Stat 221 Exam 1 October 14, 2008

1. (8 pts, 4 pts each) Suppose that the observational units in a study are restaurants in San Luis Obispo county.

a) State a *quantitative* variable that you could record on these observational units.

Examples include: number of customers that the restaurant served yesterday, square footage of the restaurant, number of items on the menu, average price of an item on the menu.

b) State a *categorical* variable that you could record on these observational units.

Examples include the type of cuisine (Chinese, Italian, etc.) that the restaurant specializes in, whether or not the restaurant serves breakfast, whether or not the restaurant is in the city of San Luis Obispo.

The most common error was to write about a summary across all restaurants, such as: how many restaurants are in SLO county. This is not a variable that can be recorded from restaurant to restaurant.

2. (12 pts) In the August 12, 2007 issue of *Parade* magazine (which comes with the Sunday newspaper for millions of Americans), readers were asked to go online and vote on the question: Should the drinking age be lowered? The results were published in the October 7 issue: more than 14,000 readers voted, and 48% said "yes."

a) (4 pts) Is this number (48%) a parameter or a statistic? Explain (in one sentence or less) how you know.

This is a statistic, because it refers to a sample of adult Americans.

b) (8 pts) Do you trust this sample to be representative of the population of all American adults? Explain.

No. One reason is that only readers of *Parade* magazine saw the question. Another is that only people who cared enough about the issue, and had easy access to the internet, actually voted.

3. (6 pts) It can be shown that when you roll a pair of fair dice and calculate the sum of the dots showing, the expected value is 7. (Do not bother to verify this.) Fill in the blanks, with a few words each:

This means that if you roll a pair of fair dice a very large number of times, the long-run average number of dots showing will be very close to 7.

4. (28 pts) An article about handwriting appeared in the October 11, 2006 issue of the *Washington Post*. The article mentioned that among students who took the essay portion of the

SAT exam in 2005-6, those who wrote in cursive style scored significantly higher on the essay, on average, than students who used printed block letters.

a) (8 pts) Identify the explanatory and response variables in this study. Classify each as categorical or quantitative.

Explanatory: type of handwriting (cursive or block letters)
Type: categorical (also binary)
Response: score on SAT essay
Type: quantitative

b) (6 pts) Is this an observational study or an experiment? Explain briefly.

This is an observational study, because students were not assigned to use a particular writing style.

c) (6 pts) Would you conclude from this study that using cursive style causes students to score better on the essay? Explain why or why not.

No. Because this is an observational study, there could be confounding variables that help to explain the relationship between handwriting style and essay score. For example, maybe better schools encourage students to use cursive writing and also provide better instruction on writing essays. Or maybe cursive writers can write more quickly and therefore express more ideas in the time allotted. Or maybe better writers simply choose to use cursive style more often than poor writers do.

The same article also mentioned a different study in which the identical essay was shown to many graders, but some graders were randomly chosen to see a cursive version of the essay and the other graders were shown a version with printed block letters. The average score assigned to the essay with the cursive style was significantly higher than the average score assigned to the essay with the printed block letters.

d) (10 pts) What conclusion would you draw from this second study? Be clear about how and why this conclusion would differ from that of the first study. Explain briefly.

This study is now a randomized experiment. The only difference between the two groups is the type of handwriting used. Because it turned out that the cursive group scored significantly higher than the other group, we can conclude that cursive writing *causes* higher scores.

Note: Even from the original study, we can conclude that those who use cursive writing tend to score higher than those who use block letters. The difference in (d) is that we can conclude that the higher scores are *because of* the cursive writing.

5. (32 pts) Is the husband older than the wife in most (more than half) marriages? To investigate this, a student went to the county courthouse and examined a sample of 94 marriage licenses. He found that the husband was older in 67 of these marriages, and the wife was older in 27 of them.

a) (3 pts) What are the observational units in this study?

Marriages (or marriage licenses, or married couples)

b) (3 pts) Calculate the proportion of these marriages in which the husband is older.

$67/94 \approx .713$

c) (4 pts) Is this number (your answer to b) a parameter or a statistic? Explain briefly.

This is a statistic, because it is based on a sample of married couples.

d) (4 pts) Describe the null model that the student could investigate through simulation here.

The null model says that neither the husband nor wife is more likely to be older, so the husband is older in half of all marriages.

Suppose that you conduct a simulation analysis to investigate whether these data provide convincing evidence that the husband is older than the wife in more than half of all marriages.

e) (4 pts) Fill in the blanks, either with a number or a word: You could conduct this simulation analysis by flipping a fair coin 94 times and then repeating that process a large number (say 1000) times. You would then count how often the simulated result produced 67 or more heads.

f) (4 pts) Fill in the blanks, with several words each: The p-value for this study turns out to be .00002. This says that it's extremely unlikely for 67 or more couples to have a husband older than the wife, assuming that there's no tendency for either person to be older more often (equivalently, that the husband is older in half of all marriages).

g) (6 pts) Based on the p-value, what conclusion would you draw? (Be sure to relate your conclusion to the context and the research question that the student started with.)

The very small p-value indicates that the sample data (husband older in 67 of 94 marriages) provides strong evidence that the husband really does tend to be older in more than half of all marriages.

h) (4 pts) Suppose that the sample size had been much smaller, and had still resulted in the same proportion of marriages in which the husband was older than the wife. Would the p-value be smaller, the same, or larger? Explain briefly.

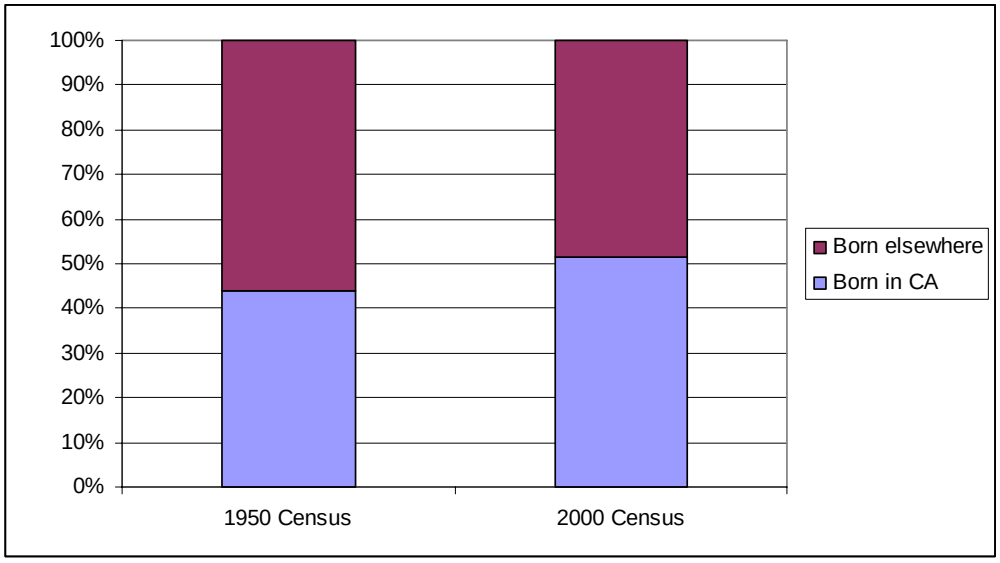
The p-value would have been larger. It would have less surprising to obtain 71% having an older husband with a smaller sample. Smaller samples produce more variability away from the null value of 50%.

6. (14 pts) Earlier this year I took a random sample of 500 California residents who filled out a U.S. Census form in 1950, and I also took a random sample of 500 California residents who filled out a U.S. Census form in 2000. For each person I recorded whether or not they had been born in California. I figured that society is more mobile now that it was 50 years ago, so I expected to see a larger proportion of California-born residents in 1950 than in 2000.

a) (3 pts) Does this study involve random sampling or random assignment? (No explanation is necessary.)

Random sampling was used to select 500 CA residents in each year.

The following segmented bar graph displays the results of this study:



b) (3 pts) About what percentage of California residents in 1950 were born elsewhere?

The graph reveals that about 43% were born in CA, so about 57% were born elsewhere.

c) (8 pts) Summarize what the segmented bar graph reveals. Be sure to comment on whether my expectation (described in the last paragraph before a) was supported by the data.

The 2000 Census saw a slightly higher proportion of California-born residents than 1950, about 51% vs. 43%. My expectation is not supported by the data, because I expected to see the opposite.

Note that this discussion is about the *proportion* or *percentage* born in CA. Many students incorrectly wrote about the *number* born in CA.