STAT 3014/3914

Semester 2	Applied Statistics	2015

Solution to Tutorial 10

1. We have $N = 5$ and $n = 2$.

   (a) The population mean μ and population variance σ^2 are

   $$\begin{aligned} \mu &= \frac{1}{5}(0+3+3+4+5) = 3 \\ \sigma^2 &= \frac{1}{5}[(0-3)^2+(3-3)^2+(3-3)^2+(4-3)^2+(5-3)^2] = 2.8 \end{aligned}$$

   (b) We have 10 different samples:

$\{y_i, y_j\}$	$\{0,3\}$	$\{0,3\}$	$\{0,4\}$	$\{0,5\}$	$\{3,3\}$	$\{3,4\}$	$\{3,5\}$	$\{3,4\}$	$\{3,5\}$	$\{4,5\}$
$\bar{y}$	1.5	1.5	2	2.5	3	3.5	4	3.5	4	4.5

   The mean and variance of the sample mean $\bar{y}$ are

   $$\begin{aligned} E(\bar{y}) &= \frac{1}{10}(1.5+\cdots+4.5) = 3 \\ Var(\bar{y}) &= \frac{1}{10}[(1.5-3)^2+\cdots+(4.5-3)^2] = 1.05 \end{aligned}$$

   Moreover

   $$\begin{aligned} \left(1-\frac{n}{N}\right)\frac{S^2}{n} &= \left(1-\frac{2}{5}\right)\frac{3.5}{2} = 1.05 \\ \text{where} \qquad S^2 &= \frac{N}{N-1}\sigma^2 = \frac{5}{4}(2.8) = 3.5. \end{aligned}$$

2. We have $N = 1000$, $S \approx \sigma = 6$, $\delta_\mu = 0.5$ and $z_{0.05} = 1.645$.

   (a) With replacement or ignoring fpc,

   $$n \geq \frac{z_{\alpha/2}^2 S^2}{\delta_\mu^2} = \frac{1.645^2 6^2}{0.5^2} = 389.6676.$$

   Take $n = 390$.

   (b) Without replacement,

   $$n \geq \frac{NS^2}{N\delta_\mu^2/z_{\alpha/2}^2 + S^2} = \frac{1000(6^2)}{1000(0.5^2)/1.645^2 + 6^2} = 280.4035.$$

   Take $n = 281$. It is smaller because sampling without replacement produces more informative samples and hence more precise estimates.

(c) For large sample size n $(n = 40)$, and small to moderate sampling fraction $f = \frac{40}{1000}$, we have the approximation

$$(\bar{y} - \mu)/\sqrt{\operatorname{Var} \bar{y}} \sim \mathcal{N}(0, 1).$$

Now $S^2 = \dfrac{N}{N-1}\sigma^2$, $\bar{y} - \mu = -1$ and

$$\operatorname{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = \left(\frac{N-n}{N}\frac{N}{N-1}\right)\frac{\sigma^2}{n} = \left(\frac{1000-40}{1000-1}\right)\frac{6^2}{40} = 0.864865.$$

Hence

$$\Pr\left(z < \frac{\bar{y} - \mu}{\sqrt{\operatorname{Var}(\bar{y})}}\right) = \Pr\left(z < \frac{-1}{\sqrt{0.864865}}\right) = \Pr(z < -1.07529) = 0.1411225$$

3. Let Y_i denote the number of decayed teeth and C_1 the subpopulation of children who have at least one decayed teeth. Let

$$Y_i' = \begin{cases} Y_i & \text{if } i \in C_1 \\ 0 & \text{if } i \notin C_1 \end{cases}$$

We have $N = 200$, $N_1 = 140$, $N_2 = 60$, $n = 20$, $n_1 = 12$, $n_2 = 8$,

$Y_i = Y_i'$, $\quad \displaystyle\sum_{i=1}^{n} y_i^2 = \sum_{i \in C_1} y_i^2 = 4(1^2) + 2(2^2) + \cdots + 1(10^2) = 252,$

$\displaystyle\bar{y}' = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{42}{20} = 2.1, \qquad \bar{y}_1 = \frac{1}{n_1}\sum_{i \in C_1} y_i = \frac{n}{n_1}\bar{y}' = \frac{42}{12} = 3.5.$

The sample variances are

$$s_1'^2 = \left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 /n\right] /(n-1) = [252 - 20 \times 2.1^2]/19 = 8.621$$

$$s_1^2 = \left[\sum_{i \in C_1} y_i^2 - \left(\sum_{i \in C_1} y_i\right)^2 /n_1\right] /(n_1 - 1) = [252 - 12 \times 3.5^2]/11 = 9.545$$

The estimates of the total number of decayed teeth are respectively

$$\begin{aligned} \widehat{Y} &= N\bar{y}' = 200(2.1) = 420 \\ \widehat{Y}_{pst,1m1} &= N_1 \frac{N}{N_1}\bar{y}' = 140(3) = 420 \quad \text{(3 is mean est. in } C_1) \\ \widehat{Y}_{pst,1m2} &= N_1 \frac{n}{n_1}\bar{y}' = N_1 \bar{y}_{C_1} = 140(3.5) = 490 \end{aligned}$$

which are done in lecture. The first two estimates $\widehat{Y} = \widehat{Y}_{pst,1m1} = 420$ because there are no decayed teeth in the other subpopulation defined by C_2.

We have

$$\begin{aligned}
\text{var}(\widehat{Y}_{pst,1m1}) &= N^2\left(1-\frac{n}{N}\right)\frac{{s_1'}^2}{n} = 200^2\left(1-\frac{20}{200}\right)\frac{8.621}{20} = 15517.89 \\
\text{var}(\widehat{Y}_{pst,1m2}) &= N^2 W_l\left(1-\frac{n}{N}\right)\frac{{s_1}^2}{n} = 200^2\frac{140}{200}\left(1-\frac{20}{200}\right)\frac{9.545}{20} = 12027.27
\end{aligned}$$

4. (a) Annual profit: $N_1 = 480$, $N_2 = 320$, $S_1 = 211$ and $S_2 = 708$.

$$\begin{aligned}
\sum_{l=1}^{L}\frac{N_l s_l}{\sqrt{C_l}} &= \frac{480\times 211}{\sqrt{0.08}} + \frac{320\times 708}{\sqrt{0.1}} = 1,074,524.5 \\
n_1 &= n\left(\frac{\frac{N_1 s_1}{\sqrt{C_1}}}{\sum_{i=1}^{L}\frac{N_i s_i}{\sqrt{C_i}}}\right) = n\left[\frac{358,078.9}{1,074,524.5}\right] = 0.333n \\
n_2 &= n\left(\frac{\frac{N_2 s_2}{\sqrt{C_2}}}{\sum_{i=1}^{L}\frac{N_i s_i}{\sqrt{C_i}}}\right) = n\left[\frac{716,445.6}{1,074,524.5}\right] = 0.667n
\end{aligned}$$

(b) Cost constraint:

$$\begin{aligned}
n_1C_1 + n_2C_2 \le 9.3 &\Rightarrow 0.333n\times 0.08 + 0.667n\times 0.1 \le 9.3 \\
&\Rightarrow 0.09334n \le 9.3 \\
&\Rightarrow n \le 99.6
\end{aligned}$$

Take $n = 99$. $n_1 = 0.333\times 99 = 33$ and $n_2 = 0.667\times 99 = 66$.

(c) The estimate of the mean annual profit of trading firms and the 95% confidence interval of the estimate are

$$\begin{aligned}
\widehat{\overline{Y}}_{st} &= \sum_{l=1}^{3}\frac{N_1}{N}\bar{y}_l \\
&= \frac{480}{800}\times 540 + \frac{320}{800}\times 1,730 = 1,016 \text{ thousands.} \\
\text{Var}(\widehat{\overline{Y}}_{st}) &= \sum_{l=1}^{L}\frac{N_l}{N}^2\left(1-\frac{n_l}{N_l}\right)\frac{S_l^2}{n_l} \\
&= \left(\frac{480}{800}\right)^2\left(1-\frac{33}{480}\right)\frac{211^2}{33} + \left(\frac{320}{800}\right)^2\left(1-\frac{66}{320}\right)\frac{708^2}{66} \\
&= 452.293 + 964.553 = 1,416.846 \\
\text{SE}(\widehat{\overline{Y}}_{st}) &= \sqrt{1416.846} = 37.64 \\
95\%\text{ CI for } \overline{Y}_{st} &= (1,016 - 1.96\times 37.64,\ 1,016 + 1.96\times 37.64) = (942.22, 1,089.77).
\end{aligned}$$

Extra exercise

1. (a) $p = 0.18 + 0.19 + 0.04 + 0.22 = 0.63$, $n = 82$ and $N = 1400$. The population proportion having *at least* 50-50 chance of advertising in the future:

$$\begin{aligned}
\widehat{P} &= p = 0.63 \\
\mathrm{var}(\widehat{P}) &= \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n-1} = \left(1 - \frac{82}{1400}\right)\frac{0.63 \times 0.37}{81} = 0.002709222 \\
\mathrm{se}(\widehat{P}) &= \sqrt{0.002709222} = 0.052050189 \\
95\% \text{ CI for } P &= (\widehat{P} - z_{0.025} \times se(\widehat{P}),\ \widehat{P} + z_{0.025} \times se(\widehat{P})) \\
&= (0.63 - 1.96 \times 0.052,\ 0.63 + 1.96 \times 0.052) = (0.528, 0.732)
\end{aligned}$$

(b) $p = 0.15 + 0.35 + 0.05 + 0.35 = 0.90$, $n = 46$ and $N = 1400$. The population proportion having *at least* 50-50 chance of advertising again:

$$\begin{aligned}
\widehat{P} &= p = 0.90 \\
\mathrm{var}(\widehat{P}) &= \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n-1} = \left(1 - \frac{46}{1400}\right)\frac{0.90 \times 0.10}{45} = 0.001934286 \\
\mathrm{se}(\widehat{P}) &= \sqrt{0.001934286} = 0.04398052
\end{aligned}$$

(c) $p = 0.63$, $n = 82$ and $N = 1400$. The total number of accountants out of the $N = 1400$ accountants having *at least* 50-50 chance of advertising in the future

$$\begin{aligned}
N\widehat{P} &= Np = 1400 \times 0.63 = 882 \\
\mathrm{var}(N\widehat{P}) &= N^2\left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n-1} = 1400^2\left(1 - \frac{82}{1400}\right)\frac{0.63 \times 0.37}{81} = 5310.07512 \\
\mathrm{se}(\widehat{P}) &= \sqrt{5310.07512} = 72.87026225
\end{aligned}$$

2. (a) We have $n = 4$, $N = 10$, $\sum_i y_i^2 = 5.5^2 + 8^2 + 12^2 + 13^2 = 407.25$.

$$\begin{aligned}
\widehat{\overline{Y}}_1 &= \bar{y} = \frac{1}{4}(5.5 + 8 + 12 + 13) = 9.625 \\
s_y^2 &= \frac{1}{n-1}\left(\sum_i y_i^2 - n\overline{y}^2\right) = \frac{1}{3}(407.25 - 4 \times 9.625^2) = 12.22916667 \\
\mathrm{var}(\widehat{\overline{Y}}_1) &= \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \left(1 - \frac{4}{10}\right)\frac{12.22916667}{4} = 1.834375001
\end{aligned}$$

(b) Now we have y_3 included and $n = 3$, $N = 9$, $\bar{y} = \frac{1}{3}(8 + 12 + 13) = 11$ and

$\sum_i y_i^2 = 8^2 + 12^2 + 13^2 = 377.$

$$\begin{aligned}
\widehat{\overline{Y}}_2 &= \frac{1}{4}(y_3 + 3\bar{y}) = \frac{1}{4}(5.5 + 3 \times 11) = 9.625 \\
s_y'^2 &= \frac{1}{n'-1}(\sum_i y_i^2 - n\bar{y}'^2) = \frac{1}{2}(377 - 3 \times 11^2) = 7 \\
\mathrm{var}(\widehat{\overline{Y}}_2) &= \left(\frac{1}{4}\right)^2 \mathrm{var}(y_3) + \left(\frac{3}{4}\right)^2 \mathrm{var}(\bar{y}) \\
&= \frac{9}{16}\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \frac{9}{16}\left(1 - \frac{3}{9}\right)\frac{7}{3} = 0.875 < 1.834375001 = \mathrm{var}(\widehat{\overline{Y}}_1)
\end{aligned}$$

because one point is fixed and hence the variability is less.

(c) We have $n_l = 2$, $n_h = 2$, $N_l = 2$, $N_h = 5$, $\bar{y}_l = \frac{1}{2}(5.5+8) = 6.75$, $\bar{y}_h = \frac{1}{2}(12+13) = 12.5$, and $\sum_i y_i^2 = 8^2 + 12^2 + 13^2 = 377$.

$$\begin{aligned}
\bar{y}_l &= \frac{1}{2}(5.5 + 8) = 6.75 \\
s_l^2 &= \frac{1}{n_l - 1}(\sum_i y_{li}^2 - n_l\bar{y}_l^2) = \frac{1}{1}(5.5^2 + 8^2 - 2 \times 6.75^2) = 3.125 \\
\bar{y}_h &= \frac{1}{2}(12 + 13) = 12.5 \\
s_h^2 &= \frac{1}{n_h - 1}(\sum_i y_{hi}^2 - n_h\bar{y}_h^2) = \frac{1}{1}(12^2 + 13^2 - 2 \times 12.5^2) = 0.5 \\
\widehat{\overline{Y}}_3 &= \frac{1}{3}\bar{y}_l + \frac{2}{3}\bar{y}_h = \frac{1}{3} \times 6.75 + \frac{2}{3} \times 12.5 = 10.583 \\
\mathrm{var}(\widehat{\overline{Y}}_3) &= \left(\frac{1}{3}\right)^2 \mathrm{var}(\bar{y}_l) + \left(\frac{2}{3}\right)^2 \mathrm{var}(\bar{y}_h) \\
&= \frac{1}{9}\left(1 - \frac{n_l}{N_l}\right)\frac{s_l^2}{n_l} + \frac{4}{9}\left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h} \\
&= \frac{1}{9}\left(1 - \frac{2}{5}\right)\frac{3.125}{2} + \frac{4}{9}\left(1 - \frac{2}{5}\right)\frac{0.5}{2} \\
&= 0.1042 + 1.0667 = 0.1709 < 1.834375001 = \mathrm{var}(\widehat{\overline{Y}}_1)
\end{aligned}$$

because the variability in each subsample is less.

3. (a) The estimate of the total number of *female* students N_1P_1 who regard 'clothing'

as important is

$$\begin{aligned}
N_1\widehat{P}_1 &= N_1 p_1 = 4800 \times \frac{9}{15} = 2880 \\
\mathrm{se}(N_1\widehat{P}_1) &= N\sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{p_1(1-p_1)}{n_1-1}} \\
&= 4800\sqrt{\left(1-\frac{15}{4800}\right)\frac{\frac{9}{15}\frac{6}{16}}{14}} = 627.484775
\end{aligned}$$

(b) The estimate of the average amount spent on clothing by male students $\overline{Y}_1$ and the standard error estimate are

$$\begin{aligned}
\widehat{\overline{Y}}_1 &= \overline{y}_1 = \frac{\sum_i y_{1i}}{n_1} = \frac{412}{12} = 34.33 \\
s_{y1}^2 &= \frac{\sum_i y_{1i}^2 - n_1\overline{y}_1^2}{n_1-1} = \frac{16390 - 12 \times 34.33^2}{11} = 204.0606 \\
\mathrm{se}(\widehat{Y}_1) &= \sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{s_{y1}^2}{n_1}} = \sqrt{\left(1-\frac{12}{4400}\right)\frac{204.0606}{12}} = 4.118091
\end{aligned}$$

The estimate of the average amount spent on clothing by female students $\overline{Y}_2$ and the standard error estimate are

$$\begin{aligned}
\widehat{\overline{Y}}_2 &= \overline{y}_2 = \frac{\sum_i y_{2i}}{n_2} = \frac{824}{12} = 54.93 \\
s_{y2}^2 &= \frac{\sum_i y_{2i}^2 - n_2\overline{y}_2^2}{n_2-1} = \frac{47604 - 15 \times 54.93^2}{14} = 167.0667 \\
\mathrm{se}(\widehat{Y}_2) &= \sqrt{\left(1-\frac{n_2}{N_2}\right)\frac{s_{y2}^2}{n_2}} = \sqrt{\left(1-\frac{15}{4800}\right)\frac{167.0667}{15}} = 3.337331
\end{aligned}$$

(c) The estimate of the gender difference in the average amount spent on clothing $\overline{Y}_1 - \overline{Y}_2$ and the standard error estimate are

$$\begin{aligned}
\widehat{\overline{Y}}_2 - \widehat{\overline{Y}}_1 &= \overline{y}_2 - \overline{y}_1 = 54.93 - 34.33 = 20.6 \\
\mathrm{se}(\widehat{Y}) &= \sqrt{(1-\frac{n_1}{N_1})\frac{s_{y1}^2}{n_1} + (1-\frac{n_2}{N_2})\frac{s_{y2}^2}{n_2}} = \sqrt{\mathrm{var}(\widehat{Y}_1) + \mathrm{var}(\widehat{Y}_1)} \\
&= \sqrt{4.118091^2 + 3.337331^2} = 5.30061 \\
95\%\ \text{CI for } \overline{Y}_1 - \overline{Y}_2 &= (20.6 - 1.96 \times 5.30061,\ 20.6 + 1.96 \times 5.30061) \\
&= (10.21081,\ 30.98919)
\end{aligned}$$

Since the CI for $\overline{Y}_1 - \overline{Y}_2$ does not contain 0, there is a significant gender difference in the average amount spent on clothing for the students in a certain university.

(d) The estimate the overall average amount spent on clothing $\overline{Y}$ using the stratified random sample estimator and the standard error estimate are

$$\begin{aligned}
\widehat{\overline{Y}}_{st} &= \frac{N_1}{N}\overline{y}_1 + \frac{N_1}{N}\overline{y}_2 = \frac{4400}{9200} \times \frac{412}{12} + \frac{4800}{9200} \times \frac{824}{15} = 45.08116 \\
\mathrm{se}(\widehat{\overline{Y}}_{st}) &= \sqrt{\frac{N_1^2}{N^2}\left(1 - \frac{n_1}{N_1}\right)\frac{s_{y1}^2}{n_1} + \frac{N_2^2}{N^2}(1 - \frac{n_2}{N_2})\frac{s_{y2}^2}{n_2}} \\
&= \sqrt{\frac{4400^2}{9200^2}\left(1 - \frac{12}{4400}\right)\frac{204.0606}{12} + \frac{4800^2}{9200^2}\left(1 - \frac{15}{4800}\right)\frac{167.0667}{15}} \\
&= 2.627047
\end{aligned}$$

4. (a) $\delta_p = 0.03$. Take $s_p^2 = p(1-p) = 0.3(1-0.3) = 0.21$. Note that $s_p^2 = 0.2(1-0.2) = 0.16$ will give a smaller n. Hence

$$n = \frac{z_{\alpha/2}^2 s_p^2}{\delta_p} = \frac{1.96^2 \times 0.21}{0.03^2} = 896.37$$

Take $n = 897$.

(b) For a more conservative estimate, take $s_p^2 = 0.5(1 - 0.5) = 0.25$.

$$n = \frac{z_{\alpha/2}^2 s_p^2}{\delta_p} = \frac{1.96^2 \times 0.25}{0.03^2} = 1067.11$$

Take $n = 1068$.

(c) $\delta_p = 1.96 \times \mathrm{se}(p) = 1.96 \times 0.015 = 0.0294$.

$$n = \frac{z_{\alpha/2}^2 s_p^2}{\delta_p} = \frac{1.96^2 \times 0.25}{0.0294^2} = 1111.11$$

Take $n = 1112$.

(d) $\delta_p = 0.03$. Take $s_{p'-p}^2 = s_{p'}^2 + s_p^2 = 0.3 \times 0.7 + 0.4 \times 0.6$.

$$n = \frac{z_{\alpha/2}^2 s_{p'-p}^2}{\delta_{p'-p}} = \frac{1.96^2 \times (0.3 \times 0.7 + 0.4 \times 0.6)}{0.03^2} = 1922.3$$

Take $n = 1923$. A more conservative estimate is

$$n = \frac{z_{\alpha/2}^2 s_{p'-p}^2}{\delta_{p'-p}} = \frac{1.96^2 \times 2 \times 0.5 \times 0.5}{0.03^2} = 2134$$