Statistics 571: Statistical Methods
Fall 2002 Final Exam
Ramón V. León

Name: ___________________________

This book is closed-book and closed-notes. However, you can use up to twenty pages of personal notes as an aid in answering your questions. Please be concise in your answers

1. Suppose you have a sample of 25 measurements and you wish to estimate the standard error of the sample median. How would you use the bootstrap procedure to do this?

2. Under what circumstances would one prefer to use the Spearman’s or Kendall’s rank correlation coefficient over the Pearson correlation coefficient?

3. For what type of statistical problem would one use the Kruskal-Wallis test? How about the Friedman test?

4. One can test that the median of distribution has a given value using either the sign test or the Wilcoxon signed rank test. What added assumption about the distribution of the observations is needed to use the Wilcoxon signed rank test?
5. An experiment was conducted to investigate the strength of a weld in a steel bar. Two factors were varied: the gauge bar setting (the distance the well die travels during the automatic weld cycle), which had three settings, and the welding time (total time of the automatic weld cycle), which was fixed at five levels. The weld strength data are shown in the JMP table below:

<table>
<thead>
<tr>
<th>Bar</th>
<th>Time</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>
Interpret the following JMP output: (Comment next to the JMP outputs)
6. Explain why the experiment described in Question 5 above is a Completely Randomized Design CRD rather than a Randomized Block Design (RBD). What assumptions is one making about the order in which the thirty experimental runs were made?

7. In the multiple comparison of means what is the difference in the type I error rate between the least significance difference (LSD) method and the Tukey method?

8. Why does a Variance Inflation Factor (VIF) greater than ten indicate the presence of multicollinearity?
9. What is the difference between an outlier and an influential observation? What methods are used to use to tell if one has an outlier or an influential observation?

10. Explain what is the prediction error sum of squares (PRESS) and how it is used?

11. Explain the difference between the two sampling models used in two-way count data.

12. In the univariate continuous distribution context what is difference between a confidence interval, a prediction interval, and a tolerance interval?
13. Hospitals are graded based on their success rates in treating different categories of patients. We want to compare two hospitals – A, which is a university affiliated research hospital, and B, which is a general community hospital – with respect to success rates for a certain complicated surgery. The data classified by low risk patients and high risk patients are shown in the following table.

<table>
<thead>
<tr>
<th>Low Risk</th>
<th></th>
<th>High Risk</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Success</td>
<td>Failure</td>
<td>Total</td>
</tr>
<tr>
<td>Hospital A</td>
<td>400</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>Hospital B</td>
<td>300</td>
<td>200</td>
<td>500</td>
</tr>
<tr>
<td>Total</td>
<td>700</td>
<td>300</td>
<td>1000</td>
</tr>
</tbody>
</table>

a. Calculate the success rates for each category of patients for both hospitals. Which hospital is better?

<table>
<thead>
<tr>
<th>Success Rates</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Aggregate the data over the two categories of patients and calculate the overall success rates for both hospitals. Now which hospital is better?

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
</tr>
</tbody>
</table>

c. Explain the discrepancy between the results obtained in (a) and (b).

d. What are the adjusted (standardized) success rates for each hospital?

<table>
<thead>
<tr>
<th>Success Rates</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk</td>
<td>Hospital A</td>
<td>Hospital B</td>
<td>Total</td>
</tr>
<tr>
<td>Low</td>
<td>80%</td>
<td>60%</td>
<td>1000</td>
</tr>
<tr>
<td>High</td>
<td>20%</td>
<td>10%</td>
<td>1000</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>2000</td>
</tr>
</tbody>
</table>
14. A sample of employed men aged 18 and 67 were asked if they had carried out work on their home in the preceding year for which they would have previously employed a craftsman. The following table gives the summary of responses of 906 homeowners.

<table>
<thead>
<tr>
<th>Work</th>
<th>Home Repair</th>
<th>Age</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skilled</td>
<td>Yes</td>
<td>&lt;30</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31-45</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46+</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>&lt;30</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31-45</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46+</td>
<td>8</td>
</tr>
<tr>
<td>Unskilled</td>
<td>Yes</td>
<td>&lt;30</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31-45</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46+</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>&lt;30</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31-45</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46+</td>
<td>51</td>
</tr>
<tr>
<td>Office</td>
<td>Yes</td>
<td>&lt;30</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31-45</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46+</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>&lt;30</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31-45</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46+</td>
<td>61</td>
</tr>
<tr>
<td>Column Totals</td>
<td></td>
<td>&lt;30</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31-45</td>
<td>427</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46+</td>
<td>306</td>
</tr>
</tbody>
</table>

How would you analyze these data if you are interested in determining if home repair and age are associated?