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

Answers

1. The terms undercoverage and overcoverage describe disagreement between the target population and the set of units included in the sampling frame.
   TRUE

2. Convenience sampling is an unbiased sampling method, as the units that are easiest to select or most likely to respond are systematically no different than the harder-to-select or nonresponding units.
   FALSE

3. As long as a questionnaire is otherwise well designed, question wording and question ordering will have very little effect on the responses obtained.
   FALSE

4. In designing survey questionnaires, it is advisable to write specific rather than general questions, and offer a selection of choices to answer rather than simply "agree or disagree."
   TRUE

5. In designing survey questionnaires, it is generally recognized that the more questions included, the better; if we're going to conduct the survey, we should learn everything we can about the respondents while we have their attention.
   FALSE

6. The usual "margin of error" reported in published surveys only accounts for sampling error; the assessment of other sources of error (such as selection bias or measurement error) generally requires additional information about how the survey was conducted.
   TRUE

7. It is conceivable that estimates based on sample surveys might be *more* accurate than those based on a census, because investigators can be more careful when collecting data.
   TRUE

8. Simple random sampling is probabilistically equivalent to drawing n balls at random from an urn containing N balls; the result is that each of the $\binom{N}{n}$ possible samples has the same chance of being the sample selected.
   TRUE

9. Under simple random sampling, every unit in the population has the same probability of being selected; for an SRS of size n from a population of size N this probability is given by $\pi_i = n/N$ for $i = 1, \ldots, N$.
   TRUE

10. In estimating a numerical property of a finite population based on probability sampling, an unbiased estimator is always preferable to a biased one.
    FALSE

11. Under simple random sampling, holding the sample size n fixed, the larger the population size N the more precise the estimator (in terms of lower standard error), owing to the finite population correction.

FALSE

12. A simple random sample *with* replacement will generally yield more precise estimation of quantities of interest than ordinary SRS (without replacement).

FALSE

13. Under simple random sampling, the true coverage rate of a nominal 95% confidence interval is equal to the percentage of the $\binom{N}{n}$ possible samples for which the CI formula, applied to that sample, would contain the true parameter value.

TRUE

14. In estimation based on simple random sampling, the larger the sample size n, the more precise the estimator (in terms of a lower standard error) and the narrower will be the confidence interval.

TRUE

15. As long as the sample size is sufficiently large, measurement bias, selection bias and nonsampling error are essentially non-issues, and can be ignored.

FALSE

1. The article "What People Buy from Fast-Food Restaurants: Caloric Content and Menu Item Selection" (Obesity [2009]; 1369-1374) reported that the average number of calories consumed at lunch in New York City fast food restaurants was 827.

   The researchers selected 267 fast food locations at random. The paper states that at each of these locations "adult customers were approached as they entered the restaurant and asked to provide their receipt when exiting and to complete a brief survey."

   (a) How might nonresponse bias manifest itself in this study?

   If people who order larger meals are less likely to participate in the survey, our estimate of the average calorie count will be negatively biased.

   (b) Might measurement bias be an issue? If not, why not; and if so, how?

   If people who agree to participate then order less than they would have otherwise, our estimate of the average calorie count will be negatively biased.

2. The following population of $N = 8$ units:

i	1	2	3	4	5	6	7	8
y_i	9	12	14	13	9	15	12	12

   has population mean and variance of $\bar{y}_U = 12$ and $S^2 = 4.57$, respectively.

   (a) Consider a sampling scheme in which the sample consists of either the first 4, 5 or 6 units, with equal probability. That is,

sample	probability
$\{1, 2, 3, 4\}$	1/3
$\{1, 2, 3, 4, 5\}$	1/3
$\{1, 2, 3, 4, 5, 6\}$	1/3

For the sampling distribution of the sample mean $\bar{y}$ we note that

$$\bar{y}_{s_1} = 12.0 \quad \textit{and} \quad \bar{y}_{s_2} = 11.4 \quad \textit{and} \quad \bar{y}_{s_3} = 12.0$$

and thus $\bar{y}$ equals 11.4 with probability 1/3, and 12.0 with probability 2/3.

(b) *For the sampling scheme in part (a),*

$$E[\bar{y}] = 11.4\left(\frac{1}{3}\right) + 12.0\left(\frac{2}{3}\right) = 11.8$$

and

$$V[\bar{y}] = (11.4 - 11.8)^2\left(\frac{1}{3}\right) + (12.0 - 11.8)^2\left(\frac{2}{3}\right) = 0.08$$

so

$$\textit{Bias}(\bar{y}) = -0.2 \quad \textit{and} \quad \textit{MSE}(\bar{y}) = 0.08 + (-0.2)^2 = 0.12$$

(c) *For a simple random sample of size $n = 4$,*

$$E[\bar{y}] = \bar{y}_U = 12 \quad \textit{and} \quad V[\bar{y}] = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{4.57}{4}\left(1 - \frac{4}{8}\right) = 0.57$$

so

$$\textit{Bias}(\bar{y}) = 0 \quad \textit{and} \quad \textit{MSE}(\bar{y}) = V[\bar{y}] = 0.57$$

3. In a simple random sample of 935 assistant nurses from a Norwegian county with 2700 assistant nurses, a total of 745 assistant nurses responded; 149 of the 745 respondents reported that bullying occurred in their department.

(a) *We estimate the proportion of assistant nurses in the county who would report bullying in their department by*

$$\hat{p} = \frac{149}{745} = 0.20$$

(b) *A standard error for vhe above point estimate is*

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{(.20)(.80)}{744}\left(1 - \frac{745}{2700}\right)} = 0.01248$$

(c) *A 95% confidence interval for the proportion of assistant nurses in the county who would report bullying in their department is*

$$\hat{p} \pm 1.96SE(\hat{p}) \quad \Rightarrow \quad .20 \pm 1.96(.0125) \quad \Rightarrow \quad [.1755, .2245]$$

Converting from proportions to total number *of nurses who would report bullying,*

$$2700(.1755) = 473.85 \quad \textit{and} \quad 2700(.2245) = 606.15$$

and we are 95% confident that between 474 and 606 of the 2700 assistant nurses would report bullying in their department.

(d) What assumptions must you make about the nonrespondents for the above analysis to be valid?

We assume that nurses who would report bullying in their department are as a group no more or less likely to participate in the survey than are the nurses who would not report bullying.