NAME: SOLUTIONS STUDENT NO:

STA 304/1003F Test 2 November 20, 2009 SS 2117 1.10 to 2 p.m.

Aids: Two sides handwritten notes (8 1/2 x 11) and one non-programmable calculator.

There are **four** questions total, and **6 pages.**
Please answer all questions **on the question paper**

1. (25 marks) Some researchers undertook a survey taken to study how many children might potentially enrol in single-sex schools, if these were made available in their neighbourhood. Questionnaires were distributed to all parents who attended selected clinics in the Chicago area during a 1-week period for well- or sick-child visits.

   (a) (15 points) Suppose the quantity of interest is the number of children interested in transferring to single-sex schools. Describe why this is a cluster sample. What is the psu? The ssu? Is it a one-stage or two-stage cluster sample? How would you estimate the total number of children who might transfer, and the standard error of the estimate?

   – first clinics were sampled: these are the psu's (2)

   – Then parents were questioned: these are the ssu's (2)

   – One stage sampling since all parents were questioned (2)

   – Clinics 1, 2, ... n
   # parents M_1, M_2, ... M_n
   # yes y_1, y_2, ... y_n (3)

   – use total formula: $\frac{N}{n} \sum y_i$ (3)

   – and s.e. $N^2(1 - \frac{n}{N})\frac{s^2}{n}$ (3)

   (b) (10 points) Do you think this sampling procedure results in a representative sample of households with children? Why, or why not?

   Probably not (5)
   - could be flu season, perhaps more affluent parents go to clinic;
   - we don't know how many parents didn't complete questionnaire;
   - more likelihod to sample larger families (3) for one reason, (2) more for another

2. (30 marks) Suppose $y_{ij}, j = 1, \ldots, m_i; i = 1, \ldots, n$ is a sample of observations from a two-stage cluster sample.

(a) (10 marks) The anova table below shows the between and within sums of squares for the data on the cost of replacing books, based on a sample of $n = 12$ shelves and $m_i = 5$ books on each shelf. The estimate of R_a^2 from this table is about 0.41.

Source	df	Sum of Squares	Mean Square
between shelves	11	25571.0	2324.6
within shelves	48	23445.2	488.4
total	59	49016.2	

For this problem, is cluster sampling likely to be more precise or less precise than simple random sampling with the same number of sampled books? Explain.

MSB = 2324.6 (2),
$S^2 = 49016.2/50 = 830.8(2)$;
Since MSB bigger than S^2 (2), cluster sampling is less precise.
Variability between shelves is larger than between randomly sampled books. Probably because books shelved by themes (4).

(b) (10 points) The variance of the estimate of the population mean will depend on the within cluster variance and the between cluster variance. Show that

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\bar{y}_{i.}-\bar{y}_{..})^2,$$

where $\bar{y}_{i.} = \sum_{j=1}^{m_i} y_{ij}/m_i$ and $\bar{y}_{..} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}/\sum_{i=1}^{n} m_i$.

$$\begin{aligned} LHS &= \sum\sum(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{y}_{..})^2 (5) \\ &= \sum\sum(y_{ij}-\bar{y}_{i.})^2 + \sum\sum(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum\sum(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..}) \end{aligned}$$

Last term is zero because $\sum_j(y_{ij}-\bar{y}_{i.}) = m_i\bar{y}_{i.} - m_i\bar{y}_{i.} = 0$ (2 for stating it is zero, 3 for proving)

(c) The sampling weight for y_{ij} is $w_{ij} = 1/\pi_{ij}$, where π_{ij} is the probability that unit j in psu i is selected.
Show that

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} \hat{t}_i$$

is an unbiased estimate of the population total.
Hint: First show that $\hat{t}_{unb} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} w_{ij} y_{ij} Z_{ij}$, where $Z_{ij} = 1$ if ssu j in cluster i is sampled, and 0 otherwise.

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} \frac{M_i}{m_i} \sum_{j \in \mathcal{S}_\rangle} y_{ij} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_\rangle} \frac{N}{n} \frac{M_i}{m_i} y_{ij} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_\rangle} w_{ij} y_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} w_{ij} y_{ij} Z_{ij}$$

because $Z_{ij} = 1$ if y_{ij} is in the sample.

$$E(\hat{t}_{unb}) = \sum \sum \frac{1}{\pi_{ij}} y_{ij} E(Z_{ij}) = \sum \sum \frac{1}{\pi_{ij}} y_{ij} \pi_{ij} = t$$

(4) for showing the hint; (1) for $w_{ij} = \frac{NM_i}{nm_i}$ (5) for expected value

NAME: SOLUTIONS STUDENT NO:

3. (30 marks) Otters are semi-aquatic mammals that live in dens in coastal areas. Scientists used a stratified sample to estimate the number of otter dens along the 1400-km coastline of Shetland, UK. The coastline was divided into 5-km sections, and each section was assigned to the stratum whose terrain type predominated. Sections were then chosen randomly from the sections in each stratum. In each section chosen, investigators counted the total number of dens in a 110-meter wide strip along the coast. The population and sample sizes are as follows:

Stratum	Total Sections N_h	Sampled Sections n_h	Total No. of Dens in sample	Average No. of Dens $\bar{y}_h$	Variance s_h^2
1 Cliffs over 10 m	89	19	33	1.737	5.427
2 Agriculture	61	20	35	1.750	6.829
3 Not 1 or 2, peat	40	22	292	13.273	58.779
4 Not 1 or 2, nonpeat	47	21	86	4.095	15.590

(a) (10 marks) Estimate the total number of otter dens along the coast in Shetland, along with a standard error for your estimate.

$$\hat{t} = \sum_h n_h \bar{y}_h = 89 \times 1.737 + 61 \times 1.750 + 40 \times 13.273 + 47 \times 4.095 = 984.7$$

(5)

$$\begin{aligned}
\widehat{V}(\hat{t}) &= \sum_{h=1}^{H} (1 - \frac{n_h}{N_h}) N_h^2 \frac{s_h^2}{n_h} \\
&= (1 - \frac{19}{89}) 89^2 \frac{5.427}{19} + (1 - \frac{20}{61}) 61^2 \frac{6.829}{20} + (1 - \frac{22}{40}) 40^2 \frac{58.779}{22} \\
&\quad + (1 - \frac{21}{47}) 47^2 \frac{15.590}{21} \\
&= 5464.317 = 73.92^2
\end{aligned}$$

(5)

(b) (10 marks) Discuss possible sources of bias in this study. Do you think it is possible to avoid all selection and measurement bias?

- sections might not neatly stratify as indicated;
- dens may be missed;
- probability of missing may depend on terrain
- probably not possible to avoid all bias

(c) (10 marks) Did the scientists use proportional allocation in deciding how many sections to sample? Explain.

no. (4)

PA means n_h/N_h is constant. Here $n_h \approx 20$ no matter what N_h is (6)

(d) **Bonus**: If the costs of sampling the four strata are $c_1 = 9, c_2 = c_3 = c_4 = 1$ and the total sample size is limited to 84, what is the optimal allocation? (Use s_h^2 for your estimate of the variance in each stratum.)

$$\begin{aligned} n_h &\propto \frac{N_h s_h}{c_h} = (\frac{89\sqrt{5.427}}{3}, \frac{61\sqrt{6.829}}{1}, \frac{40\sqrt{58.779}}{1}, \frac{47\sqrt{15.590}}{1}) \\ &\propto 69, 159, 307, 186 \end{aligned}$$

so $n_{opt} = (8, 18, 36, 22)$ as they must sum to 84

4. (15 marks)

   (a) What is the difference between a random sample and a systematic sample?

   (b) What is an advantage of a systematic sample?

   (c) Give an example of a sampling problem where systematic sampling might be useful and effective.

(5 points each)

(a) - in SRS each sampling unit has equal probability to be selected; in a systematic sample only the starting point is random, after that each kth unit is selected

(b) - systematic is cheaper, easier, close to random if the list is in random order, can be more precise if list is in decreasing or increasing order

(c) - sampling hazardous waste sites in US on grid points from a random start gives good coverage of entire area