Exam 1, STAT2331-802/3, Spring 2003

Please do all questions. There are two types of problems, multiple choice and short answer. There are 7 pages, 10 multiple-choice problems, and 2 short answer problems. Don’t forget to show all your work for partial credit on the short answer problems.

Section I: Multiple Choice

1. (5 points) With respect to quantitative and categorical variables, which statement is FALSE?
   
   a. Categorical variables may be described numerically
   b. Only some quantitative variables can be measured in units
   c. The number of siblings someone has could be used as a categorical variable
   d. The number of hours slept by STAT2331 students before writing this exam would be a quantitative variable
   e. Pie charts are a simple and effective tool that can be used to summarize categorical variables

2. (5 points) In general, which of the following statements is FALSE?
   
   a. The sample mean is more sensitive to extreme values than the median.
   b. The sample range is more sensitive to extreme values than the standard deviation.
   c. The sample standard deviation is a measure of spread around the sample mean.
   d. The sample standard deviation is a measure of central tendency around the median.
   e. If a distribution is symmetric, then the mean will be equal to the median.

3. (5 points) The stem and leaf plot below summarizes the final year averages of a graduating honor class in the Dedman College. Select the CORRECT statement.

   6 | 88
   7 | 24567
   8 | 2346
   9 | 014
   10 |

   a. The distribution is bimodal
   b. The values in the leaf are not correctly displayed
   c. The IQR is 12
   d. The distribution is unimodal and left skewed
   e. The range is 23
4. (5 points) You are given that the temperature on 7 hottest days last summer were 28, 29, 30, 31, 32, 33, 35 degrees Celsius (humidity factor not considered). Calculate D where D = median – mean;
   a. -0.23
   b. -0.14
   c. 0.14
   d. 0.23
   e. Cannot be determined

5. (5 points) Each option below gives two sets of numbers. Each set has a standard deviation. Which of the following pairs includes sets that do NOT have the same standard deviation? (You don’t need to waste time in calculating SD)
   a. {4,4,4,4} and {5,5,5,5}
   b. {4,4,5,5} and {4,4,5,5,5}
   c. {4,4,5,5} and {5,5,5,5}
   d. {4,4,5,5} and {4,5,4,5}
   e. {4,4,5,4} and {4,5,5,5,5}

6. (5 points) Which of the following is a CORRECT statement about scatterplots.
   a. A scatterplot would be a good tool to use in examining the differences in height of statistics students by sex.
   b. Form, direction and outliers provide a complete description of the overall pattern in a scatterplot.
   c. As interest rates rise, the price of bonds fall. A scatterplot would show these two variables are positively associated.
   d. A scatterplot that has a curved pattern with all of the points following the curved pattern very closely is said to be weak because we have a preference for a linear relationship.
   e. None of the above statements are correct.

7. (5 points) In the boxplot at the top of next page, the third quartile of the heights of the Oak Trees is closest to which height?
   a. 67 feet
   b. 60 feet
   c. 53 feet
   d. 48 feet
   e. 40 feet
The heights of the mature Oak Trees in Question 8 are normally distributed with a mean of 53 feet and a standard deviation of 8 feet. (Answer Questions 8-10)

8. (5 points) What must be the height of an Oak Tree so that 2.5% of all the Oak Trees are shorter than it? (Use 68-95-99.7 rule.)

   a. 37  
   b. 45  
   c. 61  
   d. 69  
   e. 77  

9. (5 points) What percentage of mature Oak Trees are taller than 37 feet and shorter than 61 feet? (Use 68-95-99.7 rule)

   a. 47.5%  
   b. 68%  
   c. 81.5%  
   d. 84%  
   e. 95%  

10. (5 points) What percentage of mature Oak Trees are taller than 63 feet? (Use Table A)

   a. 10.6%  
   b. 11.5%  
   c. 14%  
   d. 88.5%  
   e. 89.4%  

Multiple Choice Answers:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>d</td>
<td>c</td>
<td>b</td>
<td>b</td>
<td>e</td>
<td>b</td>
<td>a</td>
<td>c</td>
<td>a</td>
</tr>
</tbody>
</table>
Section II: Short Answer / Calculation

1. The height (in feet) and volume of usable lumber (in cubic feet) of 6 cherry trees are measured by a researcher.

<table>
<thead>
<tr>
<th>X: Height (in feet)</th>
<th>61</th>
<th>65</th>
<th>75</th>
<th>80</th>
<th>85</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y: Volume of usable lumber (in cubic feet)</td>
<td>7</td>
<td>15</td>
<td>27</td>
<td>38</td>
<td>38</td>
<td>55</td>
</tr>
</tbody>
</table>

Given that

\[ S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2 = 640, \quad S_{yy} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 1516, \quad S_{xy} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 967. \]

(a) (5 points) Calculate the correlation \( r \) for height and the volume of usable lumber.

\[
r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{967}{\sqrt{640 \cdot 1516}} = 0.9817
\]

(b) (7 points) Show that the equation of the regression line fitted to predict volume of usable lumber from the height of a tree is \( \hat{y} = -84.83 + 1.51x \).

\[
\bar{y} = 30, \bar{x} = 76
\]

\[
b = \frac{S_{xy}}{S_{xx}} = \frac{967}{640} = 1.51094
\]

\[
a = \bar{y} - b\bar{x} = 30 - (1.51094)(76) = -84.8313
\]
1. (Cont’d)

(c) (4 points) Predict the volume of usable lumber of a cherry tree that is 75 feet tall.

When \( x = 75 \), \( \hat{y} = -84.83 + (1.51)(75) = 28.48906 \)

(d) (3 points) What is the value of the residual in part (c)?

Residual = observed value – predicted value = 27 – 28.48906 = -1.48906

(e) (6 points) Predict the volume of usable lumber of a cherry tree that is 49 feet tall. What has gone wrong here? Explain.

When \( x = 49 \), \( \hat{y} = -84.83 + (1.51)(49) = -10.7953 \)

The problem is the volume of usable lumber cannot be negative. We have this problem since we are doing extrapolation using regression line which is calculated from the data with range \( x = 61 \) to 90, and 49 far away from this range. It is possible that the linear relationship does not hold outside the range \( x = 61 \) to 90.
2. In 1972-1994 a one-in-six survey of the electoral roll, largely concerned with thyroid disease and heart disease was carried out in Wichkham, a mixed urban and rural district near Newcastle upon Tyne, in the UK. Twenty years later, a follow-up study was conducted.

Here are the results for two age groups of females. Each table shows the twenty-year survival status for smokers and non-smokers.

<table>
<thead>
<tr>
<th>Age 55-64</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dead</td>
<td>Alive</td>
</tr>
<tr>
<td>Smokers</td>
<td>51</td>
<td>64</td>
</tr>
<tr>
<td>Non-smokers</td>
<td>40</td>
<td>81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age 65-74</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dead</td>
<td>Alive</td>
</tr>
<tr>
<td>Smokers</td>
<td>29</td>
<td>7</td>
</tr>
<tr>
<td>Non-smokers</td>
<td>101</td>
<td>28</td>
</tr>
</tbody>
</table>

(a) (3 points) Calculate the death rate of smoker and the death rate of non-smoker for age group 55-64. (Look at the first table only.)

Death rate of smoker for age group 55-64: 51/(51+64) = 44.35%

Death rate of non-smoker for age group 55-64: 40/(40+81) = 33.06%

(b) (3 points) Calculate the death rate of smoker and the death rate of non-smoker for age group 65-74. (Look at the second table only.)

Death rate of smoker for age group 65-74: 29/(29+7) = 80.56%

Death rate of non-smoker for age group 65-74: 101/(101+28) = 78.29%

(c) (4 points) Make a two-way table classifying people by smoking and whether or not they were died, that is collapsing over age group.

<table>
<thead>
<tr>
<th></th>
<th>Dead</th>
<th>Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>80</td>
<td>71</td>
</tr>
<tr>
<td>Non-smokers</td>
<td>141</td>
<td>109</td>
</tr>
</tbody>
</table>
2. (Cont’d)
   (d) (3 points) Calculate the death rate of smoker and the death rate of non-smoker. (Use your collapsed table!)

   Death rate of smoker: \( \frac{80}{80+71} = 52.98\% \)

   Death rate of non-smoker: \( \frac{141}{141+109} = 56.4\% \)

   (e) (12 points) Explain simply what is going on here. (Use the structure outlined in class for your explanation.)

   1) Explanatory variable: smoker or non-smoker
      Response variable: Dead or alive
      Confounding variable: Age group

   2) Confounding & Explanatory: There is a higher percentage of non-smoker in the age group 65-74 (78.18%) compare with that in the age group 55-64 (51.27%).

      Confounding & Response: There is a higher death rate in the age group 65-74 (78.79%) compare with that in age group 55-64 (38.56%).

   3) This is an example of Simpson’s paradox. It looks like the death rate is higher for non-smoker overall, but this is not true because most of the people in the age group 65-74 are non-smoker and a high death rate in this age group (older people has a higher death rate is reasonable).

      ---- The End ----
      Good Luck!!!