Statistics 257 Midterm – October 25, 2004 (Selected Solutions)

1. (b) We see that $N_1 = 51$, $N_2 = 33$, $N_3 = 35$, $N_4 = 55$, $N_5 = 36$, $N_6 = 53$, and $N_7 = 52$ so that $N = 315$. We also compute that

$$\overline{y}_1 = \frac{80}{32} = 2.5, \;\; \overline{y}_2 = \frac{72}{16} = 4.5, \;\; \overline{y}_3 = \frac{108}{18} = 6, \;\; \overline{y}_4 = \frac{132}{33} = 4,$$

$$\overline{y}_5 = \frac{42}{21} = 2, \;\; \overline{y}_6 = \frac{84}{24} = 3.5, \;\; \overline{y}_7 = \frac{81}{27} = 3.$$

Letting $\overline{y}_{\text{ST}}$ denote the average number of publications among those who responded across the seven strata, we see that

$$\begin{aligned}
\overline{y}_{\text{ST}} &= \frac{1}{N}(N_1\overline{y}_1 + N_2\overline{y}_2 + N_3\overline{y}_3 + N_4\overline{y}_4 + N_5\overline{y}_5 + N_6\overline{y}_6 + N_7\overline{y}_7) \\
&= \frac{1}{315}(51 \cdot 2.5 + 33 \cdot 4.5 + 35 \cdot 6 + 55 \cdot 4 + 36 \cdot 2 + 53 \cdot 3.5 + 52 \cdot 3) \\
&= \frac{1119.5}{315} \\
&\approx 3.55.
\end{aligned}$$

1. (c) By assumption, $E(y_i) = \mu$ for each i. Hence,

$$E(\overline{y}_i) = E\left(\frac{y_1 + \cdots + y_{n_i}}{n_i}\right) = \frac{E(y_1) + \cdots + E(y_{n_i})}{n_i} = \frac{n_i\mu}{n_i} = \mu.$$

Thus,

$$\begin{aligned}
E(\overline{y}_{\text{ST}}) &= \frac{1}{N}E\left(N_1\overline{y}_1 + N_2\overline{y}_2 + N_3\overline{y}_3 + N_4\overline{y}_4 + N_5\overline{y}_5 + N_6\overline{y}_6 + N_7\overline{y}_7\right) \\
&= \frac{1}{N}\left(N_1E(\overline{y}_1) + N_2E(\overline{y}_2) + N_3E(\overline{y}_3) + N_4E(\overline{y}_4) + N_5E(\overline{y}_5) + N_6E(\overline{y}_6) + N_7E(\overline{y}_7)\right) \\
&= \frac{1}{N}\left[(N_1 + N_2 + N_3 + N_4 + N_5 + N_6 + N_7)\mu\right] \\
&= \frac{N\mu}{N} = \mu.
\end{aligned}$$

Since $E(\overline{y}_{\text{ST}}) = \mu$ we conclude that $\overline{y}_{\text{ST}}$ is an unbiased estimator for μ.

1. (d) An approximate 95% confidence interval for μ is

$$\overline{y}_{\text{ST}} \pm 2\sqrt{\hat{V}(\overline{y}_{\text{ST}})}$$

where

$$\hat{V}(\overline{y}_{\text{ST}}) = \frac{1}{N^2}\sum_{i=1}^{7} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right).$$

In order to estimate the s_i, we can use the fact that the range is approximately 4 standard deviations. Thus, we find that

$$s_1 = 1, \;\; s_2 = 2, \;\; s_3 = 2, \;\; s_4 = 4, \;\; s_5 = 1, \;\; s_6 = 1, \;\; s_7 = 2.$$

Hence,

$$\begin{aligned}
&\hat{V}(\overline{y}_{\mathrm{ST}}) \\
&= \frac{1}{315^2}\left[51^2\left(\frac{51-32}{51}\right)\left(\frac{1^2}{32}\right) + 33^2\left(\frac{33-16}{33}\right)\left(\frac{2^2}{16}\right) + 35^2\left(\frac{35-18}{35}\right)\left(\frac{2^2}{18}\right)\right. \\
&\left.+55^2\left(\frac{55-33}{55}\right)\left(\frac{4^2}{33}\right) + 36^2\left(\frac{36-21}{36}\right)\left(\frac{1^2}{21}\right) + 53^2\left(\frac{53-24}{53}\right)\left(\frac{1^2}{24}\right) + 52^2\left(\frac{52-27}{52}\right)\left(\frac{2^2}{27}\right)\right] \\
&\approx 0.0116.
\end{aligned}$$

Thus, an approximate 95% CI is

$$3.55 \pm 2\sqrt{0.0116} \quad \text{OR} \quad 3.55 \pm 0.22.$$

1. (e) We can answer this question in one of two ways. The first solution is to construct separate confidence intervals for the true average number of Linguistic publications and for the true average number of History publications. If these two confidence intervals are disjoint, then there is a statistically significant difference. If these intervals are not disjoint, then there is not a statistically significant difference. The second solution is to compute a single confidence interval for the difference. If this confidence interval covers 0, then there is not a statistically significant difference, otherwise there is a statistically significant difference.

Solution 1: An approximate 95% confidence interval for the true average number of History publications is

$$\overline{y}_4 \pm 2\sqrt{\hat{V}(\overline{y}_4)} \;\text{ or }\; 4 \pm 2\sqrt{\left(\frac{55-33}{55}\right)\left(\frac{4}{33}\right)} \;\text{ or }\; 4 \pm 0.44.$$

An approximate 95% confidence interval for the true average number of Linguistics publications is

$$\overline{y}_5 \pm 2\sqrt{\hat{V}(\overline{y}_5)} \;\text{ or }\; 2 \pm 2\sqrt{\left(\frac{36-21}{36}\right)\left(\frac{1}{21}\right)} \;\text{ or }\; 2 \pm 0.28.$$

Since these two CIs are disjoint, there is a statistically significant difference in the average number of publications for faculty members in Linguistics compared to History.

Solution 2: An approximate 95% confidence interval for the true difference is

$$(\overline{y}_4 - \overline{y}_5) \pm 2\sqrt{\hat{V}(\overline{y}_4) + \hat{V}(\overline{y}_5)} \;\text{ or }\; (4-2) \pm 2\sqrt{\left(\frac{55-33}{55}\right)\left(\frac{4}{33}\right) + \left(\frac{36-21}{36}\right)\left(\frac{1}{21}\right)} \;\text{ or }\; 2 \pm 0.52.$$

Since this CI does not overlap with 0, we draw the same conclusion, namely that there is a statistically significant difference in the average number of publications for faculty members in Linguistics compared to History.

1. (f) We use equation (5.7) in order to find the approximate allocation which minimizes cost for a fixed value of $V(\overline{y}_{\mathrm{ST}})$. From that equation, the allocation proportions are

$$w_i = \frac{n_i}{n} = \frac{N_i\sigma_i/\sqrt{c_i}}{\sum_{i=1}^{7} N_i\sigma_i/\sqrt{c_i}} = \frac{N_i/\sqrt{c_i}}{\sum_{i=1}^{7} N_i/\sqrt{c_i}}$$

since $\sigma_i = 10000$ for each stratum. Thus,

$$\sum_{i=1}^{7} N_i/\sqrt{c_i} = \frac{32}{4} + \frac{16}{4} + \frac{18}{3} + \frac{33}{3} + \frac{21}{3} + \frac{24}{3} + \frac{27}{3} = 8 + 4 + 6 + 11 + 7 + 8 + 9 = 53$$

so that

$$w_1 = \frac{8}{53},\ w_2 = \frac{4}{53},\ w_3 = \frac{6}{53},\ w_4 = \frac{11}{53},\ w_5 = \frac{7}{53},\ w_6 = \frac{8}{53},\ w_7 = \frac{9}{53}.$$

Since we wish to have a bound of 1000, this tells us that $D = 1000^2/4 = 250000$ in equation (5.8) for n, namely

$$n = \frac{\left(\sum_{i=1}^{7} N_i\sigma_i/\sqrt{c_i}\right)\left(\sum_{i=1}^{7} N_i\sigma_i\sqrt{c_i}\right)}{N^2D + \sum_{i=1}^{7} N_i\sigma_i^2}.$$

Notice that

$$\sum_{i=1}^{7} N_i\sigma_i^2 = 10000^2 \sum_{i=1}^{7} N_i = 10000^2 \cdot 171$$

and

$$\sum_{i=1}^{7} N_i\sigma_i/\sqrt{c_i} = 10000 \sum_{i=1}^{7} N_i/\sqrt{c_i} = 10000 \cdot 53$$

and

$$\sum_{i=1}^{7} N_i\sigma_i\sqrt{c_i} = 10000(32 \cdot 4 + 16 \cdot 4 + 18 \cdot 3 + 33 \cdot 3 + 21 \cdot 3 + 24 \cdot 3 + 27 \cdot 3) = 10000 \cdot 561.$$

Combining, gives

$$n = \frac{(10000 \cdot 53)(10000 \cdot 561)}{(171^2 \cdot 250000) + (10000^2 \cdot 171)} \approx 122$$

so that

$$n_1 \approx 18.4 \approx 18,\ \ n_2 \approx 9.2 \approx 9,\ \ n_3 \approx 13.8 \approx 14,\ \ n_4 \approx 25.3 \approx 25,$$

$$n_5 \approx 16.1 \approx 16, \quad n_6 \approx 18.4 \approx 18, \quad n_7 \approx 20.7 \approx 21.$$

Note that $18 + 9 + 14 + 25 + 16 + 18 + 21 = 121$. Since both n_1 and n_6 have the largest fractional part, and since strata 1 (Literature) is more expensive to sample from strata 6 (Political Science), it is reasonable to add the extra data point to strata 6. Thus, take $n_6 = 19$.

1. (g) If the sociologist is constrained to have a total cost of \$1122, then it follows that she must have

$$n_1c_1 + n_2c_2 + n_3c_3 + n_4c_4 + n_5c_5 + n_6c_6 + n_7c_7 = 1122.$$

Since $n_i = w_i n$, this equation simplifes to

$$n(w_1c_1 + w_2c_2 + w_3c_3 + w_4c_4 + w_5c_5 + w_6c_6 + w_7c_7) = 1122.$$

Using the values of w_i and c_i from **(d)** gives

$$\begin{aligned} n &= \frac{1122}{w_1c_1 + w_2c_2 + w_3c_3 + w_4c_4 + w_5c_5 + w_6c_6 + w_7c_7} \\ &= \frac{1122}{\frac{1}{53}(8 \cdot 16 + 4 \cdot 16 + 6 \cdot 9 + 11 \cdot 9 + 7 \cdot 9 + 8 \cdot 9 + 9 \cdot 9)} \\ &= \frac{1122}{561/53} \\ &= 53 \cdot 2 \\ &= 106. \end{aligned}$$

2. (a) Data Step: This is the step where you create and modify SAS data sets. You can read in files containing data and compute new variables.

Proc Step: These steps use existing SAS procedures. Most SAS procedures compute statistics and prepare reports.

2. (b)

```
PROC MEANS DATA=country MEAN MEDIAN VAR;
     VAR pop92;
      CLASS cont;
RUN;
```

Note that capitalization is not important.