Biostatistical Methods I (BIOS 5710)
Breheny

Assignment 1
Due: Thursday, September 8

1. One study on slavery in America estimated that "11.9% of slaves were skilled craftsmen." This estimate was based on the records of thirty plantations in Plaquemines Parish, Louisiana. The study is making a generalization about a population based on a sample.

   (a) In this study, what is the population of interest?

   (b) In this study, what does the sample consist of?

   (c) Could the study's generalization be biased? If so, what type of bias would this be?

2. A Dutch study of rheumatoid arthritis found that patients who experienced more pain were more likely to respond to a health survey. In their questionnaire, 2% of respondents said that they experience no pain associated with their condition.

   (a) Is the "2%" number an estimate or a parameter of interest?

   (b) Is the "2%" number a biased estimate? If so, what type of bias is present?

   (c) Is the true percentage of individuals with rheumatoid arthritis who suffer no pain likely to be less than 2%, greater than 2%, or equal to 2%?

3. This question involves an article by Abramowicz *et al.* (2008) on the epidemiology of ultrasound [Link]. Read the section on the second page titled "Low Birth Weight."

   (a) The authors discuss the use of animal studies to make inferences about humans. What sampling concept that we discussed in class does this violate?

   (b) Are the observational studies and controlled experiments in agreement?

   (c) Do the authors seem to place more emphasis on the observational studies or the randomized controlled experiments?

4. A 1979 randomized trial was performed comparing a surgical treatment for angina pectoris (chest pains due to obstruction of the coronary arteries) to a placebo (non-surgical medical management). In the study, 6 of the patients randomized to receive surgery died before they could be operated on. There was a subsequent debate over how best to analyze the data. Some favored approach A, in which all patients were analyzed as they were randomized. Others favored approach B, in which those 6 patients were excluded on the grounds that they never actually received the surgical treatment that they were randomized to. Which approach do you think is better? Why?

5. [vBFHL2.3] Give an example of an experiment that would be difficult to blind (i.e., difficult to create a placebo for). This does not have to be a real study, but the example must be specific.

6. [vBFHL2.1] For each of the studies in Chapter 1.5.2, 1.5.3, and 1.5.4, is the study a controlled experiment or an observational study?

7. A study of sexual bias in admissions was conducted by the Graduate Division at the University of California, Berkeley. Admissions results from the six largest majors are listed below (university policy does not allow the departments to be identified by name):

| | Men | | Women | |
Major	Applicants	Admitted	Applicants	Admitted
A	825	512	108	89
B	560	353	25	17
C	325	120	593	202
D	417	138	375	131
E	191	53	393	94
F	373	22	341	24

(a) From the table above, calculate the overall percentage of men and the overall percentage of women who were admitted.

(b) Does your answer from part (a) suggest sexual bias? If so, against whom?

(c) Create a table listing the percentage of men and women who were admitted, broken down by department.

(d) Does the table you made in part (c) indicate sexual bias? If so, against whom?

(e) Construct a weighted average of the percentage of male and female applicants who were admitted, controlling for the effect of department (*i.e.*, report one number for men and one number for women).

(f) Does your answer from part (e) indicate sexual bias? If so, against whom?

(g) The purpose of the study was to answer the question, "Were admissions biased on the basis of sex"? The analyses above (i.e., comparing answers from (b), (d), and (f)) indicate different answers to this question. Why do the answers differ? And how would you answer the question of whether a sex bias was present?

8. The course web page contains data (`whickham.txt`) from a British population health study carried out in Whickham from 1972-1974. Twenty years later, a follow-up study was conducted. The data set contains the smoking status and age (categorized for the sake of simplicity) of 1,314 women from the original study, along with whether the woman was alive or not when the twenty-year follow-up was conducted.

(a) What percent of smokers were alive 20 years later?

(b) What percent of nonsmokers were alive 20 years later?

(c) Based on your answers from (a) and (b), does smoking appear to be beneficial or harmful?

(d) Based on this sample, describe the relationship between smoking and age in the early 1970s in Whickham.

(e) Based on your answers from (c) and (d), describe how a confounding relationship might be present.

(f) If you adjust for the possible confounder of age by carrying out comparisons in more homogeneous subgroups, does smoking appear beneficial or harmful (back up your conclusion with numbers).

(g) Make a bar chart depicting the number of women who were/were not alive vs. smoking status.

(h) Make a bar chart depicting the number of women who were/were not alive vs. smoking status, conditioning on age.

9. [vBFHL3.7] Suppose a new variable y is calculated from x according to the formula $y = a + bx$ (such a formula is called a linear transformation). For all of the following, derive the descriptive statistic for y in terms of a, b, and the corresponding statistic for x. Show your work.

   (a) Mean
   (b) Median
   (c) IQR
   (d) Variance
   (e) Standard deviation

10. [vBFHL3.17] A continuous variable Y is grouped into bins of width w (as you would do when making a histogram). Suppose you tried to calculate the mean of Y by representing all the Y values in a bin by the midpoint of the bin interval. What is the maximum error possible?