Survey Sampling
Statistics 4234/5234 — Fall 2017
Second in-class exam

November 16, 2017

NAME: UNI:

Instructions: Write your name and UNI in the spaces provided above, and on your bluebook. Do not turn over this page until instructed to do so.

You have 75 minutes to complete this examination. Read each part of each question carefully. There are a total of 60 points on this exam — you are responsible for checking that your paper is complete. You are permitted one $8\frac{1}{2}$ by 11 sheet (both sides) of original handwritten notes, and a hand-held calculator. **No other outside material or assistance is permitted.**

Please sign below to indicate your agreement with the Columbia College Honor Code, whether or not you are a student of Columbia College.

> I affirm that I will not plagiarize, use unauthorized materials, or give or receive illegitimate help on assignments, papers, or examinations. I will also uphold equity and honesty in the evaluation of my work and the work of others. I do so to sustain a community built around this Code of Honor.

SIGNATURE: _____

1. (18 points) Consider a population of six students. Suppose we know the test scores of the students to be

Student	1	2	3	4	5	6
Score	66	59	70	83	82	71

You are given that the mean and variance of this population are 71.83 and 86.17, respectively.

(a) There are 15 possible SRS's of size 4. List them. (Use shorthand—the sample consisting of the first, second, fourth and fifth units, for example, can be listed as 1245.)

(b) Let stratum 1 consist of students 1–3 and stratum 2 consist of students 4–6. List the possible stratifed random samples of size 4, in which 2 students are selected from each stratum.

(c) Find the sampling distribution of $\bar{y}_{\text{str}}$ for the stratified sampling scheme described in part (b).

(d) For simple random sampling we have

$$E(\bar{y}) = \bar{y}_U = 71.83$$

and

$$V(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{86.17}{4}\left(1 - \frac{4}{6}\right) = 7.18$$

i. Is $E(\bar{y}_{\text{str}})$ greater than, less than, or equal to 71.83?

ii. Is $V(\bar{y}_{\text{str}})$ greater than, less than, or equal to 7.18?

2. (14 points) Norwegian researchers used stratification techniques to estimate ringed seal populations in Svalbard fjords. The study area was divided into three zones, and each zone was divided into a number of plots (N_h in the table below). A random sample of plots in each zone was examined, and the number of breathing holes in each sampled plot was recorded.

Zone	N_h	n_h	$\bar{y}_h$	s_h
1	68	17	1.76	1.82
2	84	12	4.42	3.40
3	48	11	10.55	6.79

(a) Estimate the total number of breathing holes in the study region.

(b) Give the standard error of your estimate in part (a).

(c) You have been asked to design a follow-up survey to estimate the total number of breathing holes at a later date; a total of 40 plots are again to be sampled. How many plots would you sample from each of Zones 1, 2 and 3?

3. (22 points) In a simple random sample of $n = 31$ black cherry trees from a forest of $N = 2967$ trees, the mean timber volume per tree was 30.17 cubic feet, with a standard deviation of 16.44 cubic feet.

   (a) Calculate the standard error of $\bar{y} = 30.17$ as an estimate of $\bar{y}_U$, the mean timber volume per tree for *all* trees in the forest.

   (b) Give a 95% confidence interval for the *total* timber volume of all trees in the forest.

   Now suppose it is known that the sum of the diameters for all the trees in the forest is $t_x = 41{,}835$ inches. For the 31 sampled trees, the average girth (diameter) was 13.25 inches, with a standard deviation of 3.14 inches. The sample correlation between girth and volume was 0.967.

   (c) Use ratio estimation to estimate the total volume for all trees in the forest.

   (d) Use regression estimation to estimate the total volume for all trees in the forest.

   In parts (c) and (d) your answer will consist of a point estimate only.

   (e) Suppose the relative efficiency of ratio estimation versus the ordinary sample mean is about 2.5, and the relative efficiency of regression estimation versus ratio estimation is about 6; that is

$$\frac{\hat{V}(\bar{y})}{\hat{V}(\hat{\bar{y}}_r)} = 2.5 \qquad \text{and} \qquad \frac{\hat{V}(\hat{\bar{y}}_r)}{\hat{V}(\hat{\bar{y}}_{\text{reg}})} = 6.0$$

   Give the shortest possible (valid) 95% confidence interval for total timber volume you can find from these data.

4. (6 points) In Section 3.5 of the textbook, the section titled *Defining Strata*, Lohr writes

   > In a survey with more precise prior information, we will want to use more strata—many surveys are stratified to the point that only two sampling units are observed in each stratum.

   Suppose we have the resources to take a total sample size of n, and sufficiently precise information about the population that we can intelligently define n strata. We can then take a stratified random sample of just *one* unit in each stratum. Does this seem like a good idea? What practical or statistical problem might result from such a sampling scheme?

   (Keep it short! You should be able to answer this with a single sentence.)