Statistics 4234/5234 — Fall 2018
Midterm 1

October 16, 2018

Name: UNI:

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

You have 70 minutes to complete this examination. Read each part of each question carefully. There are a total of 50 points on this exam — you are responsible for checking that your paper is complete. You are permitted one $8\frac{1}{2} \times 11$ 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: ______________________________

Section 1: True or false? Circle the appropriate choice (1 point each).

1. Simple random sampling without replacement (SRS) is the most fundamental sampling design in sample surveys, because SRS guarantees that the sample will be representative of the population.

   TRUE FALSE

2. The *observation unit* is the basic unit of observation; in studying human populations, observation units are often individuals.

   TRUE FALSE

3. The *target population* is the collection of all possible observation units that might have been chosen in a sample, that is, the population from which the sample is taken.

   TRUE FALSE

4. A *sampling unit* is a unit that can be selected for a sample; for example, households may serve as the sampling units, while the observation units are the individuals living in the households.

   TRUE FALSE

5. An important consideration in survey sampling is *questionnaire design*, because confusing or poorly worded questions can lead to *selection bias* in a survey.

   TRUE FALSE

6. Failure to include all of the target population in the sampling frame leads to *undercoverage*.

   TRUE FALSE

7. When a response in the survey differs from the true value, *measurement error* has occurred; *measurement bias* occurs when the response has a tendency to differ from the true value in one direction.

   TRUE FALSE

8. *Nonsampling error* refers to error that cannot be attributed to sample-to-sample variability; selection bias and measurement error are examples of nonsampling error.

   TRUE FALSE

9. Consider a population consisting of 4 clusters of 30 units each, suppose a *one-stage cluster sample* is drawn (one of the four clusters is selected at random), and the population total is estimated by $\hat{t}_y = 4\sum_{i\in\mathcal{S}} y_i$; then $\hat{t}_y$ is a *biased* estimator of the population total t_y.

   True False

10. For a *biased* estimation method, the mean squared error of the estimator exceeds its variance; for an *unbiased* estimator, the mean squared error and variance are equal.

    True False

11. Under simple random sampling from a population $\{y_1, \ldots, y_N\}$ with mean $\bar{y}_U$ and variance $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$, the sample variance $s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2$, where $\bar{y}$ denotes the sample mean, is unbiased: $E(s^2) = S^2$.

    True False

12. Under simple random sampling of size n from a population of size N, the standard error of the sample mean is given by

    $$\mathrm{SE}(\bar{y}) = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}\,,$$

    where s denotes the sample standard deviation; ignoring the $\sqrt{1 - n/N}$ term will lead to *conservative* inference, i.e., confidence intervals that are wider than they need be.

    True False

13. Letting $\bar{y}_{\mathrm{SRS}}$ denote the sample mean from a simple random sample without replacement, and $\bar{y}_{\mathrm{SRSwR}}$ the sample mean from a simple random sample *with* replacement; the means and variances of their respective sampling distributions satisfy

    $$E(\bar{y}_{\mathrm{SRS}}) = E(\bar{y}_{\mathrm{SRSwR}}) \quad \text{and} \quad V(\bar{y}_{\mathrm{SRS}}) \leq V(\bar{y}_{\mathrm{SRSwR}}).$$

    True False

14. Letting n_{SRS} denote the sample size required to satisfy $P(|\bar{y} - \bar{y}_U| \leq e) \geq 0.95$, for a particular margin of error e, and n_{SRSwR} the analogous quantity but assuming simple random sampling *with* replacement, then $n_{\mathrm{SRS}} \geq n_{\mathrm{SRSwR}}$.

    True False

15. Stratified random sampling generally results in biased estimators of the population mean and population total; however, this bias is usually offset by a reduction in variance, and actually results in lower mean squared error.

True False

16. If the sample sizes within each stratum are sufficiently large (or the sampling design has a very large number of strata), an approximate 95% confidence interval for the population mean $\bar{y}_U$ is given by

$$\bar{y}_{\text{strat}} \pm 1.96 \text{ SE} \left(\bar{y}_{\text{strat}}\right),$$

where $\bar{y}_{\text{strat}}$ is the appropriately weighted average of the within-stratum sample means $\bar{y}_h$.

True False

17. The greater are the within-stratum variances S_h^2 relative to the overall variance S^2, the more dramatic the gain in precision from stratified sampling.

True False

18. Under simple random sampling, the inclusion probability is the same for every unit in the population, and given by $\pi_i = n/N$; the sampling weight $w_i = 1/\pi_i$ can be interpreted as the number of population units represented by unit i (if it is included in the sample).

True False

19. In the optimal allocation of observations for stratified sampling, the greater the within-stratum variance S_h^2, the less heavily stratum h should be sampled (other things being equal).

True False

20. When little or no prior information about the target population is available, there is little to be gained from stratification, and you may as well use an SRS; even then, there is generally no loss in precision from stratifying, just no substantial gain to justify additional complexity of sampling design.

True False

End of Section 1.

Section 2: Answer all questions in the space provided.

1. (12 points) Consider taking a simple random sample of size $n = 2$ from the following population of size $N = 4$:

$$\{y_1, y_2, y_3, y_4\} = \{33, 33, 30, 36\}$$

   (a) Specify the sampling distribution of $\bar{y}$, the sample mean for a SRS of size $n = 2$.

   (b) Consider estimating the population mean with the interval $\bar{y} \pm k$, based on SRS of size $n = 2$.

      i. Give the confidence level of the interval $\bar{y} \pm 1$.

      ii. Give the confidence level of the interval $\bar{y} \pm 2$.

   (c) Assuming a simple random sample of size $n = 2$, specify the sampling distribution of the sample standard deviation s.

2. (12 points) The following table summarizes the results of a stratified random sample of faculty at a large state university in the early 1980s.

Rank	Number sampled	Mean salary	SD of salaries
Assistant	18	18000	2000
Associate	14	23000	2000
Professor	20	30000	4000

Assume proportional allocation was used to determine the sample sizes.

(a) Estimate the mean salary among this university's faculty.

(b) Give the standard error of your estimate in part (a). Express your answer as the simplest possible function of N = total number of faculty at this university.

(c) A follow-up study will be conducted, this time based on a total sample size of 78 faculty. Would you recommend simple random sampling, or another stratified sample? If the latter, how many faculty members should be sampled from each rank?

3. (6 points) A common method for estimating the size of an audience is to take an SRS of n of the N rows in an auditorium, count the number of people in each of the selected rows, and multiply the total number of people in your sample by N/n.

   A small theater has $N = 5$ rows; for a recent performance there were 10 people seated in the first row, 9 in the second row, 8 in the third, 7 in the fourth, and 6 in the back row. Give the bias, variance, and mean squared error for the estimator defined above with $n = 1$.