Short Answer

Concepts to Understand: Ethics in Marketing Research

a. Briefly describe one important moral philosophy. There are several discussed in your text, you can choose any one of them.

b. What is the purpose of an Institutional Review Board? How does the process work? The IRB has the institution that has the authority to approve and disapprove specific research proposals, with a particular focus on the ethics of using people as subject in research studies, to make sure, for example, that participants in a research project are assured anonymity. The IRB then monitors the research that is conducted.

c. What is informed consent and why is it important? Informed consent means that each participant in the study is informed as to the any potential risks and/or benefits of their participation, and, given this understanding, agree to participate in the study.

d. What is a right to confidentiality and why is it important? The right to confidentiality is the right to privacy. The responses and behavior of the participant during the study are not linked with his or her specific identity.

e. What is the process of debriefing? What is its purpose? The debriefing process follows the subject’s participation in the research project. The purpose is to have the participants understand what occurred during the research, as well as for the researcher to understand the participants’ experience. Any questions the participant has are answered at this time.

Concepts to Understand: Data Analysis

1. What are deviation and squared deviations? Statistics is the study of variability, of how data values vary. How do we assess variability? The key concept to assess variability is the deviation of each data value from its own mean, that is, how far away is each value (score) from its mean? To compute the standard deviation we square these deviations.

2. Define “standard deviation” with words (not a formula). Include the concept of a deviation score in the discussion.

Standard deviation: Square root of the average squared deviation score, defining average on
the basis of the degrees of freedom for the sample standard deviation calculated using the sample mean.

3. **How is the standard deviation an indicator of variability?**

The standard deviation is based on the average squared deviation score. The more variability among the data values, the larger the deviations from the mean, the larger the squared deviations, the larger the standard deviation.

4. **What is the range of the distribution of a normally distributed variable?**

About 95% of the values of a normal distribution are within two standard deviations of the population mean. The exact number is 1.96 standard deviations. That is a defining characteristic of all normal distributions.

---

**Notation you need to know**

5. **What is n?**

The sample size, the number of different people (or whatever) in the sample.

6. **What is m?**

The mean of the variable of interest for a sample of data values for the variable. That is, the sample mean.

7. **What is s?**

The standard deviation of the variable of interest for a sample of data values for the variable. That is, the sample standard deviation.

8. **Why do we need statistical inference? That is, why is information regarding characteristics of the population, such as the population mean, $\mu$, more useful than the corresponding information that describes a sample from that population, such as the sample mean, $m$?**

Management decisions are made regarding the future, an implicit or explicit forecast as to values of variables as they will exist in the future. What projects into the future are the stable population values that underlie the sampling process by which past data were collected. Sample statistics are influenced by the population values, but also by inherently unstable sampling error. Better forecasts are made from the underlying population values.

9. **How can the sample mean, $m$, be conceptualized as a variable? What does it mean to...**
say that the value of $m$ varies?

We typically do not observe any variation in $m$, we typically only have 1 sample and so have only one sample mean. Hence the variation of $m$ is an abstraction of what *would* happen over repeated sampling, something that *would* be true if many samples were taken and the sample mean computed for each sample.

10. **What does the distribution of the sample mean over repeated samples of the same size $n$ from the same population look like?**

For a reasonably large sample size, larger than 20 or 30, the sample mean is normally distributed over many samples (though in practice only one sample is observed).

11. **What is the law of large numbers?**

The Law of Large Numbers states that, for all random samples from the same population of the same size the, larger the sample size, on average, the better the estimate of the corresponding population value.

12. **What is the purpose of a confidence interval? Give an example of what one would look like?**

The confidence interval estimates the value of a population value, such as the population mean, by providing an interval that likely contains the unknown population value centered over the sample mean. For example, we are 95% confident that the true proportion of people voting for Candidate A is between 46.3 and 49.8%.

13. **What is the standard error and its relation to statistical inference?**

The standard error is a specific type of standard deviation. A *standard error is the standard deviation of a statistic*, usually over hypothetical random samples. For example, calculate the mean for the first sample, then for the second sample, etc. The less the variability of the statistic over (usually hypothetical random samples), the better the estimate tends to be, on average, of the underlying population value. If the estimates from different samples vary all over the place, then there is little confidence that any one of them would be close to the true population value.

Remember that mathematical derivations show us what happens over repeated samples without actually having to have the multiple samples. We know that the estimated standard deviation of the sample over repeated samples the standard error, $s_y$, is the standard deviation of the data divided by the square root of the sample size, $\frac{s}{\sqrt{n}}$. That is why they repeated samples are described as hypothetical. The standard error is the standard deviation that we can compute without having to have the multiple samples. It represents the
variability of what would happen if we did have multiple samples.

14. **What is the advantage to analyze a line (run) chart of a variable with values ordered sequentially before computing the confidence interval?**

To provide a meaningful estimate of a population mean, all the data values in the sample need to come from the same population. If the data values follow a trend up or down over time, for example, then the corresponding population means may be moving up or down, in which case there are multiple population means to estimate from different populations. If the data follow a trend over time, the estimate of a single mean is meaningless because there is no single mean to estimate. IF the data values are time ordered, then they should show no systematic trend or other pattern over time. That is, they should indicate random variation about the sample mean, the pattern of data values all sampled from the same population.

15. **When is the t-distribution used in place of the distribution of the z-distribution? How often is this usually done in practice?**

We are interested in the z-distribution because the sample means over a (usually hypothetical) distribution are normally distributed, and for a normal distribution the probabilities for given z-values are easily obtained. The problem is that the standard deviation of this distribution of the sample means, its standard error, is unknown.

Use the t-distribution when the population standard deviation is unknown. That is, use the estimated standard error of the mean from the data, \( s_m \), in place of the true, population value of the standard error, \( \sigma_m \). This is the condition in virtually every single data analysis ever done. Because the standard error is an estimate, there is a price to pay, which is that a t-distribution is wider, with more variability than the corresponding normal distribution.

16. **How does the family of t-distributions resemble the normal distribution?**

As the sample size grows, the corresponding t-distribution becomes more and more similar to the normal distribution, though all t-distributions have the characteristic “bell curve” shape. For example, the 0.025 cutoff gets closer and closer to the normal curve value of 1.96 as the \( df = n - 1 \) increases. For \( df = 15 \), the 0.025 cutoff is 2.13, but for \( df = 200 \), the cutoff is 1.97.

---

**Notation you need to know**

17. **What is \( \mu \)?**
The mean of the variable of interest over the entire population.

18. What is $\sigma$?
The standard deviation of the variable of interest over the entire population.

19. What is $s_m$?
The estimated standard error of the mean, which is standard deviation of the sample mean for variable Y, $m$, over usually hypothetical multiple samples. In practice estimate the standard error from the one sample of data.

20. What is $\sigma_m$?
The actual standard error of the mean, which is standard deviation of all possible population means for variable Y, $m$, over all possible samples of the same size.

---

**Worked Problems**

**Track C**

For worked Problem #1, analyze responses to the specified item, Learn_1 according to the web address below for the location of the data and of the questions from Qualtrics (what you will use for your project).

I want to learn more about how to use the computer to perform statistical analyses of marketing research data.

a. What are the possible responses to the item from the perspective of the respondent?

b. How are the items scored for data analysis, that is what data values are in the file?
   What is coding key to translate how the items are scored for analysis vs how the respondent answers the item?
   According to the pdf, which is an excerpt of the item as presented to the respondent’s in Qualtrics, the possible responses to the item:

   Strongly Disagree
   Disagree
   Somewhat Disagree
   Neither Agree nor Disagree
   Somewhat Agree
   Agree
   Strongly Agree
Qualtrics codes the responses in the order in which they are presented, which is from 1 (Strongly Disagree) to 7 (Strongly Agree).

c. Do a bar chart of Learn_1.

```
> mydata <- Read("http://web.pdx.edu/~gerbing/460/Resources/460S14.csv")
> BarChart(Learn_1)
```

![Bar Chart](image)

**Figure 1:** I want to learn more of how to apply statistical techniques to marketing research.

d. How many responses are there to Learn_2?

<table>
<thead>
<tr>
<th>--- Learn_2 ---</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Frequencies:</td>
</tr>
<tr>
<td>Proportions:</td>
</tr>
</tbody>
</table>

There is a total of 31 responses to the item.

e. Interpret the results.

The most popular response categories were 6 and 7, that is, many students, over 65% agreed or strongly agreed of wanting to learn more about how to apply statistics to marketing research.

**Track ABC**

1. The item with the variable name **Learn2.1** on the class survey from the previous homework is completed as follows, answered on a 100 point scale, from Disagree to Agree.

   I want to learn more of how to apply statistical techniques to marketing research.
The downloaded csv file of student responses to the class survey are available on the web.  
http://web.pdx.edu/~gerbing/data/460S14.csv

First, read the data:

```r
> mydata <- Read("http://web.pdx.edu/~gerbing/data/460S14.csv")
```

From the `Read` statement, the data file `460S14.csv` contains the following variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Missing</th>
<th>Unique</th>
<th>Values</th>
<th>Values</th>
<th>Values</th>
<th>First and last values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn_1</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>3 ... 7 7 5</td>
</tr>
<tr>
<td>Learn_2</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3 ... 6 7 6</td>
</tr>
<tr>
<td>Learn_3</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>7 ... 7 7 2</td>
</tr>
<tr>
<td>Learn_4</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>3 ... 6 7 6</td>
</tr>
<tr>
<td>Feel_1</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>5 ... 6 3 4</td>
</tr>
<tr>
<td>Feel_2</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>3 ... 6 7 6</td>
</tr>
<tr>
<td>Feel_3</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>4 ... 5 6 5</td>
</tr>
<tr>
<td>Feel_4</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>1 ... 5 7 5</td>
</tr>
<tr>
<td>Past_1</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4 ... 2 2 2</td>
</tr>
<tr>
<td>Past_2</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2 ... 2 2 2</td>
</tr>
<tr>
<td>Past_3</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3 ... 2 3 3</td>
</tr>
<tr>
<td>Past_4</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2 ... 2 3 1</td>
</tr>
<tr>
<td>Past2_1</td>
<td>integer</td>
<td>27</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>7 ... NA 5 7</td>
</tr>
<tr>
<td>Gender</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2 ... 1 1 2</td>
</tr>
<tr>
<td>Class</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2 ... 1 1 1</td>
</tr>
<tr>
<td>Learn2_1</td>
<td>integer</td>
<td>30</td>
<td>1</td>
<td>18</td>
<td>59</td>
<td>30</td>
<td>NA ... 80 77 50</td>
</tr>
<tr>
<td>Learn2_2</td>
<td>integer</td>
<td>30</td>
<td>1</td>
<td>15</td>
<td>78</td>
<td>50</td>
<td>NA ... 65 78 60</td>
</tr>
<tr>
<td>Learn2_3</td>
<td>integer</td>
<td>31</td>
<td>0</td>
<td>15</td>
<td>95</td>
<td>60</td>
<td>100 ... 100 89 29</td>
</tr>
<tr>
<td>Learn2_4</td>
<td>integer</td>
<td>30</td>
<td>1</td>
<td>18</td>
<td>53</td>
<td>50</td>
<td>NA ... 87 85 85</td>
</tr>
</tbody>
</table>

Here are the summary statistics from either the `Histogram` or `SummaryStats` function for the variable `Learn2_1`.

```r
> Histogram(Learn2_1)
--- Learn2_1 ---

n  miss  mean   sd  min  mdn  max
30  1  64.63 19.02  5.00 63.50 100.00
```

a. What are the possible responses to the item from the perspective of the respondent?

The respondent responds to the item on a scale from 0 to 100, where 0 is maximum disagreement and 100 is maximum agreement, and 50 is neutral.
b. How are the items scored for data analysis, that is what data values are in the file?
   The data values are the integers from 0 to 100.

c. Specify the sample size, \( n \), the number of respondents. A total of 30 people responded to this item.

d. Specify the value of \( s \).
   The standard deviation: \( s = 19.02 \)

e. How many students had a score above 80 in terms of agreeing with the item?
   Include the full range of possible values.

   > style("sienna")
   > Histogram(Learn2_1, bin.start=0)

   ![Histogram](image)

   **Figure 2:** I want to learn more of how to apply statistical techniques to marketing research.

   From reading the histogram, 3 people had scores between 80 and 90, and one person had a score above 90. The frequency table confirms this analysis.
A total of $3 + 1 = 4$ students expressed strong agreement with the Learn2.1, with a response of greater than 80.

f. What is the deviation score for the person listed in the first row of data? What does this mean?

To get the deviation score, from the mean, we need the data value of Learn2.1. You could go right to the web and view the data file. Or, after reading into R, this data value can be obtained from the output of `read` which lists the first few data values for each variable. From the previous output, the first data value is 59 for this variable. The sample mean of this variable from the previous table of summary statistics is $m = 64.63$. The deviation score for this person:

$$\text{deviation} = Y_i - m = 59 - 64.63 = -5.63$$

That is, the first person scored below the mean. He or she was less enthusiastic than average for learning data analysis.

Also note we can use the standard R function `head` to get the first 6 lines of data and the variable names, as in `head(mydata)`.

```
> head(mydata)
     Learn_1 Learn_2 Learn_3 Learn_4 Feel_1 Feel_2 Feel_3 Feel_4 Past_1 Past_2
1      3      5      7      6      6      7      5      7      3      4
2      4      4      3      6      2      6      6      6      4      4
3      3      3      7      3      5      3      4      1      4      2
4      4      4      5      2      6      4      5      2      2      2
5      6      6      5      3      6      4      3      2      3      3
6      7      7      2      7      5      7      7      3      1
```

g. Calculate $s_m$. 

MKTG 460 Solutions #2
\[ s_m = \frac{s}{\sqrt{n}} = \frac{19.02}{\sqrt{30}} = 3.47 \]

h. Compare the sizes of \( s \) and \( s_m \). Which one is smaller and why?
\[ s = 19.46 < s_m = 3.47, \text{ which is always the case because the sample mean, } m \text{ is less variable than the values of } Y, \text{ such as the sample data values.} \]

i. Present the 95% confidence interval from a computer analysis, such as R with the lessR function \texttt{tt.brief}.

```r
> tt.brief(Learn2_1)

Learn2_1: n.miss = 1, n = 30, mean = 64.63, sd = 19.02

t-cutoff: tcut = 2.045
Standard Error of Mean: SE = 3.47

Margin of Error for 95% Confidence Level: 7.10
95% Confidence Interval for Mean: 57.53 to 71.73
```

j. Most important type of question in the entire homework:

**Interpret this confidence interval.**

The item for which the responses are analyzed:

\[ \text{I want to learn more of how to apply statistical techniques to marketing research.} \]

With 95\% confidence, the true population mean attitude of PSU Marketing Research 460 students for learning data analysis applied to marketing research, as assessed on a 100 point scale from maximum disagreement (0) to maximum agreement (100), is somewhere between 57.5 to 71.7. There is a general level of moderate agreement with a desire to learn how to do data analysis regarding marketing research.

2. *On the class survey taken at the beginning of the term the following item appeared:*

   *I am only taking this class because it is required.*

The issue of interest is if, on average, marketing majors take Marketing Research because it is a required class. Answer this question by completing the following questions. The variable name is \texttt{Learn2_3}.

The downloaded \texttt{csv} file of your responses to the class survey are available.

\[ \text{http://web.pdx.edu/~gerbing/460/Resources/460S14.csv} \]
a. Generate the histogram of the data. Interpret.

```
> Histogram(Learn2_3)
```

![Histogram](image)

**Figure 3:** I am only taking this class because it is required.

The majority of students in that section of MKTG 460 are taking the class because it is a required class for graduation. Only 3 students strongly disagree with the statement, providing scale scores from 0 to 20 on a 100 point scale. Many students, however, responded in the 40-60 range, indicating that they are neutral regarding taking the class only because it is required. Many students, however, 8, strongly agree, answering in the 80 to 100 range.

b. Given the shape of the distribution of the data, how useful is knowing the mean of this distribution to understanding the needs and motivations of the students in this class?

The mean is useful to the extent that the distribution is symmetric about the middle and not too variable. Knowing the mean of 64.67 is useful to some extent because it shows, on average, students tend to be taking the class because it is required. However there is so much variability, that the mean by itself does not provide a sufficient summary. Some students, a minority, want to take the class, and more do not want to be in the class at all.

c. Present the 95% confidence interval from a computer analysis, such as R with the lessR function `tt.brief`.

MKTG 460 Solutions #2
d. Show the calculation for the margin of error for the confidence interval of the mean.
Margin of Error is 2.045 standard errors: \( E = (t_{0.025})(s_m) = 2.042(4.65) = 9.51 \)

e. Show the calculation for the 95% confidence interval of the mean.
Lower Bound: \( m - E = 64.71 - 9.51 = 55.20 \)
Upper Bound: \( m + E = 64.71 + 9.51 = 74.22 \)

f. Most important type of question in the entire homework:

**Interpret this confidence interval.**

The item for which the responses are analyzed:

*I am only taking this class because it is required.*

With 95% confidence, the true population mean attitude of PSU Marketing Research 460 students for learning data analysis applied to marketing research, as assessed on a 100 point scale from maximum disagreement (0) to maximum agreement (100), is somewhere between 55.20 to 74.22. There is a general level of moderate agreement with a desire to learn how to do data analysis regarding marketing research.