Part 1: Multiple Choice. Circle the letter corresponding to the best answer.

1. Which of the following scatterplots would indicate that \( Y \) is growing exponentially over time?

   (a) \( \begin{array}{c|c|c|c|c|c} Time & 2 & 4 & 6 & 8 & 10 \\ \hline Y & 2 & 4 & 8 & 16 & 32 \end{array} \) 

   (b) \( \begin{array}{c|c|c|c|c|c} Time & 2 & 4 & 6 & 8 & 10 \\ \hline Y & 1 & 2 & 4 & 8 & 16 \end{array} \) 

   (c) \( \begin{array}{c|c|c|c|c|c} Time & 2 & 4 & 6 & 8 & 10 \\ \hline Y & 2 & 6 & 12 & 24 & 48 \end{array} \) 

   (d) \( \begin{array}{c|c|c|c|c|c} Time & 2 & 4 & 6 & 8 & 10 \\ \hline Y & 2 & 3 & 6 & 12 & 24 \end{array} \) 

   (e) none of these

Consider the two-way table of data at the right.

2. The percent of cars listed in the table with 4-cylinder engines that are made in Germany is

   (a) 10.5%. 

   (b) 21%. 

   (c) 50%. 

   (d) 80%. 

   (e) 91%.

3. From the table we might conclude that

   (a) there is clearly no relation between country of origin and number of cylinders. 

   (b) the correlation between country of origin and number of cylinders is likely to be about 0.5. 

   (c) a regression line fitted to these data would probably have a negative slope. 

   (d) there is evidence of some relation between country of origin and number of cylinders. 

   (e) the United States has far more cars than any of the other countries.

4. According to the 1990 census, those states with an above-average number \( X \) of people who fail to complete high school tend to have an above average number \( Y \) of infant deaths. In other words, there is a positive association between \( X \) and \( Y \). The most plausible explanation for this is

   (a) \( X \) causes \( Y \). Programs to keep teens in school will help reduce the number of infant deaths. 

   (b) \( Y \) causes \( X \). Programs that reduce infant deaths will ultimately reduce high school dropouts. 

   (c) Lurking variables are probably present. For example, states with large populations will have both larger numbers of people who don’t complete high school and more infant deaths. 

   (d) Both of these variables are directly affected by the higher incidence of cancer in certain states.
5. Recent data show that states that spend an above-average amount of money \( X \) per pupil in high school tend to have below-average mean Verbal SAT scores \( Y \) of all students taking the SAT in the state. In other words, there is a negative association between \( X \) and \( Y \). This is particularly true in states that have a large percent of all high school students taking the exam. Such states also tend to have larger populations. The most plausible explanation for this association is
(a) \( X \) causes \( Y \). Overspending generally leads to extra, unnecessary programs, diverting attention from basic subjects. Inadequate training in these subjects generally leads to lower SAT scores.
(b) \( Y \) causes \( X \). Low SAT scores creates concerns about the quality of education. This inevitably leads to additional spending to help solve the problem.
(c) Changes in \( X \) and \( Y \) are due to a common response to other variables. If a higher percent of students take the exam, the average score will be lower. Also, states with larger populations have large urban areas where the cost of living is higher and more money is needed for expenses.
(d) Changes in \( X \) and \( Y \) are due to confounding with a third variable. As an example, states in which more money is spent on education tend to be those states with more students living in poverty. As a result, it’s the students’ living conditions that lead to lower SAT Verbal scores, and not the amount spent per pupil.
(e) The association between \( X \) and \( Y \) is purely coincidental.

A business has two types of employees, managers and workers. Managers earn either $100,000 or $200,000 per year. Workers earn either $10,000 or $20,000 per year. The number of male and female managers at each salary level and the number of male and female workers at each salary level are given in the two tables below.

<table>
<thead>
<tr>
<th></th>
<th>Managers</th>
<th></th>
<th>Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>$100,000</td>
<td>80</td>
<td>20</td>
<td>$10,000</td>
</tr>
<tr>
<td>$200,000</td>
<td>20</td>
<td>30</td>
<td>$20,000</td>
</tr>
</tbody>
</table>

6. The proportion of male managers who make $200,000 per year is
(a) 0.067. (b) 0.133. (c) 0.200. (d) 0.400. (e) 0.667.

7. From these data we may conclude that
(a) the mean salary of female managers is greater than that of male managers.
(b) the mean salary of males in this business is greater than the mean salary of females.
(c) the mean salary of female workers is greater than that of male workers.
(d) this is an example of Simpson's paradox.
(e) all of the above.

\[
\text{mean salary of all females} = \frac{140,000(50) + 18,000(100)}{150} = \$5,333.33
\]
\[
\text{mean salary of all males} = \frac{84,000(64) + 14,000(40)}{104} = \$8,461.54
\]
A survey was designed to study how the operations of a group of businesses vary with their size. Companies were classified as small, medium, and large. Questionnaires were sent to 200 randomly selected businesses of each size, for a total of 600 questionnaires. Since not all questionnaires in a survey of this type are returned, it was decided to examine whether or not the response rate varied with the size of the business. The data are given in the following two-way table:

<table>
<thead>
<tr>
<th>Size</th>
<th>Response</th>
<th>No Response</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>125</td>
<td>75</td>
<td>200</td>
</tr>
<tr>
<td>Medium</td>
<td>81</td>
<td>119</td>
<td>200</td>
</tr>
<tr>
<td>Large</td>
<td>40</td>
<td>160</td>
<td>200</td>
</tr>
</tbody>
</table>

8. What percent of all small companies receiving questionnaires responded?
   (a) 50.8%  (b) 20.8%  (c) 62.5%  (d) 33.3%  (e) 12.5%

9. Which of the following conclusions seems to be supported by the data?
   (a) There are more small companies than large companies in the survey.
   (b) Small companies appear to have higher response rates than medium or big companies.
   (c) Exactly the same number of companies responded as didn't respond.
   (d) Small companies dislike larger companies.
   (e) If we combined the medium and large companies, then their response rate would be equal to that of the small companies.

An article in the student newspaper of a large university had the headline “A's swapped for evaluations?” The article included the following.

According to a new study, teachers may be more inclined to give higher grades to students, hoping to gain favor with the university administrators who grant tenure. The study examined the average grade and teaching evaluation in a large number of courses in order to investigate the effects of grade inflation on evaluations. “I am concerned with student evaluations because instruction has become a popularity contest for some teachers,” said Professor Smith, who recently completed the study.

Results showed that higher grades directly corresponded to a more positive evaluation.

10. Which of the following would be a valid conclusion to draw from the study?
    (a) A teacher can improve his or her teaching evaluations by giving good grades.
    (b) A good teacher, as measured by teaching evaluations, helps students learn better, resulting in higher grades.
    (c) Teachers of courses in which the mean grade is above average apparently tend to have above-average teaching evaluations.
    (d) Teaching evaluations should be conducted before grades are awarded.
    (e) All of the above.
Part 2: Free Response

11. Cell phones, a fairly recent innovation, have become increasingly popular with all segments of our society. According to the Strategis Group, the number of cellular and personal communications systems subscribers in the United States increased dramatically beginning in 1990, as shown in the following table.

<table>
<thead>
<tr>
<th>Year</th>
<th>No. of subscribers (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>5.3</td>
</tr>
<tr>
<td>1991</td>
<td>7.6</td>
</tr>
<tr>
<td>1992</td>
<td>11.0</td>
</tr>
<tr>
<td>1993</td>
<td>16.0</td>
</tr>
<tr>
<td>1994</td>
<td>24.1</td>
</tr>
<tr>
<td>1995</td>
<td>33.8</td>
</tr>
<tr>
<td>1996</td>
<td>43.4</td>
</tr>
</tbody>
</table>

(a) Apply a test to show that the number of subscribers is increasing exponentially.

(b) Enter the data into your calculator. Then perform an appropriate transformation to linearize the data. Find the equation of the least-squares line for the transformed data. Record it below. Be sure to define any variables you use.

\[ \log \hat{y} = 0.154x + 0.731 \]

(c) How well does the linear model you calculated in (b) fit the transformed data? Justify your answer with graphical and verbal evidence.

(d) The Strategis Group predicts 70.8 million subscribers in 1998, and 99.2 million in the year 2000. How many subscribers does your model predict for these years? Show your method.
12. Foresters are interested in predicting the amount of usable lumber they can harvest from various tree species. The following data have been collected on the diameter of Ponderosa pine trees (in inches), measured at chest height, and the yield in board feet. Note that a board foot is defined as a piece of lumber 12 inches by 12 inches by 1 inch.

(a) Use your calculator to construct a scatterplot that would help the foresters make their predictions. Describe what you see.

(b) One forester suggests that a reasonable model for predicting amount of usable lumber \((y)\) from Ponderosa pine diameter \((x)\) is exponential. Use your calculator to perform a transformation that linearizes the data (using logs) if the forester’s suggested model is correct. Then find the equation of the least-squares line for the transformed data. Record the equation below. Define any variables you use.

\[
\log y = .049x + .535 \quad x = \text{diameter} \quad y = \text{board feet}
\]

(c) How well does the linear model you calculated in (b) fit the transformed data? Justify your answer with graphical and verbal evidence.

(d) Suggest a different transformation, involving logarithms, that would help linearize the data. Find the least squares regression line for the transformed data suggested and define all variables.

(e) Perform an inverse transformation to re-express your equation from (d) in terms of \(x\) and \(y\).

\[
\hat{y} = (.0027x^{3.114})
\]

(g) Use your model from (d) or (e) to predict the amount of usable lumber from a Ponderosa pine with diameter 30 inches. Show your method. How confident are you in this prediction?

\[
\hat{y} = .0027(30)^{3.114} = 117.34 \quad \text{fairly confident - within domain of } x \& \text{ power is a good fit.}
\]

(f) Use your model from (d) or (e) to predict the amount of usable lumber from a Ponderosa pine with diameter 62 inches. Show your method. How confident are you in this prediction?

\[
\hat{y} = .0027(62)^{3.114} = 1144.74 \quad \text{not confident - 62 is well outside my domain.}
\]

<table>
<thead>
<tr>
<th>Diameter</th>
<th>Bd. Feet</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>192</td>
</tr>
<tr>
<td>28</td>
<td>113</td>
</tr>
<tr>
<td>28</td>
<td>88</td>
</tr>
<tr>
<td>41</td>
<td>294</td>
</tr>
<tr>
<td>19</td>
<td>28</td>
</tr>
<tr>
<td>32</td>
<td>123</td>
</tr>
<tr>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>38</td>
<td>252</td>
</tr>
<tr>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>31</td>
<td>141</td>
</tr>
<tr>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>25</td>
<td>86</td>
</tr>
<tr>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>39</td>
<td>231</td>
</tr>
<tr>
<td>33</td>
<td>187</td>
</tr>
<tr>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>37</td>
<td>205</td>
</tr>
<tr>
<td>23</td>
<td>57</td>
</tr>
<tr>
<td>39</td>
<td>265</td>
</tr>
</tbody>
</table>
13. There is a positive association between the number of drownings and ice cream sales. Is the association between these two variables most likely due to causation, confounding, or common response? Justify your answer.

Common response → temperature is a variable that affects both. More people swim (so more drown) and eat ice cream when the weather is warm.

14. From tax records, it is relatively easy to determine the amount of liquor consumed per person and the number of cigarettes consumed per person for each of the 10 provinces of Canada. These are plotted on a scatterplot and a high positive correlation is found. Is the association between these two variables most likely due to causation, confounding, or common response? Justify your answer.

Common response → smoking & drinking both tend to be associated with average family income.

15. A company decided to expand, so it opened a new factory with 455 available jobs. The following tables summarize the hiring decisions made by the company.

<table>
<thead>
<tr>
<th>Workers</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied</td>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>Hired</td>
<td>300</td>
<td>85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Managers</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Hired</td>
<td>30</td>
<td>40</td>
</tr>
</tbody>
</table>

(a) Calculate the percent of male and female workers that are hired. Then do likewise for male and female managers.

Workers Hired

M: 75%  F: 85%

(b) Use the tables above to create a two-way table that shows the relationship between gender and hiring decision.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied</td>
<td>400</td>
<td>300</td>
</tr>
<tr>
<td>Hired</td>
<td>300</td>
<td>125</td>
</tr>
</tbody>
</table>

(c) Calculate the percent of male and female applicants that were hired.

Males Hired 55%  higher 9%
Females Hired 42%

(d) Explain your findings in (a) and (c).

When workers & managers are considered separately, a higher percentage of female workers were hired. However, when combined, a higher percentage of males were hired. This is an example of Simpson's Paradox.
16. **Life Expectancy.** The data below list the life expectancy for white males in the United States every decade during the last century (1 = 1901 – 1910, 2 = 1911 – 1920, etc.).

<table>
<thead>
<tr>
<th>Decade</th>
<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>Life Expectancy</td>
<td>48.6</td>
<td>54.4</td>
<td>59.7</td>
<td>62.1</td>
<td>66.5</td>
<td>67.4</td>
<td>68.0</td>
<td>70.7</td>
<td>72.7</td>
<td>74.9</td>
</tr>
</tbody>
</table>

a. Letting \( x \) be the decade and \( y \) the life expectancy, find the equation of the regression line of \( \log y \) in terms of \( x \), calculate the correlation, and sketch the graph. (Label the axes and quickly sketch the graph from your calculator screen.)

\[
\log \hat{y} = 0.0186x + 1.704 \\
\hat{r} = 0.951
\]

b. Letting \( x \) be the decade and \( y \) the life expectancy, find the equation of the regression line of \( \log y \) in terms of \( \log x \), calculate the correlation, and sketch the graph. (Label the axes and quickly sketch the graph from your calculator screen.)

\[
\log \hat{y} = 0.185 \log x + 1.68 \\
\hat{r} = 0.996
\]

c. Based on your findings in (a) and (b), determine whether the exponential or power regression model is most appropriate. Find that model by re-expressing the linear regression from (a) or (b). (b) is better (higher \( r \) and more linear), so power regression is more appropriate for \( x, y \).

\[
\hat{y} = 10^{0.185 \log x + 1.68} = (10^{0.185})^{(10^{1.68})} = (10^{0.185})(10^{1.68})
\]

\[\hat{y} = 47.86 \times \]

\[\hat{y} = 48.39 \times ^{1.85} \]

\[\hat{y} = 48.39(7)^{1.85} \quad \text{yes}!!
\]

d. Use the calculator to find the appropriate exponential or power model. Does this match your answer from (c)? (It should!)

\[
\hat{y} = 48.39(15)^{1.85} = 71.86
\]

\[
\hat{y} = 48.39(7)^{1.85} = 71.86
\]

\[\text{residual} = -1.4
\]

\[\text{not confident - extrapolation}
\]

\[\text{residual} = -1.4
\]
17. Use the table below to answer the following questions.

**College Students by Gender and Age Group, 2003**

<table>
<thead>
<tr>
<th>Education</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 – 17 years</td>
<td>89</td>
<td>61</td>
<td>150</td>
</tr>
<tr>
<td>18 – 24 years</td>
<td>5,668</td>
<td>4,697</td>
<td>10,365</td>
</tr>
<tr>
<td>25 – 34 years</td>
<td>1,904</td>
<td>1,589</td>
<td>3,493</td>
</tr>
<tr>
<td>35 years or older</td>
<td>1,660</td>
<td>970</td>
<td>2,630</td>
</tr>
<tr>
<td>Total</td>
<td>9,321</td>
<td>7,317</td>
<td>16,638</td>
</tr>
</tbody>
</table>

a. Fill in the row and column totals in the margins of the table.

b. Compute (in percents) the marginal distribution of college students.

\[
\begin{align*}
15-17\text{yrs} &= \frac{150}{16,638} = .9% \\
18-24\text{yrs} &= 62.3% \\
25-34\text{yrs} &= 3.493% \\
35+\text{yrs} &= 15.82%
\end{align*}
\]

c. Compute (in percents) the conditional distribution of females for each age group.

\[
\begin{align*}
15-17\text{yrs} &= \frac{89}{9321} = .95% \\
18-24\text{yrs} &= 60.8% \\
25-34\text{yrs} &= 20.4% \\
35+\text{yrs} &= 17.87%
\end{align*}
\]

d. Use the grid to display the conditional distributions from (c) in a graph. Label your graph.