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>

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

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

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. Compute a 90% CI for the true mean carbon monoxide concentration. INTERPRET in a meaningful way.

NOTE: make sure you could do parts b, c and d on Minitab. See study notes on Minitab page 5.

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

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>1.74</td>
</tr>
<tr>
<td>SE</td>
<td>39.48</td>
</tr>
</tbody>
</table>
b. Describe the shape of this histogram.

c. Can you justify the use of Student’s T for the calculation of a 95% confidence interval for $\mu$? Why or why not?

d. Suppose researchers would like to carry out a similar study, but would like the SE to be about 1.0 at the 95% confidence level. How many observations should they sample?

3. Suppose that the diameter of a population of Douglas fir trees in the Pacific Northwest is normally distributed with mean 50 cm and standard deviation 12 cm.

a. What percentage of trees in the population have a diameter more than 75 cm?

b. A certain tree’s diameter is in the 33rd percentile. What is the diameter?

c. If 25 trees were randomly sampled from this population, what is the probability that their average diameter is less than 45 cm?

4. Two dietary instruments that are often used in studies to assess consumption of specific foods are the food-frequency questionnaire (FFQ) and the dietary record (DR). The major difference between the two methods is that the FFQ collects records of certain foods eaten over a certain period of time and then nutrient intake is estimated, while the DR calculates nutrient intake exactly from a record of every food ingested. The FFQ less expensive to administer but is considered less accurate than the DR (the gold standard of dietary analyses). To validate the FFQ, 173 nurses participated in a study to record their food intake using both methods and total caloric intake (per day) was obtained for the two methods. Descriptive statistics for each method can be found below (calor_dr for the DR and calor_ffq for the FFQ). Suppose that the data do not come from a normal distribution.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>calor_dr</td>
<td>173</td>
<td>1619.9</td>
<td>1606.0</td>
<td>1613.3</td>
<td>323.4</td>
<td>24.6</td>
</tr>
<tr>
<td>calor_ffq</td>
<td>173</td>
<td>1371.7</td>
<td>1297.6</td>
<td>1344.3</td>
<td>482.1</td>
<td>36.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>calor_dr</td>
<td>910.0</td>
<td>2518.0</td>
<td>1410.0</td>
<td>1784.5</td>
</tr>
<tr>
<td>calor_ffq</td>
<td>463.2</td>
<td>3077.3</td>
<td>1031.2</td>
<td>1591.0</td>
</tr>
</tbody>
</table>

a. Which method seems to have a more reliable estimate of the true mean caloric intake? Justify using the output.

b. Calculate a 95% confidence interval for the true mean caloric intake for the FFQ method.

c. It was mentioned above that the data do not come from a normal distribution, would you still consider the confidence interval that you calculated above valid? Why or why not?

d. (TRUE or FALSE and say why) We are 95% confident that the average caloric intake of the 173 nurses using the FFQ method is between (interval calculated in part b).

e. Suppose you calculated a 90% confidence interval instead of the 95% confidence interval in part b. Would this new interval be wider or narrower? Why (be sure to include a mathematical explanation)?

f. Suppose that researchers wanted to carry out a similar study of average caloric intake, but they want to keep the SE to about 20. How many subjects should be sampled?
5. 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 33478
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 (below) 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.

NOTE: If you had the data could you create the plots on Minitab?

6. Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after
1 second of constant effort. Assume that in 45-54 year-old nonsmoking men FEV is normally distributed with
mean 4.0 liters and standard deviation of 0.7 liter. In comparably aged group of currently smoking men, FEV is
normally distributed with a mean of 3.2 liters and a standard deviation of 0.8 liter. Suppose researchers are
interested in a FEV of less than 2.5 liters because this level is regarded as showing some functional impairment
(occasional breathlessness, inability to climb stairs, etc…)

a. What is the probability that a randomly selected currently smoking man has a functional impairment.

b. What percentage of non-smoking men have a functional impairment.

c. In a random sample of 5 smoking men, what is the probability that the average FEV is less than 2.5 liters.

d. Find the value of FEV in non-smoking men that represents the 67th percentile.

7. 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.

8. 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 (p, µ, ̂ p or ̅ y ).

c. Identify the statistic of interest (p, µ, ̂ p or ̅ y ).
IN ADDITION PLEASE STUDY THE FOLLOWING:

Your homework

Class notes

How to interpret Minitab output (bar charts, histograms, dotplots, boxplots, descriptive statistics, generating random numbers, normal probability plots, confidence intervals for the mean and difference in means)

Topics:
- Characteristics of symmetric and skewed distributions (i.e. mean, median, standard deviation)
- How to create and interpret graphs
- Population vs. sample vs. sampling distribution
- Sampling distributions for the mean
- Parameters vs. statistics
- Appropriate notation
- The BIG 3: Shape, center, spread/Dispersion (variability)
- Empirical Rule
- Resistance
- SE vs. SD
- Central Limit Theorem (CLT)
- Probabilities or percentages from Ch 4 and 5 using standardization
- How to calculate confidence intervals
- Appropriate conclusions for confidence intervals
- Assumptions for confidence intervals to be valid
- Sample size calculations
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>10.5</td>
</tr>
<tr>
<td>13.2</td>
</tr>
<tr>
<td>11.0</td>
</tr>
<tr>
<td>10.4</td>
</tr>
<tr>
<td>9.9</td>
</tr>
<tr>
<td>9.8</td>
</tr>
<tr>
<td>11.1</td>
</tr>
</tbody>
</table>

\[ \sum y = 209.5 \quad \sum y^2 = 2222.73 \]
\[ \sum (y - \bar{y})^2 = 28.2175 \]

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?

\[ \bar{y} = \frac{209.5}{20} = 10.475 \]
\[ s^2 = \frac{2222.73 - \left(\frac{209.5}{20}\right)^2}{20 - 1} = 1.485 \quad \text{OR} \quad s^2 = \frac{28.2175}{20 - 1} = 1.485 \]
\[ s = \sqrt{1.485} = 1.219 \]

c. Make a stem and leaf display.

Leaf Unit = 0.10

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>57</td>
</tr>
<tr>
<td>9</td>
<td>345789</td>
</tr>
<tr>
<td>10</td>
<td>1345</td>
</tr>
<tr>
<td>11</td>
<td>01123</td>
</tr>
<tr>
<td>12</td>
<td>05</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
</tr>
</tbody>
</table>
d. Compute a 90% CI for the true mean carbon monoxide concentration. INTERPRET in a meaningful way.

\[
\bar{y} \pm t_{0.05}(\frac{\sigma}{\sqrt{n}}) = 10.475 \pm 1.729\left(\frac{1.219}{\sqrt{20}}\right) = 10.475 \pm 0.471 = (10.004, 10.946)
\]

We are highly confident, at the 0.10 level, that the true mean carbon monoxide concentration for the Central Coast is between 10.004 and 10.946 ppm.

NOTE: make sure you could do parts b, c and d on Minitab. See study notes on Minitab page 5.

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

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

   - mean: 11.78
   - median: 42.00
   - SD: 1.74
   - SE: 39.48

b. Describe the shape of this histogram.

   Skewed to the Left.

c. Can you justify the use of Student’s T for the calculation of a 95% confidence interval for \( \mu \)? Why or why not?

   Yes, because the sample size is 46. The CLT says even if the distribution of Y is not normal, the sampling distribution of \( \bar{y} \) will be approximately normal if the sample size is large.

d. Suppose researchers would like to carry out a similar study, but would like the SE to be about 1.0 at the 95% confidence level. How many observations should they sample?

\[
1 = 11.78/\sqrt{n}
\]

\[
\sqrt{n} = 11.78
\]

\[
n = 138.8 \approx 139 \text{ observations}
\]
3. Suppose that the diameter of a population of Douglas fir trees in the Pacific Northwest is normally distributed with mean 50 cm and standard deviation 12 cm.

a. What percentage of trees in the population have a diameter more than 75 cm?

\[ Y \geq 75 = Z \geq \frac{75 - 50}{12} = Z \geq 2.08 \]

Look up 2.08 on Z table, 0.9812, so the percent more than 75 cm is 1 - 0.9812 = 0.0188

b. A certain tree’s diameter is in the 33rd percentile. What is the diameter?

Look inside the z table for 0.33. The z score associated with the 33rd percentile is -0.44, so now solve in terms of Y.

\[ -0.44 = \frac{Y - 50}{12} \]

\[ Y = 44.72 \text{ cm} \]

c. If 25 trees were randomly sampled from this population, what is the probability that their average diameter is less than 45 cm?

\[ P(\bar{Y} \leq 45) = \left( \frac{Y - 45}{12/\sqrt{25}} \right) = P(Z \leq -2.08) = 0.0188 \]

4. Two dietary instruments that are often used in studies to assess consumption of specific foods are the food-frequency questionnaire (FFQ) and the dietary record (DR). The major difference between the two methods is that the FFQ collects records of certain foods eaten over a certain period of time and then nutrient intake is estimated, while the DR calculates nutrient intake exactly from a record of every food ingested. The FFQ less expensive to administer but is considered less accurate than the DR (the gold standard of dietary analyses). To validate the FFQ, 173 nurses participated in a study to record their food intake using both methods and total caloric intake (per day) was obtained for the two methods. Descriptive statistics for each method can be found below (calor_dr for the DR and calor_ffq for the FFQ). Suppose that the data do not come from a normal distribution.

**Descriptive Statistics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>calor_dr</td>
<td>173</td>
<td>1619.9</td>
<td>1606.0</td>
<td>1613.3</td>
<td>323.4</td>
<td>24.6</td>
</tr>
<tr>
<td>calr_ffq</td>
<td>173</td>
<td>1371.7</td>
<td>1297.6</td>
<td>1344.3</td>
<td>482.1</td>
<td>36.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>calor_dr</td>
<td>910.0</td>
<td>2518.0</td>
<td>1410.0</td>
<td>1784.5</td>
</tr>
<tr>
<td>calr_ffq</td>
<td>463.2</td>
<td>3077.3</td>
<td>1031.2</td>
<td>1591.0</td>
</tr>
</tbody>
</table>

a. Which method seems to have a more reliable estimate of the true mean caloric intake? Justify using the output.

The DR method has a more reliable estimate of the mean because it’s SE (24.6) is smaller than the SE (36.6) for the FFQ method.
b. Calculate a 95% confidence interval for the true mean caloric intake for the FFQ method.

\[
\bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 1371.7 \pm t(172)_{0.05}\left(\frac{482.1}{\sqrt{173}}\right) = 1371.7 \pm \left(1.977\right)(36.653) = (1299.24, 1444.16)
\]

c. It was mentioned above that the data do not come from a normal distribution, would you still consider the confidence interval that you calculated above valid? Why or why not?

Yes the interval above would still be considered valid because the sample size is large (n = 173). The central limit theorem guarantees that when n is large the sampling distribution of \(\bar{y}\) will be approximately normal regardless of the underlying distribution.

d. (TRUE or FALSE and say why) We are 95% confident that the average caloric intake of the 173 nurses using the FFQ method is between (interval calculated in part b).

FALSE, we know that the sample average caloric intake is between the interval calculated above.

e. Suppose you calculated a 90% confidence interval instead of the 95% confidence interval in part b. Would this new interval be wider or narrower? Why (be sure to include a mathematical explanation)?

A 90% confidence interval would be narrower than a 95% confidence interval because as the confidence goes down the interval becomes smaller. This is a result of the t multiplier becoming smaller with decreased confidence, so we would be adding and subtracting a smaller number to our estimate.

f. Suppose that researchers wanted to carry out a similar study of average caloric intake, but they want to keep the SE to about 20. How many subjects should be sampled?

Using the sd from the current as a guess.

\[
DesiredSE \geq \frac{GuessSD}{\sqrt{n}}
\]

\[
20 \geq \frac{482.1}{\sqrt{n}}
\]

\[
581.05 \approx 582\text{ Subjects}
\]

5. 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 33478
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 (below) 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 inoculated 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 researchers theory.

NOTE: If you had the data could you create the plots on Minitab?
6. Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. Assume that in 45-54 year-old nonsmoking men FEV is normally distributed with mean 4.0 liters and standard deviation of 0.7 liter. In comparably aged group of currently smoking men, FEV is normally distributed with a mean of 3.2 liters and a standard deviation of 0.8 liter. Suppose researchers are interested in a FEV of less than 2.5 liters because this level is regarded as showing some functional impairment (occasional breathlessness, inability to climb stairs, etc…)

a. What is the probability that a randomly selected currently smoking man has a functional impairment.

\[ P(Y < 2.5) = P\left(Z < \frac{2.5 - 3.2}{0.8}\right) = P(Z < -0.88) = 0.1894 \]

b. What percentage of non-smoking men have a functional impairment.

\[ Y < 2.5 = Z < \frac{2.5 - 4}{0.7} = Z < -2.14 \]

\[ \mu = 2.5 \text{ liters}, \sigma = 0.7 \text{ liters} \]

\[ P(Y < 2.5) = P(Z < -2.14) = 0.0162 \]

1.62%

c. In a random sample of 5 smoking men, what is the probability that the average FEV is less than 2.5 liters.

\[ P(\bar{Y} < 2.5) = P\left(Z < \frac{2.5 - 3.2}{0.8/\sqrt{5}}\right) = P(Z < -1.96) = 0.025 \]

d. Find the value of FEV in non-smoking men that represents the 67th percentile.

The 67th percentile (0.67 below and 0.33 above) corresponds to a z score of 0.44

\[ 0.44 = \frac{Y - 4}{0.7} \]

\[ Y = 4.308 \text{ Liters} \]

7. 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.

b. 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 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.

c. 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.

9. 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 (p, \( \mu \), \( \hat{p} \) or \( \bar{y} \)).

\( \mu \)

c. Identify the statistic of interest (p, \( \mu \), \( \hat{p} \) or \( \bar{y} \)).

\( \bar{y} \)