1. **(10 points)** Government statisticians in England conducted a study of the relationship between smoking and lung cancer. The data concern 25 occupational groups and are condensed from data on thousands of individual men. Two indices are presented for each occupational group. The smoking index is the ratio of the average number of cigarettes smoked per day by men in the particular occupational group to the average number of cigarettes smoked per day by all men. The mortality index is the ratio of the rate of deaths from lung cancer among men in the particular occupational group to the rate of deaths from lung cancer among all men. (Source: Moore, David S., and George P. McCabe (1989). *Introduction to the Practice of Statistics*. Original source: *Occupational Mortality: The Registrar General’s Decennial Supplement for England and Wales, 1970-1972*, Her Majesty’s Stationery Office, London, 1978.)

From the scatterplot above, what number below appears to be closer to the correlation between the **Mortality** index and the **Smoking** index? Why?

(a) $-0.999$
(b) $0.999$
(c) $0.73$
(d) $0$
(e) $-0.73$

There appears to have a not-so-tight upward linear relationship between the **Mortality** index and the **Smoking** index.
2. Consider the following stem-and-leaf plot of the 19 sales Richard made last month (in $100) at a computer store.

```
Stem Width = 10
3  9
4  2689
5  4578999
6  15
7  03
8  6
9  0
```
(So, for example, the number under stem 9 is read as 90 (in $100).)

(a) (5 points) What is the shape of the data?

Turning the stem-and-leaf display counter-clockwise 90°, it is obvious that the data set is right skewed due to long right tail.

(b) (10 points) what is the median amount of sales (in $100) that Richard made last month?

The median is the \( \frac{n+1}{2} = \frac{20}{2} = 10 \)th ordered data value which is 58 (in $100). That is, $5,800.

3. (10 points) The chest sizes, measured in inches, of Scottish militiamen in the early 19th century were recorded (from DASL: Chest sizes of Militiamen). A histogram of chest sizes shows an approximately normal curve. It is known that the mean chest size is 39.1 inches and that approximate 95% of soldiers have chest size in between 35.7 inches and 42.5 inches. (Note that 39.1 inches is half way between 35.7 inches and 42.5 inches.) What is the standard deviation of chest size? Show your work. (Hint: Use empirical rule.)

According to the empirical rule, 95% of soldiers have chest size are within an interval of length \( 4SD \) centered at the mean. So,

\[
SD = \frac{42.5 - 35.7}{4} = \frac{6.8}{4} = 1.7
\]

4. Students in grades 4–6 in selected schools in Michigan, were asked the following question: What would you most like to do at school?

A) Make good grades. B) Be good at sports. C) Be popular.
The result follows:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Goals</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>grades</td>
<td>popular</td>
</tr>
<tr>
<td>Boys</td>
<td>117</td>
<td>50</td>
</tr>
<tr>
<td>Girls</td>
<td>130</td>
<td>91</td>
</tr>
<tr>
<td>Total</td>
<td>247</td>
<td>141</td>
</tr>
</tbody>
</table>


(a) (5 points) Calculate the degrees of freedom.

\[
df = (2 - 1) \times (3 - 1) = 2.
\]

(b) (5 points) The (partially completed) table of expected frequencies is given:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Goals</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>grades</td>
<td>popular</td>
<td>sports</td>
</tr>
<tr>
<td>Boys</td>
<td>117.3</td>
<td>???</td>
<td>42.7</td>
</tr>
<tr>
<td>Girls</td>
<td>129.7</td>
<td>74.0</td>
<td>47.3</td>
</tr>
<tr>
<td>Total</td>
<td>247</td>
<td>141</td>
<td>90</td>
</tr>
</tbody>
</table>

The expected frequency for the cell (Boys, popular) is

\[
E_{12} = \frac{(\text{row 1 total}) \times (\text{column 2 total})}{\text{grand total}} = \frac{227 \times 141}{478} = 67.0
\]

Note that \( E_{12} = 141 - 74.0 \).

(c) (5 points) What is the critical value? (That is, what is the value of \( b \))?

The critical value is \( b = 5.99 \) at 2 degrees of freedom.

(d) (5 points) It turns out that the chi-square statistic is \( \chi^2 = 21.46 \) which far exceeds the critical value. Draw the conclusion about these two categorical variables *Gender* and *Goals*.

They are dependent since the chi-square statistic is greater than the critical value.

5. Data were obtained for per capita numbers of cigarettes smoked (sold) by 43 states and the District of Columbia in 1960 together with death rates per hundred thousand population
from various forms of cancer. (Source: DASL. Also: J.F. Fraumeni, “Cigarette Smoking and Cancers of the Urinary Tract: Geographic Variations in the United States,” *Journal of the National Cancer Institute*, 41, 1205–1211.)

It is of interest to predict *Lung Cancers* (lung cancer deaths per 100K population) based on the *Cigarettes* (number of cigarettes smoked (heads per capita)). The scatter plot of the data and the summary statistics are given below:

![Scatter plot of Lung Cancers vs. Cigarettes](image)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Cigarettes</em></td>
<td>24.91</td>
<td>5.57</td>
</tr>
<tr>
<td><em>Lung Cancers</em></td>
<td>19.65</td>
<td>4.23</td>
</tr>
<tr>
<td>Correlation: $r = 0.6974$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope: $b = 0.53$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept: $a = 6.47$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**(a) (10 points)** Which variable is the response variable? Which is the explanatory variable?

*Lung Cancers* is the response variable and *Cigarettes* is the explanatory variable.

**(b) (10 points)** Write the least square regression equation and interpret the slope.

The least square equation is

$$\text{LungCancers} = 6.47 + 0.53 \times \text{Cigarettes}$$

There is 0.53 deaths increases in *Lung Cancers* per 100K population (or, equivalently, 53 deaths increase in *Lung Cancers* per 10 million population) for every head of *Cigarettes* increase.

**(c) (10 points)** For the pair of data (*Cigarettes*=24.96, *Lung Cancers*=22.72), calculate the residual.
The predicted \( \text{Mortality} \) rate is

\[ \hat{y} = 6.47 + 0.53 \times 24.96 = 6.47 + 13.23 = 19.70 \]

and hence the residual is

\[ 22.72 - 19.70 = 3.02. \]

6. \textbf{(15 points)} Why does the moon appear to be so much larger when it is near the horizon than when it is directly overhead? This question has produced a wide variety of theories from psychologists. An important early hypothesis was put forth by Holway and Boring (1940) who suggested that the illusion was due to the fact that when the moon was on the horizon, the observer looked straight at it with eyes level, whereas when it was at its zenith, the observer had to elevate his or her eyes as well as his or her head to see it. Kaufman and Rock (1962) devised an apparatus to test this hypothesis. For each subject the ratio of the perceived size of the horizon moon to the perceived size of the zenith moon was recorded with eyes elevated and with eyes level. A ratio of 1.00 would represent no illusion. If Holway and Boring were correct, there should be a greater illusion in the eyes-elevated condition than in the eyes-level condition. The results were given:

<table>
<thead>
<tr>
<th>Subject</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>Elevated</td>
<td>1.65</td>
<td>1.00</td>
<td>2.03</td>
<td>1.25</td>
<td>1.05</td>
<td>1.02</td>
<td>1.67</td>
<td>1.86</td>
<td>1.56</td>
<td>1.73</td>
</tr>
<tr>
<td>Level</td>
<td>1.73</td>
<td>1.06</td>
<td>2.03</td>
<td>1.40</td>
<td>0.95</td>
<td>1.13</td>
<td>1.41</td>
<td>1.73</td>
<td>1.63</td>
<td>1.56</td>
</tr>
</tbody>
</table>

Of the two analyses below, only one is correct:

<table>
<thead>
<tr>
<th>Analysis I. Two Independent Samples</th>
<th>Difference in Means (Level – Elevated)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis II. Paired Samples</td>
<td>–0.019</td>
<td>(–0.333, 0.295)</td>
</tr>
</tbody>
</table>

Choose the correct analysis, reason your choice, and draw conclusion.

\textbf{Analysis II} is the correct analysis since the data are paired data. The confidence interval \((-0.103, 0.065)\) includes zero and consequently there is no difference between the eyes-elevated condition and in the eyes-level condition.

7. \textbf{(5 points) Extra Credit Problem}

Demographic and crime-related data for 47 US states in 1970 showed that the correlation between \textit{crime rate} (number of offenses reported to police per million population) and \textit{mean education years} (mean number of schooling years for persons of age 25 or older) is 0.32 (Hand, D.J., et al. (1994) \textit{A Handbook of Small Data Sets}, London: Chapman & Hall, 101–103.). Which of the following statements is true about these two variables. (Answer and reason your choice.)

(a) The \textit{crime rate} causes the increase in \textit{mean education years}.  

5
(b) These two variables are negatively linearly related.
(c) These two variables are positively linearly related.
(d) There is no relationship between these two variables.
(e) The increase in mean education years causes the increase in crime rate.

There is a positive linear relationship between the two variables. However, the positive linear relationship does not establish causation.