1. To study air quality on the Central Coast 20 air samples from various areas were obtained. For each one the carbon monoxide concentration was calculated.

<table>
<thead>
<tr>
<th>Concentration (ppm)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9.4</td>
<td>12.5</td>
<td>10.4</td>
<td>10.1</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td>10.5</td>
<td>8.5</td>
<td>11.2</td>
<td>9.9</td>
<td>12.0</td>
<td></td>
</tr>
<tr>
<td>13.2</td>
<td>9.3</td>
<td>9.8</td>
<td>9.7</td>
<td>11.3</td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td>8.7</td>
<td>10.3</td>
<td>11.1</td>
<td>9.5</td>
<td></td>
</tr>
</tbody>
</table>

\[
\sum x = 209.5 \quad \sum x^2 = 2222.73
\]

\[
\sum (x - \bar{x})^2 = 28.2175
\]

a. What is the population? Sample? Variable of interest?

b. What is the mean and standard deviation of the carbon monoxide concentration?

c. Make a stem and leaf display.

d. Which of the following boxplots is the correct one for this data?

2. Examine the histogram below (n = 46).

a. Match the following descriptive statistics to their most likely value.

   \begin{align*}
   \text{mean} & \quad 11.78 \\
   \text{median} & \quad 42.00 \\
   \text{SD} & \quad 39.48
   \end{align*}

b. Describe the shape of this histogram.

c. Can you tell if the data for var1 is numerical or categorical?
d. Can you tell if the data is discrete or continuous?

e. Why would a bar chart of var1 not give the same picture as the histogram? Explain.

3. A study was conducted to demonstrate a researcher's theory that soy beans inoculated with nitrogen-fixing bacteria would yield more and grow adequately without the use of expensive synthesized fertilizers. The study was conducted under controlled conditions with uniform amounts of soil, on 30 inoculated soy bean plants. The plant yield as measured by pod weight (gm) for each plant are represented below in the stem and leaf display.

```
#1
Stem-and-leaf of I N = 30
Leaf Unit = 0.010
1   10 0
1   11
1   12
1   13
3   14 56
8   15 3478
11  16 455
14  17 689
(5)  18 56799
11  19 1689
7   20 0
6   21 25
4   22 02
2   23 04
```

a. Find the minimum and maximum observation.

b. Which graph is the appropriate relative frequency histogram for this data?

c. Suppose data were also collected for a group of 30 uninoculated soy bean plants grown under similar conditions. If the average pod weight for the uninoculated group was 1.084 gm, what is your overall impression concerning a comparison of the average pod weight between the two groups? Does this support the researcher's theory? NOTE: you should not need to make any calculations to answer this problem.

d. Is this study observational or experimental? Explain.

4. The salary of professional athlete’s receives much attention in the media. It is becoming commonplace to hear of the multimillion-dollar contracts for the few select superstar athletes on each team. Because of this team owners and players' associations spend much time with salary negotiations over additional salary and increased benefits.

a. Typically a professional sports team consists of one or two superstar players that make these multimillion-dollar salaries, while the majority of the team members only take home salaries in the high hundreds of thousands. Describe the shape of the distribution of salaries for such a team. Justify and be specific.

b. If the ownership of this team wanted to support their argument for why they are paying too much for "average" team salaries, which measure of center should they use? The mean or the median? Justify and be specific.
5. The Centers for Disease Control (CDC) conducts an annual survey of the general health of the US population. The CDC uses random dialing of phone numbers for US citizens over the age of 18. After permission is obtained a series of questions are asked and the data is recorded. Among these questions was the following:

1) Count the **number of days** during the previous month where your physical health not good because of stress or emotional problems?

a. Identify the variable of interest for this survey question as being categorical or quantitative.

b. Identify the parameter of interest \( (\pi, \mu, \rho, \bar{x}) \).

c. Identify the statistic of interest \( (\pi, \mu, \rho, \bar{x}) \).

6. The following graph represents the divorce rate per 1,000 women and the percent of female population in the work force for the last 9 decades.

![Graph](image)

a. Suppose that the correlation coefficient for this regression was calculated to be 0.924. What can you say about the relationship between percent of women in the work force and the divorce rate?

b. Calculate and interpret the coefficient of determination.

c. Can we say that an increase in percent of women in the work force causes an increase in the divorce rate?

d. Sketch a picture of the corresponding residual plot.

e. Suppose researchers wanted to know if percent of women in the work force is useful for predicting divorce rate what statistical information would you consult to determine this? Explain.
7. Can the maximum oxygen uptake (mm/kg), a measure often used by physiologists to indicate an individual's state of cardiovascular fitness, be used to predict the performance of distance runners? Six long-distance runners submitted treadmill tests for determination of their maximum oxygen uptake. The results along with the runner's best mile time (min), are shown below:

<table>
<thead>
<tr>
<th>Athlete</th>
<th>Maximum Oxygen Uptake (X)</th>
<th>Mile Time (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63.3</td>
<td>241.5</td>
</tr>
<tr>
<td>2</td>
<td>60.1</td>
<td>249.8</td>
</tr>
<tr>
<td>3</td>
<td>53.6</td>
<td>246.1</td>
</tr>
<tr>
<td>4</td>
<td>58.8</td>
<td>232.4</td>
</tr>
<tr>
<td>5</td>
<td>67.5</td>
<td>237.2</td>
</tr>
<tr>
<td>6</td>
<td>62.5</td>
<td>238.4</td>
</tr>
</tbody>
</table>

a. Assuming a linear model is appropriate, find the equation of the regression line.
b. Interpret the slope in the context of this example.
c. Estimate the mean and standard deviation of the mile time for a long-distance runner with maximum oxygen uptake of 65 mm/kg.
d. Determine the coefficient of determination and interpret it's meaning in the context of the setting.
e. Determine the correlation coefficient and interpret it's meaning in the context of the setting.
f. Describe how would you check to be sure that it is reasonable to fit a line to the data?
g. What is the response variable?
h. What is the explanatory variable?

8. In a study of acupuncture, patients with headaches are randomly divided into two groups. One group is given acupuncture and the other group is given aspirin. The acupuncturist evaluates the effectiveness of the acupuncture and compares it to the results from the aspirin group.

a. Is this an observational or experimental study?
b. Should this study be carried out as single blind, double blind, or neither? Explain who and why.
c. Could this study be biased in favor of acupuncture? Explain.
d. Explain how blocking by subject could be used to help reduce the bias in this study.

9. Childhood lead poisoning is a public health concern in the US. In fact, one child in eight has a high blood lead level (defined as 30 ug/dLi or more). Suppose we were to randomly choose a group of three US children.

a. Are these dependent or independent events? Explain.
b. What is the probability that none of the three children have a high lead level?
c. What is the probability that exactly one of the three children have a high lead level?
d. What is the probability that at least one of the three children have a high lead level?
IN ADDITION PLEASE STUDY THE FOLLOWING:

Your homework

Class notes

Assigned Reading

How to interpret Minitab output (stem and leaf, bar charts, pie charts, histograms, scatterplots, dotplots, boxplots, descriptive statistics, regression output, residual plots)

Topics:
- Characteristics of symmetric and skewed distributions (ie. mean, median, standard deviation)
- How to create and interpret graphs
- population vs. sample
- parameters vs. statistics
- appropriate notation
- The BIG 3: Shape, center, spread
- Empirical and Chebyshev’s Rule
- resistance
- Sampling techniques and designs
- Experimental vs. observational studies
- Confounding and extraneous variables
- Correlation and coefficient of determination
- Regression, equation of a line, fit of the model, making predictions
- Scatterplots and residual plots
- Regression outliers and influential observations
- Curvelinear models and transformations
- Probability, dependence vs. independence, empirical probability
ANSWER KEY

1. To study air quality on the Central Coast 20 air samples from various areas were obtained. For each one the carbon monoxide concentration was calculated.

<table>
<thead>
<tr>
<th>Concentration (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.4</td>
</tr>
<tr>
<td>12.5</td>
</tr>
<tr>
<td>10.4</td>
</tr>
<tr>
<td>10.1</td>
</tr>
<tr>
<td>11.1</td>
</tr>
<tr>
<td>10.5</td>
</tr>
<tr>
<td>8.5</td>
</tr>
<tr>
<td>11.2</td>
</tr>
<tr>
<td>9.9</td>
</tr>
<tr>
<td>12.0</td>
</tr>
<tr>
<td>13.2</td>
</tr>
<tr>
<td>9.3</td>
</tr>
<tr>
<td>9.8</td>
</tr>
<tr>
<td>9.7</td>
</tr>
<tr>
<td>11.3</td>
</tr>
<tr>
<td>11.0</td>
</tr>
<tr>
<td>8.7</td>
</tr>
<tr>
<td>10.3</td>
</tr>
<tr>
<td>11.1</td>
</tr>
<tr>
<td>9.5</td>
</tr>
</tbody>
</table>

\[
\sum x = 209.5 \quad \sum x^2 = 2222.73
\]

\[
\sum (x - \bar{x})^2 = 28.2175
\]

a. What is the population? Sample? Variable of interest?

Population - All locations on the Central Coast.
Sample - 20 air samples from various locations
Variable of interest - carbon monoxide concentration (ppm)

b. What is the mean and standard deviation of the carbon monoxide concentration?

\[
\bar{x} = \frac{209.5}{20} = 10.475
\]

\[
s^2 = \frac{28.2175}{20-1} = 1.485
\]

\[
s = \sqrt{1.485} = 1.219
\]

c. Make a stem and leaf display.

<table>
<thead>
<tr>
<th>Leaf Unit = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
</tbody>
</table>

d. Which of the following boxplots is the correct one for this data?

The correct boxplot is #2 because the median is at 10.35 and the observation 13.2 would not be considered an outlier in this data set.
2. Examine the histogram below (n = 46).

![Histogram](image)

a. Match the following descriptive statistics to their most likely value.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>11.78</td>
</tr>
<tr>
<td>median</td>
<td>42.00</td>
</tr>
<tr>
<td>SD</td>
<td>39.48</td>
</tr>
</tbody>
</table>

b. Describe the shape of this histogram.

Negatively skewed (to the Left).

c. Can you tell if the data for var1 is numerical or categorical?

Histograms are used for numerical data that has been grouped into classes (meaningful intervals). Not to mention that we have summary statistics for the mean, median and standard deviation which are all calculated for numerical data.

d. Can you tell if the data is discrete or continuous?

No we cannot tell from the histogram if the data is discrete or continuous. The bars range from 5-<10, 10-<15, 15-<20 etc….We do not know if the data inside the interval 5-10 is a 5,6,7,8, or a 9 (all would support discrete) not to mention that the data could be 5.1, 5.2, etc…(which would support continuous).

e. Why would a bar chart of var1 not give the same picture as the histogram? Explain.

A bar chart would not give the same picture as the histogram because the bar chart is used for categorical data. If we were to create a bar chart with this data it would give a bar for every distinct value in the data set. This would not be a meaningful display of the data, and therefore the histogram handles the data better.
3. A study was conducted to demonstrate a researcher’s theory that soy beans inoculated with nitrogen-fixing bacteria would yield more and grow adequately without the use of expensive synthesized fertilizers. The study was conducted under controlled conditions with uniform amounts of soil, on 30 inoculated soy bean plants. The plant yield as measured by pod weight (gm) for each plant are represented below in the stem and leaf display.

Stem-and-leaf of I         N  = 30
Leaf Unit = 0.010
   1  10 0
   1  11
   1  12
   1  13
   3  14 56
   8  15 3478
  11  16 455
  14  17 689
  (5) 18 56799
  11 19 1689
   7  20 0
   6  21 25
   4  22 02
   2  23 04

a. Find the minimum and maximum observation.
   minimum is 1.00 gm and maximum is 2.34 gm
   (notice the leaf unit is 0.01)

b. Which graph is the appropriate relative frequency histogram for this data?
   #2 - the units are correct and we have percent (relative frequency) on the y-axis

c. Suppose data were also collected for a group of 30 uninoculated soy bean plants grown under similar conditions. If the average pod weight for the uninoculated group was 1.084 gm, what is your overall impression concerning a comparison of the average pod weight between the two groups? Does this support the researcher’s theory? NOTE: you should not need to make any calculations to answer this problem.

   The average pod weight for the inocculated group is approximately 1.75 gm, this is higher than the average for the uninoculated group. Therefore, on average it appears that the pod weight will be higher with soy bean plants inoculated with nitrogen-fixing bacteria; this supports the researcher’s theory.

d. Is this study observational or experimental? Explain.

   This study is experimental because there was a planned intervention by inoculating or not inoculating the soy bean plants. Also the soy bean plants were manipulated to all grown under the same growing conditions.
4. The salary of professional athlete’s receives much attention in the media. It is becoming commonplace to hear of the multimillion-dollar contracts for the few select superstar athletes on each team. Because of this team owners and players’ associations spend much time with salary negotiations over additional salary and increased benefits.

a. Typically a professional sports team consists of one or two superstar players that make these multimillion-dollar salaries, while the majority of the team members only take home salaries in the high hundreds of thousands. Describe the shape of the distribution of salaries for such a team. Justify and **be specific**.

   The distribution would be positively skewed (to the right). The majority of the players will cluster in the high 100 thousands while the one or two superstars in the millions will pull the distribution of salaries slightly to the right.

b. If the ownership of this team wanted to support their argument for why they are paying too much for “average” team salaries, which measure of center should they use? The mean or the median? Justify and **be specific**.

   In a skewed right distribution the mean is larger than the median. If the owners want support for their argument that they are paying too much money they should use the mean (because it is larger). This way it will appear that the team is on "average" being paid more.

5. The Centers for Disease Control (CDC) conducts an annual survey of the general health of the US population. The CDC uses random dialing of phone numbers for US citizens over the age of 18. After permission is obtained a series of questions are asked and the data is recorded. Among these questions was the following:

1) Count the number of days during the previous month where your physical health not good because of stress or emotional problems?

   a. Identify the variable of interest for this survey question as being categorical or quantitative.

      Number of days - quantitative

   b. Identify the parameter of interest (π, μ, p or x).

      μ

   c. Identify the statistic of interest (π, μ, p or x).

      x

6. The following graph represents the divorce rate per 1,000 women and the percent of female population in the work force for the last 9 decades.
a. Suppose that the correlation coefficient for this regression was calculated to be 0.924. What can you say about the relationship between percent of women in the work force and the divorce rate?

We can say that there is a strong LINEAR relationship between divorce rate and women in the work force.

b. Calculate and interpret the coefficient of determination.

\[ R^2 = 0.924^2 = 0.854 \]

85.3% of the total variability in divorce rate can be explained by a linear regression with number of women in the workforce.

c. Can we say that an increase in percent of women in the work force causes an increase in the divorce rate?

No, just because there is an association between the two variables we would never say that an increasing number of women in the workforce cause more people to become divorced. There is probably some extraneous variability that was left uncontrolled for here. For example, there might actually be a time dependent relationship (ie. more women and more divorce simply because of time) that is actually driving the significance.

d. Sketch a picture of the corresponding residual plot.

![Residual Plot](image)

I did this by hand on the computer so it is literally just a sketch!

e. Suppose researchers wanted to know if percent of women in the work force is useful for predicting divorce rate what statistical information would you consult to determine this? Explain.

We could consult R² which at 85% would be indicates a strong linear relationship, as does the correlation. However we would want to caution researchers strongly to examine for potential confounding or extraneous variables that could be driving the relationship. One such variable might be time, in other words over time not only has the number of women in the workforce gone up, but also the divorce rate
7. Can the maximum oxygen uptake (mm/kg), a measure often used by physiologists to indicate an individual’s state of cardiovascular fitness, be used to predict the performance of distance runners? Six long-distance runners submitted treadmill tests for determination of their maximum oxygen uptake. The results along with the runner’s best mile time (min), are shown below:

<table>
<thead>
<tr>
<th>Athlete</th>
<th>Maximum Oxygen Uptake (X)</th>
<th>Mile Time (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63.3</td>
<td>241.5</td>
</tr>
<tr>
<td>2</td>
<td>60.1</td>
<td>249.8</td>
</tr>
<tr>
<td>3</td>
<td>53.6</td>
<td>246.1</td>
</tr>
<tr>
<td>4</td>
<td>58.8</td>
<td>232.4</td>
</tr>
<tr>
<td>5</td>
<td>67.5</td>
<td>237.2</td>
</tr>
<tr>
<td>6</td>
<td>62.5</td>
<td>238.4</td>
</tr>
</tbody>
</table>

\[ \bar{x} = 60.967 \]
\[ \bar{y} = 240.9 \]

a. Assuming a linear model is appropriate, find the equation of the regression line.

Find the slope first

\[ b = \frac{-54.21}{110.193} = -0.4920 \]

Find the y-intercept

\[ a = 240.9 \times (-0.4919550244)(60.967) = 270.893 \]

\[ \hat{y} = 270.893 - 0.492x \]

b. Interpret the slope in the context of this example.

For every 1 mm/kg increase of oxygen uptake, mile time decreases by 0.492 minutes. A more meaningful interpretation might be for every 10 mm/kg increase in oxygen uptake, mile time decreases by 4.92 minutes.

c. Estimate the mean and standard deviation of the mile time for a long-distance runner with maximum oxygen uptake of 65 mm/kg.

mean when \( x = 65 \) is : \( 270.893 - (0.492)(65) = 238.913 \)

\[ s_e = \sqrt{\frac{SS(resid)}{n-2}} = \sqrt{\frac{172.13}{4}} = 6.56 \]

d. Determine the coefficient of determination and interpret it’s meaning in the context of the setting.

\[ SS(total) = 26.67 + 172.13 = 198.8 \]
$r^2 = 1 - \frac{172.13}{198.8} = 0.134$

13.4% of the total variability in mile time can be explained by the linear relationship of mile time and maximum oxygen uptake. This is a poor amount of variability that can be explained.

e. Determine the correlation coefficient and interpret its meaning in the context of the setting.

$$r = \sqrt{r^2 \text{ (SignOfSlope)}} = \sqrt{0.134(-1)} = -0.366$$

There is virtually no linear relationship between mile time and maximum oxygen uptake. A correlation of –0.366 is an extremely weak negative linear relationship.

f. Describe how would you check to be sure that it is reasonable to fit a line to the data?

To check that a line reasonably fits the data we would create a residual plot. We would want to see a random pattern about the 0 line. If a curve pattern exists we should consider quadratic regression. If a funnel shape exists we have violated the constant variance assumption and we should consider a transformation.

NOTE: Also if outliers are present in either plot we would consider using a transformation.

g. What is the response variable?

The response is mile time

h. What is the explanatory variable?

The explanatory variable is maximum oxygen uptake.

8. In a study of acupuncture, patients with headaches are randomly divided into two groups. One group is given acupuncture and the other group is given aspirin. The acupuncturist evaluates the effectiveness of the acupuncture and compares it to the results from the aspirin group.

a. Is this an observational or experimental study?

This is an experimental study because the researcher intervenes and imposes treatment conditions (acupuncture vs. aspirin).

b. Should this study be carried out as single blind, double blind, or neither? Explain who and why.

Actually it seems that as described it would be impossible to do a blinded study at all. The subjects obviously know that they are getting acupuncture or an aspirin. The acupuncturist, who is also evaluating the outcome, also knows what treatment they got.

c. Could this study be biased in favor of acupuncture? Explain.

The acupuncturist expects acupuncture to work better than aspirin so she might “see” more improvement in someone given acupuncture than in someone given aspirin, even if the two groups are truly equivalent to each other in their response to the treatment.
d. Explain how blocking by subject could be used to help reduce the bias in this study.

If we blocked by subject this would mean that a person would get both aspirin and acupuncture. Not at the same time, rather, they would be randomized to one treatment, and then come back at a later time to get the other. This design could help in that people’s headaches could vary and by allowing the subject to experience both treatments this would control for extraneous variability due to severity of headache.

It should be noted that some bias could still exist because the acupuncturist is still not blinded. A way to solve this issue would be to ask an impartial person to observe the effectiveness and not tell them whether it was acupuncture or aspirin.

9. Childhood lead poisoning is a public health concern in the US. In fact, one child in eight has a high blood lead level (defined as 30 ug/dL or more). Suppose we were to randomly choose a group of three US children.

a. Are these dependent or independent events? Explain.

Assuming that none of the children are related, we can treat this as independent events. Knowing the probability that one child has a high lead level does not change the probability that the next child will have a high lead level.

b. What is the probability that none of the three children have a high lead level?

\[ P(\text{none}) = \left(\frac{7}{8}\right) \left(\frac{7}{8}\right) \left(\frac{7}{8}\right) = 0.6699 \]

c. What is the probability that exactly one of the three children have a high lead level?

\[ P(\text{exactly one}) = \left(\frac{1}{8}\right) \left(\frac{7}{8}\right) \left(\frac{7}{8}\right) + \left(\frac{7}{8}\right) \left(\frac{1}{8}\right) \left(\frac{7}{8}\right) + \left(\frac{7}{8}\right) \left(\frac{7}{8}\right) \left(\frac{1}{8}\right) = 0.2871 \]

d. What is the probability that at least one of the three children have a high lead level?

\[ P(\text{at least one}) = 1 - P(\text{none}) = 1 - 0.6699 = 0.3301 \]