STAT22000 Autumn 2017 Homework 2

All page, section, and exercise numbers below refer to the course text (OpenIntro Statistics, 3rd edition, by Diez, Barr, and Cetinkaya-Rundel.).

Reading: Section 1.7, 1.1, 1.3, 1.4, 1.5 (in this order)

Problems for Self-Study: (Do Not Turn In)

• Exercise 1.13, 1.15, 1.21, 1.23, 1.25, 1.27, 1.37 on p.59-64

• Answers can be found at the end of the book (p.405-407).

• Self-study problems are as important as Turn-In problems. We don’t require submission because we think you can learn from those problems by doing them yourself and checking the answers, without grading feedbacks. If having questions about those problems, you are welcome to ask the instructor or CAs.

Problems to Turn In: due 3 pm on Friday, Oct. 13, on Canvas.

1. Refer to the mosaic plot in Exercise 1.68 on p. 73 and the background information therein, for each of the following statements below, determine whether it is TRUE or FALSE, and explain briefly.

(a) More than half of the subjects in the sample identified themselves as Democrats.

(b) More than half of the Democrat in the sample thought it’s better to raise taxes on the rich.

(c) Based on the mosaic plot, views on raising taxes and political affiliation appear to be independent

2. Refer to the CDC data in Lab #3: http://www.stat.uchicago.edu/~yibi/s220/labs/lab03.html.

We want to investigate the relationship between whether the subjects had exercised in the past month (exerany) and their self-rated general health (genhlth). Notice genhlth is an ordinal categorical variable with five ordered levels: excellent, very good, good, fair, and poor. We need to tell R the order of the five levels, or R is going to order them alphabetically as: excellent, fair, good, poor and then very good.

cdc = read.table("cdc.dat", header=TRUE)
cdc = transform(cdc, genhlth = ordered(genhlth, levels=c("excellent", "very good", "good", "fair", "poor")))

(a) Obtain the two-way contingency table between exerany and genhlth.

(b) What proportion of the sample had exercised in the past month? What proportion of the sample reports being in excellent health?

(c) Among those who had exercised in the past month, what proportion of them reported being in excellent health.
(d) Make a segmented bar chart that represents the two-way table in part (a).
(e) Make a standardized segmented bar chart that compares the self-rated general health between those who had exercised in the past month and those who hadn’t.
(f) Make a mosaic plot that compares the self-rated general health between those who had exercised in the past month and those who hadn’t. Based on the plots, which group had better self-rated general health?
(g) Are the two variables \textit{exerany} and \textit{genhlth} independent?

3. According to a study done at Kaiser Permanente in Walnut Creek, California, users of oral contraceptives have a higher rate of cervical cancer than non-users, even after adjusting for age, education, and marital status. Investigators concluded that the pill causes cervical cancer.

Some facts about cervical cancer: Current research suggests that certain strains of HPV (human papilloma virus) can cause cervical cancer, and HPV can be transmitted sexually. Women that are more active sexually than others, and have more partners are more likely to be exposed to the viruses.

(a) Is this a controlled experiment or an observational study?
(b) What are the two groups compared in the study? What’s the response variable?
(c) The passage mentioned the result had been adjusted for age, education, and marital status. How these three factors may confound the result if they were not adjusted.
(d) The passage mentioned “users of oral contraceptives have a higher rate of cervical cancer than non-users, even after adjusting for age, education, and marital status.” Which of the following statements is the correct explanation of the sentence above?
   i. The conclusions of the study might not be valid because users and non-users of oral contraceptives might differ in age, education, and marital status, and women with different age, education, and marital status have different risk of getting cervical cancer.
   ii. The difference in the rates of getting cervical cancer between users and non-users can be explained by many confounding factors but the investigators had ruled out three confounding factors: age, education, and marital status, by comparing users of oral contraceptives only with non-users of similar age, education level and the same marital status.
(e) Women using the pill were likely to differ from non-users on another factor which affects the risk of cervical cancer. What factor is that?
(f) Were the conclusions of the study justified by the data? Answer yes or no, and explain briefly.

4. A study published in 2010 showed that city dwellers have a 21% higher risk of developing anxiety disorders and a 39% higher risk of developing mood disorders than those who live in the country. A follow-up study published in 2011 used brain scans of city dwellers and country dwellers as they took a difficult math test. To increase the stress of the participants, those conducting the study tried to humiliate the participants by telling them how poorly
they were doing on the test. The brain scans showed very different levels of activity in stress centers of the brain, with the urban dwellers having greater brain activity than rural dwellers in the areas that react to stress\(^1\).

(a) Is the 2010 study an experiment or an observational study?
(b) Can we conclude from the 2010 study that living in a city increases a person’s likelihood of developing an anxiety or mood disorder?
(c) Is the 2011 study an experiment or an observational study?
(d) In the 2011 study, what is the explanatory variable and what is the response variable?
(e) Can we conclude from the 2011 study that living in a city increases activity in stress centers of the brain when a person is under stress?

5. Smokers may have a more difficult time quitting smoking if they live with another smoker. A study comparing bupropion (an antidepressant and smoking cessation aid) with placebo tried to take this into account in their design. The researchers first split the subjects based on whether they lived with another smoker. The subjects who live with another smoker were randomly assigned to take bupropion or a placebo, and those who didn’t live with smokers were also randomly assigned to take bupropion or a placebo. The figure shows a flow chart of the design, when 250 of the 429 study subjects lived with nonsmokers and 179 lived with another smoker.

\[\begin{align*}
\text{All Subjects:} & \quad \text{429 Smokers} \\
& \quad \text{Live with Smoker (} n = 179 \text{)} \\
& \quad \text{Bupropion (} n = 90 \text{)} \\
& \quad \text{Placebo (} n = 89 \text{)} \\
& \quad \text{Not Live with Smoker (} n = 250 \text{)} \\
& \quad \text{Bupropion (} n = 125 \text{)} \\
& \quad \text{Placebo (} n = 125 \text{)}
\end{align*}\]

The percentages of subjects relapsed were then compared among the four groups.

(a) If the two groups of subjects that received bupropion had substantially lower relapse rates than the two placebo groups, can we claim that bupropion is effective as a smoking cessation aid?
(b) Has blocking been used in this study? If so, identify them.
(c) If the placebo group that the subjects lived with another smoker had a higher relapse rate than the placebo group that the subjects did not live with smokers, can we claim that living with smoker(s) makes it harder to quit smoking?

6. The table shows results of whether the death penalty was imposed in murder trials in Florida between 1976 and 1987. For instance, the death penalty was given in 53 out of 467 cases in which a white defendant had a white victim\(^2\).

\(^1\)“A New York state of mind,” *The Economist*, June 25, 2011, p.94.
<table>
<thead>
<tr>
<th>Defendant’s Race</th>
<th>Victim’s Race</th>
<th>Death Penalty Yes</th>
<th>Death Penalty No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>White</td>
<td>53</td>
<td>414</td>
<td>467</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>0</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Black</td>
<td>White</td>
<td>11</td>
<td>37</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>4</td>
<td>139</td>
<td>143</td>
</tr>
</tbody>
</table>

(a) First, consider only those cases in which the victim was white. Find the conditional proportions that got the death penalty when the defendant was white and when the defendant was black. Was a white or a black defendant more likely to receive death penalty when the victim was white?

(b) Repeat part a for cases in which the victim was black. Was a white or a black defendant more likely to receive death penalty when the victim was black?

(c) Now add these two tables together to get a summary contingency table that describes the association between the death penalty verdict and defendant’s race, ignoring the information about the victim’s race. What proportion of white defendants received death penalty? What proportion of black defendants receive death penalty? Which proportion is higher?

(d) How can you explain the association in part (c), whereby white defendants were more like to receive death penalty? How can this association be so different from the ones you found in part (a) and (b)? Your answer must address the following.

   i. When the defendant was white, was the victim more likely to be black or white?
   ii. When the defendant was black, was the victim more likely to be black or white?
   iii. Was the death penalty more likely to be given when the victim was white or black, or no apparent difference?

Use the three observations above to explain how the association in (c) is so different from in the one in part (a) and (b).