Stat 257: Solutions to Assignment #2

(2.1) In order to calculate $\overline{y}$, we must *pool* the averages from each of the two simple random samples remembering to weight by the appropriate sample sizes. That is,

$$\overline{y} = \frac{n_1}{n_1+n_2}\cdot\overline{y}_1 + \frac{n_2}{n_1+n_2}\cdot\overline{y}_2 = \frac{300}{800}\cdot 2.98 + \frac{500}{800}\cdot 3.42 = 3.255.$$

An approximation of the pooled sample standard variance is

$$s^2 = \frac{n_1-1}{n_1+n_2-1}\cdot s_1^2 + \frac{n_2-1}{n_1+n_2-1}\cdot s_2^2 = \frac{299}{799}\cdot 4.27 + \frac{499}{799}\cdot 3.68 \approx 3.896.$$

This gives an estimated variance of $\overline{y}$ as

$$s^2(\overline{y}) = \frac{(1-f)}{n}\cdot s^2 \approx \frac{(1-800/2500)}{800}\cdot 3.896 \approx 0.003312.$$

Thus, an approximate 99% confidence interval for the true average distance from the college that its non-residential students live is

$$3.255 \pm 2.576\sqrt{0.003312} \quad \text{or} \quad 3.255 \pm 0.148 \quad \text{or} \quad (3.107, 3.403).$$

(2.2) Suppose that Y_T denotes the total number of 'casual days' taken over the six month period of interest. Since 49.82% of the workforce of 36 000 missed no workdays, we know that $0.5018 \times 36\ 000 \approx 18\ 065$ workers missed at least one day. Let y_i, $i = 1, \ldots, 500$, denote the number of workdays missed by the ith individual in the simple random sample of size 500 selected from among the 18 065 workers who missed at least one day, and let $\overline{y}$ denote the sample average number of workdays missed by these 500 workers. Hence,

$$\sum_{i=1}^{500} y_i = 1\cdot 157 + 2\cdot 192 + 3\cdot 90 + 4\cdot 41 + 5\cdot 18 + 6\cdot 5 + 7\cdot 2 + 8\cdot 4 + 9\cdot 0 + 10\cdot 1 = 1111$$

so that

$$\overline{y} = \frac{1}{500}\sum_{i=1}^{500} y_i = \frac{1111}{500} = 2.222.$$

Furthermore,

$$\sum_{i=1}^{500} y_i^2 = 1^2\cdot 157 + 2^2\cdot 192 + 3^2\cdot 90 + 4^2\cdot 41 + 5^2\cdot 18 + 6^2\cdot 5 + 7^2\cdot 2 + 8^2\cdot 4 + 9^2\cdot 0 + 10^2\cdot 1 = 3315$$

so that

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{(n-1)}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) = \frac{1}{499}\left(3315 - 500\cdot 2.222^2\right) \approx 1.696.$$

Thus, the estimate of the total number of 'casual days' taken by those who missed at least one workday is

$$y_{T1} \approx 18\ 065 \cdot 2.222 \approx 40\ 140$$

and the estimate of the total number of 'casual days' taken by those who missed no workdays is obviously $y_{T2} = 0$. This gives an estimate of the total number of 'casual days' taken by those in the entire workforce as

$$y_T = y_{T1} + y_{T2} \approx 40\ 140.$$

Since there is no variability among the workers who missed no workdays, we see that the estimated variance of y_T is given by

$$s^2(y_T) = s^2(y_{T1}) = N^2 \cdot \frac{(1-f)}{n} \cdot s^2 \approx 18\ 065^2 \cdot \frac{(1 - 500/18\ 065)}{500} \cdot 1.696 \approx 1\ 076\ 390$$

so that the estimated standard error of y_T is

$$s(y_T) \approx \sqrt{1\ 076\ 390} \approx 1037.$$

If we do the same calculation using the data in Exercise 2.3, we find that

$$y_T = 36\ 000 \cdot 1.296 = 46\ 656$$

and

$$s^2(y_T) = N^2 \cdot \frac{(1-f)}{n} \cdot s^2 = 36\ 000^2 \cdot \frac{(1 - 1000/36\ 000)}{1000} \cdot 2.397 = 3202220$$

so that in this case the estimated standard error of y_T is

$$s(y_T) = \sqrt{3\ 202\ 220} \approx 1738.$$

(2.3) Let y_i denote the number of books observed on the ith shelf, $i = 1, \ldots, 16$, so that $\overline{y}$ denotes the estimated average number of books per shelf. From the data, we calculate that

$$\sum_{i=1}^{16} y_i = 447, \quad \sum_{i=1}^{16} y_i^2 = 12\ 683, \quad \text{and} \quad \overline{y} = \frac{1}{n} \cdot \sum_{i=1}^{16} y_i = \frac{447}{16} \approx 27.94.$$

We also find that the sample variance is

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{(n-1)} \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \right) = \frac{1}{15} \left(12\ 683 - 16 \cdot \left(\frac{447}{16} \right)^2 \right) \approx 12.996.$$

Hence, we approximate the total number of books in the library Y_T by

$$y_T = N\overline{y} = 170 \cdot \frac{447}{16} = 4749.375 \approx 4749.$$

The estimated variance of $\overline{y}$ is

$$s^2(\overline{y}) = \frac{(1-f)}{n} \cdot s^2 = \frac{(1 - 16/170)}{16} \cdot 12.996 \approx 0.7358$$

so that the estimated variance of y_T is

$$s^2(y_T) = s^2(N\overline{y}) = N^2 s^2(\overline{y}) \approx 170^2 \cdot 0.7358 \approx 21264.$$

This gives the estimated standard deviation of y_T as $s(y_T) \approx 146$ so that an approximate 95% confidence interval for Y_T, the true number of books in the library, is

$$4749 \pm 1.96(146) \quad \text{or} \quad (4463, 5035).$$

If we want to be 95% confident that a simple random sample estimate of Y_T is within 100 books of its true value, then we need to sample n shelves, where

$$n \geq N\left[1+\frac{1}{N}\left(\frac{d}{z_\alpha s}\right)^2\right]^{-1}$$

and $N = 170$, $d = 100$, $z_\alpha = 1.96$, and $s^2 = 12.996$. Substituting these values gives a required sample size of $n \geq 78$. Hence, the minimum number of shelves we need to sample in order to be 95% confident that a simple random sample estimate of Y_T is within 100 books of its true value is 78.

(2.4) We know that $E(s^2) = S^2$ since the sample variance s^2 is always an unbiased estimator for S^2 (see bottom of page 37). Since the sample size is $2n$ this means that

$$s^2 = \frac{1}{2n-1}\sum_{i=1}^{2n}(y_i - \overline{y})^2$$

is an unbiased estimator of S^2. However, we do not a priori know the sample mean $\overline{y}$; we are only given the sub-sample means $\overline{y}_1$ and $\overline{y}_2$. But, we do know that

$$\overline{y} = \frac{n}{2n}\cdot\overline{y}_1 + \frac{n}{2n}\cdot\overline{y}_2 = \frac{\overline{y}_1+\overline{y}_2}{2}.$$

Thus, a simple unbiased estimator for S^2 is given by

$$s^2 = \frac{1}{2n-1}\sum_{i=1}^{2n}\left(y_i - \frac{\overline{y}_1+\overline{y}_2}{2}\right)^2.$$

(2.5) Let P_1 denote the proportion for (a), and let P_2 denote the proportion for (b). We want $\mathrm{SE}(P_1) \leq 0.01$ and $\mathrm{SE}(P_2) \leq 0.02$, and we believe that $0.35 \leq P_1 \leq 0.55$ and $0.80 \leq P_2 \leq 0.90$. We also know that the population size is $N = 5000$. Since the two proportions are to be estimated from a single random sample, we simply need to find n_1, the minimum sample size required to the satisfy the constraints of P_1, and n_2, the minimum sample size required to the satisfy the constraints of P_2. Our total sample size required to simultaneous satisfy all constraints will then be $n = \max\{n_1, n_2\}$. Since $0.35 \leq P_1 \leq 0.55$, we see that $0.2275 \leq P_1Q_1 \leq 0.25$, and since $0.80 \leq P_2 \leq 0.90$ we see that $0.64 \leq P_1Q_1 \leq 0.81$. (These bounds can be found by optimizing the function $f(P_i) = P_i(1-P_i)$ subject to the constraints on P_i.) Now, we know that n_i satisfies

$$n_i \geq N\left[1+\frac{N-1}{P_iQ_i}\left(\frac{d_i}{z_\alpha}\right)^2\right]^{-1}$$

where $N = 5000$, $d_i/z_\alpha = \mathrm{SE}(P_i)$, and P_iQ_i is as large as possible. Thus,

$$n_1 \geq 5000\left[1+\frac{4999}{0.25}(0.01)^2\right]^{-1} \approx 1666.9$$

and

$$n_2 \geq 5000\left[1+\frac{4999}{0.81}(0.02)^2\right]^{-1} \approx 1441.5.$$

Hence, the minimum sample size requied to satisfy all of the accuracy requirements is of size $n = 1667$.