BIOS 500H U1 Homework Solutions

Book Problems (Chapter 22)

1. NOT ASSIGNED When you are conducting a survey, how is the study population related to the target population? What is the sampling frame?

   The study population is a subset of the target population we would like to describe, but it is more accessible and practical to sample from. The sampling frame is "a list of the elements in the study population."

2. NOT ASSIGNED How does the finite version of the central limit theorem differ from the more commonly used version, in which the underlying population is assumed to be infinite?

   The commonly used central limit theorem suggests "that the distribution of the mean of the sample values was approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$ ". The finite version suggests that the sample mean $\bar{x}$ has mean μ and standard deviation $\sqrt{1-(n/N)}(\sigma/\sqrt{n})$.

3. When might you prefer to use systematic sampling rather than simple random sampling? When might you prefer a multistage sample over an SRS? Briefly say why.

   Stratified may be preferable over SRS when you want to ensure sufficient sample size in each group or stratified may be more efficient when the groups are homogeneous. Multistage may be preferable over SRS when it is logitistically easier to collect data within a smaller number of clusters.

4. How can nonresponse result in a biased sample? What could you do to attempt to minimize nonresponse? Is nonresponse classified as a sampling error or a non-sampling error? Explain briefly.

   If subjects who do not respond are different from the subjects who do respond, then the bias may result. For example, suppose a survey is performed to study the general health of individuals living in Orange County. The survey is performed by calling local telephone numbers and asking a set of questions. The survey finds

that Orange County residents tend to be in declining health. Here, Orange County residents may not have responded because they do not own a local telephone number or because they work during the day and are unavailable to respond. The remaining participants are likely to be elderly, retired individuals, who are not in the prime of their health. In other words, those who respond are different from the non-responders. To minimize nonresponse, a few examples of what you could do include using multiple forms of contact, employing randomized responses, carefully wording questions, and offering incentives. Nonresponse is nonsample error. It doesn't have to do with sampling.

5. A study was conducted to examine the effects of maternal marijuana and cocaine use on fetal growth. Drug exposure was assessed in two different ways: the mothers were questions directly during an interview, and urinanalysis was performed.

   (a) Suppose that it is necessary to rely entirely on the information provided by the mothers. How might nonresponse affect the results of the study?

   Noneresponse might bias the results of the study. It is conceivable, for example, that mothers who use either marijuana or cocaine might be more likely to refuse to respond to questions about drug use in this case, an estimate of drug use prevalence based on the survey would underestimate the true population prevalence.

   (b) An alternative strategy might be to interview only those women who agree to be questioned. Do you feel that this method would provide a representative sample of the underlying population of expectant mothers? Why or why not?

   Interviewing only those mothers who agree to be questioned is unlikely to provide a representative sample from the population of expectant mothers. Again, we might expect that women who agree to be interviewed are less likely to use either marijuana or cocaine, while those who refuse are more likely to use drugs.

6. NOT ASSIGNED Each year, the United States Department of Agriculture uses the revenue collected from excise taxes to estimate the number of cigarettes consumed in this country. Over the 11-year period 1974 to 1985, however, repeated surveys of smoking practices can account for only about 72% of the total consumption.

(a) How would you explain this discrepancy between the estimates of cigarette consumption?

It is possible that some of the individuals questioned in the surveys about smoking practices lie; they claim to be nonsmokers when they really are not, or underestimate the number of cigarettes they actually smoke per day.

(b) Which source are you more likely to believe, the excise tax revenue or the surveys of smoking practices?

The excise tax revenue is more believable; if excise taxes were paid, the cigarettes must have been purchased by someone.

7. The data set **low_b_wt** contains information describing 100 low birth weight infants born in Boston, Massachusetts. Assume that these infants constitute a finite population. Their measures of systolic blood pressure are saved under the variable name sbp; the mean systolic blood pressure is $\mu = 47.1$ mm Hg. Suppose that we do not know the true population mean and wish to estimate it using a sample of 20 newborns.

Bios 500H - Unit 1, Sampling: Unit 1 HW - Low Birth Weight Infants Born in Boston, MA
Descriptive Values for the Entire Low Birth Weight Data Set (N=100)
The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Median
sbp	Systolic Blood Pressure (mmHg)	100	47.08	11.40	47.00
sex	Gender	100	0.44	0.50	0.00
tox	Tox	100	0.21	0.41	0.00
grmhem	Germinal Matrix Hemorrhage	100	0.15	0.36	0.00
gestage	Gestational Age (weeks)	100	28.89	2.53	29.00
apgar5	APGAR Score (5 minutes)	100	6.25	2.43	7.00
ID	ID	100	50.50	29.01	50.50

There are N = 100 observations in the population, and their average systolic blood pressure $\mu = 47.1$ mmHg.

(a) What is the sampling fraction of the population? What is the sampling weight?

The sampling fraction is $n/N = 20/100 = 0.2$. The sampling weight is $N/n = 100/20 = 5$.

(b) Modify SAS code given in `500_U1_Sampling.sas` to produce SRS of size 20 from `low_b_wt.sas7bdat` as requested using PROC SURVEYSELECT. State the IDs selected and the average SBP from the sample.

Results will vary by choice of SEED value. (Here, SEED=3107). The mean SBP value for this sample is 46.05. The IDs selected were 7, 12, 14, 23, 28,

29, 34, 35, 36, 37, 44, 54, 59, 61, 67, 84, 88, 93, 95, and 98.

Bios 500H - Unit 1, Sampling: Unit 1 HW - Low Birth Weight Infants Born in Boston, MA
SRS of n=20 observations, using PROC SURVEYSELECT

Obs	sbp	sex	tox	grmhem	gestage	apgar5	ID
1	31	Male	Yes	No	27	7	7
2	47	Female	No	Yes	30	6	12
3	56	Female	No	No	29	1	14
4	54	Male	No	No	27	4	23
5	39	Male	No	No	28	7	28
6	29	Female	No	No	29	4	29
7	34	Male	No	No	29	9	34
8	62	Female	No	No	27	7	35
9	59	Female	No	No	27	8	36
10	36	Male	No	No	27	9	37
11	48	Female	No	Yes	31	7	44
12	40	Male	No	No	27	7	54
13	40	Female	No	Yes	26	3	59
14	53	Male	Yes	No	29	9	61
15	58	Female	Yes	No	33	7	67
16	46	Female	Yes	No	34	9	84
17	48	Male	No	No	30	5	88
18	49	Female	Yes	No	32	8	93
19	45	Male	No	No	31	9	95
20	47	Male	Yes	No	32	5	98

(c) NOT ASSIGNED Use SAS (or Excel) to produce initial random number. You can select the systematic sample of size 20 by hand (if more convenient). State the IDs selected. Then calculate the mean SBP for your systematic sample. Use whatever is most convenient to calculate the mean SBP for the sample. This statement may be helpful if you'd like to compute the mean SBP for the sample in SAS:

```
PROC MEANS DATA=lowbwt MAXDEC=3;
WHERE id IN ( ); *separate selected ids by commas;
RUN;
```

Is the systematic sample EPSEM? Is it a SRS?

First, we need to output a random number between 1 and 5 (because $100/20 = 5$). My results are based on an output of 2, but any value between 1 and 5 may occur. Then we simply use IDs $2 + 5k, k = 0, \ldots, 19$. For 2, the mean SBP is 46.4.

To randomly generate a number in excel, simply use =RANDBETWEEN(1,5).

Yes, the systematic sample is EPSEM because everyone has a probability of 1/5. However, it the systematic sample is not SRS because not every sample is equally likely.

(d) Sort the data on TOX. Select a stratified sample with sample size = 10 in each strata using PROC SURVEYSELECT. What is the average SBP in each sample strata (each level of TOX)? Estimate of population mean SBP by calculating a weighted average of the strata sample averages as described in the lecture notes. Check your answer using PROC SURVEYMEANS.

With SEED = 3107, the average SBP for Tox = No is 45.4 mm Hg while that for Tox = Yes is 49.9 mm Hg.

With SEED = 500, the average SBP for Tox = No is 51.2 mm Hg while that for Tox = Yes is 44.8 mm Hg.

(e) What are the sampling fractions in each of the two strata?

For the Tox=No strata, the sampling fraction is $10/79 = 0.127$. For the Tox=Yes strata, the sampling fraction is $10/21 = 0.476$.

Bios 500H - Unit 1, Sampling: Unit 1 HW - Low Birth Weight Infants Born in Boston, MA
Select a Stratified Random Sample, using Proc Surveyselect
Strata: Tox (yes/no), 10 observations selected in each category = 20 total observations
data=stratified_sample_low_b_wt
Descriptive Statistics for Each Age Category SEPARATELY from Stratified Sample
The MEANS Procedure

Tox=No

Analysis Variable : sbp Systolic Blood Pressure (mmHg)	
N	Mean
10	51.20

Tox=Yes

Analysis Variable : sbp Systolic Blood Pressure (mmHg)	
N	Mean
10	44.80

Supplemental Problems

1. (a) Select a SRS of size 15 units from a population of 820 units using "Method 1" from notes and Random Number Table starting at Row 30.

   212 641 117 052 676 625 399 722 220 500 645 689 140 242 416

   (b) Select a SRS of size 6 units from a population of 620 units using "Method 2" from the notes and Random Number Table starting at Row 10.

   530 229 331 335 329 370

2. A public health firm is interested in determining the proportion of residents who have access to "Improved Sanitation" in a district in Uganda. The district was divided into three Regions, A,B,C which vary in population density, proximity to water sources, household income and access to septic systems.

   There are 1550 households in Region A,

   620 in Region B and

   930 in Region C.

   The firm decides to randomly select 10 households from Region A, 20 households from Region B and 15 households from Region C.

   (a) Which of the following best describes this study design?

   iii. A Stratified Sample

(b) State the sampling fraction in each region. State the sampling weight in each region. In English, what does the sampling weight represent in, say, Region A?

Sampling fraction of A: 0.006; sampling weight: 155

sampling fraction of B: 0.03; sampling weight: 31

sampling fraction of C: 0.02; sampling weight: 62

For every person sampled from region A, they represent 155 people of that population (region A).