Statistics 257–Applied Survey Techniques
Fall 2005 (200530)
Final Exam Solutions

Instructor: Michael Kozdron

1. **(a)** The Petersen estimator of N is given by

$$\tilde{N} = \frac{nm}{r} = \frac{59 \times 45}{9} = 295$$

and has estimated variance

$$s^2(\tilde{N}) = \frac{nm(n-r)(m-r)}{r^3} = \frac{59 \times 45 \times (59-9) \times (45-9)}{9^3} \approx 6555.556.$$

Thus, an approximate 95% confidence interval for the true number of pronghorns in Cypress Hills is $295 \pm 2\sqrt{6555.556}$ or $(133, 457)$.

(b) The Chapman estimator of N is given by

$$\hat{N} = \frac{(n+1)(m+1)}{(r+1)} - 1 = \frac{60 \times 46}{10} - 1 = 275$$

and has estimated variance

$$s^2(\hat{N}) = \frac{(n+1)(m+1)(n-r)(m-r)}{(r+1)^2(r+2)} = \frac{60 \times 46 \times 50 \times 36}{10^2 \times 11} \approx 4516.364.$$

Thus, an approximate 95% confidence interval for the true number of pronghorns in Cypress Hills is $275 \pm 2\sqrt{4516.264}$ or $(141, 409)$.

(c) The total number of yellow-bellied sapsuckers in Cypress Hills is estimated by

$$\tilde{T} = A\tilde{\delta} = \frac{An}{2\ell w} = \frac{400 \times 258}{2 \times 30 \times 0.1} = 17\ 200.$$

Note that $\ell = 30$ km, and $w = 100$ m $= 0.1$ km.

2. From the data presented, we find that

$$\Pr(\text{match}) = \frac{38}{50} \cdot (1-p) + \frac{12}{50} \cdot p$$

where $\Pr(\text{match})$ is estimated by $n_1/n = 198/280$ (and $\theta = 12/50$). Thus, p is estimated by

$$\hat{p} = \frac{n_1/n - (1-\theta)}{\theta - (1-\theta)} = \frac{198/280 - 38/50}{12/50 - 38/50} = \frac{37}{364} \approx 0.102$$

(just do the easy algebra and solve for $\hat{p}$) and has estimated variance

$$s^2(\hat{p}) = \frac{1}{(2\theta-1)^2} \cdot \frac{1}{n} \cdot \frac{n_1}{n} \cdot \left(1 - \frac{n_1}{n}\right) = \frac{1}{(2 \cdot 12/50 - 1)^2} \cdot \frac{1}{280} \cdot \frac{198}{280} \cdot \left(1 - \frac{198}{280}\right) \approx 0.00274.$$

Thus, an approximate 95% confidence interval for the proportion of undergrads who have plagiarized on a final exam is $0.102 \pm 2\sqrt{0.00274}$ or $(-0.008, 0.212)$. However, since p is a proportion, it must be between 0 and 1. Therefore, we discount the negative part of the interval above, and conclude that the 95% confidence interval for p is $(0, 0.212)$.

3. **(a)** The population of interest consists of the $N = 62$ *snowplow operators who work for Physical Plant Operations*. Thus, the average amount of time taken to plow the sidewalks is estimated by

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{12+11+17+12+13+9+13+9}{8} = \frac{96}{8} = 12.$$

The sample variance s^2 is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{0^2+1^2+5^2+0^2+1^2+3^2+1^2+3^2}{7} = \frac{46}{7}$$

so that $\overline{y}$ has estimated variance

$$s^2(\overline{y}) = \frac{(1-f)}{n}s^2 = \frac{(1-8/62)}{8} \cdot \frac{46}{7} = \frac{621}{868} \approx 0.715.$$

Thus, an approximate 95% confidence interval for the true average amount of time taken to plow the sidewalks in front of the College West building uring the winter is $12 \pm 2\sqrt{0.715}$ or $(10.308, 13.692)$ minutes.

(b) Since the confidence interval for plowing the snow from **(a)** overlaps with the confidence interval 10 ± 1 for cutting the grass, there is no significant difference in the amount of time taken for these two tasks.

4. The most common approach was to estimate each range by the midpoint of that range. This gives a point estimate of

$$\frac{3 \cdot 50 + 17 \cdot 150 + 7 \cdot 250 + 28 \cdot 350 + 31 \cdot 450 + 14 \cdot 550}{100} = 359.$$

(In and of itself, however, this solution is inadequate. By estimating each range by the midpoint of that range, it is implicitly being assumed that the distribution is uniform on that range. This is a very strong assumption, and one which needs to be discussed.)

5. **(a)** In order to determine the ratio estimate $\overline{y}_R$, we given by computing the ratio

$$r = \frac{\sum y_i}{\sum x_i} = \frac{897.6}{12} = 6.8$$

so that

$$\overline{y}_R = r\overline{X} = 6.8 \cdot \frac{5436}{453} = 6.8 \times 12 = 81.6.$$

This has estimated variance given by

$$\begin{aligned} s^2(\overline{y}_R) = \frac{(1-f}{n(n-1)}\sum(y_i - rx_i)^2 &= \frac{(1-f}{n(n-1)}\left(\sum y_i^2 - 2r\sum y_i x_i + r^2\sum x_i^2\right) \\ &= \frac{(1-12/453)}{12 \cdot 11} \cdot (73560 - 2 \cdot 6.8 \cdot 9820 + 6.8^2 \cdot 1473) \\ &\approx 59.882. \end{aligned}$$

Therefore, an approximate 95% confidence interval for the average annual salary of University of Regina full-time faculty members is $81.6 \pm 2\sqrt{59.882}$ or $(66.123, 97.077)$ (in \$1000s of dollars); that is, (\$66 123, \$97 077).

(b) In order to determine the regression estimate, we given by computing the estimated slope of the regression line, namely

$$\tilde{b} = \frac{\sum y_i x_i - n\,\overline{x}\,\overline{y}}{\sum x_i^2 - n\,\overline{x}^2} = \frac{9820 - 12 \cdot 11 \cdot 74.8}{1473 - 12 \cdot 11^2} \approx -2.55.$$

This gives the regression estimate as

$$\overline{y}_L = \overline{y} + \tilde{b}(\overline{X} - \overline{x}) \approx 74.8 - 2.55(12 - 11) = 72.248.$$

We also find that

$$s_y^2 = \frac{1}{n-1}\left(\sum y_i^2 - n\overline{y}^2\right) = \frac{1}{11}\left(73560 - 12 \cdot 74.8^2\right) \approx 583.59$$

and

$$s_{xy} = \frac{1}{n-1}\left(\sum x_i y_i - n\,\overline{x}\,\overline{y}\right) = \frac{1}{11}(9820 - 12 \cdot 11 \cdot 74.8) \approx -4.87$$

so that $\overline{y}_L$ has estimated variance

$$s^2(\overline{y}_L) = \frac{1-f}{n} \cdot \left(s_y^2 - \tilde{b}s_{xy}\right) = \frac{1 - 12/453}{12}(583.59 - (-2.55)(-4.87)) \approx 46.336.$$

Therefore, an approximate 95% confidence interval for the average annual salary of University of Regina full-time faculty members is $72.248 \pm 2\sqrt{46.336}$ or $(58.634, 85.862)$ (in \$1000s of dollars); that is, (\$58 634, \$85 862).

(c) The relative efficency of the two estimators is simply the ratio of their estimated variances. Depending on which you chose for the numerator, there are two equivalent solutions.

Solution 1: We easily compute that

$$\text{RelEff}(\overline{y}_R, \overline{y}_L) = \frac{s^2(\overline{y}_R)}{s^2(\overline{y}_L)} \approx \frac{59.882}{46.336} \approx 1.29.$$

Since $1.29 \gg 1$, there is sufficient evidence to suggest that the variance of $\overline{y}_R$ is greater than the variance of $\overline{y}_L$. This implies that the regression estimator is preferrable to the ratio estimator in this particular problem.

Solution 2: We easily compute that

$$\text{RelEff}(\overline{y}_L, \overline{y}_R) = \frac{s^2(\overline{y}_L)}{s^2(\overline{y}_R)} \approx \frac{46.336}{59.882} \approx 0.77.$$

Since $0.77 \ll 1$, there is sufficient evidence to suggest that the variance of $\overline{y}_L$ is less than the variance of $\overline{y}_R$. This implies that the regression estimator is preferrable to the ratio estimator in this particular problem.

6. **(a)** Since N is known, and since we are explicitly asked to compute an unbiased estimator of the average score that would be achieved by all Central High School students on this test, we find $\overline{y}_{c(b)}$, which is given by

$$\overline{y}_{c(b)} = \frac{M}{N} \cdot \frac{1}{m} \sum_{i=1}^{M} y_{iT} = \frac{40}{850} \cdot \frac{1}{5} \cdot [21 \cdot 13 + 18 \cdot 14 + 22 \cdot 17 + 32 \cdot 11 + 17 \cdot 15] \approx 14.174,$$

as the required estimator.

(b) The estimated variance of $\overline{y}_{c(b)}$ is

$$s^2(\overline{y}_{c(b)}) = \frac{(M-m)M}{mN^2(m-1)} \sum_{i=1}^{m} (y_{iT} - \frac{N}{M}\overline{y}_{c(b)})^2 \approx \frac{35 \cdot 40}{5 \cdot 850^2 \cdot 4} 13230.8 \approx 1.282.$$

Therefore the estimated standard error is

$$s(\overline{y}_{c(b)}) \approx \sqrt{1.282} \approx 1.132.$$

7. **(a)** For this stratified sample, an estimator of $\overline{Y}$ is given by

$$\overline{y}_{st} = \sum_{i=1}^{2} W_i \overline{y}_i = \frac{N_1}{N}\,\overline{y}_1 + \frac{N_2}{N}\,\overline{y}_2 = \frac{350}{350+450} \cdot 12 + \frac{450}{350+450} \cdot 16 = 14.25.$$

The estimated variance is given by

$$\begin{aligned} s^2(\overline{y}_{st}) &= \sum_{i=1}^{k} W_i^2 (1-f_i) \frac{s_i^2}{n_i} = \frac{1}{N^2} \left[N_1 \left(N_1 - n_1\right) \frac{s_1^2}{n_1} + N_2^2 \left(N_2 - n_2\right) \frac{s_2^2}{n_2} \right] \\ &= \frac{1}{(350+450)^2} \left[350 \left(350 - 45\right) \frac{4}{45} + 450 \left(450 - 55\right) \frac{9}{55} \right] \approx 0.0603. \end{aligned}$$

In other words, an approximate 95% confidence interval for the true average score that would be achieved by all Eastern High School students on this test is $14.25 \pm 2\sqrt{0.0603}$ or 14.25 ± 0.491.

(b) For proportional allocation, the sample fractions are the same as the population fractions. Thus,

$$\frac{n_1}{n} = W_1 = \frac{N_1}{N} = \frac{350}{800} \quad \text{and} \quad \frac{n_2}{n} = W_2 = \frac{N_2}{N} = \frac{450}{800}.$$

For a fixed bound of $\text{Var}(\overline{y}_{st}) = V$, the optimal sample size is given by

$$n = \frac{\frac{1}{V} \sum_{i=1}^{2} W_i S_i^2}{1 + \frac{1}{NV} \sum_{i=1}^{2} W_i S_i^2} = \frac{\frac{1}{0.02} \left(\frac{350}{800} \cdot 4 + \frac{450}{800} \cdot 9 \right)}{1 + \frac{1}{800 \cdot 0.02} \left(\frac{350}{800} \cdot 4 + \frac{450}{800} \cdot 9 \right)} \approx 238.9 \approx 239$$

where we approximated S_i^2 by s_i^2. Thus, the proportional allocation gives $n_1 \approx 104.563$ and $n_2 \approx 134.438$. Since we can't have fractional people we allocate $n_1 = 105$ and $n_2 = 134$.

(c) For the Neyman allocation, the allocation ratios include the standard deviations:

$$\frac{n_i}{n} = \frac{W_i S_i}{\sum_{i=1}^{2} W_i S_i} = \frac{N_i S_i}{N_1 S_1 + N_2 S_2}.$$

Since we do not know S_i we approximate by s_i. Hence,

$$\frac{n_1}{n} = \frac{350 \cdot 2}{350 \cdot 2 + 450 \cdot 3} = \frac{70}{205} \approx 0.341 \quad \text{and} \quad \frac{n_2}{n} = \frac{450 \cdot 3}{350 \cdot 2 + 450 \cdot 3} = \frac{135}{205} \approx 0.659.$$

For a fixed bound of $\text{Var}(\overline{y}_{st}) = V$, the optimal sample size is given by

$$n = \frac{\frac{1}{V}\left(\sum_{i=1}^{2} W_i S_i\right)^2}{1 + \frac{1}{NV}\sum_{i=1}^{2} W_i S_i^2} = \frac{\frac{1}{0.02}\left(\frac{350}{800} \cdot 2 + \frac{450}{800} \cdot 3\right)^2}{1 + \frac{1}{800 \cdot 0.02}\left(\frac{350}{800} \cdot 4 + \frac{450}{800} \cdot 9\right)} \approx 230.274 \approx 231$$

where we approximated S_i^2 by s_i^2. (Note that we must round 230.274 up to 231 because if we were to round down to 230, the resulting variance would be *greater* that 0.02.) Thus, the Neyman allocation gives $n_1 \approx 78.771$ and $n_2 \approx 152.229$. Since we can't have fractional people we allocate $n_1 = 79$ juniors and $n_2 = 152$ seniors.

8. **(a)** The population of interest consists of all *passengers who will fly on either Canadian airline (namely Air Canada or WestJet) in January 2006.* The variable of interest is *complaints received by either Canadian airline in January 2006.* The best possible sampling frame is the *list of Air Canada flights offered in January 2006.* This is public knowledge; in fact, it is easily found on the Air Canada website. Of course, this information *must* be public because Air Canada wants to sell seats on these flights! Note that all Air Canada has agreed to tell you is the total number of passengers in January 2006, and not the individual passengers' names, or even the number of passengers on each flight. The sampling units are the *individual passengers of Air Canada in January 2006.* Each Air Canada passenger in January 2006 will receive your questionnaire.

(b) Here is a brief outline of one possible method. There will be a total of N passengers who fly on Air Canada in January 2006. The number N will be known to you because Air Canada will tell you it. You will distribute N questionnaires to these passengers. However, not all the questionnaires will be returned to you; only a certain number n_1 will be. If y_1 denotes the total number of complaints received on the n_1 returned questionnaires, then an estimate y_T for the total number of complaints received by Air Canada in January 2006 is

$$y_T = N \cdot \frac{y_1}{n_1}.$$

(c) You are explicitly being asked to extrapolate your results from Air Canada to WestJet. Therefore, you should use your questionnaire as an opportunity to learn why individuals chose to fly on Air Canada and not on WestJet. The problem indicates that there are only two Canadian airlines, namely Air Canada and WestJet. With this dichotomy, it is reasonable to assume that if someone chose to fly with Air Canada, then they also chose *NOT* to fly with WestJet.

You should ask why they made this decision. Does the passenger use Air Canada exclusively? or did the passenger have a particular reason to chose Air Canada over WestJet? You should also ask if the Air Canada passenger has ever flown with WestJet. If so, ask why that person is not flying with WestJet this time.

(d) In part **(c)**, we explicitly asked if an individual had any reasons for not flying with WestJet. Suppose that n_2 of the sampled customers indicated that they had flown with WestJet, of which y_2 had a complaint against WestJet. This suggests that y_2/n_2 is the proportion of previous WestJet customers currently flying with Air Canada who had a prior complaint against WestJet. Comparing this estimate with our estimate y_1/n_1 from part **(b)** might give us an indication of how Air Canada and WestJet compare with respect to complaints received. In order to estimate the total number of complaints received by WestJet in January 2006, we need to estimate n, the total number of passengers that will fly with WestJet in January 2006. The easiest way to estimate n is to determine Air Canada's average passengers per flight (obtained from our knowledge of N and the public knowledge of the number of Air Canada flights) and multiply this by the number of West Jet flights (also public knowledge) in January 2006. Thus, the estimate of the total number of WestJet complaints in January 2006 is

$$n \cdot \frac{y_2}{n_2}.$$