum, significant difference in sample standard deviation between the two counties (while the number of the sample collected are the same) is biggest reason why the SE's of these counties are either biggest or smallest.

Problem 6

Revisit problem 2 above. Compute and interpret the design effect using the survey package for each of four estimates computed for exercise 9. Which estimate has the smallest DEff and which has the largest? Use the EDA (graphs) you produced for problem 2 to explain why these variables have the smallest and largest DEff.

```
> svymean(~acres87, design.strat, deff = T)
          mean     SE   DEff
acres87 298547  16293 0.8008
> svymean(~farms92, design.strat, deff = T)
           mean      SE   DEff
farms92 637.164  24.278 0.9751
> svymean(~largef92, design.strat, deff = T)
            mean      SE  DEff
largef92 56.6980  3.5577 0.865
> svymean(~smallf92, design.strat, deff = T)
            mean      SE   DEff
smallf92 56.8628  7.2014 0.9789
```

$$Var(\hat{t}_{str}) = (1 - \frac{n}{N})\frac{N}{n}(SSW + \sum_h s_h^2)$$

$$Var(\hat{t}_{ssr}) = (1 - \frac{n}{N})N^2\frac{SSB + SSW}{n(N-1)}$$

$$DEff = \frac{Var(\hat{t}_{str})}{Var(\hat{t}_{ssr})} = \frac{(1 - \frac{n}{N})\frac{N}{n}(SSW + \sum_h s_h^2)}{(1 - \frac{n}{N})N^2\frac{SSB + SSW}{n(N-1)}} = \frac{SSW + \sum_h s_h^2}{\frac{N(SSB + SSW)}{N-1}} = \frac{SSW + \sum_h s_h^2}{NS^2}$$

The equation above suggests that DEff gets bigger as the proportion of SSW + $\sum_h s_h^2$ over SST (approximately, as SST = $(N-1)S^2$) gets bigger. In other words, DEff gets larger if the relative portion of SSB in SST gets smaller (SSW + SSB = SST).
`acres87` has the smallest DEff, and `smallf92` has the largest DEff. EDB illustrates that the relatitve portion of SSB on SST in the case of `acres87` is big because we can observe huge difference in values among different strata; however, relative portion of SSB on SST is small in `smallf92` as we only see small difference in values among different strata.

```
> acres.aov <- aov(acres87 ~ region, data = agstrat)
> anova(acres.aov)
Analysis of Variance Table

Response: acres87
           Df     Sum Sq    Mean Sq F value    Pr(>F)
region      3 6.9528e+12 2.3176e+12  26.449 3.481e-15 ***
Residuals 296 2.5937e+13 8.7625e+10
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> smallf.aov <- aov(smallf92 ~ region, data = agstrat)
> anova(smallf.aov)
Analysis of Variance Table

Response: smallf92
           Df  Sum Sq Mean Sq F value   Pr(>F)
region      3  217384   72461  4.2593 0.005769 **
Residuals 296 5035679   17012
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA output also supports the claim that the relative size of SSB to SST in `acres87` case is significantly bigger than `smallf92` case. Approximately, SSB/SST in `acres87` case is $\frac{6.9528e+12}{(6.9528e+12)+(2.5937e+13)} = 0.2113968$ but only $\frac{217384}{217384+5035679} = 0.04138233$ in `smallf92` case.

Problem 7

Lohr textbook ch. 3 exercise 35 parts (a)-(d).

(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")
>
> table(pop$team)  # all roughly the same size populations

ANA ARI ATL BAL BOS CHA CHN CIN CLE COL DET FLO HOU KCA LAN MIL MIN MON
 26  28  28  25  27  26  29  27  28  27  26  26  25  27  24  25  25  28
NYA NYN OAK PHI PIT SDN SEA SFN SLN TBA TEX TOR
 29  26  27  25  27  26  27  28  26  26  27  26
> pop$logsal <- log(pop$salary)
>
> pop$N <- recode(pop$team, ANA = 26, ARI = 28, ATL = 28, BAL = 25,
+     BOS = 27, CHA = 26, CHN = 29, CIN = 27, CLE = 28, COL = 27,
+     DET = 26, FLO = 26, HOU = 25, KCA = 27, LAN = 24, MIL = 25,
+     MIN = 25, MON = 28, NYA = 29, NYN = 26, OAK = 27, PHI = 25,
+     PIT = 27, SDN = 26, SEA = 27, SFN = 28, SLN = 26, TBA = 26,
+     TEX = 27, TOR = 26)
>
> head(pop %>% group_by(team) %>% summarize(min(N), max(N)))
# A tibble: 6 x 3
  team  `min(N)` `max(N)`
  <fct>    <dbl>    <dbl>
1 ANA         26       26
2 ARI         28       28
3 ATL         28       28
```

```
4 BAL         25       25
5 BOS         27       27
6 CHA         26       26
>
> set.seed(30)  # put your favorite large integer here
> baseball.strat <- pop %>% group_by(team) %>% sample_n(size = 5) %>%
+     ungroup()
> str(baseball.strat)
Classes 'tbl_df', 'tbl' and 'data.frame':   150 obs. of  32 variables:
 $ team    : Factor w/ 30 levels "ANA","ARI","ATL",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ league  : Factor w/ 2 levels "AL","NL": 1 1 1 1 1 2 2 2 2 2 ...
 $ player  : Factor w/ 791 levels "aardsda0","abbotpa0",..: 154 291 252 270 198 118 669 125 725 115 ...
 $ salary  : int  375000 575000 9900000 301500 5750000 335000 500000 2750000 325000 325750 ...
 $ POS     : Factor w/ 9 levels "1B","2B","3B",..: 5 3 3 7 7 9 7 1 7 7 ...
 $ G       : int  108 46 58 5 1 154 27 20 17 69 ...
 $ GS      : int  27 22 19 0 33 125 18 1 0 0 ...
 $ InnOuts: int  743 640 495 263 625 3297 362 27 54 152 ...
 $ PO      : int  75 26 11 2 16 141 8 7 0 3 ...
 $ A       : int  1 46 27 5 24 383 21 1 4 10 ...
 $ E       : int  0 10 2 0 0 15 2 0 0 0 ...
 $ DP      : int  1 2 2 1 1 61 0 0 0 1 ...
 $ PB      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ GB      : int  108 46 58 5 1 154 27 20 17 69 ...
 $ AB      : int  285 114 207 0 2 564 31 27 0 1 ...
 $ R       : int  41 10 47 0 0 56 0 1 0 0 ...
 $ H       : int  79 23 52 0 0 148 4 3 0 0 ...
 $ SecB    : int  11 5 11 0 0 31 0 0 0 0 ...
 $ ThiB    : int  4 0 1 0 0 7 0 0 0 0 ...
 $ HR      : int  7 4 18 0 0 4 0 0 0 0 ...
 $ RBI     : int  34 13 42 0 0 49 1 1 0 0 ...
 $ SB      : int  18 1 2 0 0 3 0 0 0 0 ...
 $ CS      : int  3 1 3 0 0 3 0 0 0 0 ...
 $ BB      : int  46 7 31 0 0 31 0 1 0 0 ...
 $ SO      : int  54 30 52 0 0 59 14 5 0 1 ...
 $ IBB     : int  2 0 3 0 0 2 0 0 0 0 ...
 $ HPB     : int  0 0 3 0 0 2 0 0 0 0 ...
 $ SH      : int  1 0 0 0 0 12 1 0 0 0 ...
 $ SF      : int  5 0 1 0 0 4 0 0 0 0 ...
 $ GIDP    : int  2 3 6 0 0 11 0 0 0 0 ...
 $ logsal  : num  12.8 13.3 16.1 12.6 15.6 ...
 $ N       : num  26 26 26 26 26 28 28 28 28 28 ...
```

I used 5 samples from each strata using SRS (number of sample is constant (5) because population size is roughly same across the strata).

(b)

```
> baseball.strat$wts <- baseball.strat$N/5
>
> design7.strat <- svydesign(id = ~1, fpc = ~N, weights = ~wts,
+     strata = ~team, data = baseball.strat)
>
> svymean(~logsal, design7.strat)
         mean     SE
logsal 13.883 0.0879
```

```
> confint(svymean(~logsal, design7.strat), df = degf(design7.strat))
           2.5 %   97.5 %
logsal 13.70935 14.05757
```

$\hat{\bar{logsal}} = 13.883$
95% CI: (13.70935, 14.05757)

(c)

```
> pitcher <- ifelse(baseball.strat$POS == "P", 1, 0)
> update(design7.strat, pitcher = pitcher)
Stratified Independent Sampling design
update(design7.strat, pitcher = pitcher)
>
> svymean(~pitcher, design7.strat)
           mean     SE
pitcher 0.44918 0.0364
> confint(svymean(~pitcher, design7.strat), df = degf(design7.strat))
            2.5 %    97.5 %
pitcher 0.3771607 0.5212082
```

$\hat{p}_{pitcher} = 0.44918$
95% CI: (0.3771607, 0.5212082)

(d)

```
> knitr::kable(data.frame(logsal = c("Mean", "SE", "CI"), ssr = c(13.982,
+     0.095, "(13.79421 14.16963)"), str = c(13.883, 0.0879, "(13.70935, 14.05757)")))
```

logsal	ssr	str
Mean	13.982	13.883
SE	0.095	0.0879
CI	(13.79421 14.16963)	(13.70935, 14.05757)

```
>
> knitr::kable(data.frame(pitcher = c("Mean", "SE", "CI"), ssr = c(0.493333,
+     0.0369, "(0.420412591, 0.56625408)"), str = c(0.44918, 0.0364,
+     "(0.3771607, 0.5212082)")))
```

pitcher	ssr	str
Mean	0.493333	0.44918
SE	0.0369	0.0364
CI	(0.420412591, 0.56625408)	(0.3771607, 0.5212082)

The estimates from stratified sampling is smaller than those from SSR. Also the SE's are smaller than SSR. So, the CI's from stratified sampling is narrower than those from SSR.