Case-based Social Statistics I  
CSSS 321  
Winter 2007  
Solutions to First Examination

Exam: Thursday, February 15 and Friday 16, 10:30am - 11:20am

Professor: Mark S. Handcock

Name: Ronald Aylmer Fisher

1. Please write your name in the above space.

2. **You need to do all 4 questions.** All questions are of equal value (but not necessarily of equal difficulty).

3. Do not turn the page until so instructed. (You will have 90 minutes to work after the examination has been discussed with you.)

4. You may use your crib sheet and your calculator. Otherwise this is a closed book examination.

5. If you do not have enough room for your work in the place provided, use the back of a nearby page. (However, be sure to mark clearly which problem the material on the back of any page refers to.) If you pull the pages apart, sign all pages.

6. Answers should unambiguously state, in words, the approach taken. You should show your work so that partial credit can be given. Poorly described solutions will be penalized. unsupported answers

7. Good luck!

<table>
<thead>
<tr>
<th>Question</th>
<th>Subject</th>
<th>Points Available</th>
<th>Points Earned</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>‘Walkable’ Communities May Make Elders Healthier</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Diffusion of Biodiesel</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Surge and Class</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Poverty and Fertility</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>100</strong></td>
<td></td>
</tr>
</tbody>
</table>

Remember: DO ALL FOUR QUESTIONS
Question 1) ’Walkable’ Communities May Make Elders Healthier (25 points)

Read the attached article that appeared on ScienceDaily a couple of days ago on “Walkable Communities May Make Elders Healthier” and answer the questions posed below. An extract from the study in “The American Journal of Public Health” referred to in the ScienceDaily article is also attached.

Note that your answers may be brief and mainly address the article’s reporting of the attached study, not the details of the study itself. You do not have to answer all questions, but should address the most pertinent issues: If there is not information on a question, leave it blank.

Here are the questions:

a) (12 points)

Briefly address the following basic questions about the article and the underlying study. You only need to address the questions you think are most important to the study.

1. What are the major claims made in the article?

The title of the article implies that “‘Walkable’ Communities May Make Elders Healthier”. This appears to suggest that there is a causal link between being in a walkable community and health. The major claim is made in the third paragraph: “the chief factors contributing to an area’s walkability were higher residential density and clusters of destinations such as grocery stores, restaurants and other services.” Another claim is that “we’d be better if we had more of these kinds of (walkable) neighborhoods”.

2. Are the claims supported in the article? If so, how?

The claims are supported based on the results of the attached study published in the American Journal of Public Health, and interviews with the primary researcher.

3. Is there data in the article used to support the claims?

Yes, there is some data in the article used to support the claims. It reports basic statistical facts from the AJPH article.

4. Is there data in the reference for the claims?

Yes. The article is closely bound to the paper describing the original research. It reports on the walking habits of a randomly selected group of 65 to 97 year olds and the characteristics of their neighborhoods.

The results show the “Higher walkability scores were associated with significantly more walking for exercise ...” They also conclude that “Findings suggest that neighborhood characteristics are associated with the frequency of walking for physical activity in older people.”

5. Can you think of alternative explanations for the data in the article? Is there data in the article to support these alternatives?

The AJPH article goes to some length to emphasize that they are measuring associations between neighborhood characteristics and the frequency of walking, and
that the neighborhood characteristics are not causing the people to walk more. The ScienceDaily article suggests a causal link in the title.

Although the study is randomized, this was done in 1994. So the actual participants in the study from 2003 were a selection of those originally in the study. Many moved out of King County, passed away or were otherwise incapacitated. So the actual participants were not randomly selected. In particular less healthy older people will be underrepresented in the survey.

The study focuses on walking, but people may choose complimentary exercise programs. For example, those that work out in a gym more may choose to walk less.

It is possible that less healthy people choose less walkable neighborhoods (as they plan to walk less anyway). For example, urban areas are more costly and generally have higher socio-economic status people, and this is associated with better health care and health outcomes. A natural way to test the causal link suggested by the ScienceDaily article is to randomly place people in more and less walkable neighborhoods and follow their habits. However this is not feasible.

6. Does the data presented support the claims?

The data presented supports the claims of association (or correlation), but not causation.

b) (13 points)

Briefly describe each of the following components of the study. You only need to address the questions you think are most important to comment on.

Component 1: The individuals or objects studied and how they were selected

936 older people in King County were selected in 2001 based on the remains of a randomly chosen group of 2500 in 1994. This is a good approach, although “drop outs” from the original group are a concern. The final group is likely to be less healthy than the start group.

Component 2: The exact nature of the measurements made and the questions asked

The degree of walking was based on self-reports. That is, we only know what the participants claimed to have walked rather than the actual amounts. For example, people who move to highly walkable neighborhoods in part do so to walk more. They may report walking more than they did to reflect that cognitive bias.

Component 3: The setting or context in which the measurement were made

This is a survey collected in person so it should be pretty good. The neighborhood measurement was based on GIS and should be good.

Component 4: The extraneous differences between groups being compared

Because the 936 people measured were only those still in the study after 7 years, there was a degree of selection from the original group of 2500. So there could be differences. The major concern is that the people (quiet rightly!) choose were to live rather than being randomized to neighborhoods. So there are quite likely large differences between the groups in each neighborhood.
Component 5: The magnitude of any claimed effects of differences

The odds ratio for the largest effect is 2.26 and could be close to 1, so the effects are not huge. However, they are not tiny either.

Component 6: The source of the research and the funding

The source of research was our very own University of Washington. The funding is from the national Institutes of Health. These are both reputable.

Component 7: The researchers who had contact with the participants

This might have been a factor, but it is typical for a survey of this type. The survey is typically undertaken by trained professionals.
Question 2) Diffusion of Biodiesel (25 points)

Biodiesel is a automobile fuel that is a replacement for regular diesel but is derived from biological sources (such as vegetable oils). Biodiesel is biodegradable and non-toxic, and typically produces about 60% less net carbon dioxide emissions than regular diesel. Like many innovations it takes time to be adopted by drivers. One study considered the spread of the use of biodiesel in Seattle in 2005. The study randomly surveyed 36 current biodiesel drivers to see how many days after the advertisement of the availability of biodiesel did the drivers first start using it in their cars. The measurement is the number of days until biodiesel was adopted by the driver. The diffusion process leads to great variation in time. You are asked to describe and summarize the times for the 36 times listed below:

60 62 70 71 74 77
77 82 84 86 87 88
88 90 90 93 94 94
96 96 98 99 100 104
104 106 106 108 108 111
113 116 117 120 133 174

a) (6 points)

Construct a stem-and-leaf plot of the times. Be sure to include a “legend” describing the units of the plot.

Solution:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>02</td>
</tr>
<tr>
<td>7</td>
<td>01477</td>
</tr>
<tr>
<td>8</td>
<td>246788</td>
</tr>
<tr>
<td>9</td>
<td>003446689</td>
</tr>
<tr>
<td>10</td>
<td>046688</td>
</tr>
<tr>
<td>11</td>
<td>1367</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>4</td>
</tr>
</tbody>
</table>

6 | 0 = 60 days

More or less stems could be used, although this would obscure the features.
b) (4 points)

Give two numerical summary measures of the typical time to adoption. Which do you think is the more appropriate in this case?

**Solution:** The median is 95. There are multiple modes (77, 88, 90, 94, 96, 104, 106, 108). The mean turns out to be 96.56 (although you did not need to calculate this necessarily). As there is an outlier, the median is better than the mean. The mode is inappropriate.

c) (5 points)

Compute the inter-quartile range of the times to adoption. Is it a better measure of variability than the standard deviation in this case (you do not need to compute the standard deviation)?

**Solution:** The lower quartile is 85.50 and the upper quartile is 106.50. The interquartile range is thus 21. The IQR is a better measure of variability than the standard deviation in this case because the latter is affected by the outlier.

d) (5 points)

The Deputy Mayor of Seattle, unable to grasp the significance of the stem-and-leaf plot, asks you for a familiar table of relative frequencies for the classes 60-69, 70-79, ... Convert the stem-and-leaf plot into a table of relative frequencies.

**Solution:** The percentage table for these data is:

<table>
<thead>
<tr>
<th>Time (in days)</th>
<th>Percentage Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>60-69</td>
<td>6</td>
</tr>
<tr>
<td>70-79</td>
<td>14</td>
</tr>
<tr>
<td>80-89</td>
<td>17</td>
</tr>
<tr>
<td>90-99</td>
<td>25</td>
</tr>
<tr>
<td>100-109</td>
<td>19</td>
</tr>
<tr>
<td>110-119</td>
<td>11</td>
</tr>
<tr>
<td>120-129</td>
<td>3</td>
</tr>
<tr>
<td>130-139</td>
<td>3</td>
</tr>
<tr>
<td>140-149</td>
<td>0</td>
</tr>
<tr>
<td>150-159</td>
<td>0</td>
</tr>
<tr>
<td>160-169</td>
<td>0</td>
</tr>
<tr>
<td>170-179</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

e) (5 points)

The Mayor, hearing of your fine work, requests an additional written summary. Describe, briefly, the features of the times to adoption of biodiesel.

**Solution:** The median time of adoption is 95 days. The distribution is slightly positively skewed, with an outlying largest time of 174 days. The lowest time is 60 days, with quartiles 85.5 days and 106.5 days. An additional notable feature is the outlying value of a driver who waited 174 days. The lower outlier cutoff is at
$54 = 85.5 - 21 \times 1.5$ days and the upper outlier cutoff is at $138 = 106.5 + 21 \times 1.5$ days.
Question 3) Surge and Class (25 points)

A polling company has conducted a random survey to solicit the public’s feelings towards President Bush’s plan for a troop “surge” in Iraq. A breakdown of the 400 responses is shown by self-reported economic class in the accompanying table.

<table>
<thead>
<tr>
<th></th>
<th>Working Class</th>
<th>Middle Class</th>
<th>Upper-Middle Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>For</td>
<td>134</td>
<td>64</td>
<td>22</td>
</tr>
<tr>
<td>Against</td>
<td>126</td>
<td>36</td>
<td>18</td>
</tr>
</tbody>
</table>

Suppose that a citizen is selected at random from this group; the relevant events are defined as follows:

\[ W = \text{the citizen selected is working class} \]
\[ M = \text{the citizen selected is middle class} \]
\[ U = \text{the citizen selected is upper-middle class} \]
\[ F = \text{the citizen selected favors the plan} \]

a) (3 points)

Define a sample space for this response of the citizen (in terms of the above definitions)?

**Solution**: The sample space must represent the fact that each citizen has two attributes: economic and attitude to the proposed plan. In terms of the above definitions

\[ S = \{ W \cap F, M \cap F, U \cap F, W \cap \overline{F}, W \cap \overline{F}, W \cap \overline{F} \} \]

b) (2 points)

List the simple events for this experiment belonging to the event F.

**Solution**:

\[ F = \{ W \cap F, M \cap F, U \cap F \} \]
c) (3 points)
Express the event $F \cap M$ in words and calculate its probability.

**Solution**: This is the event that the citizen selected favors the plan and is middle class.

$$P(F \cap M) = \frac{64}{400} = 0.16$$

d) (8 points)
Find $P(F)$, $P(M)$ and $P(F | M)$.

**Solution**: We use the theoretical concept of probability to count the number of possibilities divided by the total number of possibilities.

$$P(F) = \frac{220}{400} = 0.55, \quad P(M) = \frac{100}{400} = 0.25, \quad P(F \cap M) = \frac{64}{400} = 0.16$$

Thus

$$P(F | M) = \frac{P(F \cap M)}{P(M)} = \frac{0.16}{0.25} = 0.64$$

e) (4 points)
Express the event $W \cup \bar{M}$ in words and calculate its probability.

**Solution**: This is the event that the citizen selected is either working or not middle class.

$$P(W \cup \bar{M}) = P(W \cup U) = \frac{260 + 40}{400} = 0.75$$

f) (2 points)
Are the events “the citizen is middle class” and “the citizen favors the plan” mutually exclusive? Explain your answer.

**Solution**: The events are mutually exclusive if the event

$$\{\text{the citizen is middle class}\} \cap \{\text{the citizen favors the plan}\}$$

is empty. But 64 citizens fall into both categories, so the events are **not** mutually exclusive.

g) (3 points)
Are the events “the citizen is middle class” and “the citizen favors the plan” independent? Explain your answer.

**Solution**: The events are independent if:

$$P(\text{the citizen favors the plan given the citizen is middle class}) = P(\text{the citizen favors the plan})$$

$$P(F|M) = P(F)$$

That is, if

$$0.64 = 0.55.$$

As this is not true the events are **not** independent. This also means that the attitude to the plan is not independent of the person’s class. That is proportionately more middle class people favor the plan than favor the plan overall.
Question 4) Poverty and Fertility (25 points)

Pregnancy among 15-17 year olds is associated with many negative “life outcomes” for the parents and children. Sociologists and demographers have studied the relationship between such teenage pregnancy and poverty. In this question we consider the relationship between the two for the 50 states and the District of Columbia. The poverty rate is measured by the percent of a state’s population in 2000 living in households with incomes below the federally defined poverty level. The teenage pregnancy rate is measured by the number of births per 1000 females aged 15 to 17 years.

The scatter plot below relates the poverty rate to the teenage pregnancy rate.

Is there a relationship between the poverty rate and the teenage pregnancy rate?

a) (3 points)

Estimate the regression line by eye and draw it on the above scatter plot.

Solution: The line is sketched on the above graph. The two points used were the value at \( x = 5 \) of about 10 and the value at \( x = 20 \) of about 30.

b) (6 points)

Based on the line you have drawn on the plot, estimate the coefficients of the regression line?

Solution: Based on the two points, the slope is then about \( (30 - 10)/(20 - 5) = 1.33 \) births per percent poverty (per 1000 people). The intercept is the birth rate at 0% poverty rate. This is off the left-hand-side of the graph, but we can estimate it from the equation:

\[
y = a + b \times x
\]

or \( a = y - b \times x \). Using the first value from above: \( a = 10 - 1.33 \times 5 = 3.35 \), or a little above zero.
The least-squares regression slope is actually 1.37 and the intercept is 4.27. Your values should be close to these.

c) (4 points)

Estimate the correlation coefficient between the teenage pregnancy rate and the poverty rate. Does it appear that the relationship between the two variables is approximately linear?

Solution: This is hard to estimate, but it seems to have a magnitude of about 0.7 and a positive sign as the relationship is decreasing. So 0.7 is a realistic estimate. The true value is 0.73.

d) (2 points)

Does the correlation coefficient indicate a positive or negative association between the two variables? Does it indicate a strong, or weak linear relationship between the two variables?

Solution: It is a positive relationship and indicate a strong linear relationship between the teenage pregnancy rate and the poverty rate.

e) (3 points)

What is the interpretation of the intercept of the regression line?

Solution: This is the average teenage pregnancy rate in states where there is a poverty rate of zero. In this case it indicates that this average is a little over 4 (per 1000 people).

f) (3 points)

The U.S. overall poverty rate in 2000 was 11.3% and it has increased slightly to 12.7% today. How large a change in the teenage pregnancy rate is expected to be associated with this increase in poverty rate?

Solution: This is just 12.7-11.3=1.4 times the slope - that is, about $1.4 \times 1.73 = 2.42$. That is, we expect it to decrease by about 2.42 more births per 1000 persons.

g) (4 points)

Predict the teenage pregnancy rate for a state where the poverty rate is 0%. Do you think this prediction is likely to be valid? Predict the teenage pregnancy rate for a state where the poverty rate is 100% (i.e., everybody is in poverty). Do you think this prediction is likely to be valid?

Solution: Both 0% and 100% are outside the range of the data, one is “below” whereas the other is “above”.

The first prediction is just the intercept (4.26). As it is just below the range of the data it is probably valid. Note that if the slope were larger the intercept could be negative (which is not realistic).

The prediction for 100% poverty is way off the right-hand-side of the graph, so we must use the equation:

$$y = a + b \times x = 4.27 + 1.73 \times 100 = 177.27 \text{births per 1000 women}$$

That is, on average a birth for every 5.6 women (1000/177.27). This seems very high. We are extrapolating way off the end of the graph and so this prediction
must be taken with a pound of salt (that is much more than a grain of salt). Without additional information we can not be confident that it is valid.