Math 255 - Homework 5

Colin Pi

Due in class, Monday April 29

Problem 1

$N_1 = 1000,\ N_2 = 1000,\ S_1 = 10,\ S_2 = 20,\ c_1 = 2,\ c_2 = 1$

(a)

c({a_h},n) $= c_0 + \sum_{h=1}^{H} c_h(n \cdot a_h) = 0 + 2 \cdot 50 + 1 \cdot 50 = 150$

V({a_h},n) $= \sum_{h=1}^{H} (\frac{N_h}{N})^2 (1 - \frac{n \cdot a_h}{N_h}) \frac{S_h^2}{n \cdot a_h} = (\frac{1000}{2000})^2 (1 - \frac{50}{1000}) \frac{100}{50} + (\frac{1000}{2000})^2 (1 - \frac{50}{1000}) \frac{1}{50} = 0.47975$

(b)

$$a_1 = \frac{N_1 S_1/\sqrt{c_1}}{\sum_{k=1}^{H} N_k S_k/\sqrt{c_k}} = \frac{1000 \cdot 10/\sqrt{2}}{1000 \cdot 10/\sqrt{2} + 1000 \cdot 1/\sqrt{1}} = 0.8761007$$

$$a_2 = \frac{N_2 S_2/\sqrt{c_2}}{\sum_{k=1}^{H} N_k S_k/\sqrt{c_k}} = \frac{1000 \cdot 1/\sqrt{1}}{1000 \cdot 10/\sqrt{2} + 1000 \cdot 1/\sqrt{1}} = 0.1238993$$

(c)

$150 = \sum_{h=1}^{H} c_h(n \cdot a_h) = 2(n \cdot 0.8761007) + 1(n \cdot 0.1238993) = 1.876101n$

n $= \frac{150}{1.876101} = 79.95305 = 79$

(d)

V({a_h},n) $= \sum_{h=1}^{H} (\frac{N_h}{N})^2 (1 - \frac{n \cdot a_h}{N_h}) \frac{S_h^2}{n \cdot a_h} = (\frac{1000}{2000})^2 (1 - \frac{69}{1000}) \frac{100}{69} + (\frac{1000}{2000})^2 (1 - \frac{10}{1000}) \frac{1}{10} = 0.3620688$

The budget is the same with (a), but the variance is lower than (a).

(e)

```
> c <- c(2, 1)
> a <- c(0.8761007, 0.1238993)
> N <- c(1000, 1000)
> S <- c(10, 1)
>
> n_0 <- qnorm(0.975)^2 * sum((N/sum(N))^2 * (S^2/a))/1.36^2
> n_0
[1] 63.45656
>
> n <- 64/(1 + 64/2000)
> n
[1] 62.0155
>
> cost <- sum(c * round(63 * a))
> cost
[1] 118
```

```
>
> var <- sum((N/sum(N))^2 * (1 - round(63 * a)/N) * S^2/round(63 *
+     a))
> var
[1] 0.4605455
```

$$n_0 = \frac{1.96^2 \cdot \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{S_h^2}{a_h}}{1.36^2} = 63.45656 = 64$$
$$n = \frac{n_0}{1 + \frac{n_0}{2000}} = \frac{64}{1 + \frac{64}{2000}} = 62.0155 = 63$$
$$\text{c} = \sum_{h=1}^{H} c_h (n \cdot a_h) = 118$$
$$Var[\bar{y}_{str}] = \sum_h (\frac{N_h}{N})^2 (1 - \frac{n \cdot a_h}{N_h})(\frac{S_h^2}{n \cdot a_h}) = 0.4605455$$

Both budget and variance is smaller than the scenario (a)

Problem 2

Lohr textbook ch. 3 exercise 8.

$N_{phone} = 0.9N, N_{nonphone} = 0.1N, S_{phone} \approx S_{nonphone} = S$

(a)

$$a_{phone} = \frac{\frac{0.9NS}{\sqrt{30}}}{\frac{0.9NS}{\sqrt{30}} + \frac{0.1NS}{\sqrt{40}}} = \frac{0.9\sqrt{40}}{0.9\sqrt{40} + 0.1\sqrt{30}} = 0.9122215$$
$$a_{nonphone} = \frac{\frac{0.1NS}{\sqrt{40}}}{\frac{0.9NS}{\sqrt{30}} + \frac{0.1NS}{\sqrt{40}}} = \frac{0.1\sqrt{30}}{0.9\sqrt{40} + 0.1\sqrt{30}} = 0.08777855$$
$$n = \frac{C - c_0}{\sum_{h=1}^{H} c_h a_h} = \frac{20000 - 5000}{30 \cdot 0.9122215 + 40 \cdot 0.08777855} = 485.7861 \approx 485$$
$$n_{phone} = n \cdot a_{phone} = 442.42 \approx 442$$
$$n_{nonphone} = n \cdot a_{nonphone} = 42.5726 \approx 43$$

(b)

$$a_{phone} = \frac{\frac{0.9NS}{\sqrt{10}}}{\frac{0.9NS}{\sqrt{10}} + \frac{0.1NS}{\sqrt{40}}} = \frac{0.9\sqrt{40}}{0.9\sqrt{40} + 0.1\sqrt{10}} = 0.9473684$$
$$a_{nonphone} = \frac{\frac{0.1NS}{\sqrt{40}}}{\frac{0.9NS}{\sqrt{10}} + \frac{0.1NS}{\sqrt{40}}} = \frac{0.1\sqrt{10}}{0.9\sqrt{40} + 0.1\sqrt{10}} = 0.05263158$$
$$n = \frac{C - c_0}{\sum_{h=1}^{H} c_h a_h} = \frac{20000 - 5000}{10 \cdot 0.9473684 + 40 \cdot 0.05263158} = 1295.455 \approx 1295$$
$$n_{phone} = n \cdot a_{phone} = 1226.842 \approx 1227$$
$$n_{nonphone} = n \cdot a_{nonphone} = 68.1579 \approx 68$$

Problem 3

Lohr textbook ch. 3 exercise 22(a-b)

$c_1 = c_2, \dfrac{N_1}{N} = 0.4, n = 2000$

(a)

Since $c_1 = c_2$, it is Neyman allocation.

$a_1 = \dfrac{0.4N\sqrt{0.9 \cdot 0.1}}{0.4N\sqrt{0.9 \cdot 0.1} + 0.6N\sqrt{0.97 \cdot 0.03}} = 0.539684$

$a_2 = \dfrac{0.6N\sqrt{0.97 \cdot 0.03}}{0.4N\sqrt{0.9 \cdot 0.1} + 0.6N\sqrt{0.97 \cdot 0.03}} = 0.460316$

$n_1 = n \cdot a_1 = 1079.368 \approx 1079$

$n_2 = n \cdot a_2 = 920.632 \approx 921$

(b)

In that N is a huge number, let's assume $1 - \dfrac{n_h}{N_h} \approx 1$.

- Proportional Allocation

$V[\hat{p}_{str}] = \sum_{h=1}^{H} (\dfrac{N_h}{N})^2 \dfrac{p_h(1-p_h)}{n_h - 1} = 0.4^2(\dfrac{0.9 \cdot 0.1}{0.4 \cdot 2000 - 1}) + 0.6^2(\dfrac{0.97 \cdot 0.03}{0.6 \cdot 2000 - 1}) = 2.675981 \cdot 10^{-5}$

- Optimal Allocation

$V[\hat{p}_{str}] = \sum_{h=1}^{H} (\dfrac{N_h}{N})^2 \dfrac{p_h(1-p_h)}{n_h - 1} = 0.4^2(\dfrac{0.9 \cdot 0.1}{1079 - 1}) + 0.6^2(\dfrac{0.97 \cdot 0.03}{921 - 1}) = 2.675981 \cdot 10^{-5} = 2.474503 \cdot 10^{-5}$

- SRS

$\hat{p} = \sum_h \dfrac{N_h}{N} p_h = 0.4 \cdot 0.10 + 0.6 \cdot 0.03 = 0.058$

$V[\hat{p}_{ssr}] = (1 - \dfrac{n}{N})\dfrac{\hat{p}(1-\hat{p})}{n-1} \approx \dfrac{\hat{p}(1-\hat{p})}{n-1} = 2.733167 \cdot 10^{-5}$

Problem 4

Lohr textbook ch. 4 exercise 1.

(a)

x_i = television news broadcasts time in day i

y_i = time in television news broadcasts devoted to sports in day i

$\hat{p}_r = \hat{B} = \dfrac{t_y}{t_x}$

(b)

x_i = number of fish an angler i caught in August

y_i = number of fish caught by angler i in a lake in August

$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = \bar{y}\dfrac{\bar{x}_U}{\bar{x}}$, where B $= \dfrac{\bar{y}}{\bar{x}} = \dfrac{t_y}{t_x}$

(c)

x_i = spending of undergraduate student i in fall term

y_i = spending of undergraduate student i in fall term for textbook

$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = \bar{y}\dfrac{\bar{x}_U}{\bar{x}}$, where B $= \dfrac{\bar{y}}{\bar{x}} = \dfrac{t_y}{t_x}$

(d)

x_i = weight of chicken i
y_i = weight of usable meat of chicken i
$t_{yr} = \hat{B} t_x = \frac{\bar{y}}{\bar{x}} t_x = \bar{y} \frac{t_x}{\bar{x}}$

Problem 5

Lohr textbook ch. 4 exercise 2

```
> y <- c(10, 7, 13, 17, 8, 1, 15, 7, 4)
> x <- c(13, 7, 11, 12, 4, 3, 11, 3, 5)
> n <- 3
> N <- 9
```

(a)

```
> t.x <- sum(x)
> t.y <- sum(y)
> s.x <- sd(x)
> s.y <- sd(y)
> r <- cor(x, y)
> b <- t.y/t.x
```

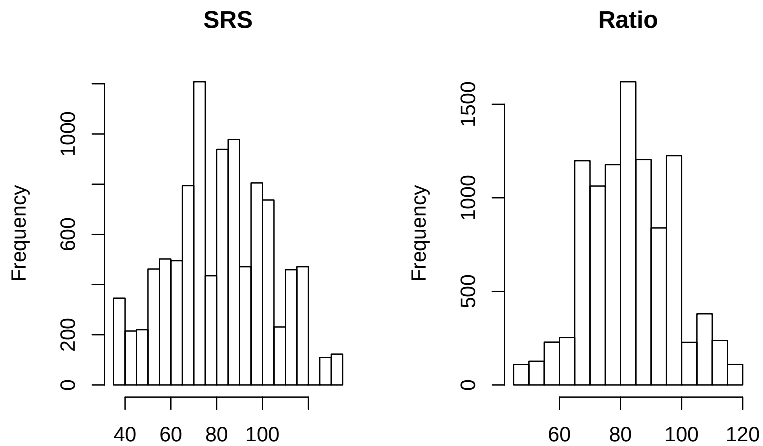
$t_x = 69$
$t_y = 82$
$S_x = 4.0926764$
$S_y = 5.1827706$
$R = 0.8152062$
$B = 1.1884058$

(b)

```
> reps <- 10000
> results <- data.frame(run = 1:reps, t.srs = NA, t.ratio = NA)
> set.seed(124)
>
> for (i in 1:reps) {
+     s <- sample(1:N, n, replace = F)
+     y.samp <- y[s]
+     x.samp <- x[s]
+     results$t.srs[i] <- N * mean(y.samp)
+     results$t.ratio[i] <- sum(y.samp)/sum(x.samp) * t.x
+ }
```

(c)

```
> par(mfrow = c(1, 2))
> hist(results$t.srs, main = "SRS", xlab = "")
> hist(results$t.ratio, main = "Ratio", xlab = "")
```




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

(d)

```r
> knitr::kable(data.frame(Statistics = c("Mean", "Variance", "Bias"),
+     yr = c(mean(results$t.ratio, na.rm = T), var(results$t.ratio,
+         na.rm = T), mean(results$t.ratio, na.rm = T) - t.y),
+     ysrs = c(mean(results$t.srs, na.rm = T), var(results$t.srs,
+         na.rm = T), mean(results$t.srs, na.rm = T) - t.y)))
```

Statistics	yr	ysrs
Mean	82.8569796	81.9306
Variance	194.7119383	475.0133
Bias	0.8569796	-0.0694

$\hat{t}_{yr}$ has expected value close to t_x, so the bias is close to 0 (0.8569796) . But the expected value of $\hat{t}_{y,\ SRS}$ is closer than that that of $\hat{t}_{yr}$ with bias of -0.0694. The variance of $\hat{t}_{yr}$ is almost 1/2 of $\hat{t}_{y,\ SRS}$.

(e)

$\bar{x}_U = 7.6666667$

```r
> (1 - 3/10) * (1/(3 * 7.666667)) * (b * s.x^2 - r * s.x * s.y)
[1] 0.07956049
```

$$Bias(\hat{\bar{y}}_r) = (1 - \frac{n}{N})\frac{1}{n\bar{x}_U}(BS_x^2 - RS_xS_y)$$

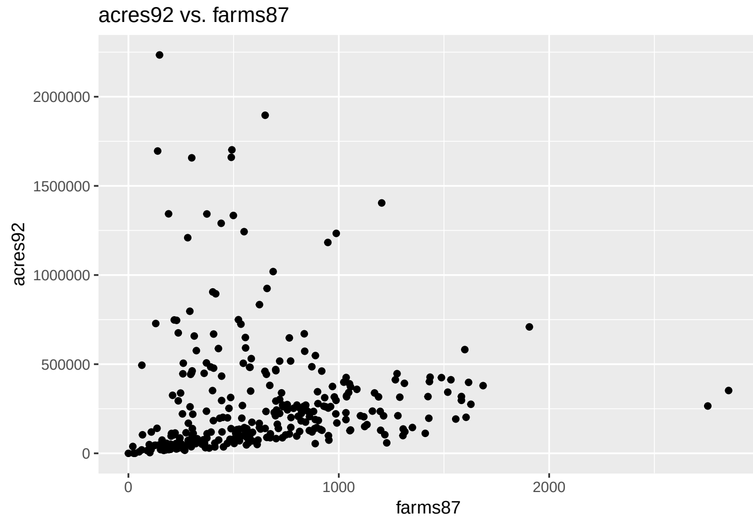
$Bias(\hat{\bar{y}}_r) = (1 - 3/10)\dfrac{1}{3 \cdot 6.9}(1.188406 \cdot 4.092676^2 - 0.8152062 \cdot 4.092676 \cdot 5.182771) = 0.08840055$
$Bias(\hat{t}_{yr}) \approx NBias(\hat{\bar{y}}_r) = 0.8840055$
It is close to the bias calculated in (c) (0.8569796).

Problem 6

Lohr textbook ch. 4 exercise 8(a,b,d). For part (d), ignore the regression estimator when answering the question.

(a)

```r
> ggplot(data = agsrs, aes(x = farms87, y = acres92)) + geom_point() +
+     ggtitle("acres92 vs. farms87")
```



(b)

y_i = number of acres devoted to farming in county i in 1992
x_i = number of farms in county i in 1987
$\hat{t}_{yr} = Bt_x = \frac{\bar{y}}{\bar{x}}t_x = \frac{t_x}{\bar{x}}\bar{y} = 960{,}155{,}061$

```r
> agsrs$n <- nrow(agsrs)
> agsrs$N <- 3078
> agsrs$wts <- agsrs$N/agsrs$n
> design6.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = agsrs)
>
> ratio.farms <- svyratio(~acres92, ~farms87, design6.srs)
> tx.farms <- 2087759
> ty.farms <- predict(ratio.farms, tx.farms)
> ty.farms
$total
          farms87
acres92 960155061

$se
         farms87
acres92 68446406
> confint(ratio.farms, df = degf(design6.srs)) * tx.farms
                      2.5 %     97.5 %
acres92/farms87 825457349 1094852773
```

(d)

```r
> ratio.acres <- svyratio(~acres92, ~acres87, design6.srs)
> tx.acres <- 964470625
> ty.acres <- predict(ratio.acres, tx.acres)
> ty.acres
$total
          acres87
acres92 951513191
```

```
$se
        acres87
acres92 5546162
> confint(ratio.acres, df = degf(design6.srs)) * tx.acres
                        2.5 %    97.5 %
acres92/acres87 940598734 962427648
>
> cor(agsrs$acres92, agsrs$acres87)
[1] 0.995806
> cor(agsrs$acres92, agsrs$farms87)
[1] 0.05964677
```

Ratio estimation with auxiliary variable `acres87` gives the most precision (SE = 5546162). SE[$\hat{B}$] decreases as $\hat{R}$ increases, or in other words, we can get more precise ratio estimation as the auxiliary variable gets more correlated with our variable of interest. As noted above, `acres92` exhibits a strong correlation with `acres87` ($\hat{R} = 0.995806$) while not with `farms87` ($\hat{R} = 0.05965$), and this difference accounts why using `acres87` gives more precise result than using `farms87`.

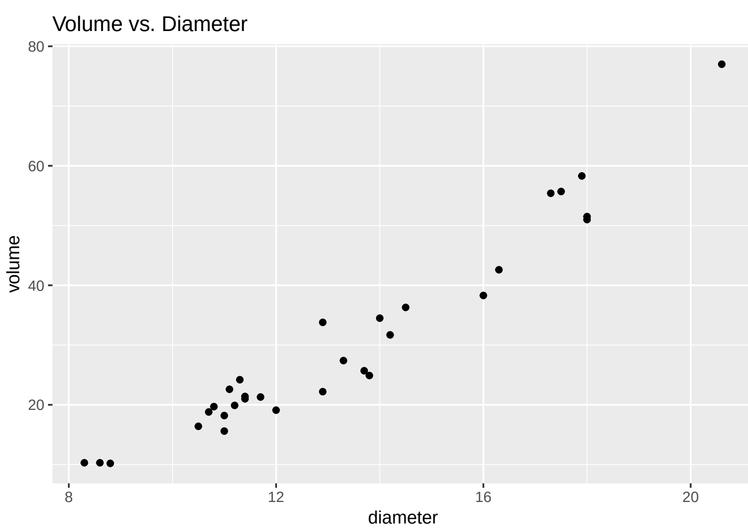
Problem 7

Lohr textbook ch. 4 exercise 10(a,b). Data is found:

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

(a)

```
> ggplot(data = cherry, aes(x = diameter, y = volume)) + geom_point() +
+     ggtitle("Volume vs. Diameter")
```



(b)

y_i = volume of black cherry tree i
x_i = diameter of black cherry tree i
$\hat{t}_{yr} = Bt_x = \frac{\bar{y}}{\bar{x}}t_x = \frac{t_x}{\bar{x}}\bar{y} = \frac{41,835}{13.24839} \cdot 30.17097 = 95272.16$
$SE[\hat{t}_{yr}] = N\sqrt{(1-\frac{n}{N})(\frac{\bar{x}_U}{\bar{x}})^2\frac{s_e^2}{n}} = 2967\sqrt{(1-\frac{31}{2967})(\frac{14.1001}{13.24839})^2\frac{s_e^2}{31}}$

$= 2967\sqrt{(1-\frac{31}{2967})(\frac{14.1001}{13.24839})^2\frac{94.05287}{31}} = 5471.434$
$s_e^2 = s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B}\hat{R}s_y s_x = 270.2028 + 2.277331^2 \cdot 9.847914 - 2 \cdot 2.277331 \cdot 0.9671194 \cdot 16.43785 \cdot 3.138139 = 94.05287$

95% CI: $\hat{t}_{yr} \pm qt_{0.975,\ df=30} \cdot SE[\hat{t}_{yr}] = (84098.0,\ 106446.3)$

```
> t.x.7 <- 41835
> b.7 <- sum(cherry$volume)/sum(cherry$diameter)
> t.y.hat.7 <- b.7 * t.x.7
> t.y.hat.7
[1] 95272.16
>
> xbar_u <- t.x.7/2967
>
> s.e.7_2 <- var(cherry$volume) + b.7^2 * var(cherry$diameter) -
+     2 * b.7 * cor(cherry$volume, cherry$diameter) * sd(cherry$volume) *
+         sd(cherry$diameter)
>
> se.7 <- 2967 * sqrt((1 - 31/2967) * (xbar_u/mean(cherry$diameter))^2 *
+     s.e.7_2/31)
>
> t.y.hat.7 - qt(c(0.975, 0.025), df = 30) * se.7
[1]  84098.0 106446.3
>
> cherry$n <- nrow(cherry)
> cherry$N <- 2967
> cherry$wts <- cherry$N/cherry$n
> design7.srs <- svydesign(id = ~1, fpc = ~N, weights = ~wts, data = cherry)
>
> ratio.cherry <- svyratio(~volume, ~diameter, design7.srs)
> tx.cherry <- t.x.7
> ty.cherry <- predict(ratio.cherry, tx.cherry)
> ty.cherry
$total
       diameter
volume 95272.16

$se
       diameter
volume 5471.434
> confint(ratio.cherry, df = degf(design7.srs)) * tx.cherry
                 2.5 %   97.5 %
volume/diameter 84098 106446.3
```

Problem 8

Assuming $\frac{\bar{x}_U}{\bar{x}}$,
$SE[\bar{y}_r] = \sqrt{(1-\frac{n}{N})(\frac{S_y^2 + B^2 S_x^2 - 2BRS_xS_y}{n})} < \sqrt{(1-\frac{n}{N})\frac{S_y^2}{n}}$
If we cancel out some terms,
$BS_x = \frac{\bar{y}}{\bar{x}}S_x < 2RS_y$

We know that CV(x) = $\frac{S_x}{\bar{x}}$, CV(y) = $\frac{S_y}{\bar{y}}$. If we rearrange the equation above,
$\frac{S_x}{\bar{x}} < 2R\frac{S_y}{\bar{y}} \rightarrow \frac{CV(x)}{2CV(y)} < R$