Referred Solution to Assignment 03

Part I : Problems from Chapter 3

3.2. (a) (i) Stratum 1:

Unit number	(1,2)	(1,3)	(1,8)	(2,3)	(2,8)	(3,8)
y	(1,2)	(1,4)	(1,8)	(2,4)	(2,8)	(4,8)
$\hat{t}_1 = 4\bar{y}_{1S}$	6	10	12	18	20	24
$p(\hat{t}_1)$	1/6	1/6	1/6	1/6	1/6	1/6

(ii) Stratum 2:

Unit number	(4,5)	(4,6)	(4,7)	(5,6)	(5,7)	(6,7)
y	(4,7)	(4,7)	(4,7)	(7,7)	(7,7)	(7,7)
$\hat{t}_2 = 4\bar{y}_{2S}$	22	22	22	28	28	28
$p(\hat{t}_2)$	1/6	1/6	1/6	1/6	1/6	1/6

Additionally, we also can have the following result:

y	(4,7)	(7,7)
$\hat{t}_2 = 4\bar{y}_{2S}$	22	28
$p(\hat{t}_2)$	1/2	1/2

(b) For example,

Unit number	(1,2);(4,5)	(1,2);(4,6)	(1,2);(4,7)
$\hat{t}_{\text{str}} = \sum_h \hat{t}_{hS}$	28	28	28
$p(\hat{t}_{\text{str}})$	1/36	1/36	1/36

So, we can find the sampling distribution of $\hat{t}_{\text{str}}$ as follows:

$\hat{t}_{\text{str}} = \sum_h \hat{t}_{hS}$	28	32	34	38	40	42	46	48	52
$p(\hat{t}_{\text{str}})$	1/12	1/12	1/6	1/12	1/6	1/12	1/6	1/12	1/12

(c) $E(\hat{t}_{\text{str}}) = \sum \hat{t}_{\text{str}} p(\hat{t}_{\text{str}}) = 40$; $Var(\hat{t}_{\text{str}}) = E(\hat{t}^2_{\text{str}}) - [E(\hat{t}_{\text{str}})]^2 = 47.33$.
Recall Example 2.2 and 2.3, we have $E(\hat{t}_{\text{SRS}}) = 40$ and $Var(\hat{t}_{\text{SRS}}) = 54.86$.
Therefore, the stratified sampling produces much precision.

3.5. (a) Scholars in selected ACLS societies in seven disciplines.

(b) $\hat{p}_{\text{str}} = \sum_h \frac{N_h}{N}\hat{p}_h = 0.3337$,

$$S.E(\hat{p}_{\text{str}}) = \sqrt{\sum_h \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}} = 0.0079.$$

3.6. (a) Since the sample size is fixed and $S_1 = 2S_2 = 2S_3$ is provided, we can use the optimal allocation, Neyman allocation.

Let $S_2 = S_3 = S$ and $S_1 = 2S$. Based on $n_{h,\text{Neyman}} = \dfrac{N_h S_h}{\sum_l N_l S_l} n$, we can obtain

Stratum	house	apt.	condo
Sample Size	504	324	72

(b) (i) $p = p_{\text{str}} = \sum_h \dfrac{N_h}{N} p_h = 0.3033$.

(ii) Based on the proportional allocation, $\dfrac{N_h}{N} = \dfrac{n_h}{n}$, we can determine the sample size of each stratum as follows:

Stratum	(1) house	(2) apt.	(3) condo
Sample Size	350	450	100

$$V_{\text{prop}}(\hat{p}_{\text{str}}) = \sum_h \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{N_h}{N_h - 1} \frac{p_h(1-p_h)}{n_h} \approx 0.000213;$$

$$V_{\text{SRS}}(\hat{p}_{\text{SRS}}) = \left(1 - \frac{n}{N}\right) \frac{N}{N-1} \frac{p(1-p)}{n} \approx 0.000233.$$

So, $\dfrac{V_{\text{prop}}(\hat{p}_{\text{str}})}{V_{\text{SRS}}(\hat{p}_{\text{SRS}})} \approx 0.914$.

3.7. (a)

Stratum	Biological	Physical	Social	Humanities
N_h	102	310	217	178
n_h	7	19	13	11
$\hat{t}_h$	320.57143	652.63158	267.07692	80.90909
s_h^2	6.80952	8.21053	4.35897	0.87273

$$\hat{t}_{\text{str}} = \sum_h \hat{t}_h = \sum_h N_h \bar{y}_h = 1321.189,$$

$$S.E(\hat{t}_{\text{str}}) = \sqrt{\sum_h \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}} = 256.146.$$

(b) $\hat{t}_{\text{SRS}} = N\bar{y} = 1436.46$,

$$S.E(\hat{t}_{\text{SRS}}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}} = 296.504.$$

So, the standard error of the estimate from SRS is greater than that from the stratified sampling.

(c)

Stratum	Biological	Physical	Social	Humanities
$\hat{p}_h$	0.14286	0.52632	0.69231	0.72727

$$\hat{p}_{\text{str}} = \sum_h \frac{N_h}{N} \hat{p}_h = 0.5668,$$

$$S.E(\hat{p}_{\text{str}}) = \sqrt{\sum_h \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}} = 0.0658.$$

(d) From the sampled result, we know that the distributions of referred publications are quite different among these four areas. So, the stratified sampling would decrease the standard error of estimates, as shown in (b). In other words, it increases precision.

3.11. (a)

Stratum	Zone 1	Zone 2	Zone 3
N_h	68	84	48
n_h	17	112	11
$\bar{y}_h$	1.76471	4.41667	10.54546
s_h^2	3.31618	11.53788	46.07273

$\hat{t}_{\text{str}} = \sum_h \hat{t}_h = \sum_h N_h \bar{y}_h = 997.1818,$

$S.E(\hat{t}_{\text{str}}) = \sqrt{\sum_h \left(1 - \dfrac{n_h}{N_h}\right) N_h^2 \dfrac{s_h^2}{n_h}} = 118.0264.$

(b) If estimate the total number of breathing holes, we may prefer the optimal allocation; if compare the density in three zones, we may prefer equal sample size for each zone.

3.15. (a) (i) Advantage: it is easy to operate.

(ii) Disadvantage: it may not achieve the highest level of precision when estimating population parameters.

(b) (i) $\bar{y}_{\text{U}}$ can be estimated by $\bar{y}_{\text{str}} = \sum_h \dfrac{N_h}{N}\bar{y}_h = 3.93855.$

(ii) Since $n - H = 44$ is greater than 30, we may not consider the Satterthwaite correction. We form 95% confidence interval for $\bar{y}_{\text{U}}$ by

$$\bar{y}_{\text{str}} \pm z_{\alpha/2} S.E(\bar{y}_{\text{str}}) \Rightarrow (3.9201, 3.9570),$$

where $S.E(\bar{y}_{\text{str}}) = \sqrt{\sum_h \left(1 - \dfrac{n_h}{N_h}\right)\left(\dfrac{N_h}{N}\right)^2 \dfrac{s_h^2}{n_h}}.$

(c) We should use ANOVA test to test $H_0 : \bar{y}_{1\text{U}} = \bar{y}_{2\text{U}} = \bar{y}_{3\text{U}}$ and refer to page 86 in the textbook. When using s_h^2, $\bar{y}_h$, and $\bar{y}_{\text{str}}$ to estimate S_h^2, $\bar{y}_{h\text{U}}$, and $\bar{y}_{\text{U}}$, respectively, we have the ANOVA table as follows:

Source	df	SS	MS	F	p-value
Between strata	2	0.0447	0.02233	5.89	0.0028
Within strata	1045	5.3256	0.00379		

At significant level α=0.05, we reject H_0. It is evidence that prices are different in the three strata.

3.21. We construct an example where there is much less or no variation between strata by $SSB < \sum_h \left(1 - \dfrac{N_h}{N}\right) S_h^2$. For instance, there is a population with two strata where one stratum is (1,4) and the other is (2,3) and then, we sample one from each stratum.

Stratum	N_h	$\bar{y}_{h\text{U}}$	S_h^2	n_h
1	2 (1,4)	2.5	2.12	1
2	2 (2,3)	2.5	0.71	1

$\bar{y}_\text{U} = \frac{1}{4}(1+2+3+4) = 2.5 \Rightarrow \bar{y}_\text{U} = \bar{y}_{1\text{U}} = \bar{y}_{2\text{U}} = 2.5.$

So, $SSB = \sum_h N_h(\bar{y}_{h\text{U}} - \bar{y}_\text{U})^2 = 0$ and $0 < \sum_h \left(1 - \frac{N_h}{N}\right) S_h^2$.

3.22. In this problem, we have the following information:

Stratum	N_h	p_h	n_h
1	$N_1 = 0.4N$	0.1	n_1
2	$N_2 = 0.6N$	0.03	n_2
	N		2000

(a) Under the optimal allocation, we can use

$$n_{h,\text{Neyman}} = \frac{N_h S_h}{\sum_l N_l S_l} n$$

where $S_h = \sqrt{\frac{N_h}{N_h - 1} p_h(1-p_h)} \approx \sqrt{p_h(1-p_h)}$ when N_h is large.

So, we have $S_1 = 0.3$ and $S_2 = 0.1706$ and obtain $n_1 = 1079$ and $n_2 = 921$.

(b) $V(\hat{p}_\text{str}) = \sum_h \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{N_h}{N_h - 1}\frac{p_h(1-p_h)}{n_h} \approx \sum_h \left(\frac{N_h}{N}\right)^2 \frac{p_h(1-p_h)}{n_h}$ when N_h is large.

(i) Under proportional allocation, $n_1 = 800$ and $n_2 = 1200$,
$V_\text{prop}(\hat{p}_\text{str}) \approx 2.673 \times 10^{-5}$.

(ii) Under optimal allocation, $n_1 = 1079$ and $n_2 = 921$,
$V_\text{optimal}(\hat{p}_\text{str}) \approx 2.482 \times 10^{-5}$.

$V_\text{SRS}(\hat{p}_\text{SRS}) = \left(1 - \frac{n}{N}\right)\frac{N}{N-1}\frac{p(1-p)}{n} \approx \frac{p(1-p)}{n}$ when N is large. Since $p = \frac{(0.4N)(0.1) + (0.6N)(0.03)}{N} = 0.058$, we have $V_\text{SRS}(\hat{p}_\text{SRS}) \approx 2.732 \times 10^{-5}$.

Part II : Extra Problems

(a) $p(1-p)$ is a quadratic function with maximum value at $p = 0.5$. When p is between 0.3 and 0.7, $\sqrt{p(1-p)}$ is between .46 and .5 which are very close.

(b) Similar.

(c) $S.E(\hat{p}_\text{str}) = 0.0069$ and $S.E(\hat{p}_\text{SRS}) = 0.0071$ are close to each other.