11.220
Final exam solutions
Spring 2000

1. [8] Turn to the section entitled “Data analysis” on page 331. Identify the parent and target population for this study (some information on the latter can be found in the conclusions section that begins on page 336). In a few sentences, identify any sampling biases that you think might exist in the sample data as a result of the authors’ sampling strategy. (Don’t use a “kitchen sink” approach here—think a bit and include just two or three potential biases that you think are important and are the most likely to result from the authors’ sampling approach.)

On page 331 (in the section titled “Data analysis”), we learn that the authors’ data come from the Neighborhood Reinvestment Corporation through its affiliated NeighborWorks organizations, as well as from the Census Bureau and the Bureau of Labor Statistics. The parent population (the group from which the final sample was drawn) includes the 204 NeighborWorks Organizations that submitted annual surveys, staff reports, and quarterly economic reports during the period 1990-1995. The target population is more difficult to identify—sometimes the authors make statements about CDCs focused on investment-related activities, and at other times they appear to be generalizing to all CDCs in the US.

The sampling strategy employed was simply to identify all CDCs from the parent population who satisfied 2 criteria: (1) they were “in existence and affiliated with the NRC for at least three of the study years” (1990-1995); and (2) they filed an annual survey, a staff report, and quarterly economic reports for at least three of the study years.

Sampling biases fall into two principal categories: (1) Differences in the target and parent populations and (2) Systematic differences between CDCs selected into the final sample and those not selected. As for the first category of bias, we have no information about the differences between CDCs that are members of NeighborWorks and those that are not (or even the % of all CDCs that are members). If the authors do intend to generalize their findings to CDCs throughout the country, it would be important to compare the characteristics of these two groups and think carefully about any biases that might be introduced by excluding non-NeighborWorks organizations from the sample. Moreover, to the extent that “young” CDCs differ systematically from older ones, excluding CDCs who have not been in existence for at least 3 years could introduce bias into our research. The authors also note themselves that the study does not include CDCs that once existed but do not currently; a similar caveat should be raised about them. Finally, it seems reasonable to suspect that CDCs who regularly file documentation with their parent organization are systematically different from those that do not. For example, we might suspect that regular filers have more staff, in particular that they have a staff position dedicated to gathering, organizing, and submitting information such as this. They might thus also be larger organizations. We would like to see some analysis that compares the 57 (204-147) excluded CDCs to those included in the sample.
2. [10] Using the information from this same section on page 331, determine (a) the probability that a randomly selected CDC from the 204 mentioned in the first paragraph was selected for the final sample in the study; (b) the probability that a randomly selected CDC was selected for the final sample given that it had been in existence and affiliated with the NRC for at least three of the study years; and (c) the probability that a randomly selected CDC has both failed to file reports for 3 of the study years and has been open and affiliated with the NRC for at least 3 years. Show your work. Draw a probability tree or table if this helps you “see” the problem.

The probability tree would look like this:

![Probability Tree](image)

The first question is a simple probability problem. The probability that a randomly selected CDC from the 204 mentioned in the first paragraph was selected for the final sample in the study is simply given by the total number selected for the sample divided by the total number that could have been selected:

\[ p = \frac{147}{204} = 0.721 \]

The second question is a conditional probability problem. The probability that a randomly selected CDC was selected for the final sample given that it had been in existence and affiliated with the NRC for at least three of the study years is given by the number of CDCs selected for the final sample divided by the total number of CDCs in existence and affiliated with the NRC for at least three of the study years:

\[ p = \frac{147}{161} = 0.913 \]

Alternatively, \( P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \)

Let \( P(A) = P(\text{Selected for final sample}) \)

Let \( P(B) = P(\text{In existence 3 yrs. + NRC member}) \)

\( P(A \text{ and } B) = \frac{147}{204} = 0.721 \)

\( P(B) = \frac{161}{204} = 0.789 \)
Thus, \( P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.721}{0.789} = 0.913 \)

The third question is a joint probability problem. The probability that a randomly selected CDC (from the 204) has both failed to file reports for 3 of the study years and has been open and affiliated with the NRC for at least 3 years is given by the number of CDCs that were open for 3 years and failed to file reports divided by the total number of CDCs that could have been included in the sample:

\[ p = \frac{14}{204} = 0.069 \]

Alternatively: \( P(A \mid B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B) \)

Let \( P(A) = P(\text{Open and affiliated with NRC}) = \frac{161}{204} = 0.789 \)
Let \( P(B) = P(\text{Failed to file reports}) = \frac{14+63}{204} = 0.279 \)
\( P(B \mid A) = \frac{14/161} = 0.087 \)

Thus, \( P(A \mid B) = P(A) \times P(B \mid A) = 0.789 \times 0.087 = 0.069 \)

3. [8] Turn to page 333, and consider Table 2, which presents values related to each explanatory (independent) variable in the authors’ regression model. Devise a test of the null hypothesis that the proportion of minority population in the service areas of CDCs is greater than or equal to 0.50. (You know more than one way to do this, and you can use the method you prefer.) Show your work (including the formulas for your null and alternative hypotheses) and interpret your results in one or two clear statements that someone with no training in statistics could understand. Use a significance level of .01 for your analysis.

Using a hypothesis testing approach:

\[ H_0: \ p \geq 0.5 \  \ (p=0.5 \text{ is also fine}) \]
\[ H_\alpha: \ p < 0.5 \]

Then: \[ z = \frac{\hat{p} - p_o}{\sqrt{p_o(1-p_o)/n}} = \frac{0.408 - 0.50}{\sqrt{(0.50 \times 0.50)/147}} = -2.23 \]

We find in the M&B standard normal table that the p-value associated with a test statistic of 2.23 is 0.4871. We subtract this from 0.5 and find that our p-value is 0.5-0.4871= 0.0129 (remember this is a one-tailed test). Technically, this is greater than our determined p-value threshold of .01; we should thus not reject the null hypothesis of \( p \geq 0.5 \) in this case. In practice, we would most likely report \( p=.01 \) and reject our null hypothesis that \( p \geq 0.50 \). That is, we have evidence suggesting that the proportion of non-white population in the service areas of CDCs is less than 0.50. More specifically, assuming the null hypothesis is true, the probability that we would obtain a sample proportion value of 0.408 by chance alone is only .01.

Alternatively, using a critical value approach:
We determine the critical value from the standard normal table. For a one-tailed test, this is -2.33 (remember the negative comes from our alternative hypothesis). We compute the test statistic in the same way and compare its value to our critical value. We see that -2.23 does not pass the critical value into the rejection region; thus we would not reject our null hypothesis that $p \geq 0.5$ at the .01 significance level.

This is a good example of the limitations of the critical value approach. With the p-value approach above, we could see that we were very close to the threshold we’d set for statistical significance—so close, in fact, that we would likely be willing to “round down” and thus reject our null hypothesis. In the critical value approach, we can also see that -2.23 is close to the critical value -2.33, but we’d have to take the extra step of finding the associated p-value to make a final judgment.

Note that we have not learned to use a confidence interval approach for a one-tailed hypothesis test as this question requires. If you wanted to go this route, you would need to construct a confidence interval that was **twice as wide** as the one you would construct for a two-tailed test (i.e., you would multiply $\alpha$ by 2).

4. [8] The authors discuss the age of CDCs in a number of different parts of their analysis. Suppose that you wanted to do some follow up research on this topic; in particular, you wanted to work with the youngest 2.5% and the oldest 2.5% of CDCs. What would be the cutoff age for each group? Which of these two values should be considered the 97.5th percentile? (Hint: Employ a simple rule of thumb here and avoid longer analysis and computations.)

We know that, for normally distributed variables, (roughly) 95% of observations fall within two standard deviations on either side of the mean value. For the CDC age variable, the mean and median values are quite close, so we proceed under the assumption that this variable is distributed normally. This means that by adding and subtracting two times the standard deviation value to and from the mean value, we find the cutoff points beyond which a total of 5% (2.5% in each tail) of the area under the curve lies. Thus:

$$14.6 + (2 \times 5.0) = 24.6$$

$$14.6 - (2 \times 5.0) = 4.6$$

If you forgot the rule, you could also employ the standard z-score approach (which actually yields more precise values). This strategy would be to find the z value associated with the p-value $\text{.50-.025}=.475$ in the normal table. This value is 1.96. Then we proceed as follows:

$$1.96 = \frac{x_U - 14.6}{5.0}$$

$$-1.96 = \frac{x_L - 14.6}{5.0}$$

$$1.96 \times 5 = x_U - 14.6$$

$$-1.96 \times 5 = x_L - 14.6$$

$$9.8 + 14.6 = x_U$$

$$-9.8 + 14.6 = x_L$$
24.4 = x_U 
4.8 = x_L

Finally, 24.4 is the 97.5th percentile, because 97.5% of all data values fall below it.

5. [8] Turn to page 328 and consider the following sentence in paragraph 3:

“Total staff compensation is highly correlated with staff size (Pearson’s r=0.98) as is total operating budget, but there is no ambiguity about what is included in the reported figures for total staff compensation.”

If you conducted a simple linear regression analysis in which total staff compensation were your dependent variable and staff size were your independent variable, what would the \( r^2 \) value of your model be? In two clear sentences, explain what the values \( r \) and \( r^2 \) mean with respect to these two variables. (Be clear whether the statements you make are referring to the sample or the population.) In another 1-2 sentences, explain which of these measures you think is more useful from a planning/policy perspective and why.

The \( r^2 \) value would be \((0.98 \times 0.98)\), or 0.96. Pearson’s correlation coefficient \( r \) is a measure of the strength of the linear relationship between the two variables total staff compensation and staff size. The \( r^2 \) value, by contrast, would indicate that variation in staff size explains 96% of variation in total staff compensation. The latter is generally more useful for policy making and analysis, as we can quantify the influence that one policy variable has on another.

6. [9] On pages 327-328 the authors discuss the indicator used as the dependent variable in the authors’ regression model in paragraph 4. Identify two points they make that speak to the validity of the indicator (you will need to define what we mean by validity in measurement in order to do this). The authors also admit that the indicator may be biased against CDCs with a particular characteristic. In two or three clear sentences, explain what this characteristic is and how the bias arises.

Validity in measurement refers to how close we come to measuring the construct we’re interested in with our measurement approach. Remember our example from class: shoe size is a reliable measure (doesn’t vary from one measurement to the next given that my foot size has remained the same) of athletic ability, but not a very valid one (shoe size doesn’t capture the construct of athletic ability very well).

In order to assess validity, you have to know what you’re trying to measure. The authors here say that “an appropriate measure of efficiency should capture the results achieved by an organization, undistorted by the characteristics of the environment in which it operates.” They then make several points regarding why they think their measure does in fact capture the essence of efficiency. Any two of these is fine.

Bias in measurement refers to a process that gives us values that are systematically different from the true values. The authors acknowledge that, if the construct of efficiency can be described as the ratio of outputs to inputs within a CDC, their indicator captures only those inputs and outputs that can be measured in investment dollars. Thus, the value of the variable will be systematically lower for CDCs who
engage in above-average levels of activities that do not result in direct investment (e.g., social services and community organizing). This happens because the indicator is a fraction, in which the denominator is a measure of all staff compensation, not just compensation for work involved in investment-related activities. The numerator is a measure of total direct investment resulting from CDC activities. Thus, for a CDC that engages in relatively more non-investment activities, the numerator will be smaller and the denominator larger (in a relative sense), thus resulting in a low efficiency value, as compared to CDCs that focus on investment-related activities.

9. [8] Consider Table 3 on page 333. The parameter estimate $b_{13}$ is associated with the variable “Activities.” In the fourth column of the table, we see a p-value associated with a test of the hypothesis $\beta_{13} = 0$ of 0.019.

As was pointed out in a recent QR e-mail message, no standard error is provided in this table (or any table in the article). However, assuming a two-tailed test of the hypothesis and $n-13=134$ degrees of freedom, you have everything you need find the correct standard error value associated with $b_{13}$. Find it. Show your work. (Hints: Look at your degrees of freedom and use your table carefully.)

We know that the formula for finding the p-value associated with this hypothesis is given by

$$ t = \frac{b_5 - \beta_0}{s.e.} = \frac{\beta_0}{s.e.} $$

We also know that the p-value found by the authors for this hypothesis test is 0.019, and that our degrees of freedom is large (thus allowing us to use a normal table for computations). We know how to use the standard normal table to find a z value associated with a particular probability value. We have a p-value of 0.019, but this refers to a two-tailed test. To use the M&B normal table, which deals with one tail at a time, we split this value in half (0.019 / 2 = 0.0095). Moreover, this value refers to the area in the tail of the curve; to use the M&B table we need the area between the mean and the z-value associated with 0.019. We obtain this by $0.5 - 0.0095 = 0.4905$. Looking in the body of the table, we find this p-value falls on the row with the z value 2.3; moreover, it is exactly halfway between the values for 2.34 and 2.35. Thus, we’ve found our z-value: 2.345.

Now, we have everything we need to solve for the standard error. Substituting:

$$ z = \frac{b_5}{s.e.} $$

$$ 2.345 = \frac{0.263}{s.e.} $$

$$ s.e. = \frac{0.263}{2.345} = 0.112 $$
10. Rather than using a ratio variable for staff size in their analysis, the authors break the sample into quartiles and create a categorical variable for staff size using full-time equivalents (FTEs). Suppose you were interested in knowing whether there were a relationship between the size of a CDC and the gender of its director. Suppose that, in this study, 49% of directors of the CDCs in the first quartile (≤2.5 FTEs) were women; 41% of directors in the second quartile were women; 32% in the third quartile; and 24% in the fourth quartile.

Devise a test of the hypothesis that there is no association between staff size and the gender of a CDC’s director using these data. Show your work and state clearly what you conclude about your hypothesis. Use a .05 significance level.

First, we recognize that the use of quartiles automatically provides us with a percentage distribution for the sample. We generate the following contingency table (note that the categories the authors use are not mutually exclusive, so we’ve made a small adjustment for that):

<table>
<thead>
<tr>
<th></th>
<th>1st quartile (≤2.5 FTEs)</th>
<th>2nd quartile (2.5-3.39 FTEs)</th>
<th>3rd quartile (3.4-5.67 FTEs)</th>
<th>4th quartile (&gt;5.67 FTEs)</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female director</td>
<td>49%</td>
<td>41%</td>
<td>32%</td>
<td>24%</td>
<td>100%</td>
</tr>
<tr>
<td>Male director</td>
<td>51%</td>
<td>59%</td>
<td>68%</td>
<td>76%</td>
<td>100%</td>
</tr>
<tr>
<td>Column totals</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Remember that we cannot conduct the appropriate test—a $X^2$ test—with percentage data. We next have to convert these to count data. We know that each column represents 25% of the sample total, so .25 x 147 = 36.75. Then we find the observed count for males and females in each column by multiplying the relevant percentage by 36.75:

<table>
<thead>
<tr>
<th></th>
<th>1st quartile (≤2.5 FTEs)</th>
<th>2nd quartile (2.5-3.39 FTEs)</th>
<th>3rd quartile (3.4-5.67 FTEs)</th>
<th>4th quartile (&gt;5.67 FTEs)</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female director</td>
<td>18.01</td>
<td>15.07</td>
<td>11.76</td>
<td>8.82</td>
<td>53.66</td>
</tr>
<tr>
<td>Male director</td>
<td>18.74</td>
<td>21.68</td>
<td>24.99</td>
<td>27.93</td>
<td>93.35</td>
</tr>
<tr>
<td>Column totals</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>148</td>
</tr>
</tbody>
</table>

These are whole organizations (we’re dealing with observed values), so let’s round off. We can see that this rounding has given us a sample size of 148, which is one more than the researchers actually have:

<table>
<thead>
<tr>
<th></th>
<th>1st quartile (≤2.5 FTEs)</th>
<th>2nd quartile (2.5-3.39 FTEs)</th>
<th>3rd quartile (3.4-5.67 FTEs)</th>
<th>4th quartile (&gt;5.67 FTEs)</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female director</td>
<td>18</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>54</td>
</tr>
<tr>
<td>Male director</td>
<td>19</td>
<td>22</td>
<td>25</td>
<td>28</td>
<td>94</td>
</tr>
<tr>
<td>Column totals</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>148</td>
</tr>
</tbody>
</table>

Next we need to compute expected values under the assumption that the null hypothesis of no association between gender and staff size is true. For each cell, we multiply the
row total by the column total, then divide by the sample size. This gives us the following (think about why all the values in a given row are identical!):

<table>
<thead>
<tr>
<th></th>
<th>1st quartile (≤2.5 FTEs)</th>
<th>2nd quartile (2.5-3.39 FTEs)</th>
<th>3rd quartile (3.4-5.67 FTEs)</th>
<th>4th quartile (&gt;5.67 FTEs)</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female director</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>54</td>
</tr>
<tr>
<td>Male director</td>
<td>23.5</td>
<td>23.5</td>
<td>23.5</td>
<td>23.5</td>
<td>94</td>
</tr>
<tr>
<td>Column totals</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>148</td>
</tr>
</tbody>
</table>

Then, we generate the test statistic. This is given by:

\[ X^2 = \sum \frac{(O-E)^2}{E} \]

\[ X^2 = \frac{(18-13.5)^2}{13.6} + \frac{(15-13.5)^2}{13.5} + \cdots + \frac{(28-23.5)^2}{23.5} \]

\[ X^2 = 5.248 \]

The degrees of freedom for our example is given by \((r-1) \times (c-1) = 1 \times 3 = 3\). The critical value for the test at a .05 significance level is 7.81. Our test statistic value falls short of this, indicating that we do not have sufficient evidence to reject the null hypothesis of no association between gender and staff size. In fact, the p-value associated with our test statistic value is between .15 and .20 (because our test statistic falls between the critical values 4.64 and 5.32).

11. [6] With respect to the analysis you conducted in Question #10, what would making a Type I and a Type II error mean in the context of this example? Be specific. How could you reduce the probability of making a Type I error? What about a Type II error?

A Type I error refers to rejecting the null hypothesis when it is, in fact, true—in this case, rejecting the hypothesis that there is no association between gender and staff size when in fact there is no such association. To reduce the probability of this type of error, we simply adjust our chosen significance level downward, from .05 to, say, .01.

A type II error refers to accepting the null hypothesis when it is, in fact, not true—in this case, accepting the hypothesis that there is no association between gender and staff size when in fact there is an association. To reduce the probability of this type of error, we must increase our sample size. Unfortunately, we cannot have it all—as we reduce the probability of making one type of error, we simultaneously increase the probability of making the other type.

12. Turn to page 326. In the third paragraph in the section headed “Measures of CDC performance,” the authors cite a study by Avis Vidal (1992) in which the mean output of CDCs in his sample (21) was “similar to the average output of private, for-profit housing developers.”
Suppose Vidal arrived at this conclusion using a hypothesis testing approach. What would his null and alternative hypothesis have been? Express your answer in words and in formulas.

The most likely null hypothesis that the mean annual output by private developers and CDCs is equal, i.e.,

\[ H_0: \mu_p = \mu_C \quad H_A: \mu_p \neq \mu_C \]

We don’t know definitively whether Vidal suspected a particular direction of effect, so we have an alternative hypothesis that the mean annual output by private developers and CDCs is not equal (i.e., we care about significant differences in either direction).

Next, suppose that the information Vidal used for this hypothesis test were as follows:

<table>
<thead>
<tr>
<th>Private developers</th>
<th>CDCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean annual output (units) (( \bar{x} ))</td>
<td>22.5</td>
</tr>
<tr>
<td>Variance (( s^2 ))</td>
<td>11.4</td>
</tr>
<tr>
<td>Sample size (( n ))</td>
<td>32</td>
</tr>
</tbody>
</table>

Show how a test of the hypothesis you supplied above would be conducted. Use a .05 significance level. Show your work, including your equations, computations, and the p-value associated with your result. You may use any hypothesis testing approach that you prefer and that is appropriate for answering the question.

Because the sample variance values appear quite different and we have no reason to justify the equal variance approach, let’s use a hypothesis test for two means assuming unequal variances in the population. The procedure is:

Find the standard error for each group:

\[ s_{e_p} = \frac{s_p}{\sqrt{n_p}} = \frac{3.38}{5.66} = 0.60 \quad s_{e_C} = \frac{s_C}{\sqrt{n_C}} = \frac{3.70}{5.10} = 0.73 \]

Then find the overall standard error:

\[ s_{e_{overall}} = \sqrt{s_{e_1}^2 + s_{e_2}^2} = \sqrt{0.36 + 0.53} = 0.94 \]

Then compute a t statistic in the usual way:

\[ t = \frac{\bar{x}_p - \bar{x}_C}{s_{e_{overall}}} = \frac{22.5 - 21.0}{0.94} = 1.60 \]

We can use a normal table to find our p-value for the test statistic value of 1.60—it’s 0.4452. As usual, we are interested in the area in the tail of the distribution, which is given by 0.5-0.4452 = 0.0548. We have set this up as a two-tailed test, so we multiply this value by 2 to obtain the total probability value: .0548 x 2 = .1096 or .11. We
compare this to our significance level of .05 and we **cannot** reject our null hypothesis that mean annual output of CDCs is equivalent to that of private sector developers.

Alternatively, we could employ a critical value approach and identified –1.96 and 1.96 as our critical values. Our test statistic of 1.60 does not exceed 1.96, so again we fail to reject the null hypothesis of $\mu_P = \mu_C$.

13. [8] Suppose you were on the board of a CDC that had experienced problems with rapid turnover of its executive directors. You would like to devise a strategy to improve director longevity in your CDC; that is, you want to think of ways to lengthen the current director’s tenure. If you could only choose one of the following strategies, which would you use and why? (Use the information in Table 4 to answer this question.)

(a) Increase the director’s salary by $1000
(b) Send the director to a 10-hour training course
(c) Hire an additional staff member to support the director

*In order to compare the predicted impact each of these three strategies on director tenure, we simply multiply the value mentioned in each case by the relevant parameter estimate:*

(a) Increase the director’s salary by $1000

\[0.068 \times 1 = 0.068 \text{ years expected increase in tenure (less than 1 month)}\]

(b) Send the director to a 10-hour training course

\[-0.017 \times 10 = -0.170 \text{ years expected decrease in tenure (about 2 months)}\]

(c) Hire an additional staff member

\[0.078 \times 1 = 0.078 \text{ years expected increase in tenure (less than 1 month)}\]

*Thus, of these three strategies, hiring an additional staff member is predicted to have the largest positive impact on the director’s tenure.*