Daniel Carpenter
ID: 113009743

Warm Up: Appendix C11:

a. Yes, the average change in GPA in the sample was greater than zero.

Exercise 1.1

Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved student performance of fourth graders.

1. **If you could conduct any experiment you want, what would you do? Be specific.**
    a. First, I would create two groups, the first being the control group, which are those in larger classes; the second being the test group, which are those in smaller classes. Ideally the teacher and course contents would be identical.
    b. Next, I would measure the performance before the study, and then measure the performance at a given time in the future.
    c. Finally, I would compare the results.

2. **More realistically, suppose you can collect observational data on several thousand fourth graders in a given state. You can obtain the size of their fourth-grade class and a standardized test score taken at the end of fourth grade. Why might you expect a negative correlation between class size and test score?**
    a. Schools with greater admittance may receive greater funding per pupil. The resources that are fed into the large classes may cause the students to perform better on tests compared to the underfunded classrooms.
3. **Would a negative correlation necessarily show that smaller class sizes cause better performance? Explain**
    a. Correlation does not indicate causation.
    b. There may be an unknown variable that is causing this relationship to occur, such as funding per pupil or funding per classroom.

C2

Use the data in BWGHT to answer this question.

1. **How many women are in the sample, and how many report smoking during pregnancy?**
    a. Women in sample: n = 665
    b. Women smoking during pregnancy: n = 112

2. **What is the average number of cigarettes smoked per day? Is the average a good measure of the "typical" woman in this case? Explain.**
    a. Avg. number of cigarettes smoked per day: 2.09 cigarettes
    b. The average is most likely not a great measure. It includes all women in the calculation, even those who do not smoke. When the sample size increases, the average number will decrease because there are many who did not smoke while pregnant.

3. **Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part (ii), and why?**
    a. Avg. smoked per day while pregnant: 12.41 cigarettes
    b. In the above measure, it filters out all women who do not smoke; therefore, the measure better shows how severe the problem is for those who are smoking.

4. **Find the average of fatheduc in the sample. Why are only 1,192 observations used to compute this average?**
    a. Avg. father's education: 13.19 years.
    b. Within the dataset, there are many "NA" values within the "fatheduc" column. Initially, I hypothesized that it excluded "motheduc;" however, my hypothesis was not true. It is unclear why there are null values within this column.

5. **Report the average family income and its standard deviation in dollars.**
    a. Avg. family income: $29,030
    b. Standard Deviation: $18,740

C3

The data in MEAP01 are for the state of Michigan in the year 2001. Use these data to answer the following questions.

1. **Find the largest and smallest values of math4. Does the range make sense? Explain.**
    a. Max: 100%, min: 0%
    b. Theoretically, these values do make sense. This range measures the "percent" of "students" who passed a 4th grade math exam (per district per building). If the values equal "0%," then the school had no students with passing scores.

2. **How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?**
    a. Perfect pass rate number: 38 schools
    b. Perfect pass rate (% of total): 2.08% of total schools

3. **How many schools have math pass rates of exactly 50%?**
    a. Pass rate of 50%: 17 schools

4. **Compare the average pass rates for the math and reading scores. Which test is harder to pass?**
    a. Avg. Math Score: 71.91%
    b. Avg. Reading Score: 60.06%
    c. The average conveys that more students pass the math exam than the reading exam.

5. **Find the correlation between math4 and read4. What do you conclude?**
    a. Correlation: 87.82%
    b. We can assume that schools with higher math scores are more likely to have higher reading scores. However, there may be an unobservable variable that is causing this relationship to occur, such as funding per student.

6. **The variable exppp is expenditure per pupil. Find the average of exppp along with its standard deviation. Would you say there is wide variation in per pupil spending?**
    a. Avg. expenditure per pupil: $5,194.87
    b. SD of expenditure per pupil: $1,091.89
    c. It appears that the distribution of expenditures per pupil is wide. Considering that the average is around five thousand dollars, a standard deviation of one thousand is high.

7. **Suppose School A spends $6,000 per student and School B spends $5,500 per student. By what percentage does School A's spending exceed School B's? Compare this to , which is the approximation percentage difference based on the difference in the natural logs. (See Section A-4 in Appendix A.)**
    a. $5,500 vs $6,000: 9.09% increase in expenditures per pupil

b. [log(6,000)-log(5,500)]: 8.70% increase in expenditures per pupil
c. There is a difference of 0.38% between the two approaches.

Cool Down: 2.1

Let kids denote the number of children ever born to a woman, and let educ denote years of education for the woman. A simple model relating fertility to years of education is

$$kids = B_0 + B_1 educ + u$$

where u is the unobserved error.

1. **What kinds of factors are contained in u? Are these likely to be correlated with level of education?**
    a. Factors included in u are unobservable forces that influence the number children born by women. For example, an unobservable force may be the level of income that the household has. If a household has higher incomes, then they may be able to afford raising more children.
    b. These variables likely are correlated with the level of education. Unobservable forces are hard to identify due to them potentially having a positive correlation with the independent variable.

2. **Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.**
    a. It will not underline the ceteris paribus effect of education of fertility. The only way that we may be able to see this effect is if u is mean independent of education. Through establishing this relationship, we can make assumptions that and increase in education leads to increased fertility; however, until we identify its independence from other variables, it is impossible to infer causality.