Statistics 257–Applied Survey Techniques
Fall 2004 (200430)
Final Exam Solutions

Instructor: Michael Kozdron

1. A simple random sample of size n from a population of size N is a probability sample in which *any* sample of size n drawn from the population has the same chance of being selected as any other sample of size n.

2. The principal reasons for using stratified random sampling rather than simple random sampling are:

- Stratificiation may produce a smaller bound on the error of estimation than would be produced by a SRS of the same size. This result is particularly true if the measurements within strata are homogeneous.
- The cost per observation may be reduced by stratification of the population elements into convenient groupings.
- Estimates of population parameters may be desired for subgroups of the population. These subgroups should then be identifiable strata.

3. Cluster sampling is an effective design for obtaining a specified amount of information at minimum cost under the following conditions:

- A good frame listing population elements either is not available or is very costly to obtain, while a frame listing clusters is easily obtained.
- The cost of obtaining observations increases as the distance separating elements increases.

4. As is clearly stated in the problem, the target population is *residents of both Regina and Saskatoon*. The variable of interest in *inter-provincial travel patterns of Regina and Saskatoon residents*. However, since the survey respondents were limited to heads of households, it could be argued that the theoretical population to which the inferences can be applied is the population of *heads of households in Regina and Saskatoon*. Since random digit dialing was employed, there is, in fact, *no formal frame*. Theoretically, the frame consists of all possible combinations of 7 digit telephone numbers, excluding ones that do not correspond to Regina or Saskatoon telephone exchanges, and excluding ones that do not belong to an appropriate head of household. It is extremely important to note that the telephone directories of Regina and Saskatoon *do not constitute the frame* since unlisted numbers may be randomly dialed. The sampling units are those listed by the frame. Hence, the sampling units in this case consist of *all of those phone numbers that belong to heads of households in Regina and Saskatoon*. A shortcoming with this sampling scheme is that some heads of households may have multiple phone numbers (cell phones, land lines, office phones), while some heads of households may not have a single phone. One other drawback to this scheme is that by randomly dialing digits, many numbers will be generated that do not correspond to heads of households such as business numbers or children's phones. This will lead to a loss of time.

5. This problem does not apply for Fall 2005.

6. **(a)** We easily compute that the mean of the control group is

$$\overline{y}_1 = \frac{11+12+11+7+9}{5} = 10$$

and the sample standard deviation of the control group is

$$s_1^2 = \frac{1}{n_1 - 1}\sum (y_{1i} - \overline{y}_1)^2 = \frac{1^2+2^2+1^2+3^2+1^2}{4} = 4.$$

Therefore, the estimated variance of the control group is

$$s^2(\overline{y}_1) = \frac{N-n_1}{N}\cdot\frac{s_1^2}{n_1} = \frac{100-5}{100}\cdot\frac{4}{5} = 0.76,$$

so that an approximate 95% confidence interval is $10 \pm 2\sqrt{0.76}$ or $(8.26, 11.74)$.

(b) We easily compute that the mean of the injection group is

$$\overline{y}_2 = \frac{15+11+14+9+11}{5} = 12$$

and the sample standard deviation of the injection group is

$$s_2^2 = \frac{1}{n_2 - 1}\sum (y_{2i} - \overline{y}_2)^2 = \frac{3^2+1^2+2^2+3^2+1^2}{4} = 6.$$

Therefore, the estimated variance of the injection group is

$$s^2(\overline{y}_2) = \frac{N-n_2}{N}\cdot\frac{s_2^2}{n_2} = \frac{100-5}{100}\cdot\frac{6}{5} = 1.14$$

so that an approximate 95% confidence interval is $12 \pm 2\sqrt{1.14}$ or $(7.86, 12.14)$.

(c) Since the confidence interval computed in **(a)** and **(b)** overlap, there is no statistically significant difference between the mean number of pizza slices eaten by the control group versus the injection group. Hence, there is no evidence to conclude that *Pizza-X* increases the ability of Sociologist 101 students to eat more pizza.

7. From the problem we immediately find this is a 1-in-k systematic sample with $k = 50$, $N = 15\,200$, $n = 304$. We find y_T as noted in the hint:

$$y_T = N\overline{y} = N\cdot\frac{\sum y_i}{n} = 15\,200\cdot\frac{76}{304} = 3800.$$

Furthermore, the estimated variance $s^2(y_T)$ is found to be

$$s^2(y_T) = N^2 s^2(\overline{y}) = N^2\cdot\left(\frac{1-f}{n}\right)s^2$$

where

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

Important: Since y_i can only equal 1 or 0, we see that y_i^2 can only equal 1 or 0. Therefore,

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i = 76$$

so that

$$s^2 = \frac{1}{304-1}\left(76 - 304 \cdot \left(\frac{76}{304}\right)^2\right) \approx 0.188.$$

Thus, the estimated variance is given by

$$s^2 = N^2 \cdot \left(\frac{1-f}{n}\right) s^2 = (15\,200)^2 \cdot \left(\frac{1 - 304/15\,200}{304}\right) \cdot 0.188 \approx 140110.89.$$

Thus an approximate 95% confidence interval for the total number of Moose Jaw families who rent is given by

$$y_T \pm 2\sqrt{s^2(y_T)} \approx 3800 \pm 749.$$

8. Since we do not know the population size N, and we do not suspect that the cluster sizes N_i are the same, we use the estimator $\overline{y}_{c(a)}$. Thus,

$$\overline{y}_{c(a)} = \frac{\sum y_{iT}}{\sum n_i} = \frac{182}{546} = \frac{1}{3}$$

and

$$\begin{aligned} s^2(\overline{y}_{c(a)}) &= \frac{(M-m)m}{M(m-1)} \sum_{i=1}^{15} \left(\frac{n_i}{n}\right)^2 \left(\overline{y}_i - \overline{y}_{c(a)}\right)^2 = \frac{(M-m)m}{M(m-1)n^2} \sum_{i=1}^{15} n_i^2 \left(\overline{y}_i^2 - 2\overline{y}_i\overline{y}_{c(a)} + \overline{y}_{c(a)}^2\right) \\ &= \frac{(M-m)m}{M(m-1)n^2}\left(\sum_{i=1}^{15} n_i^2\overline{y}_i^2 - 2\overline{y}_{c(a)}\sum_{i=1}^{15} n_i^2\overline{y}_i + \overline{y}_{c(a)}^2 \sum_{i=1}^{15} n_i^2\right) \\ &= \frac{(170-15)\cdot 15}{170\cdot(15-1)\cdot 546^2}\left(2103 - 2\cdot\frac{1}{3}\cdot 4571 + \left(\frac{1}{3}\right)^2 \cdot 9981\right) \\ &= 0.000540. \end{aligned}$$

Hence, an approximate 95% confidence interval for $\overline{Y}$ is $0.333 \pm 1.96 \cdot 0.023$ or $(0.288, 0.379)$.

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

$$\overline{y}_{st} = \sum_{i=1}^{2} W_i\overline{y}_i = \frac{N_1}{N}\,\overline{y}_1 + \frac{N_2}{N}\,\overline{y}_2 = \frac{24}{24+54}\cdot 13 + \frac{54}{24+54}\cdot 26 = 22.$$

The estimated variance is given by

$$\begin{aligned} s^2(\overline{y}_{st}) &= \sum_{i=1}^{k} W_i^2(1-f_i)\frac{s_i^2}{n_i} = \frac{1}{N^2}\left[N_1\left(N_1 - n_1\right)\frac{s_1^2}{n_1} + N_2^2\left(N_2 - n_2\right)\frac{s_2^2}{n_2}\right] \\ &= \frac{1}{(24+54)^2}\left[24\left(24-6\right)\frac{9}{6} + 54\left(54-12\right)\frac{16}{12}\right] = \frac{3672}{6084} \approx 0.604. \end{aligned}$$

In other words, an approximate 95% confidence interval for the population mean $\overline{Y}$ is given by $22 \pm 2\sqrt{0.604}$ or 22 ± 1.554.

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

$$\frac{n_1}{n} = W_1 = \frac{N_1}{N} = \frac{24}{78} \quad \text{and} \quad \frac{n_2}{n} = W_2 = \frac{N_2}{N} = \frac{54}{78}.$$

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

$$n = \frac{\frac{1}{V}\sum_{i=1}^{2} W_i S_i^2}{1 + \frac{1}{NV}\sum_{i=1}^{2} W_i S_i^2} \approx \frac{\frac{1}{9/16}\left(\frac{24}{78}\cdot 9 + \frac{54}{78}\cdot 16\right)}{1 + \frac{1}{78\cdot 9/16}\left(\frac{24}{78}\cdot 9 + \frac{54}{78}\cdot 16\right)} \approx 18.7 \approx 19$$

where we approximated S_i^2 by s_i^2. Thus, the proportional allocation gives $n_1 = 5.9$ and $n_2 = 13.2$. Since we can't have fractional people we allocate $n_1 = 6$ and $n_2 = 13$.

(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,

$$\frac{n_1}{n} = \frac{24\cdot 3}{24\cdot 3 + 54\cdot 4} = \frac{72}{288} = \frac{1}{4} \quad \text{and} \quad \frac{n_2}{n} = \frac{54\cdot 4}{24\cdot 3 + 54\cdot 4} = \frac{216}{288} = \frac{3}{4}.$$

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

$$n = \frac{\frac{1}{V}\left(\sum_{i=1}^{2} W_i S_i\right)^2}{1 + \frac{1}{NV}\sum_{i=1}^{2} W_i S_i^2} \approx \frac{\frac{1}{9/16}\left(\frac{24}{78}\cdot 3 + \frac{54}{78}\cdot 4\right)^2}{1 + \frac{1}{78\cdot 9/16}\left(\frac{24}{78}\cdot 9 + \frac{54}{78}\cdot 16\right)} \approx 18.4 \approx 19.$$

(Note that we must round 18.4 up to 19 because if we were to round down to 18, the resulting variance would be *greater* that 9/16.) Thus, the Neyman allocation gives $n_1 = 4.75$ and $n_2 = 14.25$. Since we can't have fractional people we allocate $n_1 = 5$ and $n_2 = 14$.

10.

(a) We find the ratio estimator r is given by

$$r = \frac{\sum y_i}{\sum x_i} = \frac{6744}{2248} = 3.$$

Since $\overline{X} = 45$, we find that our estimate of $\overline{Y}$ is given by $\overline{y}_R = r\overline{X} = 3 \cdot 45 = 135$. This has estimated variance given by

$$\begin{aligned} s^2(\overline{y}_R) &= \frac{(1-f)}{n(n-1)} \sum_{i=1}^{50} (y_i - r x_i)^2 = \frac{(1-f)}{n(n-1)} \left(\sum_{i=1}^{50} y_i^2 - 2r \sum_{i=1}^{n} x_i y_i + r^2 \sum_{i=1}^{50} x_i^2 \right) \\ &= \frac{(1 - 50/1000)}{50(50-1)} \left(928436 - 2 \cdot 3 \cdot 305125 + 3^2 \cdot 104384 \right) \\ &= \frac{19}{49\ 000} \cdot 37142 \approx 14.402. \end{aligned}$$

In other words, an approximate 95% confidence interval for $\overline{Y}$ is $135 \pm 2(3.79)$.

(b) We find that the regression estimator is given by

$$\overline{y}_L = \overline{y} + \tilde{b}(\overline{X} - \overline{x})$$

where

$$\tilde{b} = \frac{\sum (y_i - \overline{y})(x_i - \overline{x})}{\sum (x_i - \overline{x})^2} = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2} = \frac{305125 - 50 \cdot (2248/50) \cdot (6744/50)}{104384 - 50 \cdot (2248/50)^2} = \frac{191476}{331392} \approx 0.578.$$

Hence,

$$\overline{y}_L \approx (6744/50) + 0.578 \cdot [45 - (2248/50)] \approx 134.9.$$

This has estimated variance given by

$$s^2(\overline{y}_L) = \frac{1-f}{n} \left(s_Y^2 - \tilde{b} s_{YX} \right)$$

where

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \right) = \frac{1}{50-1} \left(928436 - 50 \cdot \left(\frac{6744}{50} \right)^2 \right) \approx 383.8$$

and

$$\begin{aligned} s_{YX} &= \frac{1}{n-1} \sum_{i=1}^{50} (y_i - \overline{y})(x_i - \overline{x}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \right) = \frac{1}{50-1} \left(305125 - 50 \cdot \left(\frac{6744}{50} \right) \left(\frac{2248}{50} \right) \right) \\ &\approx 39.1. \end{aligned}$$

Thus,

$$s^2(\overline{y}_L) = \frac{1-f}{n} \left(s_Y^2 - \tilde{b} s_{YX} \right) \approx \frac{1 - 50/1000}{50} (383.8 - 0.578 \cdot 39.1) \approx 6.863$$

so that an approximate 95% confidence interval for $\overline{Y}$ is given by $134.9 \pm 2(2.6)$.

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

Solution 1: We easily compute that

$$\text{RelEff}(\overline{y}_R, \overline{y}_L) = \frac{s^2(\overline{y}_R)}{s^2(\overline{y}_L)} \approx \frac{6.863}{14.402} \approx 0.477.$$

Since $0.477 \ll 1$, we can conclude that there is sufficient evidence to suggest that the variance of $\overline{y}_L$ is smaller than the variance of $\overline{y}_R$. This implies that the regression estimator is preferrable to the ratio estimator in this particular problem.

Solution 2: We easily compute that

$$\text{RelEff}(\overline{y}_R, \overline{y}_L) = \frac{s^2(\overline{y}_L)}{s^2(\overline{y}_R)} \approx \frac{14.402}{6.863} \approx 2.099.$$

Since $2.099 \gg 1$, we can conclude that 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.

11. **(a)** We find that for undergraduates, $n_1 = 723$, $n = 900$, $\theta = 4/30$. Thus, an estimator of p is given by

$$\hat{p} = \frac{n_1/n}{2\theta - 1} - \frac{1-\theta}{2\theta - 1} = \frac{723/900}{-22/30} - \frac{26/30}{-22/30} \approx 0.086.$$

The estimated variance is given by

$$s^2(\hat{p}) = \frac{1}{(2\theta-1)^2} \cdot \frac{1}{n} \cdot \frac{n_1}{n} \cdot \left(1 - \frac{n_1}{n}\right) = \frac{1}{(-22/30)^2} \cdot \frac{1}{900} \cdot \frac{723}{900} \cdot \left(1 - \frac{723}{900}\right) \approx 0.000326.$$

In other words, an approximate 95% confidence interval for p is given by $0.086 \pm 2(0.018)$.

(b) We find that for graduates, $n_1 = 117$, $n = 150$, $\theta = 4/30$. Thus, an estimator of p is given by

$$\hat{p} = \frac{n_1/n}{2\theta - 1} - \frac{1-\theta}{2\theta - 1} = \frac{117/150}{-22/30} - \frac{26/30}{-22/30} \approx 0.118.$$

The estimated variance is given by

$$s^2(\hat{p}) = \frac{1}{(2\theta-1)^2} \cdot \frac{1}{n} \cdot \frac{n_1}{n} \cdot \left(1 - \frac{n_1}{n}\right) = \frac{1}{(-22/30)^2} \cdot \frac{1}{150} \cdot \frac{117}{150} \cdot \left(1 - \frac{117}{150}\right) \approx 0.0021.$$

In other words, an approximate 95% confidence interval for p is given by $0.118 \pm 2(0.046)$.

(c) Since the confidence intervals computed from **(a)** and **(b)** overlap, there is no statistically significant difference in cocaine use between undergraduates and graduates. ETC.

12. For this problem, you earned full points no matter which sampling scheme and method of data collection you selected provided that you had a full discussion of your proposal including some potential limitations.

The most commonly chosen proposal was cluster sampling by direct observation. For this proposal, the variable of interest is the proportion of homeowners who band their trees to prevent *Dutch Elm Disease*. Therefore, the population in question is either all Regina homeowners, or all Regina homes. In either case, a street map of the City of Regina easily reveals all of the possible city blocks which will be used as the clusters (i.e., the sampling units are the city blocks and the frame is the list of those blocks). A simple random sample of blocks may now be conducted to decide which blocks to observe. It is quite easy for the investigator to drive along

those blocks and note both the number of banded trees and unbanded trees. One limitation to this scheme is that it may be difficult to observe the quantities of banded/unbanded trees in backyards; people may not appreciate the investigator peering over fences. However, it may be argued that direct observation will at least provide complete, and accurate, results for front yard trees which should lead to reasonable estimates of the proportion who band since it seems likely that homeowners will either band all trees (both front and back) or no trees (neither front nor back). Direct observation is to be contrasted with either phone surveys or questionnaires (in which people may either refuse to answer or make mistakes in their own counts). At least with direct observation, there is no issue with non-response. Perhaps, if instead of estimating simply the proportion of homeowners who band, it was desired to know the proportion of banded trees, and if there was enough money to do so, then direct observation could be combined with either a phone survey or questionnaire to try and deal with the issue of backyard trees.