STAC50: Assignment 2 Solution (Total : 66 points)

Q 1. (Total: 14 points)

(a) 2 points $\bar{y}_U$ =71.83333, S^2=86.16667

(b) 1 point 6C4 = 15 possible samples

(c) 4 points: 2 points for correct table, and 2 points for variance of sample mean

Possible Sample Index (S)	Sample mean
1, 2, 3, 4	69.5
1, 2, 3, 5	69.25
1, 2, 3, 6	66.5
1, 2, 4, 5	72.5
1, 2, 4, 6	69.75
1, 2, 5, 6	69.5
1, 3, 4, 5	75.25
1, 3, 4, 6	72.5
1, 3, 5, 6	72.25
1, 4, 5, 6	75.5
2, 3, 4, 5	73.5
2, 3, 4, 6	70.75
2, 3, 5, 6	70.5
2, 4, 5, 6	73.75
3, 4, 5, 6	76.5

We can check that E($\bar{y}$)= $\bar{y}_U$ =71.83333

Var($\bar{y}$) = sum(($\bar{y}$ - $\bar{y}_U$)^2)/15 = 7.180556 = $\frac{S^2}{n}(1-\frac{n}{N})$

(d) (1 point) 3C2*3C2 = 9

(e) (2 points)

Possible Sample Index (S)	Not Possible Sample Index
1, 2, 4, 5	1, 2, 3, 4
1, 2, 4, 6	1, 2, 3, 5
1, 2, 5, 6	1, 2, 3, 6
1, 3, 4, 5	1, 4, 5, 6
1, 3, 4, 6	2, 4, 5, 6
1, 3, 5, 6	3, 4, 5, 6
2, 3, 4, 5	
2, 3, 4, 6	
2, 3, 5, 6	

(f) (4 points): 2 points for table and 2 points for variance of sample mean.

Possible Sample Index (S)	Stratified sample mean

1, 2, 4, 5	72.5
1, 2, 4, 6	69.75
1, 2, 5, 6	69.5
1, 3, 4, 5	75.25
1, 3, 4, 6	72.5
1, 3, 5, 6	72.25
2, 3, 4, 5	73.5
2, 3, 4, 6	70.75
2, 3, 5, 6	70.5

Var($\bar{y}_{str}$) = sum(($\bar{y}_{str}$ - $\bar{y}_U$)^2)/9=3.138889, you can use the variance formula for stratified sampling to obtain the same answer.

Q. 2 (total: 16 points)

(1) total: 10 points

(a)-(c) each 2 points: total 6 points $\sum_{h=1}^{4} s_h N_h = 7072.607$, $\sum_{h=1}^{4} s_h N_h / \sqrt{c_h} = 3668.191$

Stratum	N_h	$s_h = \sqrt{\hat{p}(1-\hat{p})}$	Prop	Neyman	Optimal
1 (fr)	6812	0.4828043	92	93	104
2 (so)	4586	0.4995998	62	65	59
3 (Jr)	2714	0.4664762	37	36	33
4(Sr)	618	0.3666061	8	6	5
	14730		199	200	201

Note that I rounded the number to get the sample size for each stratum. If the calculations are correct then assign 93 instead of 92 for proportional allocation to make n=200 should be okay too (and same thing for making the optimal allocation as 200).

(d) 4 points

We have \$680 to allocate the sample size.

680 = $3n_1 + 4.5n_2+4.5n_3+6n_4$, and each n_i is computed by Optimal allocation which is $\frac{s_h N_h / \sqrt{c_h}}{\sum_{h=1}^{4} s_h N_h / \sqrt{c_h}} \times n$.

So (3*0.51764642+4.5*0.29444091+4.5*0.16269758 +6*0.02521509)*n = 680

We can find n= 680/3.761353 = 180.786 so n=181.

(2) total: 6 points

(a) 2 points

Apply SRS formula for determining the sample size for estimating the mean in stratum 4.

$$e_4 = 0.05, \qquad z_{0.025} = 1.96$$

$$n_4 = \frac{z^2_{\alpha/2} s^2_4}{e^2_4 + \frac{z^2_{\alpha/2} s^2_4}{N_4}} = 160 \text{ seniors}$$

(b) 4 points: 2 points for calculating n4 = 56 and 2 points for computing the margin of error.

Since we are willing to spend $340 on seniors and it cost $6/senior to collect data, we can afford to collect data on 56 seniors. This would give us a margin of error of

$$e_4 = z_{\frac{\alpha}{2}} SE(\widehat{p_4}) = z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p_4}(1-\widehat{p_4})}{n_4 - 1}} = 1.96 \sqrt{\frac{0.84(0.16)}{55}} \approx 0.1$$

I would relax my margin of error constraint to be 0.1 and keep the 95% confidence level.

Q. 3 (Total: 14 points)

(a) 6 points (2 points for getting the summary statistics for each stratum, and estimated total and variance of each stratum and finally 2 points for correct estimated total and standard error).

Here are summary statistics for each stratum:

	Stratum			
	Biological	Physical	Social	Humanities
Average	3.142857	2.105263	1.230769	0.4545455
Variance	6.809524	8.210526	4.358974	0.8727273

Using the summary statistics given above, total number of publications and estimated total for each of four strata can be calculated.

Stratum	Estimated total number of publications	Estimated variance of estimated total
Biological	320.571	9,426.33
Physical	652.632	38,932.71
Social	267.077	14,843.31
Humanities	80.909	2,358.43
Total	1,321.189	65,610.78

We estimated the total number of refereed publications for the college by adding the totals for each strata; which is 1,321.1 with the estimated standard error = $\sqrt{65{,}610.78} = 256.15$.

(b) 2 points

Here, stratified sampling ensures that each division of the college is represented in the sample and it produces an estimate with a smaller estimated standard error. Observations within strata tend to be more homogeneous than observations in the population as a whole, and the reduction in the variance in the individual strata often lead to a reduced variance for population estimate.

(c) 6 points

Following is the summary table:

Stratum	N_h	n_h	$\hat{p}_h$	$(N_h/N)\hat{p}_h$	$(N_h/N)^2 V[\hat{p}_h]$
Biological	102	7	1/7	0.018	0.0003
Physical	310	19	10/19	0.202	0.0019
Social	217	13	9/13	0.186	0.0012
Humanities	178	117	8/11	0.160	0.0009
Total	807	50		0.567	0.0043

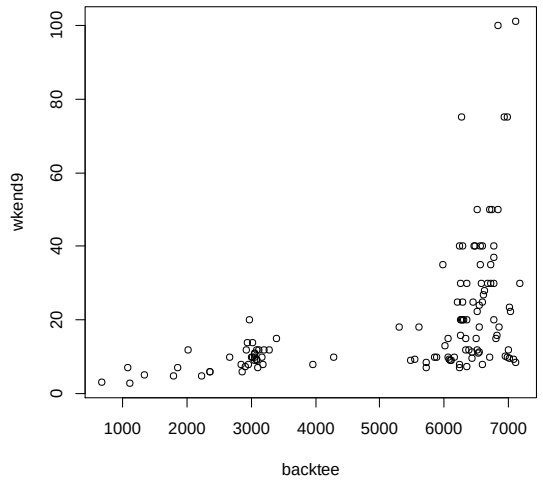
$$\hat{p}_{str} = 0.567 \quad , \quad \widehat{SE}\,[\hat{p}_{str}] = 0.0656$$

Q. 4 (total: 14 points)

(a) (2 points) $\bar{y} = 20.23667$

(b) (total: 6 points)

(i) (2 points) scatter plot



(ii)(2 points) correlation = 0.4621688

(iii) (2 points) $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = -4.04537 + 0.00456X$

(c) (2 points) No. The regression estimator of the population mean is $\widehat{\bar{y}_{reg}} = \bar{y} + \widehat{B_1}(\overline{X_u} - \overline{x})$. If we want to estimate the regression estimator, then we need to know the information of the population mean of x ($\overline{X_u}$).

(d) (4 points)

Although the relationship between backtee and wkend9 is not exactly linear, we could see the positive correlations between these two variables and expect the regression estimator is a better estimator. The standard error of sample mean is SE($\bar{y}$) = 1.632673. And the standard error of regression estimator is . $\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}$ =1.447842. The regression estimator is better than the estimator of sample mean. N=16883, s_e^2 = 253.3502.

R code for Question 4

```
> golf = read.csv("golfsrs.csv", header=T)
> attach(golf)
> mean(wkend9)
[1] 20.23667
>cor(backtee, wkend9)
> fit = lm(wkend9~backtee)
> summary(fit)
> var(fit$resid)
```

Q. 5 (total: 8 points)

(a) (2 points)

We are sampling apartment complexes(=clusters) and observing data on the elements(=residents), so this is a cluster sampling.

(b) (6 points)

N = 124 complexes, M_i = number of residents in complex i, n=20

Data: t_i = number of hotdogs eaten by residents last night in complex i

Using an unbiased estimator or ratio estimator, both accepted as a correct answer.

(i) Ratio Estimator

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{20} t_i}{\sum_{i=1}^{20} M_i} = \frac{340}{1532} = 0.221932115.$$

$$\hat{V}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}_s^2} \frac{\sum_{i=1}^{20} (t_i - M_i \hat{\bar{y}}_r)^2}{n-1} = 0.000150509.$$

$$\widehat{SE}(\hat{\bar{y}}_r) = 0.012268204.$$

(ii) Unbiased estimator

$\hat{\bar{y}}_{unb} = \frac{1}{K}\hat{t}_{unb}$ = 2108/(9498.4) = 0.2219321, note that $\hat{t}_{unb} = \frac{N}{n}\sum_{i=1}^{n} t_i = \frac{124}{20} \times 340 = 2108$

$\hat{V}\left(\hat{t}_{unb}\right) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}$, s_t^2 = 74.10526, $\hat{V}\left(\hat{t}_{unb}\right)$ = 47783.07

$\hat{V}\left(\hat{\bar{y}}_{unb}\right) = \frac{1}{K^2} \hat{V}\left(\hat{t}_{unb}\right)$ = 0.0005296307, SE($\hat{\bar{y}}_{unb}$) = 0.02301371