STATISTICS 210 – Fall 2018

Midterm Exam: Tuesday November 6, 2018

• The exam is closed book, closed notes. You may use a calculator.

• Please do not write your solutions on the exam paper.

• Write your solutions on the separate blank paper provided.
  Put your student identification number (and only that) on each sheet.

• Please turn in your exam paper at the end of the exam. A copy of the exam will be posted later.

• Tables are provided separately.
  We will collect the tables at the end of the exam so that they can be reused.

• Problem values are written to the left of each problem. The total number of points is equal to 120.

• Two important reminders:
  – show your work so that you can receive partial credit;
  – budget your time so that you don’t miss problems you know how to do.

• Good luck!
1. A cognitive scientist carried out a study to explore the effect of listening to music on memory. Sixty college students were enrolled in the study. They were randomly split into a control group and a music group. Each student studied a list of words for one hour. The music group listened to music (of their own choice) while studying the list; the control group studied in silence. After a one hour interval, the students were asked to recall as many words from the list as they could. The scientist recorded the number of words recalled by each student and also recorded the type of music the student chose. The scientist reported the following data summaries.

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>30</td>
<td>13.2</td>
<td>3.7</td>
</tr>
<tr>
<td>Music - Rap</td>
<td>10</td>
<td>9.8</td>
<td>3.3</td>
</tr>
<tr>
<td>Music - Rock</td>
<td>13</td>
<td>9.7</td>
<td>3.1</td>
</tr>
<tr>
<td>Music - Classical</td>
<td>7</td>
<td>11.7</td>
<td>3.4</td>
</tr>
</tbody>
</table>

17 (a) As an initial analysis the scientist would like to evaluate the hypothesis that all four groups have equal mean word recall. Perform an appropriate statistical test. State the null and alternative hypotheses, carry out the test, and interpret the result. (Note: The overall sample mean is 11.7.)

17 (b) A more focused analysis would compare the music groups to the control group.
   i. Provide a 95% confidence interval for the contrast $\mu_{\text{control}} - \frac{1}{3}\mu_{\text{rap}} - \frac{1}{3}\mu_{\text{rock}} - \frac{1}{3}\mu_{\text{class}}$.
   ii. Interpret the confidence interval and indicate what conclusion you draw about the effect of music on memory.

7 (c) It might also be of interest to compare memory results in the three music groups. Suggest a contrast that might be relevant and indicate whether your contrast is orthogonal to the one tested in the previous part. (Note that you are not asked to perform any data analysis here.)

10 (d) In reviewing the results of the study, a journal referee questions the design. The referee notes that it is part experiment and part observational study.
   i. Explain what the referee means.
   ii. How does the referee’s concern impact the conclusions that we can draw from this study?
2. Companies are increasingly allowing sales employees to work on a flextime schedule (four 10-hour work days per week) rather than traditional schedule (five 8-hour work days per week). One motivation is that this allows for more efficient travel by the sales force and hence less mileage driven. One Illinois company decided to carry out an experiment to judge whether this was in fact true. Fifty volunteers were randomly assigned into two groups - 25 sales people used the flextime schedule for three months while the other 25 sales people continued on the usual schedule. At the end of three months the average monthly mileage for each employee is recorded. A display (regular schedule on the left, flextime schedule on the right) and some numerical summaries of the 25 measurements in each of the two groups are provided below.

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular schedule</td>
<td>25</td>
<td>521</td>
<td>353</td>
</tr>
<tr>
<td>Flextime schedule</td>
<td>25</td>
<td>449</td>
<td>348</td>
</tr>
</tbody>
</table>

(a) Give a 90% confidence interval for the difference in the mean monthly mileage under the two work schedules.

(b) The histograms of the data in the two groups suggest that the monthly mileage measurements are not normally distributed. Would you be concerned about non-normality in this study? Explain.

(c) The consulting statistician does a post hoc (i.e., after the fact) power analysis. The statistician establishes that with this standard deviation ($\sigma = 350$) the power to detect a 75 mile difference at the .05 level with only 25 sales people in each group is 0.20.

i. Explain what the statistician is saying in words that company executives might understand.

ii. The company intends to repeat the study. Give them at least two ideas for improving the design of the study so as to have more statistical power.
3. An ecologist studying the effects of acid rain on wildlife diversity collected data from 153 streams in the Adirondack Mountains. The ecologist recorded the pH of the water in the stream (a measure of acidity) and the BCI (a measure of biological diversity of the stream). A scatterplot of the BCI measurements vs the pH measurements is provided below.

(a) Write down the usual linear regression model for analyzing BCI as a function of pH. Define and interpret all of the terms that appear in the model.

(b) The scatterplot suggests some difficulty with the usual regression assumptions. Discuss.

(c) It is decided to regress log(BCI) (i.e., log diversity) on pH. A small portion of the regression output is provided here.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Std.Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.43</td>
<td>0.18</td>
</tr>
<tr>
<td>pH</td>
<td>-0.17</td>
<td>0.03</td>
</tr>
</tbody>
</table>

\[ s_c = \text{std error of regression} = 0.15 \]
\[ R^2 = 0.25 \]

Use the provided output to test the hypothesis that the slope of the regression line is zero. State the null and alternative hypotheses, perform the test, report and interpret the p-value.

(d) Interpret the estimated slope for the biological community (i.e., try to explain it in a way that non-quantitative people will understand).

(e) We are interested in making a prediction about the diversity of a stream given its pH. Recall that the standard error for the prediction interval is \( s_c \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \). Also recall that our regression results are for the logarithm of diversity. Provide a 95% prediction interval for the diversity of a lake with pH equal to 7.3, the mean pH of the 153 lakes.

(f) Now suppose we are interested in inferring the pH of a stream based on its biological diversity. This is known as an inverse prediction problem.

i. Use the estimated regression to find an estimate the pH of a stream with BCI = 1200 (log diversity = 7.09).

ii. Using any approach that you can think of, provide an estimate of the precision of your estimated pH.