Question 1: Focus on Exploring Data — 4 points

General Scoring Notes

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

	Model Solution	Scoring
A	The distribution of gas mileage for the sample of cars manufactured in Country A has a lower center than the distribution of gas mileage for the sample of cars manufactured in Country B. The median gas mileage for the sample of cars manufactured in Country A (18 mpg) is less than the median gas mileage for the sample of cars manufactured in Country B (32 mpg). The range of the gas mileages for the sample of cars manufactured in Country A (24 mpg) is slightly greater than the range of the gas mileages for the sample of cars manufactured in Country B (22 mpg). However, the IQR of the gas mileages for the sample of cars manufactured in Country A (8 mpg) is less than the IQR of the gas mileages for the sample of the cars manufactured in Country B (12 mpg). The car manufactured in Country A with 38 mpg (the maximum of the sample of cars manufactured in Country A) is an outlier, while the distribution of gas mileage for the sample of cars manufactured in Country B has no outliers.	**Essentially correct (E)** if the response satisfies at least three of the following four components: 1. Directly compares the center for the two distributions 2. Directly compares the spread (either IQR or range) for the two distributions 3. Indicates that the gas mileage of one of the cars manufactured in Country A is an outlier 4. Provides sufficient context, which includes the manufacturing countries ("Country A" and "Country B") AND the dependent variable ("gas mileage" or "mpg") **Partially correct (P)** if the response satisfies only two of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- The response need not include specific numerical values to satisfy any given component or to score E on part A.
- A response that only uses the "means" in the comparison of center would not satisfy component 1.
- Any acceptable mention of shape in the response should be ignored because complete shape information cannot be determined from a boxplot. Acceptable mentions of shape include:
  - The shape of the distribution of the gas mileage for cars manufactured in Country A can be described as skewed, positively skewed, or right skewed.
  - The shape of the distribution of the gas mileage for cars manufactured in Country B can be

- If the response describes the shape of either distribution as just "symmetric," "normal," "unimodal," or an incorrect shape (e.g., "the distribution of gas mileages for Country A is left skewed" or "the distribution of gas mileages for Country B is right skewed"), then part A cannot be scored E.
- A response that only lists values for center and/or spread and does not directly compare them does not satisfy components 1 and/or 2.
- A response that just refers to "A" or "B" AND the dependent variable ("gas mileage" or "mpg") may satisfy component 4.

	Model Solution	Scoring
B	The mean of the distribution of gas mileage for the sample of cars manufactured in Country A is expected to be greater than 18 mpg, the median of the distribution. Because the distribution of gas mileage for the sample of cars manufactured in Country A has an outlier to the right (or is skewed to the right), the mean of the distribution (which is not resistant) is expected to be pulled above the median (which is resistant) toward the higher values of gas mileage.	**Essentially correct (E)** if the response satisfies the following three components: 1. Indicates that the mean of the distribution of gas mileage for the sample of cars manufactured in Country A is expected to be greater than 18 mpg 2. Indicates 18 as the value of the median 3. Provides a reasonable justification for the relationship between the mean and the median based on a right-skewed distribution or the expectation that an outlier in the right tail will pull the mean above the median **Partially correct (P)** if the response satisfies component 1 AND either component 2 or component 3. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- A response that states the median is 18 mpg in either part A or part B satisfies component 2.
- A response that provides a weak justification (e.g., "because the distribution is skewed") does not satisfy component 3. However, a response that provides a justification based on the distribution being right skewed may satisfy component 3.
- Component 3 may be satisfied with an appropriate numerical argument that shows that the mean must be greater than 18 mpg.
  - An example of an inappropriate numerical argument would be averaging portions of the five-number summary.

Model Solution	Scoring
C i. The maximum value in the combined data is 40 mpg because 40 mpg is the maximum gas mileage for the sample of cars manufactured in Country B, and as shown in the boxplot, all the gas mileages for the sample of cars manufactured in Country A are less than 40 mpg. The minimum value in the combined data is 14 mpg, because 14 mpg is the minimum mpg for the sample of cars manufactured in Country A, and as shown in the boxplot, all the gas mileages for the sample of cars manufactured in Country B are greater than 14. Thus, the range of the combined data set is $40 - 14 = 26$ mpg. ii. In the combined data, there are 200 gas mileages. The median is a value where at least half, or 100, of the gas mileages in the combined data are less than or equal to the median value and at least half, or 100, of the gas mileages in the combined data are greater than or equal to the median value. From the boxplot for the sample of cars manufactured in Country A, the third quartile, Q3, is 24 mpg indicating there are at least 75 gas mileages less than or equal to 24 mpg and at least 25 gas mileages greater than or equal to 24 mpg. From the boxplot for the sample of cars manufactured in Country B, the first quartile, Q1, is 24 mpg indicating there are at least 25 gas mileages less than or equal to 24 mpg and at least 75 gas mileages greater than or equal to 24 mpg. Thus, in the combined data set, there are at least 100 gas mileages less than or equal to 24 mpg and at least 100 gas mileages greater than or equal to 24 mpg, which implies 24 is the value of the median of the combined data set.	**Essentially correct (E)** if the response satisfies the following four components: 1. Correctly calculates 26 mpg as the range of the combined data set 2. Provides a justification for the range with an argument based on identifying 40 as the maximum of the combined data set and identifying 14 as the minimum of the combined data set 3. Calculates 24 mpg as a possible value of the median of the combined data set 4. Provides a justification for the median by indicating that at least half of the combined data values are at most 24 and at least half of the combined data values are at least 24 **Partially correct (P)** if the response satisfies only two or three of the four components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- A response that solely indicates a range of values (e.g., “the gas mileages range from 14 mpg to 40 mpg”) does not satisfy component 1 but may satisfy component 2.
- A response that states the range as a value of 25 or 27 from misinterpreting the minimum value in the distribution of gas mileages for Country A as 13 or 15 may be considered a minor error and satisfies components 1 and 2.

- A response that gives the correct value for the median (24) and justifies the value by looking at the quartiles on one side of 24 (e.g., identifying that one-fourth of Country B cars and three-fourths of Country A cars have mpg less than or equal to 24) satisfies component 4.
- A response that provides justification of the median based on a value near the 100th data value satisfies component 4.
- A response that provides justification of the median based on counting the segments in the boxplots may satisfy component 4. The response is not required to specify equal sample sizes.
    - For example, one-half of the boxplot segments are no greater than 24 mpg (four of eight segments—three segments from Country A and one segment from Country B) and one-half of the boxplot segments are at least 24 mpg (four of eight segments—one segment from Country A and three segments from Country B), which indicates that 24 mpg is the median of the combined data set.

Scoring for Question 1	Score
Complete Response Three parts essentially correct	**4**
Substantial Response Two parts essentially correct and one part partially correct	**3**
Developing Response Two parts essentially correct and no part partially correct *OR* One part essentially correct and one or two parts partially correct *OR* Three parts partially correct	**2**
Minimal Response One part essentially correct and no parts partially correct *OR* No part essentially correct and two parts partially correct	**1**

Question 6: Investigative Task — 4 points

General Scoring Notes

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Model Solution	Scoring
(a) (i) The inference procedure that should be used to estimate the mean price, in dollars ($), of this type of whistle at all stores that sell the whistle is a one-sample *t*-interval for a population mean. (ii) The parameter of interest is the mean whistle price, in dollars ($), of this type of whistle at all stores that sell the whistle.	**Essentially correct (E)** if the response satisfies the following three components: 1. In part (a-i) the response identifies a one-sample *t*-interval for a population mean by name (e.g., "one sample *t*-interval") 2. In part (a-ii) the response identifies the correct parameter by including reference to the mean 3. In part (a-ii) the response provides sufficient context for the parameter by including reference to the population (all stores that sell the whistle) and the variable of interest (the price of this type of whistle) **Partially correct (P)** if the response only satisfies one or two of the three components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- In part (a-i) a response that refers to a "test" does not satisfy component 1.
- In part (a-i) a response that refers to "a mean," singular, satisfies the one-sample aspect of component 1. If the response states "one-sample," it is not required to also include "for a mean" to satisfy component 1.
- A response that addresses "mean" in part (a-i) does not satisfy component 2 unless it is also addressed in part (a-ii).
- In part (a-ii) a response may satisfy the variable of interest aspect of component 3 with a minimum of "price" and "whistle."
- In part (a-ii) a response may satisfy the population aspect of component 3 by using words such as: "population," "all," or "true," or by using the symbol μ.
- In part (a-ii) if the response clearly refers to the sample mean instead of the population mean using words or symbols (e.g., $\bar{x}$), then component 3 is not satisfied unless the symbol used is defined as the population mean.

- For responses that are scored P, the number of components satisfied should be considered if holistic scoring is required.

Model Solution	Scoring
(b) (i) The distribution of the sample of whistle prices appears slightly skewed to the right, because the mean is slightly higher than the median. (ii) Based on the $1.5 \times \text{IQR}$ rule, there are not whistle prices in this sample that would be considered outliers. A whistle price is an outlier using this method if it is more than $1.5 \times \text{IQR}$ below the first quartile (Q_1) or more than $1.5 \times \text{IQR}$ above the third quartile (Q_3). Because $Q_1 - 1.5 \times \text{IQR}$ $= 4.51 - 1.5(5.475 - 4.51)$ $= 3.0625,$ and the minimum value (4.25) is greater than $3.0625,$ there are no outliers to the left. Because $Q_3 + 1.5 \times \text{IQR}$ $= 5.475 + 1.5(5.475 - 4.51)$ $= 6.9225,$ and the maximum value (6.58) is less than $6.9225,$ there are no outliers to the right.	**Essentially correct (E)** if the response satisfies the following four components: 1. In part (b-i) the response indicates the distribution is skewed to the right or approximately symmetric 2. In part (b-i) the response justifies the response to component 1 with appropriate reasoning based on the summary statistics 3. In part (b-ii) the response provides a justification for stating there are no outliers by correctly calculating the lower and upper outlier criteria with work shown 4. In part (b-ii) the response states whether there are outliers based on the calculated lower and upper outlier criteria **Partially correct (P)** if the response satisfies three of the four components required for E. *OR* satisfies components 1 and 2 *OR* components 3 and 4 of components 1–4 required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

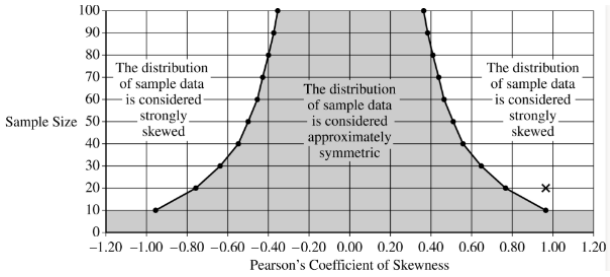
Additional Notes:

- If the response states skewed right, to satisfy component 2 the response should refer to the median being less than the mean. If the response states approximately symmetric, the response should refer to the median and mean being close together.
- A response that indicates skewed left does not satisfy components 1 or 2.
- If the response says the distribution is skewed to the right, examples of responses that satisfy component 2 include:
  - the mean is higher than the median,
  - the ratio between the mean and median is greater than 1,
  - the difference between the maximum and median is greater than the difference between the median and minimum, and
  - the difference between the maximum and third quartile is greater than the difference between the first quartile and minimum.
- For responses that score P, the number of components satisfied should be considered if holistic scoring is required.

Model Solution	Scoring

(c) (i) Pearson's coefficient of skewness is $\frac{3(5.12 - 4.885)}{0.743} \approx 0.949$.

(ii)



Essentially correct (E) if the response satisfies the following two components:

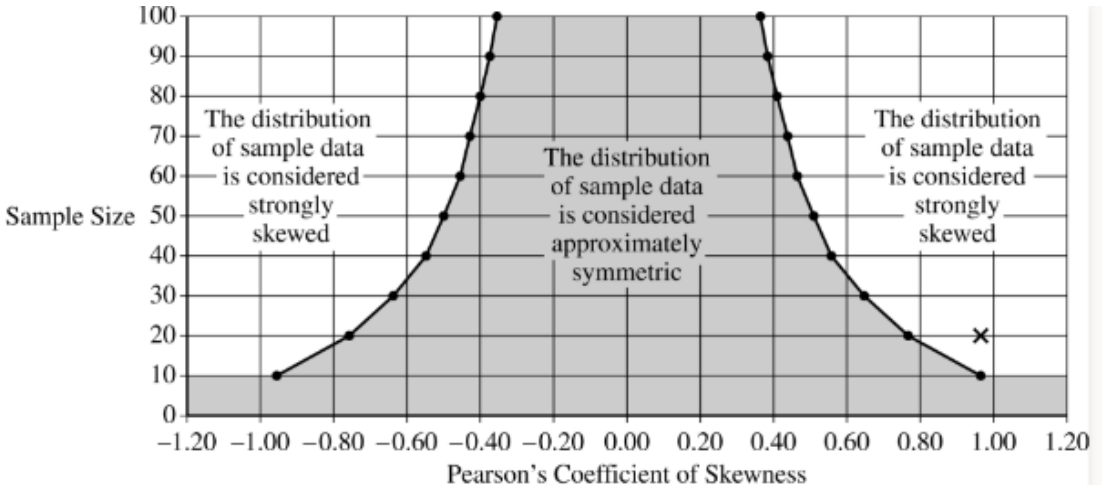
1. In part (c-i) the response calculates Pearson's coefficient of skewness with work shown
2. In part (c-ii) the response includes an x at approximately $(0.949, 20)$ or an x consistent with the calculation in part (c-i)

Partially correct (P) if the response satisfies only one of the two components required for E.

Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- An arithmetic or transcription error in a response can be ignored if correct work is shown.
- Responses must include an x that is located on the line for sample size of 20 and located between 0.90 and 1.00 on the horizontal axis or be consistent with the value given in part (c-i) to satisfy component 2.



Model Solution	Scoring
(d) (i) Looking at the graph in part (c), for a sample size of 20, and a skewness coefficient of 0.949, this point falls in "the distribution of sample data is considered strongly skewed" region. Therefore, we would consider the shape of the distribution of the sample of whistle prices to be strongly skewed. (ii) No, based on the response to part (d-i), Julio's data would not satisfy the normality condition because neither of the criteria listed are met. Julio only has a sample size of 20, which is less than 30, and Pearson's coefficient of skewness indicates the distribution of sample data is strongly skewed.	**Essentially correct (E)** if the response satisfies the following five components: 1. In part (d-i) the response indicates the sample is strongly skewed or makes a statement consistent with the graph in part (c) 2. In part (d-i) the response uses or refers to the graph or the Pearson coefficient of skewness to make the correct determination of skewness 3. In part (d-ii) the response states that the condition of normality is not met or makes a statement consistent with the explanation in part (d-i) 4. In part (d-ii) the response explains the sample size of 20 is less than 30 5. In part (d-ii) the response explains that the distribution of sample data is strongly skewed or makes a statement consistent with the explanation in part (d-i) **Partially correct (P)** if the response does not meet the criteria for E but satisfies three or four of components 1–5 required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response that determines the sample is strongly skewed, but says the normality condition is satisfied, should be scored I.
- In part (d-i) a response that indicates the distribution is strongly skewed to the left does not satisfy component 1.
- If a response in part (c-ii) has an x on the curve, the response in part (d-i) must indicate the sample is skewed (not strongly skewed or approximately symmetric), or indicate it is not possible to choose whether it is strongly skewed or approximately symmetric to satisfy component 1.
- In part (d-i) a response that indicates the distribution is skewed, but not strongly skewed, because the Pearson coefficient of skewness is not that far from the curve on the graph satisfies component 1 and component 2.
- In part (d-i) a response that indicates the distribution is skewed, but not strongly skewed, may satisfy components 3–5 if in part (d-ii) the response indicates the conditions are met because the distribution is not very skewed and there are no outliers. The response does not need to indicate the sample size is less than 30.
- In part (d-ii) a response that indicates the sample size is small without explicitly referring to the values 20 and 30 may satisfy component 4.
- The quality of communication, or the number of components satisfied for responses with score P, should be considered if holistic scoring is required.

Scoring for Question 6

Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as ½ point.

	Score
Complete Response	**4**
Substantial Response	**3**
Developing Response	**2**
Minimal Response	**1**

If a response is between two scores (for example, 2 ½ points), use a holistic approach to decide whether to score up or down, depending on the strength of the response and quality of the communication.

Question 1: Focus on Exploring Data — 4 points

General Scoring Notes

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

	Model Solution	Scoring
(a)	The five-number summary of the distribution of length of stay is: Minimum = 5 days Lower quartile (Q_1) = 6 days Median = 7 days Upper quartile (Q_3) = 8 days Maximum = 21 days	**Essentially correct (E)** if the response provides correct values for ALL FIVE of the summary statistics with labels (minimum, lower quartile, median, upper quartile, and maximum). **Partially correct (P)** if the response provides correct values for only THREE or FOUR of the summary statistics with labels. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- Any discussion of the mean, IQR, or the standard deviation of length of stay should be ignored in scoring.
- Inclusion or omission of units of measurement (days) has no bearing on scoring.
- If the response includes exactly 5 unlabeled numbers expressed together as a vertical or horizontal list, interpret the numbers as being labeled as the minimum, lower quartile, median, upper quartile, and maximum, respectively.
- A response that includes only five numbers that are correct values for the five-number summary without providing a complete set of labels or not putting them in an ordered list may be scored P.

Model Solution	Scoring
(b) **(i)** The patients who stayed for 12 days and 21 days are considered outliers using method A. An outlier using method A is a value greater than $1.5 \times \text{IQR}$ above the third quartile (Q_3) or more than $1.5 \times \text{IQR}$ below the first quartile (Q_1). Because $Q_1 - 1.5 \times \text{IQR} = 6 - 1.5(8 - 6) = 3,$ then any values below 3 are considered outliers. There are no such values. Because $Q_3 + 1.5 \times \text{IQR} = 8 + 1.5(8 - 6) = 11,$ then any values above 11 are considered outliers. **(ii)** The patient who stayed for 21 days is the only outlier using method B. An outlier using method B is a value located 2 or more standard deviations above, or below, the mean. Because $\text{Mean} \pm 2 \times \text{SD} = 7.42 \pm 2(2.37),$ then any value that is outside of the interval $(2.68, 12.16)$ is considered an outlier.	**Essentially correct (E)** if the response satisfies the following four components: 1. Correctly identifies the two outliers in part (b-i) as the patients who stayed for 12 days and 21 days 2. Provides a justification for part (b-i) by calculating the lower and upper outlier criteria for the $1.5 \times \text{IQR}$ rule (e.g., "using method A, an outlier is any value below 3 days or above 11 days") 3. Correctly identifies the one outlier in part (b-ii) as the patient who stayed for 21 days 4. Provides a justification for part (b-ii) by calculating the lower and upper outlier criteria for the 2 standard deviations rule (e.g., "using method B, an outlier is any value below 2.68 days or above 12.16 days") **Partially correct (P)** if the response satisfies only two or three of the four components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response for part (b-ii) that manually computes the standard deviation as 2.374 and then uses it to construct an interval of $(2.672, 12.168)$ satisfies component 4.
- Component 1 and component 2 are satisfied if the response to part (b-i) uses correct calculations with incorrect values of summary statistics reported in the response to part (a).

Model Solution	Scoring
(c) Quartiles and the IQR are less sensitive to extreme values in strongly skewed distributions than the mean and standard deviation. Relative to the quartiles, the mean is pulled more toward the extreme values in the longer tail of a strongly skewed distribution. For a distribution that is strongly skewed to the right, the sample mean will be pulled more toward the extreme values in the longer right tail of the distribution than the sample median, and the ratio of the standard deviation to the IQR will tend to be larger than that for more nearly symmetric distributions. As a result, this pulls the value of the outlier criterion for method B, $\text{Mean} + 2 \times \text{SD}$, more toward the extreme values in the right tail of the distribution than the outlier criterion for method A, $Q_3 + 1.5 \times \text{IQR}$. This decreases the ability of method B to identify outliers relative to method A, which means that method A may identify more outliers than method B for a distribution that is strongly skewed to the right.	**Essentially correct (E)** if the response satisfies the following two components: 1. Indicates that the mean is pulled more toward the extreme values in the longer right tail for a strongly right-skewed distribution than the quartiles (or median) OR indicates that the ratio of the standard deviation to the IQR tends to be larger for strongly skewed distributions than for more nearly symmetric distributions 2. Provides an explanation that links effects of skewness on an increased ability of method A to detect outliers relative to method B (e.g., "the larger shift in the mean relative to the shift in the median (or quartiles) has a greater effect on decreasing the ability of method B to detect outliers compared to method A" OR "the larger increase in the standard deviation, relative to the IQR, results in a greater increase in the range of non-outlier values for method B compared to method A") **Partially correct (P)** if the response satisfies only one of the two components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring for Question 1	Score
Complete Response Three parts essentially correct	**4**
Substantial Response Two parts essentially correct and one part partially correct	**3**
Developing Response Two parts essentially correct and no part partially correct *OR* One part essentially correct and one or two parts partially correct *OR* Three parts partially correct	**2**
Minimal Response One part essentially correct and no part partially correct *OR* No part essentially correct and two parts partially correct	**1**

2019

AP CollegeBoard

AP® Statistics

Scoring Guidelines

2019 SCORING GUIDELINES

Question 1

Intent of Question

The primary goals of this question were to assess a student's ability to (1) describe features of a distribution of sample data using information provided by a histogram; (2) identify potential outliers; (3) sketch a boxplot; and (4) comment on an advantage of displaying data as a histogram rather than as a boxplot.

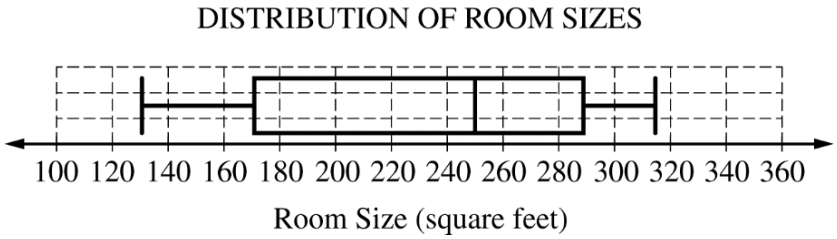
Solution

Part (a):

The distribution of the sample of room sizes is bimodal and roughly symmetric with most room sizes falling into two clusters: 100 to 200 square feet and 250 to 350 square feet. The center of the distribution is between 200 and 300 square feet. The range of the distribution is between 150 and 250 square feet. There are no apparent outliers.

Part (b):

The interquartile range is $\text{IQR} = 292 - 174 = 118$ square feet. There are no potential outliers because the minimum room size of 134 square feet does not fall below $Q_1 - 1.5(\text{IQR}) = -3$ square feet, and the maximum room size of 315 square feet does not exceed $Q_3 + 1.5(\text{IQR}) = 469$ square feet.



Part (c):

The histogram clearly shows the bimodal nature of the distribution of room sizes, but this is not apparent in the boxplot.

2019 SCORING GUIDELINES

Question 1 (continued)

Scoring

This question is scored in three sections. Section 1 consists of part (a); Section 2 consists of the outlier determination in part (b); Section 3 consists of the boxplot sketch in part (b) and part (c). Each section is scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the description of the distribution of room sizes satisfies the following four components:

1. The shape is bimodal OR there are two peaks OR there are two clusters.
2. The center is between 200 and 300 square feet.
3. The spread is addressed by stating the range is a value between 150 and 250 square feet OR the interquartile range is a value between 50 and 150 square feet OR all room sizes are between 100 and 350 square feet.
4. The response includes context.

Partially correct (P) if the response satisfies two or three of the four components.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Notes:

- Shape: Component 1 cannot be satisfied if a response describes the histogram as unimodal or describes the entire histogram as normal or approximately normal.
- Shape: A response that addresses symmetry, while appropriate, does not impact the scoring of section 1.
- Center: A response that states one cluster of the distribution is centered between 150 and 200 square feet and the other cluster is centered between 250 and 300 square feet satisfies both components 1 and 2.
- Center:
    - Responses that address center using interval language such as "the mean of the distribution is *between* 200 and 300" must, for any single measure of center, provide an interval with lower endpoint not below 200 square feet, and with upper endpoint not above 300 square feet to satisfy component 2.
    - Responses that address center using approximate language such as "the median of the distribution is *approximately* 225" must, for any single measure of center, specify a numeric value that is not less than 200 square feet, and that is not greater than 300 square feet to satisfy component 2.
    - Responses that use definitive language such as "the mean of the distribution *is* 231.4" must identify the corresponding numeric value correctly to satisfy component 2. Specifically, the median of the distribution can be correctly identified as any value between 250 and 253.5 square feet, inclusive; the mean of the distribution is 231.4 square feet; and the center (or average) of the distribution can be any value that is a correct median or mean.

2019 SCORING GUIDELINES

Question 1 (continued)

- Spread: A response recognizing all values in the sample fall between 100 and 350 square feet (or between 134 and 315 square feet) satisfies component 3 *only* for these exact endpoints and need not appeal to a specific measure of spread such as range or interquartile range (IQR).
- Spread:
    - Responses that appeal to a specific measure of spread using interval language, such as "the IQR is *between* 50 and 150," must provide bounds appropriate to the corresponding measure of spread. For range, the lower endpoint must not be below 150 square feet and the upper endpoint cannot exceed 250 square feet; for IQR, the lower endpoint must not be below 50 square feet, with upper endpoint not to exceed 150 square feet; for standard deviation, the lower endpoint must not be below 25 square feet, with upper endpoint not to exceed 100 square feet.
    - Responses that appeal to a specific measure of spread using approximate language, such as "the range is *approximately* 250," must specify a numeric value within the bounds appropriate to that measure of spread. For range, the value must be between 150 and 250 square feet(inclusive); for IQR, the value must be between 50 and 150 square feet (inclusive); for standard deviation, the value must be between 25 and 100 square feet (inclusive). Responses that appeal to a specific measure of spread using definitive language, such as "the range of the distribution *is* 181," must identify the corresponding numeric value correctly to satisfy component 3. Specifically, the range of the distribution is 181 square feet; the IQR of the distribution is 118 square feet; and the standard deviation of the distribution is 68.12 square feet.

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Computation of both upper and lower outlier boundary fences that also shows the fences formulas either in words, symbols $Q_1 - 1.5(\text{IQR})$ and $Q_3 + 1.5(\text{IQR})$, or with values substituted from the table $174 - 1.5(118)$ and $292 + 1.5(118)$, or $(174 - 177)$ and $(292 + 177)$.
2. A correct decision regarding the presence of outliers.
3. Correct justification that compares the data with the fences.

Partially correct (P) if the response satisfies only two of the three components OR if the response omits exactly one of the fences but otherwise satisfies all three components.

Incorrect (I) if the response does not satisfy the requirements for E or P.

Notes:

- A response that identifies both fence formulas using symbols, but does not substitute values for all symbols, must also include the correct fence values of -3 and 469 to satisfy component 1.
- In place of an appeal to fences, a response may compute outlier bounds representing k standard deviations from the sample mean, where k is a number from 2 to 3 (inclusive), and must include formulas for both endpoints either in words, symbols $\bar{x} \pm k(\text{standard deviation})$, or with values substituted from the table. When $k = 2$ the outlier bounds are $(95.16, 367.64)$; when $k = 3$ the bounds are $(27.04, 435.76)$.

2019 SCORING GUIDELINES

Question 1 (continued)

- A response that identifies the standard deviation bounds using symbols, but that does not substitute values for all symbols, does not satisfy component 1 unless the correct numeric bounds are provided.
- Component 3 is satisfied if the response states the outlier decision criterion: any data values falling outside of the interval from -3 to 469 are potential outliers.

Section 3 is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

1. A correct sketch of the boxplot.
2. A response for part (c) that indicates the bimodal shape of the room size distribution is apparent in the histogram but not in the boxplot.

Partially correct (P) if the response satisfies only one of the two components.

Incorrect if the response does not meet the criteria for E or P.

Notes:

- The boxplot must be completely correct to satisfy component 1. Specifically:
  - The minimum is positioned between grid lines at 120 and 140 square feet.
  - Q_1 is positioned between grid lines at 160 and 180 square feet.
  - The median is positioned between grid lines at 240 and 260 square feet.
  - Q_3 is positioned between grid lines at 280 and 300 square feet.
  - The maximum is positioned between grid lines at 300 and 320 square feet.
- If a *mean* is included as a part of the boxplot, component 1 cannot be satisfied.
- A response based on skewness or symmetry does not satisfy component 2.
- A response stating the unimodal OR normal shape of the histogram of room sizes is apparent in the histogram but not in the boxplot will satisfy component 2 *only* if the shape description in section 1 component 1 was also unimodal OR normal, respectively.

2019 SCORING GUIDELINES

Question 1 (continued)

4 Complete Response

Three sections essentially correct

3 Substantial Response

Two sections essentially correct and one section partially correct

2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

1 Minimal Response

One section essentially correct

OR

No sections essentially correct and two sections partially correct

2019 SCORING GUIDELINES

Question 2

Intent of Question

The primary goals of this question were to assess a student's ability to (1) identify components of an experiment; (2) determine if an experiment has a control group; and (3) describe how experimental units can be randomly assigned to treatments.

Solution

Part (a):

Treatments: Sprays with four different concentrations of the fungus (0 ml/L, 1.25 ml/L, 2.5 ml/L, and 3.75 ml/L)

Experimental units: 20 containers, each containing the same number of insects

Response variable: Number of insects that are still alive in each container one week after spraying

Part (b):

Yes. Because the 0 ml/L concentration contains no fungus, the containers that are sprayed with the 0 ml/L concentration form the control group.

Part (c):

Label each container with a unique integer from 1 to 20. Then use a random number generator to choose 15 integers from 1 to 20 without replacement. Use the first five of these numbers to identify the five containers that will receive the 0 ml/L treatment. Use the second five of these numbers to identify the five containers that will receive the 1.25 ml/L treatment. Use the third five of these numbers to identify the five containers that will receive the 2.5 ml/L treatment. The remaining five containers will receive the 3.75 ml/L treatment.

(Alternative solution) Using 20 equally sized slips of paper, label five slips with 0 ml/L, five slips with 1.25 ml/L, five slips with 2.5 ml/L, and five slips with 3.75 ml/L. Mix the slips of paper in a hat. For each container, select a slip of paper from the hat (without replacement) and spray that container with the treatment selected.

2019 SCORING GUIDELINES

Question 2 (continued)

Scoring

Parts (a), (b), and (c) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. Identifies the 4 concentrations (or mixtures or sprays) as the treatments
2. Identifies the 20 containers as the experimental units
3. Identifies the number of insects that are still alive in each container as the response variable

Partially correct (P) if response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:
- Listing the four treatments satisfies component 1 (including ml/L is not required). However, if the list does not include all four treatments, component 1 is not satisfied.
- To satisfy component 1, the response must refer to plural concentrations/mixtures/sprays (e.g., the mixtures, the levels of the concentration). Referring only to the explanatory variable (concentration) does not satisfy component 1.
- The following responses satisfy component 2: "the 20 containers"; "the containers"; "the 20 groups of insects"; or "the groups of insects in each container." References to only "groups of insects" do not satisfy component 2 because it is unclear if these groups are formed by treatment or by container.
- To satisfy component 3, it must be clear that the response variable is being measured separately for each experimental unit. A response that says only "number of insects alive" does not satisfy component 3 because it could be referring to the total number of insects alive.
- To satisfy component 3, the response must be stated as a variable by using "number of" or equivalent. For example, "insects alive in each container" is not a variable and would not satisfy component 3.
- If the response states that the insects are the experimental units, then component 3 can still be satisfied by providing a binary response variable for each insect (e.g., whether the insect lived or died, survival status).

2019 SCORING GUIDELINES

Question 2 (continued)

Part (b) is scored as follows:

Essentially correct (E) if the response indicates that there is a control group and justifies this claim by identifying the control group or by explaining that there is a treatment which contains no fungus.

Partially correct (P) if the response indicates that there is no control group because every container is sprayed with some mixture

OR

if the response states that there is a control group but implies that 0 ml/L is not a treatment (e.g., "the containers with 0 ml/L form a control group because they don't receive a treatment"; "yes, there is a group that got no treatment").

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- The response does not need to explain the purpose of a control group.
- The response does not need to explicitly say "yes"—it can be implied by stating that there is a control group or saying "the control group is …."

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Creates appropriate labels for the units/treatments (e.g., label the containers from 1 through 20, label 20 slips of paper with five for each treatment)
2. Describes how to correctly implement the random assignment process
3. The random assignment process results in an equal number of experimental units assigned to each treatment

Partially correct (P) if response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- If the response states that insects are the experimental units in part (a), the response in part (c) can be in terms of insects or containers. In either case, the same three components are used to determine the score.
- If the response states that the containers are the experimental units in part (a), but only describes how to assign insects to treatments in part (c), component 1 is not satisfied.
- For responses that use slips of paper:
    - If the number of slips of paper is not equal to the number of experimental units, then component 1 is not satisfied. The slips of paper do not need to be specifically identified as equally-sized.
    - If the slips of paper are not mixed/shuffled or the slips are not "selected at random," component 2 is not satisfied. Sampling without replacement is implied when using slips of paper, unless the response specifies sampling with replacement.

2019 SCORING GUIDELINES

Question 2 (continued)

- For responses that use random number generators (or a 20-sided die):
  - If the initial assignment of numbers to units does not give each unit the same probability of being assigned to each treatment (e.g., units are represented by different numbers of integers), then component 1 is not satisfied.
  - If the response does not indicate that the numbers are selected without replacement or that different numbers must be used, the response does not satisfy component 2. The response does not need to specify the interval of numbers from which they are selecting (e.g., randomly generate a number from 1 to 20).
- For responses that use a table of random digits:
  - If the initial assignment of numbers to units does not give each unit the same probability of being assigned to each treatment, component 1 is not satisfied. For example, responses that use the labels 1 to 20 (not 01 to 20) do not satisfy component 1 because label 1 has a $\frac{1}{10}$ probability of being selected but label 20 has a $\frac{1}{100}$ probability of being selected.
  - If the response does not indicate that the numbers are selected without replacement or that different numbers must be used, the response does not satisfy component 2. The response does not need to specify the interval of numbers from which they are selecting or state that the numbers corresponding to unused labels will be skipped (e.g., skip numbers 00 and 21 to 99).
- For responses that use a 4-sided die (or random integers from 1 to 4):
  - If the die is rolled for each experimental unit, then component 3 is not satisfied because an equal number of units per treatment is not guaranteed.
  - If the die is rolled for each experimental unit until treatments are "full," then component 1 is not satisfied because this setup doesn't allow for all possible random assignments to be equally likely (unless the order of the units is randomized initially).
- If a response groups the experimental units before any random assignment (e.g., forms five groups of four containers or four groups of five containers), and then randomly assigns treatments to the groups or randomly assigns treatments within each group, component 1 is not satisfied. However, if a response forms groups in the context of a randomized block design with a reasonable blocking variable, component 1 can be satisfied.
- If a response describes two different random assignment processes in detail (e.g., how to randomly assign insects to containers and how to assign containers to treatments), both descriptions are scored according to the three components and the lower score is used.
- Responses that assign experimental units only to groups and not to treatments (e.g., randomly select five containers and put them in group 1) do not satisfy component 3.
- If the response randomly assigns insects to containers, the containers must be assigned to a treatment to satisfy component 3. In this case, the assignment of treatment to container does not need to be at random to satisfy component 3.

2019 SCORING GUIDELINES

Question 2 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

2019 SCORING GUIDELINES

Question 3

Intent of Question

The primary goals of this question were to assess a student's ability to (1) use information from a two-way table of relative frequencies to compute joint, marginal, and conditional probabilities; (2) recognize whether two events are independent; and (3) compute a probability for a binomial distribution.

Solution

Part (a):

(i) $P(\text{never and woman}) = 0.0636$

(ii)

$$\begin{aligned} P(\text{never or woman}) &= P(\text{never}) + P(\text{woman}) - P(\text{never and woman}) \\ &= 0.12 + 0.53 - 0.0636 \\ &= 0.5864 \end{aligned}$$

(iii) $P(\text{never} \mid \text{woman}) = \dfrac{P(\text{never and woman})}{P(\text{woman})} = \dfrac{0.0636}{0.53} = 0.12$

Part (b):

Yes, the event of being a person who responds never is independent of the event of being a woman because

$P(\text{never} \mid \text{woman}) = P(\text{never}) = 0.12.$

Part (c):

Define X as the number of people in a random sample of five people who always take their medicine as prescribed. Then X has a binomial distribution with $n = 5$ and $p = 0.54,$ and

$$\text{P}(X \geq 4) = \binom{5}{4}(0.54)^4(0.46)^1 + \binom{5}{5}(0.54)^5(0.46)^0 \approx 0.19557 + 0.04592 \approx 0.24149.$$

Scoring

Parts (a), (b), and (c) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response reports correct values of the probabilities for (i), (ii), and (iii).

Partially correct (P) if only one or two of the probabilities are correct.

Incorrect (I) if none of the probabilities are correct.

Notes:

- Assuming independence for events never and woman in (i) without referencing the result in part (b) does not satisfy (i).
- Alternative solutions for (ii) include $0.0564 + 0.0636 + 0.1384 + 0.3280 = 0.5864$ and $0.0564 + 0.53 = 0.5864$.

2019 SCORING GUIDELINES

Question 3 (continued)

Part (b) is scored as follows:

Essentially correct (E) if the response indicates that the events are independent, gives an explanation of independence using the events in the problem, *AND* provides appropriate justification using numbers from the table.

Note: Examples of valid explanations with appropriate justifications include:

- $P(\text{woman and never}) = 0.0636$ is the same as $P(\text{woman}) \times P(\text{never}) = (0.53)(0.12) = 0.0636$.
- $P(\text{never} \mid \text{woman}) = \dfrac{0.0636}{0.53} = 0.12$ is the same as $P(\text{never}) = 0.12$.
- $P(\text{woman} \mid \text{never}) = \dfrac{0.0636}{0.12} = 0.53$ is the same as $P(\text{woman}) = 0.53$.

Partially correct (P) if the response indicates that the events are independent *AND* gives an explanation of independence using the events in the problem but does not provide justification using numbers from the table

OR

if the response uses a correct method of illustrating that events are independent but makes an arithmetic mistake or a transcription mistake that results in concluding that these two events are not independent.

Incorrect (I) if the response does not satisfy requirements for E or P.

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Clearly indicates a binomial distribution with $n = 5$ and $p = 0.54$.
2. Indicates the correct boundary value and direction of the event.
3. Reports the correct probability.

Partially correct (P) if the response satisfies component 1 but it does not satisfy one or both of the other two components

OR

if the response does not satisfy component 1 but both of the other two components are satisfied.

Incorrect (I) if the response does not meet the criteria for E or P.

2019 SCORING GUIDELINES

Question 3 (continued)

Notes:

- The response $B(5, 0.54)$ satisfies component 1.
- Components 1 and 2 are satisfied by displaying the correct formula for computing the binomial probability using the correct values for n and p, e.g.,

  $$\binom{5}{4}(0.54)^4(0.46)^1 + \binom{5}{5}(0.54)^5(0.46)^0$$

- Only component 1 is satisfied if the correct binomial distribution is used in an incorrect probability formula, e.g.,

  $$\binom{5}{4}(0.54)^4(0.46)^1$$

- For component 2, the boundary value and direction may be described in words, e.g., *P*(at least four people).
- Component 2 may be satisfied by displaying a bar graph of a binomial distribution with the appropriate bars shaded.
- The response of $1 - \text{binomcdf}(n = 5,\ p = 0.54,\ \text{upper bound} = 3) = 0.24$ is scored E since n, p, and the boundary value are clearly identified.
  The response of $1 - \text{binomcdf}(n = 5, p = 0.54, 3) = 0.24$ is scored P since n and p are clearly identified and the boundary value is not identified.
  The response of $1 - \text{binomcdf}(5, 0.54, 3) = 0.24$ is scored I.
- A normal approximation to the binomial is not appropriate since $np = 5 \times 0.54 = 2.7$ and $2.7 < 5$.
  A response using the normal approximation can score at most P. To score P, the response must include all of the following:
  - an indication that the probability calculated is a normal approximation for the binomial probability
  - a correct mean and standard deviation based on the binomial parameters
  - clear indication of boundary and direction with a *z*-score or diagram
  - the probability computed correctly.

  An example of a response which meets these four criteria is $P\left(Z \geq \frac{4 - np}{\sqrt{np(1-p)}}\right) = P\left(Z \geq \frac{4 - (5)(0.54)}{\sqrt{(5)(0.54)(0.46)}}\right) \approx 0.1217,$ and the binomial distribution is mentioned.

2019 SCORING GUIDELINES

Question 3 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

2019 SCORING GUIDELINES

Question 4

Intent of Question:

The primary goals of this question were to assess a student's ability to perform an appropriate hypothesis test to address a particular question. More specific goals were to assess students' ability to state appropriate hypotheses, identify the appropriate statistical test procedure, check appropriate assumptions/conditions for inference; calculate a correct test statistic and *p*-value; and draw a correct conclusion, with justification, in the context of the study.

Solution

Section 1:

Let p_{14} represent the proportion of the population of kochia plants in the western United States that were resistant to glyphosate in 2014. Let p_{17} represent the proportion of the population of kochia plants in the western United States that were resistant to glyphosate in 2017.

The null hypothesis $H_0 : p_{17} - p_{14} = 0$ is to be tested against the alternative hypothesis $H_a : p_{17} - p_{14} > 0$.

An appropriate inference procedure is a two-sample *z*-test for a difference in proportions. The formula for the test statistic is:

$$z = \frac{\hat{p}_{17} - \hat{p}_{14}}{\sqrt{\left(\frac{\hat{p}_c(1-\hat{p}_c)}{n_{17}} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_{14}}\right)}}$$

where $\hat{p}_c = \frac{n_{14}\hat{p}_{14} + n_{17}\hat{p}_{17}}{n_{14} + n_{17}}$ is a pooled estimate of the proportion of resistant plants for 2014 and 2017 combined.

Section 2:

The first condition for applying the test is that the data are gathered from independent random samples from the populations of kochia plants in the western United States in 2014 and 2017. The question indicates that a random sample of 61 kochia plants was taken in 2014 and a second random sample of 52 kochia plants was taken in 2017. It is reasonable to assume that the 2017 sample of plants was in no way influenced by the 2014 sample of plants.

The second condition is that the sampling distribution of the test statistic is approximately normal. This condition is satisfied because the expected counts under the null hypothesis are all greater than 10. The pooled estimate of the proportion of resistant plants is $\hat{p}_c = \frac{(61)(0.197) + (52)(0.385)}{61 + 52} \approx 0.2835$. The estimates of the expected counts are
$61(0.2835) \approx 17.29,\ 61(1 - 0.2835) \approx 43.71,\ 52(0.2835) \approx 14.74,\ 52(1 - 0.2835) \approx 37.26,$ all of which are greater than 10.

Because sampling must have been done without replacement, the independence condition for each sample should be checked. Information on the population sizes of kochia plants is not given for either 2014 or 2017, but it is reasonable to assume that each population has millions of plants. Therefore it is reasonable to assume that the sample sizes are less than 10 percent of the respective population sizes.

2019 SCORING GUIDELINES

Question 4 (continued)

Using the pooled estimate of the proportion of resistant plants, $\hat{p}_c \approx 0.2835$, the value of the test statistic is:

$$z = \frac{0.385 - 0.197}{\sqrt{\left(\frac{(0.2835)(0.7165)}{61} + \frac{(0.2835)(0.7165)}{52}\right)}} \approx 2.21$$

The p-value is 0.0135.

Section 3:

Because the p-value is less than $\alpha = 0.05$, there is convincing statistical evidence to conclude that the proportion of resistant plants in the 2017 population of kochia plants is greater than the proportion of resistant plants in the 2014 population of kochia plants.

Scoring

Sections 1, 2, and 3 are each scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response satisfies components 1 and 4 *AND* at least one of the remaining components:

1. Hypotheses imply equality of proportions in the null hypothesis and correct direction in the alternative hypothesis, which utilize an appropriate population parameter in words or symbols.
2. Identifies parameters that are population proportions.
3. Both parameters are correctly defined as proportions of resistant plants in 2014 and 2017.
4. The two-sample z-test for proportions is identified by name or formula

Partially correct (P) if the response does not meet the requirement for E, but at least two of the components are satisfied.

Incorrect if the response does not meet the criteria for E or P.

Notes:

- Correct ways to state the null hypothesis that satisfy component 1:
  $H_0 : p_{17} = p_{14}$ or $H_0 : p_{17} - p_{14} = 0$
  $H_0 : p_{17} \leq p_{14}$ or $H_0 : p_{17} - p_{14} \leq 0$
  $H_0 : p_{14} \geq p_{17}$ or $H_0 : p_{14} - p_{17} \geq 0$
  Correct ways to state the alternative hypothesis that satisfy component 1:
  $H_a : p_{17} > p_{14}$ or $H_a : p_{17} - p_{14} > 0$
  $H_a : p_{14} < p_{17}$ or $H_a : p_{14} - p_{17} < 0$
  Incorrect ways to state the null hypothesis that do not satisfy component 1:
  $H_0 : p_{17} < p_{14}$ or $H_0 : p_{17} - p_{14} < 0$
  $H_0 : p_{14} > p_{17}$ or $H_0 : p_{14} - p_{17} > 0$
  Incorrect ways to state the alternative hypothesis that do not satisfy component 1:
  $H_a : p_{17} \neq p_{14}$ or $H_a : p_{17} - p_{14} \neq 0$
  $H_a : p_{17} < p_{14}$ or $H_a : p_{17} - p_{14} < 0$

2019 SCORING GUIDELINES

Question 4 (continued)

- Examples for components 2 and 3:
  - Satisfies both components 2 and 3:
    - p_{17} is the proportion of resistant plants
      p_{14} is the proportion of resistant plants
  - Satisfies component 2 but not component 3:
    - p_1 is the proportion of resistant plants
      p_2 is the proportion of resistant plants
    - p_{17} is the proportion of plants
      p_{14} is the proportion of plants
    - p_{17}, p_{14}
    - p_1, p_2
- If the test is correctly identified by name, but then an incorrect formula is stated, this is considered to be a parallel response and component 4 is not satisfied.
- If the test identifies an unpooled two sample *z*-test for a difference in proportions as the correct test or formula, component 4 is satisfied.

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies components 1 and 2 *AND* at least two of the remaining components:

1. Notes that the use of random samples in 2014 and 2017 satisfies the randomness condition.
2. Checks for approximate normality of the test statistic by showing that the expected numbers of resistant and non-resistant kochia plants are both larger than some commonly accepted criterion (e.g. 5 or 10) for both samples.
3. Notes that the populations of kochia plants must be extremely large in both years, thus satisfies the independence (10%) conditions.
4. Reports a correct value of the *z*-test statistic.
5. Reports a *p*-value that is consistent with the stated alternative hypothesis and reported test statistic.

Partially correct (P) if the response does not meet the criteria for E, but at least two of the five components are satisfied.

Incorrect if the response does not meet the criteria for E or P.

Notes:

- For the randomness component it is minimally acceptable to say "random samples—check" or "SRSs—check." The important concept is that the study used two independent random samples. Although it is not known if a SRS was taken versus another type of random sample, it is minimally acceptable to indicate SRSs since the sampling method is unknown. If the response implies that random assignment was used, the randomness component is not satisfied.

2019 SCORING GUIDELINES

Question 4 (continued)

- To satisfy component 2, the response must include actual numbers, or a formula with numbers plugged in, as well as a clear indication of comparison of the four quantities to some standard criterion, such as 5 or 10, or the statement that each such quantity is large enough. If a formula with numbers is used, simplification is NOT required.

  Examples of acceptable quantities (comparison still must be made):
  - 12, 49, 20, 32
  - 12.017, 48.983, 20.02, 31.98
  - $61(0.197)$, $61(1 - 0.197)$, $52(0.385)$, $52(1 - 0.385)$

  Examples of unacceptable quantities:
  - $n_{17}\hat{p}_{17}$, $n_{17}(1 - \hat{p}_{17})$, $n_{14}\hat{p}_{14}$, $n_{14}(1 - \hat{p}_{14})$
  - $n_{17}p_{17}$, $n_{17}(1 - p_{17})$, $n_{14}p_{14}$, $n_{14}(1 - p_{14})$
  - $n_{17}\hat{p}_c$, $n_{17}(1 - \hat{p}_c)$, $n_{14}\hat{p}_c$, $n_{14}(1 - \hat{p}_c)$
  - $61\hat{p}_c$, $61(1 - \hat{p}_c)$, $52\hat{p}_c$, $52(1 - \hat{p}_c)$
- The test statistics for the pooled and unpooled *z*-tests are 2.21 and 2.22 respectively, thus they are close to the same value. If the response provides the unpooled formula but then states a pooled test statistic, component 4 is satisfied. If the response provides the pooled formula but then states an unpooled test statistic, component 4 is satisfied.
- If the response uses a critical value approach rather than a *p*-value approach, then the correct critical value of -1.645 or 1.645, that is consistent with the alternative hypothesis, satisfies component 5.
- If the response did not satisfy component 1 in section 1 because a two-tailed alternative was stated or the direction of the alternative was incorrect, then the *p*-value in component 5 should be consistent with the stated alternative. If the response omits hypotheses or other incorrect hypotheses are stated, assume the correct alternative hypothesis is provided when scoring component 5.

Section 3 is scored as follows:

Essentially correct (E) if the response includes the following three components:
1. Provides justification of the conclusion based on a correct comparison between a stated *p*-value and an alpha value of 0.05.
2. Provides a correct conclusion consistent with the alternative hypothesis.
3. The conclusion is stated in context.

Partially correct (P)
if the response satisfies components 1 and 2

OR

if the response satisfies components 2 and 3

OR

if the response satisfies components 1 and 3 *AND*, based on the p-value from section 2, either
- the conclusion correctly rejects the null hypothesis but does not state that there is convincing evidence for the alternative hypothesis
  OR
- the conclusion correctly fails to reject the null hypothesis but does not state there is not convincing evidence for the alternative hypothesis.

2019 SCORING GUIDELINES

Question 4 (continued)

Incorrect (I) if the response does not satisfy the criteria for E or P.

Notes:

- If the conclusion is consistent with a reasonable, but incorrect, *p*-value from section 2, and is presented in context with justification based on comparison of the *p*-value to the level of significance, then section 3 is scored E.
- If the response implies that the outcome of the hypothesis test is a "proof" of either a true or false null, the score is lowered one level (that is, from E to P, or from P to I).
- If an incorrect interpretation of the *p*-value is given, the score is lowered one level (that is, from E to P, or from P to I).
- If the response uses a critical value approach rather than a *p*-value approach, then the correct critical value of -1.645 or 1.645 replaces the *p*-value in section 2, and comparison of the test statistic from section 2 to the critical value (e.g. $2.21 > 1.645$) satisfies component 1.
- If the response clearly states a reasonable level of significance that differs from 0.05 and provides a justification and conclusion in context based on that justification, the response is scored E.
- If the response provides the incorrect comparison between the stated *p*-value and the level of significance, but the conclusion is consistent with the given comparison and the alternative hypothesis, then component 2 is satisfied.
- If the response did not satisfy component 1 in section 1 because a two-tailed alternative was stated or the direction of the alternative was incorrect, then the conclusion component 2 should be consistent with the stated alternative. If the response states other incorrect hypotheses or omits hypotheses, assume the correct alternative hypothesis is provided when scoring component 2.

2019 SCORING GUIDELINES

Question 4 (continued)

Alternative Approach
Two-Sided Confidence Interval for Difference in Two Population Proportions

Section 1 is scored E, P, or I according to the guidelines in section 1 for a 2 sample *z*-test for proportions.

Notes:

- To satisfy component 4, the two-sample confidence interval for a difference in two proportions should be identified by name or formula by referring to the *z*-distribution and two proportions.

Section 2 is scored as follows:
Essentially correct (E) if the response satisfies components 1 and 2 *AND* at least one of the remaining components:

1. Notes that the use of random samples in 2014 and 2017 satisfies the randomness condition.
2. Checks for approximate normality of the test statistic by showing that the observed numbers of resistant and non-resistant kochia plants are both larger than some commonly accepted criterion (say 5 or 10) for both samples.
3. Notes that the populations of kochia plants must be extremely large in both years, thus satisfies the independence (10%) conditions.
4. Reports the correct confidence interval that is consistent with the stated alternative hypothesis.

Partially correct (P) if the response does not meet the criteria for E, but at least two of the four components are satisfied.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Notes:

- Examples of correct 90% confidence intervals to address a one-sided alternative for $\alpha = 0.05$ are:
  $(-0.327, -0.049)$
  $(0.049, 0.327)$
- Examples of correct 95% confidence intervals to address a two-sided alternative for $\alpha = 0.05$ are:
  $(-0.353, -0.022)$
  $(0.022, 0.353)$

Section 3 is scored E, P or I according to the guidelines in section 3 for a 2 sample *z*-test for proportions.

Notes:

- Component 1 is satisfied if a confidence interval that is consistent with the alternative hypothesis is given and the appropriate interval endpoint(s) are compared to zero.

Overall Notes:

- If the response constructs two separate one-proportion *z*-intervals for 2014 and 2017, then sections 1 and 2 are scored as above, and section 3 is scored I.

2019 SCORING GUIDELINES

Question 4 (continued)

Alternative Approach
Chi-square test for homogeneity

The value of the Pearson chi-square test statistic (uncorrected) is 4.8821 with 1 degree of freedom and a *p*-value of 0.02714. This value is the same as the square of the pooled *z*-statistic, so it is the same test, but the *p*-value is for a two-sided alternative. This *p*-value could be divided by 2 to obtain an appropriate *p*-value for the one-sided alternative, but the sample data needs to be examined to determine the correct direction of the alternative.

Section 1 is scored E, P, or I according to the guidelines in section 1 for a 2 sample *z*-test for proportions.

Notes:

- Examples of unacceptable hypotheses:
  $H_0 : p_{17}$ and p_{14} are independent or $H_0 : p_{17}$ and p_{14} have no association
  $H_a : p_{17}$ and p_{14} are dependent or $H_a : p_{17}$ and p_{14} have an association
  $H_a : p_{17} \neq p_{14}$ or $H_a : p_{17} \geq p_{14}$ or $H_a : p_{14} \leq p_{17}$
- To satisfy component 4, a chi-square test for homogeneity is identified by name or formula.

Section 2 is scored E, P, or I according to the guidelines in section 2 for a 2 sample *z*-test for proportions.

Notes:

- To satisfy component 4, the reported value of the chi-square test statistic is 4.8821.
- If a one sided alternative hypothesis is given in section 1, then to satisfy component 5, the reported value of the *p*-value is 0.01357.
- If a two sided alternative hypothesis is given in section 1, then to satisfy component 5, the reported *p*-value is 0.02714.

Section 3 is scored E, P or I according to the guidelines in section 3 for a 2 sample *z*-test for proportions.

Notes:

- If the response clearly indicates that the two sample proportions were used to determine the correct one-sided direction and a *p*-value of 0.0135 was used as justification, section 3 is scored E.
- If a correct conclusion is reached based on the *p*-value of 0.02714 and a two-sided alternative, then section 3 is scored at most P.
- If the final response justifies the conclusion based on the *p*-value of 0.0135 but does not explicitly indicate how the correct direction was determined, section 3 is scored at most P.

2019 SCORING GUIDELINES

Question 4 (continued)

4 Complete Response

Three sections essentially correct

3 Substantial Response

Two sections essentially correct and one section partially correct

2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

1 Minimal Response

One section essentially correct

OR

No section essentially correct and two sections partially correct

2019 SCORING GUIDELINES

Question 5

Intent of Question

The primary goals of this question were to assess a student's ability to (1) evaluate a percentile of a normal distribution; (2) evaluate a probability for a normal distribution; and (3) compute an expected value for a random variable with two possible outcomes.

Solution

Part (a):

The 25th percentile of the standard normal distribution is -0.6745. Consequently the 25th percentile of a normal distribution with mean 30 months and standard deviation 8 months is $30 + 8(-0.6745) = 24.6$ months.

Part (b):

The probability that a randomly selected customer will need to request a replacement because the battery fails within 24 months from the date of purchase is

$$P(\text{life span} \le 24 \text{ months}) = P\left(Z \le \frac{24 - 30}{8}\right) = P(Z \le -0.75) \approx 0.2266.$$

Part (c):

The company's expected gain for each warranty purchased is

$$\begin{aligned}&(\$50) \times P(\text{life span} > 24 \text{ months}) + (-\$150) \times P(\text{life span} \le 24 \text{ months})\\ &= (\$50) \times (0.7734) + (-\$150) \times (0.2266) \approx \$4.68.\end{aligned}$$

Scoring

Parts (a), (b), and (c) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Indicates the use of a normal distribution with a mean of 30 and a standard deviation of 8.
2. Sets up a correct approach for finding the 25th percentile of the battery life span distribution.
3. Reports the correct value of the 25th percentile (24.6) or a percentile value that is consistent with components 1 and 2.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- Component 1 may be satisfied in either part (a) or (b).
- Incorrect statistical notation in specifying the distribution mean or standard deviation (for example, $\bar{x} = 30$ or $s = 8$) results in Component 1 not being met the first time it appears.

2019 SCORING GUIDELINES

Question 5 (continued)

- Component 1 may be satisfied by one of the following:
  - Graphical: Displaying a graph of a normal density function with the horizontal axis clearly labeled using the mean and standard deviation for the battery life span distribution.
  - Calculator syntax: Labeling the mean and standard deviation in an inverse normal cdf calculator statement, for example, invNorm$(0.25, \mu = 30, \sigma = 8)$.
  - z-score: Showing correct components in a standard z-score calculation, for example, $z = \frac{x - 30}{8}$, or $-0.6745 = \frac{x - 30}{8}$, or $x = 30 + 8(-0.6745)$.
  - Notation: Using standard notation for a normal distribution, for example, $N(30, 8)$ or Normal$(30, 64)$.
- Component 2 may be satisfied by one of the following:
  - Graphical: Identifying the lower-tail area corresponding to the probability of 0.25 in a graph of a normal density function.
  - Calculator syntax: Stating the correct percentile in an inverse normal cdf calculator statement, for example, invNorm$(0.25, \mu = 30, \sigma = 8)$.
  - z-score: Equating the z-score of the 25th percentile of the battery life span distribution to -0.6745, for example, $-0.6745 = \frac{x - 30}{8}$ or $x = 30 + 8(-0.6745)$.

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. Indicates the use of a normal distribution with a mean of 30 and a standard deviation of 8.
2. Specifies the correct event, including the correct boundary value and direction.
3. Reports the correct probability (0.2266) or a probability consistent with components 1 and 2.

Partially correct (P) if response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:
- Component 1 may be satisfied in either part (a) or (b).
  - Incorrect statistical notation in specifying the distribution mean or standard deviation (for example, $\bar{x} = 30$ or $s = 8$) results in Component 1 not being met the first time it appears.
- Component 1 may be satisfied by one of the following:
  - Graphical: Displaying a graph of a normal density function with the horizontal axis clearly labeled using the mean and standard deviation for the battery life span distribution.
  - Calculator syntax: Labeling the mean and standard deviation in a normal cdf calculator statement, for example, normalcdf$(0, \text{upper} = 24, \mu = 30, \sigma = 8)$.
  - z-score: Showing correct components in a standard z-score calculation, for example, $z = \frac{x - 30}{8}$, or $z = \frac{24 - 30}{8}$, or $\frac{24 - 30}{8} = -0.75$.
  - Notation: Using standard notation for a normal distribution, for example, $N(30, 8)$ or Normal$(30, 64)$

2019 SCORING GUIDELINES

Question 5 (continued)

- Component 2 may be satisfied by one of the following:
  - Graphical: A normal density graph labeled with boundary and direction corresponding to the region of interest.
  - Calculator syntax: Labeling the upper bound in a normal cdf calculator statement. The lower bound does not need to be labeled but must be less than or equal to 0, for example, normalcdf(0, upper = 24, $\mu = 30$, $\sigma = 8$).
  - In words: Specifying the correct event, boundary value, and direction, in words: for example, P(battery life span $\leq$ 24 months), or "probability requires battery replacement within 24 months."
  - Random variable: Specifying the boundary value and direction using a random variable, for example, $P(X \leq 24)$ or $P(Z \leq -0.75)$.

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

1. Provides an expected value calculation with two appropriate probabilities that add to 1 and are paired with the correct outcomes.
2. Reports the correct expected value (4.68), or a reasonable expected value consistent with work shown.

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Notes:

- For component 1, appropriate probabilities that add to 1 are:
  - probabilities consistent with the value computed in part (b); or
  - the correct probabilities; or
  - if there is no probability provided in part (b), probabilities explicitly defined in context in part (c).
- If the correct probability is used, due to rounding considerations, any number between 4.00 and 4.70 is acceptable as a correct expected value for component 2.
- A reasonable expected value for component 2 can be any number between $-\$150$ and $+\$50$ that is consistent with work shown.

2019 SCORING GUIDELINES

Question 5 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

2019 SCORING GUIDELINES

Question 6

Intent of Question

The primary goals of this question were to assess a student's ability to (1) recognize the population to which results from a random sample may be generalized; (2) describe a disadvantage of using a sample mean rather than a sample median to indicate typical values when the sample distribution is skewed; (3) describe how the theoretical sampling distribution of the sample median could be constructed; (4) construct an approximate confidence interval for a population median using results from a bootstrap procedure; and (5) interpret a confidence interval.

Solution

Part (a):

Because random sampling was used, the results of the sample may be generalized to the population of rental prices for one-bedroom apartments in the city that are listed on this particular website at the time the sample was taken.

Part (b):

Because the distribution of the 50 rental prices in the sample is skewed to the right, the sample median provides a better indicator of typical rental prices than the sample mean. Some very large rental prices results in a sample mean that is substantially larger than the more typical rental prices. As a result the sample mean would overestimate the typical rental price, whereas the sample median would be a more accurate representation of typical rental prices.

Part (c):

To determine the sampling distribution of median rental prices for random samples of 50 one-bedroom apartments from this population, Emma would need to obtain every possible sample of 50 one-bedroom apartments from this website and compute the median of each sample. The collection of all possible sample medians is the theoretical sampling distribution for sample median.

Part (d):

(i) $(0.05)(15,000) = 750$ and $(0.95)(15,000) = 14,250$. The 5th percentile is a value, say $x_{0.05}$, such that at least 750 values in the table are less than or equal to $x_{0.05}$ and at least 14,250 are greater than or equal to $x_{0.05}$. Cumulate frequencies starting with the smallest sample median listed in the table and going toward the largest (going down columns) until you first reach 750 values, to obtain $x_{0.05} = \$2,500$.

(ii) Similarly, $x_{0.95} = \$2,950$.

Part (e):

The percentage of bootstrap medians between (and including) the values found in part (d) for the 5th and 95th percentiles is

$$\frac{14,404}{15,000} \times 100\% \approx 96.03\%$$

2019 SCORING GUIDELINES

Question 6 (continued)

Part (f):

From the results in part (d) and part (e), an approximate 96 percent confidence interval for the median rental price of all one-bedroom apartments listed on this website for this city is (\$2,500, \$2,950). We are approximately 96 percent confident that the median rental price of all one-bedroom apartments listed on this website for this city is between \$2,500 and \$2,950.

Scoring

This question is scored in four sections. Section 1 consists of parts (a) and (b), section 2 consists of part (c), section 3 consists of parts (d) and (e), and section 4 consists of part (f). Sections 1, 2, 3, and 4 are each scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. The correct population (listings of one-bedroom apartments on the website) is identified in part (a).
2. In part (b), identifying that using the sample mean instead of the sample median overestimates the typical rental price.
3. The disadvantage of using the sample mean that is reported in part (b) is correctly linked to some feature of the distribution (e.g. skewness) that is evident in the histogram.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: Responses that refer to the mean being larger than the median in a skewed right distribution alone is not sufficient to satisfy component 2.

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies the following two components:
1. Indicates that Emma would need to obtain every possible sample of 50 one-bedroom apartments.
2. Indicates that Emma would need to compute the median rental price for each sample.

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Section 3 is scored as follows:

Essentially correct (E) if the response satisfies the following two components:
1. Correct values for the 5th percentile and the 95th percentile are reported in part (d).
2. The correct percentage of bootstrap samples that produced sample medians at or between the two values, if they are plausible, reported in part (d) is reported in part (e).

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Note: Plausible values for part (d) will be considered values between 2,345 and 3,062.5.

2019 SCORING GUIDELINES

Question 6 (continued)

Section 4 is scored as follows:

Essentially correct if the response in part (f) satisfies the following three components:

1. Uses \$2500 and \$2950 or the values of the percentiles reported in part (d) as the endpoints of the confidence interval.
2. Indicates an approximate 90 or 96 percent level of confidence or a level of confidence consistent with part (e).
3. Makes a correct statement in context indicating that the confidence interval is for the median.

Partially correct if the response satisfies only two of the three components.

Incorrect if the response does not satisfy the criteria for E or P.

Note: Since rental prices from the population are discrete values, the true confidence level of the interval from part (d) is unknown. A correctly calculated part (e) is a way to estimate the confidence level; from Emma's sample the confidence level is estimated to be approximately 96 percent. The process described in part (d) for calculating the interval will result in a confidence level of at least 90 percent. For these reasons, confidence levels of either 90 or 96 percent satisfy component 2.

Each essentially correct (E) section counts as 1 point, and a partially correct (P) section counts as ½ point.

4 **Complete Response**

3 **Substantial Response**

2 **Developing Response**

1 **Minimal Response**

If a response is between two scores (for example, 2½ points), use a holistic approach to decide whether to score up or down depending on the strength of the response and communication.

2018

AP CollegeBoard

AP Statistics

Scoring Guidelines

AP® STATISTICS
2018 SCORING GUIDELINES

Question 1

Intent of Question

The primary goals of this question were to assess a student's ability to (1) identify various values in regression computer output; (2) interpret the intercept of a regression line in context; (3) interpret the coefficient of determination (r^2) in context; and (4) identify an outlier from a scatterplot.

Solution

Part (a):

The estimate of the intercept is 72.95. It is estimated that the average time to finish checkout if there are no other customers in line is 72.95 seconds.

Part (b):

The coefficient of determination is $r^2 = 73.33\%$. This value indicates that 73.33% of the variability in the times it takes customers to finish checkout, including time waiting in line, can be explained by knowing how many customers are in line in front of the selected customer.

Part (c):

The outlier is the point with $x = 3$ and y close to 0. This point is considered an outlier because the combination of x and y values differs from the pattern of the rest of the data. Specifically, the value of y (time to finish checkout) is much lower than would be expected when there are $x = 3$ customers in line in front of the selected customer, given the remaining data.

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. Correctly identifies 72.95 as the intercept.
2. Communicates the concept of a y-intercept in a context that includes both time and zero customers.
3. Indicates that the value of the intercept is a prediction by using language such as "predicted," "estimated," or "average" value of y.

Partially correct (P) if the response includes only two of the three components.

Incorrect (I) if the response includes at most one of the three components.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 1 (continued)

Notes:

- Regression equations (such as $\hat{y} = 72.95 + 174.40x$) cannot be used to satisfy identification of the intercept in component 1, unless the intercept is explicitly labeled as such.
- A regression equation cannot be used to satisfy component 3.
- Incorrect regression equations are treated as extraneous and do not affect the scoring of any component.
- A response that interprets 72.95 as a slope does not satisfy components 1 or 2.

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Correctly identifies 73.33% as the coefficient of determination.
2. Provides a correct (possibly generic) interpretation of r^2.
3. Interpretation includes context.

Partially correct (P) if the response satisfies only two of the three components;
OR
if the response satisfies the three components, but reverses the roles of number of customers in line and time to finish checkout in the interpretation.

Incorrect (I) if the response satisfies at most one of the three components.

Notes:

- In component 2 the correct interpretation of the coefficient of determination can take any of several equivalent forms, such as:
  - The percent variability in y that is attributed to the linear relationship between y and x or between x and y.
  - The proportion of the total variability in the dependent variable y that is explained by the independent variable x.
  - The proportion of variation in y that is accounted for by the linear model.
  - The proportionate reduction of total variation of the y values that is associated with the use of the independent variable x.
  - The proportionate reduction in the sum of the squares of vertical deviations obtained by using the least-squares line instead of the naïve prediction of $\overline{y}$.
- In component 2 common *incorrect* interpretations of the coefficient of determination include:
  - The percent variability in the *predicted* y values that is explained by the linear relationship between y and x.
  - The percent variability in the *data* that is explained by the linear relationship between y and x.
  - The percent variability that is explained by the linear relationship between y and x.
  - The percent variability in y that is *on average* explained by the linear relationship between y and x.
- For component 3 context must include mention of time or customers.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 1 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

1. Correctly identifies the outlier.
2. Describes an unusual feature of the identified scatter plot point, relative to the remaining data points, that is sufficient to identify it as the outlier. Examples include:
    - The combination of x and y values is unusual compared to the other points.
    - The value of y is much lower than would be expected (or predicted), given the remaining data.
    - The residual for the point is unusually large relative to the other residuals.

Partially correct (P) if the response satisfies component 1 but does not satisfy component 2.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In the absence of any point being circled on the graph, component 1 can still be satisfied by explicitly referring to the coordinates of the outlier. Valid coordinates for outlier identification must specify an x value of 3 and a y value that is strictly between 0 and 250.
- A response that does not make a comparison to the remaining data points, such as stating the outlier has a large residual or is nowhere near the regression line, does not satisfy component 2.
- A response that makes a comparison to the remaining data points based upon an unusual feature that is *insufficient* for outlier identification, such as stating the point is the only point with that particular y value, does not satisfy component 2.
- In the absence of explicit numerical calculation, a response that appeals to the influence that the outlier has on the regression coefficient estimates or on the sample correlation coefficient does not satisfy component 2.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 1 (continued)

4 **Complete Response**

Three parts essentially correct

3 **Substantial Response**

Two parts essentially correct and one part partially correct

2 **Developing Response**

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 **Minimal Response**

One part essentially correct

OR

No parts essentially correct and two parts partially correct

AP® STATISTICS
2018 SCORING GUIDELINES

Question 2

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate the sample size when given the endpoints of a confidence interval for a proportion; (2) explain how bias could be present in a particular survey method; and (3) estimate a proportion from sample data collected using a method designed to decrease bias.

Solution

Part (a):

Using the standard formula for a confidence interval for one proportion, the interval (0.584 to 0.816) is found as follows. $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p} = \frac{0.584 + 0.816}{2} = 0.7,$ the margin of error is $0.816 - 0.7 = 0.116,$ and $z^* = 1.96.$

Solving $1.96\sqrt{\frac{0.7(1-0.7)}{n}} = 0.116$ yields $n = \frac{(1.96)^2(0.7)(1-0.7)}{(0.116)^2} \approx 59.95.$ The sample size was 60.

Part (b):

Bias might have been introduced because students responded directly to the environmental science teacher. Because the students would know that an environmental science teacher cares about the environment, they might say yes when they actually don't recycle. This would result in a point estimate that is greater than the proportion of all students who would respond yes to the question.

Part (c):

(i) The expected number is $(300)\left(\frac{1}{2}\right) = 150.$

(ii) The point estimate is based on expecting 150 students to be required to say no and 150 students to truthfully answer the question. Of the 213 answers of no, we expect that $213 - 150 = 63$ were from students who truthfully answered the question. That means we expect that the remaining $150 - 63 = 87$ students truthfully answered the question and responded yes. So the point estimate for the proportion of all students at the high school who would respond yes to the question is $\frac{87}{150} = 0.58.$

AP® STATISTICS
2018 SCORING GUIDELINES

Question 2 (continued)

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following five components:

1. Uses a standard error in the form $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p}$ is between 0 and 1.
2. Shows evidence that $\hat{p} = 0.7$ was correctly used in the standard error.
3. Shows evidence that 0.116 was correctly used as the margin of error in the calculation.
4. Shows evidence that $z^* = 1.96$ was correctly used as the critical value in the calculation.
5. Includes a single, positive whole-number answer.

Partially correct (P) if the response satisfies only three or four of the five components.

Incorrect (I) if the response satisfies at most two of the five components.

Notes:

- Using an equation in the form $n = \frac{z^2\hat{p}(1-\hat{p})}{\text{MOE}^2}$ satisfies component 1.
- A value of 0.21 in the numerator of the standard error implies that $\hat{p} = 0.7$ was correctly used in the standard error and satisfies component 2.
- An equation such as $0.816 = 0.7 + \text{MOE}$ implies that 0.116 was correctly used for the margin of error and satisfies component 3.
- Statements that suggest a whole-number answer is approximate (such as, "about 60" or "$\approx$ 60") satisfy component 5.
- Algebraic work between the set-up and final answer does not need to be shown to satisfy component 5.
- When calculating the values 0.7, 0.116, or 1.96, ignore minor arithmetic errors or transcription errors if they can be identified by the work shown.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 2 (continued)

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Explains why the responses to the survey might differ from the truth about student recycling in this context (for example, the survey was not anonymous, the question was asked by an authority figure).
2. Explains how the responses to the survey might differ from the truth about student recycling (for example, "students might say yes when they actually don't recycle," "students lie and say yes," "students don't recycle but lie to the teacher").
3. Describes the effect of the bias on the point estimate (or the proportion, percentage, number of yes responses in the sample) and doesn't contradict the bias described.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response satisfies at most one of the three components.

Notes:

- To satisfy component 1 the response must provide a reason that is based on a bias created by the teacher asking students in person. For example, a response that addresses the wording of the question, voluntary response, or sampling variability does not satisfy component 1.
- To satisfy component 2 the response needs to explicitly contrast what the students say with what they do.
- Evidence used to address component 3 cannot also be used to address component 2. For example, a response that says "Students might lie, producing an estimate that is too high" addresses the effect of the bias on the point estimate but should not be combined with the statement about students lying to infer that students do not actually recycle. However, a response that says "Students may lie and say yes, producing an estimate that is too high" satisfies both components 2 and 3.
- If the response is clearly about the population proportion and not about the point estimate, component 3 cannot be satisfied.
- Statements such as "the interval will be too high" do not satisfy component 3 because they don't specifically address the point estimate.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 2 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response gives an answer of 150 in (c-i) and gives an answer of 0.58 (or equivalent) in (c-ii).

Partially correct (P) if the response gives an answer of 150 in (c-i) and gives an answer of 0.42 (or equivalent) in (c-ii);
OR
if the response does not give an answer of 150 in (c-i) but gives an answer of 0.58 (or equivalent) with supporting work in (c-ii).

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In part (c-i) the answer must be a single number. Responses such as "at least 150" or "147–153" are incorrect. However, responses such as "about 150" or "$\approx$ 150" are acceptable.
- In part (c-ii) the proportion can be described verbally (e.g., "87 out of 150").
- In part (c-ii) if the response clearly indicates that 0.58 (or 0.42) is the population proportion, lower the overall score in part (c) by one level (that is, from E to P, or from P to I). Using probability notation such as P(yes) does not clearly indicate a population proportion.
- In part (c-ii) if the response includes a point estimate of 0.58 or 0.42 but uses a confidence interval as the final answer, lower the overall score in part (c) by one level (that is, from E to P, or from P to I).
- If the answer is incorrect in part (c-i) and the answer in part (c-ii) uses numerator = 87 and denominator = 300 – answer to (c-i), the response should be scored P.

2018 SCORING GUIDELINES

Question 2 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and one or two parts partially correct

AP® STATISTICS
2018 SCORING GUIDELINES

Question 3

Intent of Question

The primary goals of this question were to assess a student's ability to (1) compute a probability based on a weighted mixture of two populations; (2) compute a conditional probability; and (3) recognize a binomial random variable and compute the probability associated with it.

Solution

Part (a):

Let L denote left-handed, M denote multiple birth, and S denote single birth.
The probability that a randomly selected child born in the region is left-handed is:

$$P(L) = P(M)P(L \mid M) + P(S)P(L \mid S) = (0.035)(0.22) + (0.965)(0.11) = 0.0077 + 0.10615 = 0.11385.$$

Part (b):

From part (a), $P(L) = 0.11385$. Therefore,

$$P(M \mid L) = \frac{P(L \text{ and } M)}{P(L)} = \frac{(0.035)(0.22)}{0.11385} = \frac{0.0077}{0.11385} \approx 0.0676.$$

Part (c):

Let X represent the number of children who are left-handed in a random sample of 20 children from the region. X has a binomial distribution with $n = 20$ and $p = 0.11385$ (found in part (a)). Using the binomial distribution,

$$\begin{aligned} &P(X \geq 3) = 1 - P(X \leq 2) \\ &= 1 - \binom{20}{0}(0.11385)^0(0.88615)^{20} - \binom{20}{1}(0.11385)^1(0.88615)^{19} - \binom{20}{2}(0.11385)^2(0.88615)^{18} \\ &\approx 1 - 0.598 \approx 0.402. \end{aligned}$$

AP® STATISTICS
2018 SCORING GUIDELINES

Question 3 (continued)

Scoring

Parts (a), (b), and (c) each scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the probability is computed correctly, *AND* work is shown that includes correct numerical values using a formula, end results from a tree diagram, or some other appropriate strategy.

Partially correct (P) if the response provides a reasonable strategy for finding the probability, such as a formula or tree diagram, but uses one or more inappropriate values;
OR
if the response gives the correct probability but not enough work is shown to determine how that probability was found.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: A reasonable strategy needs to include summing the results of two multiplications.

Part (b) is scored as follows:

Essentially correct (E) if the probability is computed correctly, with work shown that includes appropriate numerical values for both the numerator and denominator.

Partially correct (P) if the response includes a numerator and denominator in calculating the conditional probability, with one appropriate term (numerator or denominator) and the other inappropriate.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: Appropriate values include incorrectly calculated values from part (a).

AP® STATISTICS
2018 SCORING GUIDELINES

Question 3 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following five components:

1. Uses a calculation based on the binomial distribution to find the probability of the number of children in the sample who are left-handed.
2. Specifies appropriate values for n and p.
3. Uses correct endpoint value for the probability.
4. Uses correct direction to calculate the probability of at least three left-handed children.
5. Correctly calculates a binomial probability consistent with the previous work.

Partially correct (P) if the response satisfies component 1 and only two or three of the other four components;
OR
if components 2, 3, 4, and 5 are met, and the response does not explicitly indicate the binomial distribution is used by name or formula.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- "Appropriate" values include incorrectly calculated values from part (a) or a recalculated probability from part (b).
- An unlabeled numerical value in a calculator statement cannot be used to satisfy a component.
- A response which calculates $P(X \leq 3)$ satisfies component 3 but does not satisfy component 4.
- A normal approximation to the binomial is not appropriate because $np = 20 \times 0.11385 = 2.277 < 5$.
  A response using the normal approximation can score at most P. To earn a score of P, the response must include all of the following:
  - a correct mean and standard deviation based on the binomial parameters
  - clear indication of boundary and direction with a z-score or diagram
  - the probability computed correctly

Notes for all parts:

- If the resulting probability or part of the calculation of the probability uses a value that is not between 0 and 1, inclusive, the score for that part is lowered by one level (that is, from E to P, or from P to I).
- An arithmetic or transcription error in a response can be ignored if correct work is shown. For example, $0.0077 + 0.10615 = 0.1385$.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 3 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

AP® STATISTICS
2018 SCORING GUIDELINES

Question 4

Intent of Question

The primary goals of this question were to assess a student's ability to (1) determine whether a cause-and-effect conclusion can be made based on how a study was conducted and (2) set up, perform, and interpret the results of a hypothesis test, in the context of the problem.

Solution

Part (a):

Yes, it would be reasonable to conclude that the new procedure causes a reduction in recovery time, for patients similar to those in the study. The patients in the study were randomly assigned to the two procedures, which reduces the chance that confounding variables will affect the results. Therefore the statistically significant reduction in mean recovery time can be attributed to the new procedure being superior to the standard procedure.

Part (b):

Step 1: State a correct pair of hypotheses.

Let μ_S represent the mean recovery time among all patients similar to those in the study if they were to receive the standard treatment.

Let μ_N represent the mean recovery time among all patients similar to those in the study if they were to receive the new treatment.

The hypotheses to be tested are $H_0 : \mu_S = \mu_N$ versus $H_a : \mu_S > \mu_N$.

Step 2: Identify a correct test procedure (by name or by formula) and check appropriate conditions.

The appropriate procedure is a two-sample *t*-test for a difference between means.

Because this is an experiment, the first condition is that subjects were randomly assigned to one treatment group or the other. In this case the condition is satisfied because we were told that the subjects were randomly assigned to either the standard or new procedure.

The second condition is that the recovery times of the two populations are normally distributed or the sample sizes are sufficiently large to presume that the distribution of the difference in the sample means is approximately normal. In this case the condition is met because the sample sizes of 110 and 100 are both sufficiently large.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 4 (continued)

Step 3: Correct mechanics, including the value of the test statistic, degrees of freedom, and *p*-value (or rejection region).

The test statistic is $t = \dfrac{\bar{x}_S - \bar{x}_N}{\sqrt{\dfrac{s_S^2}{n_S} + \dfrac{s_N^2}{n_N}}} = \dfrac{217 - 186}{\sqrt{\dfrac{34^2}{110} + \dfrac{29^2}{100}}} \approx 7.13$.

The *p*-value is the area greater than 7.13 for a *t*-distribution with $df = 207.18,$ which is essentially 0 $\left(8.36 \times 10^{-12}\right)$.

Step 4: State a correct conclusion in the context of the problem, using the result of the statistical test.

Because the *p*-value is very small, we have sufficient evidence to conclude that for patients similar to the ones in the study, those receiving the new procedure would have less recovery time, on average, than those receiving the standard procedure.

Scoring

This question is scored in three sections. Section 1 consists of part (a); section 2 consists of step 1, step 2, and the test statistic in step 3 in part (b); and section 3 consists of the *p*-value in step 3 and step 4 in part (b). Sections 1, 2, and 3 are each scored essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. Correctly states that it is reasonable to make a causal conclusion.
2. Justifies the causal conclusion based on random assignment of patients to procedures (or procedures to patients);
   OR
   justifies the causal conclusion by stating that a randomized experiment was conducted.
3. Includes the context of the situation.

Partially correct (P) if the response satisfies component 1 *AND* provides WEAK justification of the causal conclusion by stating that there was random assignment or a randomized experiment was conducted, but with no context;
OR
by stating that an experiment was conducted or there was assignment (without the word "randomized") *AND* the response includes context of the situation;
OR
by stating that the study design reduces the chance of confounding variables or balances the effects of uncontrolled variables across both groups in context without explicitly referring to the random assignment.

Incorrect (I) if the response does not meet the criteria for E or P.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 4 (continued)

Notes:

- If the response states that it is *not* reasonable to make a causal conclusion because the result could have been due to random chance *AND* explains that there is evidence for a causal conclusion based on random assignment of patients to procedures or by stating that a randomized experiment was conducted, then the response is scored E.
- If the response discusses aspects of an experiment other than random assignment (such as, control, replication, or large samples), then those aspects are considered extraneous and the response can be scored E unless those aspects are incorrect for this study (such as, blocking is a requirement, or the study used blocking, or the study used a placebo) in which case the score should be lowered one level (that is, from E to P, or from P to I).
- If the response correctly states in context that it is reasonable to make a causal conclusion but includes incorrect or contradictory justification (such as, random selection of patients), then the response is scored I.

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

1. Parameters are defined correctly.
2. Hypotheses imply equality in the null and correct direction in the alternative.
3. Correct test is identified by name or formula.
4. Correct test statistic for a difference in means is calculated.

Partially correct (P) if the response satisfies only two or three of the four components.

Incorrect (I) if the response satisfies at most one of the four components.

Notes:

- If standard symbols are used for the parameters with appropriate group labels (such as, μ_S, μ_N), component 1 is satisfied.
- If the correct test is identified, but the response states an incorrect formula or uses incorrect notation in the formula, component 3 is not satisfied.
- A pooled two-sample *t*-test is acceptable for component 3, but the student must also state and comment on the plausibility of the equal population variances assumption.
- If the response identifies a *z*-test for equal means as the correct test identification, component 3 is not satisfied but component 4 could be satisfied.

Confidence Interval approach:

- If a single two-sample *t*-interval for the difference in means is used, components 3 and 4 can be satisfied. Component 3 is satisfied if the *t*-interval is correctly identified by name or formula. Component 4 is satisfied if the correct interval is calculated. If an alpha level is stated, then an appropriate adjustment to the confidence level must be made because the appropriate test is one-sided.
- If two one-sample *t*-intervals are used, while not a recommended approach, component 3 is not satisfied but component 4 could be satisfied. Component 4 is satisfied if both intervals are calculated correctly.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 4 (continued)

Section 3 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Makes reference to an approximately correct *p*-value that is consistent with the test statistic and alternative hypothesis for a difference in means.
2. Correctly justifies the conclusion based on the size of the *p*-value or the test statistic.
3. Correctly states the conclusion in context.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P or includes a justification not based on the inferential results.

Notes:
Component 1:

- Is satisfied if the response makes reference to a large test statistic without referring to a *p*-value.

Component 2:

- No alpha level is needed to provide justification of the conclusion based on the size of the *p*-value.
- Is satisfied if the response states the *p*-value without reference to size, but it is contiguous to the conclusion and clearly indicates a continuous train of thought.
- A correct interpretation of the *p*-value with a complete explanation that obtaining a test statistic at least this extreme is unlikely due to chance alone is considered justification based on the size of the *p*-value.
- If an incorrect interpretation of the *p*-value is given, the score is lowered one level (that is, from E to P, or from P to I).
- A decision about the null hypothesis (reject H_0 or fail to reject H_0) is not required, but if an incorrect decision is stated based on the given *p*-value then component 2 is not satisfied.
- If a rejection region approach is used, a reasonable critical value replaces the *p*-value.

Component 3:

- A correct conclusion must be related to the alternative hypothesis in order to satisfy component 3.
- The following responses do not satisfy component 3:
  - States or implies that the null hypothesis is *accepted*
  - States or implies that the alternative hypothesis has been *proven*
  - States the conclusion in past tense (unless the response did not satisfy a component of section 2 for the use of past tense)

AP® STATISTICS
2018 SCORING GUIDELINES

Question 4 (continued)

Confidence Interval approach:

- If a single two-sample *t*-interval for the difference in means is used:
  - Component 1 is satisfied if the response indicates that zero is either included or not included in the calculated interval.
  - Component 2 is satisfied if the response indicates that the bounds are either both above or both below zero (consistent with alternative hypothesis) and uses that as justification for the conclusion.
  - Component 3 is satisfied if the conclusion is stated in context.
- If two one-sample *t*-intervals are used (which is not recommended) the response is scored at most P if all three components are satisfied, otherwise scored I:
  - Component 1 is satisfied if the response states that the intervals do not overlap.
  - Component 2 is satisfied if the conclusion indicates that the confidence interval for the new procedure lies below the confidence interval for the standard procedure.
  - Component 3 is satisfied if the conclusion is stated in context.

Note: If the three sections of the response are scored as E, to earn a score of 4 as a complete response, both conditions in step 2 must be correctly stated and justified. Additional condition(s) inappropriate for a two-sample *t*-test must not be stated. Otherwise, the response earns a score of 3 a substantial response.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 4 (continued)

4 Complete Response

Three sections essentially correct with conditions for inference

3 Substantial Response

Three sections essentially correct without conditions for inference

OR

Two sections essentially correct and one section partially correct

2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

1 Minimal Response

One section essentially correct

OR

No sections essentially correct and one or two sections partially correct

AP® STATISTICS
2018 SCORING GUIDELINES

Question 5

Intent of Question

The primary goals of this question were to assess a student's ability to (1) determine which of two histograms represents data with a larger median; (2) calculate the mean of a combined data set when the separate means and sample sizes are known; and (3) calculate the probability that an individual randomly chosen from a finite population will have a value within one standard deviation of the mean, when provided with values for the mean, standard deviation, and all members of the population.

Solution

Part (a):

The median teaching year for High School A is any value with 100 data values at or below it and 100 data values at or above it. The median teaching year for High School B is the 111th value in the ordered list of values. For High School A the median is in the interval that starts at 7 and ends just before 10, because there are only 94 data values below 7 and 106 data values of at least 7. Therefore the median cannot be less than 7. For High School B the median is in the interval that starts at 4 and ends just before 7 because there are more than half (113) of the data values less than 7. Therefore the median must be less than 7. So High School A must be the one with a median of 7, and High School B must be the one with a median of 6.

Another way to determine which school has the median of 7 is to notice that the distribution for High School B is highly skewed to the right, whereas the distribution for High School A is bimodal with a few possible outliers on the right. A distribution that is highly right-skewed is likely to have a substantially larger mean than median. The mean of both distributions is given as 8.2 years, so it makes sense that the highly right-skewed distribution (High School B) is the one with the bigger gap between the mean and median and, therefore, the one with the lower median of 6.

Part (b):

The mean for the original 200 teachers was given as 8.2 years, and the mean for the additional 18 teachers is 2.5 years. Therefore the mean for the combined data set is:

$$\frac{(200)(8.2) + (18)(2.5)}{200 + 18} = \frac{1{,}640 + 45}{218} \approx 7.73 \text{ years.}$$

Part (c):

The interval mean plus or minus 1 standard deviation on either side of the mean is 8.2 ± 7.2, or from 1.0 year to 15.4 years. Because teaching year is recorded as an integer, the interval includes teaching years 1 to 15. The number of teachers in that interval can be found by adding the heights of the five bars in the histogram for the intervals from 1 to 16, which includes $79 + 34 + 28 + 29 + 19 = 189$. Therefore the probability is $\frac{189}{221} \approx 0.8552$.

2018 SCORING GUIDELINES

Question 5 (continued)

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. States that the median is 6 for High School B and the median is 7 for High School A.
2. Provides a reasonable explanation of how the decision was made.
3. Provides the definition of the median or explicitly applies the definition of a median as a criterion in reaching their decision.

OR

Essentially correct (E) if the response satisfies the following three components:
1. States that the median is 6 for High School B and the median is 7 for High School A.
2. States that High School B shows a skewed distribution (or High School A shows a less skewed distribution).
3. Provides a reasonable explanation of how the more skewed distribution (High School B) would be the one with a larger separation between the mean and median.

Partially correct (P) if the response satisfies the first component and only one of the other two components required for E.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: An incorrect statistical statement in the response will result in E being lowered to P, but not P being lowered to I. For example,
- If either distribution is described as left skewed, normal, or approximately normal;
- If the discussion would indicate a median different than 7 for High School A or a median different than 6 for High School B.

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:
1. The correct answer that the mean is 7.73.
2. Enough work to show that the answer was obtained as a weighted average of the two individual means.

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not satisfy the requirements for E or P.

2018 SCORING GUIDELINES

Question 5 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Calculates that the appropriate interval is 1 to 15.4 or 1 to 15 teaching years.
2. Correctly sums the counts of data values in the numerator based on the intervals provided.
3. Computes the probability using 221 as the denominator.

Partially correct (P) if the response satisfies only two of the three components;
OR
if the response reports the correct probability (0.8552) without supporting work.

Incorrect (I) if the response satisfies at most one of the three components.

Notes:

- If the response attempts to use the Empirical Rule or normal distribution to provide the desired probability, the response is scored I.
- If an incorrect count is shown in component 2, for instance by including the interval from 16 to 19, then component 3 is satisfied if that incorrect count is divided by 221 to find the reported probability.
- It is acceptable if the count is slightly off because of difficulty reading the exact heights of the bars in the histogram.
- If only one of component 2 or component 3 is missing, but the correct probability (0.8552) is reported, the response can be scored E.
- If the response recognizes that all values in the histogram bins up to 16 fall within one standard deviation of the mean and reports the interval as 1 to 16, component 1 is satisfied.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 5 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

OR

Part (a) essentially correct and two parts partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

Part (b) or part (c) essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and one or two parts partially correct

AP® STATISTICS
2018 SCORING GUIDELINES

Question 6

Intent of Question

The primary goals of this question were to assess a student's ability to (1) describe what constitutes a Type II error for a specific hypothesis test; (2) specify a rejection region in terms of values of the sample mean; (3) compute the power of a test for a specific value in the alternative hypothesis; (4) recognize the definition of power; and (5) understand the impact of increasing the sample size on the power of a test.

Solution

Part (a):

A Type II error occurs when the alternative hypothesis is true, but the null hypothesis is not rejected. In this situation a Type II error would happen if the mean systolic blood pressure of the population of employees is greater than 122 mmHg, but the null hypothesis that it is 122 mmHg is not rejected. In other words a Type II error would happen if the mean blood pressure for the population of employees is higher than the national average, but the test does not conclude that it is higher.

Part (b):

The test is one-sided and the standard deviation is known, so the null hypothesis will be rejected if the test statistic $z = \dfrac{\overline{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} > 1.645$. With $\mu = 122,\ \sigma = 15,$ and $n = 100,$ we get $\dfrac{\overline{x} - 122}{1.5} > 1.645$. Therefore, $\overline{x} > 124.4675$.

Part (c):

If the actual population mean is 125, with $\sigma = 15$ and $n = 100,$ then the sampling distribution of $\overline{x}$ is approximately normal with mean of 125 and standard deviation $\dfrac{15}{\sqrt{100}} = 1.5$. Therefore,

$$P(\overline{x} > 124.4675) = P\left(\frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} > \frac{124.4675 - 125}{1.5}\right) = P(z > -0.355) = 0.64.$$

Part (d):

The probability found in part (c) is called the power of the test.

Part (e):

If the sample size is increased from 100 to something larger, the probability of rejecting the null hypothesis when the population mean is 125 will be higher than it is for a sample of size 100. Intuitively, more data provide a higher probability of a correct conclusion. The technical explanation is that the rejection region will still be $z > 1.645,$ but the sampling distributions of the sample mean will have a smaller standard deviation; therefore, the minimum value of $\overline{x}$ for which we would reject the null hypothesis would be lower and, in return, the probability the null hypothesis is rejected will increase.

AP® STATISTICS
2018 SCORING GUIDELINES

Question 6 (continued)

Scoring

This question is scored in three sections. Section 1 consists of part (a) and part (b), section 2 consists of part (c), and section 3 consists of part (d) and part (e). Sections 1, 2, and 3 are scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

1. Part (a) includes in the description of a Type II error the fact that the alternative hypothesis is true, either generically or in context of the situation.
2. Part (a) includes in the description of a Type II error the fact that the null hypothesis is not rejected, either generically or in context of the situation.
3. Part (b) includes a correct *z*-score for the upper 5 percent tail and indicates the correct direction for the rejection region.
4. Part (b) includes $\mu_{\bar{x}} = 122$, $\sigma_{\bar{x}} = 1.5$, and the resulting $\bar{x}$ value.

Partially correct (P) if the response satisfies only two or three of the four components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- If the response in part (a) does not include context, the number of components satisfied is reduced by one (that is, from four to three, or from three to two, and so on). Context includes a reference to units, to blood pressure, to employees, etc.
- If a response in part (a) is clearly referring to an individual's blood pressure as opposed to the mean blood pressure of all employees, neither components 1 nor 2 are satisfied.

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Recognizes that the null hypothesis will be rejected when $\bar{x} \geq 124.4675$, as found in part (b).
2. Provides the correct sampling distribution for the sample mean when the true mean is 125, including correct values for the mean and standard deviation, either explicitly or by plugging them into the test statistic formula.
3. Provides evidence of using the normal curve and finds the correct probability value.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- Components 1 and 3 can still be satisfied if errors made in finding the rejection region in part (b) are carried into part (c).
- A calculator statement that does not include labels for input values does not satisfy component 2 but may still satisfy components 1 and 3.

2018 SCORING GUIDELINES

Question 6 (continued)

Section 3 is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

1. Part (d) specifies power as the name of the probability.
2. Part (e) correctly states that the probability would be greater.
3. Part (e) correctly implies that the standard deviation of the sampling distribution decreases, either explicitly or by substituting values into a formula.
4. Part (e) indicates the minimum value of $\bar{x}$ for which the null hypothesis is rejected decreases, either explicitly or by substituting values into a formula.

Note: Component 4 can still be satisfied if a response indicates that a maximum value of $\bar{x}$ for which the null hypothesis is rejected increases if this direction is consistent with answers in parts (b) and (c).

Partially correct (P) if the response satisfies only two or three of the four components.

Incorrect (I) if the response does not meet the criteria for E or P.

2018 SCORING GUIDELINES

Question 6 (continued)

4 Complete Response

Three sections essentially correct

3 Substantial Response

Two sections essentially correct and one section partially correct

2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

1 Minimal Response

One section essentially correct

OR

No sections essentially correct and two sections partially correct

OR

Sections 1 and 2 incorrect, and section 3 partially correct with exactly three of the four components satisfied

OR

Section 1 partially correct with exactly three of the four components satisfied, and sections 2 and 3 incorrect

AP® Statistics
2016 Scoring Guidelines

SCORING GUIDELINES

Question

Intent of Question

The primary goals of this question were to assess a student's ability to (1) describe the distribution of a quantitative variable based on a histogram and (2) determine the effect of changing one data value on mean and the median.

Solution

Part (a):

The distribution of Robin's tip amounts is skewed to the right. There is a gap between the largest amount (in the \$20 to \$22.50 interval) and the second largest tip amount (in the \$12.50 to \$15 interval), and the largest tip amount appears to be an outlier. The median tip amount is between \$2.50 and \$5.00. Robin's tip amounts vary from a minimum of between \$0 and \$2.50 to a maximum of between \$20.00 and \$22.50. About 78 percent of the tip amounts are between \$0 and \$5.

Part (b):

ɘ mean: If the \$8 tip had been \$18, the mean would increase by \$10 divided by 60, or $\$\frac{1}{6}$, or about 17 cents.

ɘ median: If the \$8 tip had been \$18, the median would not change because the current median is between \$2.50 and \$5.00, and both \$8 and \$18 are greater than that.

Scoring

Parts (a) and (b) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response includes reasonable comments on the following five components:

1. Shape (skewed right)
2. Outlier (at least one) *OR* gap (one tip amount greater than \$20, next highest at most \$15)
3. Center between \$2.50 and \$5.00 (median) or between \$2.62 and \$5.13 (mean)
4. Variability, by noting that the tip amounts vary from about \$0 to at most \$22.50, or that a majority of tip amounts are between \$0 and a value greater than or equal to \$5, or by providing a correct numerical approximation of a measure of variability
5. Context (tip amounts)

Partially correct (P) if the response includes only three or four of the five components.

Incorrect (I) if the response includes at most two of the five components.

SCORING GUIDELINES

Question 1 (continued)

Part (b) is scored as follows:

Essentially correct (E) if the response includes the following four components:
1. Comments that the mean will increase
2. Correctly justifies why the mean will increase
3. Comments that the median will not change
4. Correctly justifies why the median will not change

Partially correct (P) if the response includes only two or three of the four components.

Incorrect (I) if the response includes at most one of the four components.

4 Complete Response

Both parts essentially correct

3 Substantial Response

One part essentially correct and one part partially correct

2 Developing Response

One part essentially correct and one part incorrect

OR

Both parts partially correct

1 Minimal Response

One part partially correct and one part incorrect

SCORING GUIDELINES

Question

Intent of Question

The primary goals of this question were to assess a student's ability to (1) identify, set up, perform, and interpret the results of an appropriate hypothesis test to address a particular question and (2) assess the effectiveness of treatments in a controlled experiment.

Solution

Part (a):

Step 1: States a correct pair of hypotheses.

H_0 : The proportion of children who would choose each snack is the same regardless of which type of ad is viewed.
H_a : The proportion of children who would choose each snack differs based on which type of ad is viewed.

Step 2: Identifies a correct test procedure (by name or formula) and checks appropriate conditions.

The appropriate procedure is a chi-square test of homogeneity.

The conditions for this test are satisfied because (1) the question states that the children were randomly assigned to groups, and (2) expected counts for the six cells of the table are all at least 5, as seen in the following table that lists expected counts beside observed counts.

Group	Choco-Zuties	Apple-Zuties	Total
A	21 (18.67)	4 (6.33)	25
B	13 (18.67)	12 (6.33)	25
C	22 (18.67)	3 (6.33)	25
Total	56	19	75

Step 3: Calculates the appropriate test statistic and p-value.

The test statistic is calculated as $\chi^2 = \sum \frac{(O-E)^2}{E}$, which is

$$\chi^2 \approx 0.292 + 0.860 + 1.720 + 5.070 + 0.595 + 1.754 \approx 10.291.$$

The p-value is $P(\chi^2_{df2} \geq 10.291) \approx 0.006$.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test

Because the value is very small (for instance, much smaller than $\alpha = 0.05$ we reject the null hypothesis at the 05 level (and at the 01 level). The data provide convincing statistical evidence that the proportions who would choose each snack differ based on which ad is viewed.

SCORING GUIDELINES

Question 2 (continued)

Part (b):

When neither ad was viewed, $\frac{22}{25}$ or 88 percent of the children chose Choco-Zuties whereas only
)ercent chose Apple-Zuties.

When the Choco-Zuties ad was viewed, 84 percent of the children chose Choco-Zuties, which was very similar to the percentage that chose them without viewing any ad. So watching the Choco Zuties ad did not affect the snack choice very much.

:n the Apple-Zuties ad was viewed, only $\frac{13}{25}$ or 52 percent of the children chose Choco-Zuties, and
)ercent chose Apple-Zuties. Watching the Apple-Zuties ad seemed to increase the proportion of children choosing Apple-Zuties.

Scoring

This question is scored in four sections. Section 1 consists of steps 1 and 2 in part (a); section 2 consists of step 3 in part (a); section 3 consists of step 4 in part (a); and section 4 consists of part (b). Sections 1, 2, 3 and 4 are scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response includes the following three components:
1. Both hypotheses are stated correctly with at least one in context.
2. Identifies the correct test procedure (by name or by formula).
3. The technical conditions are checked (all expected counts are greater than or equal to 5).

Partially correct (P) if the response includes only two of the three components.

Incorrect (I) if the response correctly includes at most one of the three components.

Notes for component 1:
- It is acceptable if hypotheses are stated in terms of association/independence instead of proportions.
- Context must be included in at least one of the hypotheses, but is not required to be in both.
- If the hypotheses contain language that suggests that the response refers to the sample data, the component is not satisfied.

Note for component 2: It is acceptable if the test is identified as a test of independence instead of a test of homogeneity.

Notes for component 3:
- The random assignment condition was stated so need not be explicitly mentioned.
- Stating the expected count condition is met is not sufficient for component 3. The condition must be checked by reporting the expected counts and either:
    - noting that all are greater than or equal to 5;

SCORING GUIDELINES

Question 2 (continued)

- noting that all are greater than or equal to 1 and at most 20 percent are less than 5.

- Noting that the smallest expected count is 6.33 and that it is greater than or equal to 5 is sufficient to satisfy this component.
- Component 3 is not satisfied if the expected counts are reported as integers.
- This component is not satisfied if the response includes any of the following inappropriate conditions:
  - The response implies that a random sample was taken, e.g., "SRS – check."
  - The response refers to independence of groups or independence of ads as a required condition.
  - The response indicates that a sample size greater than 30 ensures normality or the response implies normality as a condition.
- A response stating that children are independent can be ignored in the scoring of this component.

Section 2 is scored as follows:

Essentially correct (E) if the response includes the following three components:
1. Correct chi-square test-statistic
2. Correct degrees of freedom
3. Correct *p*-value

Partially correct (P) if the response includes only two of the three components.

Incorrect (I) if the response correctly includes at most one of the three components.

Notes:
- If the response makes an error in one calculation, future calculations are considered correct if they follow correctly from the initial miscalculation.
- A chi-square critical value approach is acceptable: The critical value for $\alpha = 0.05$ is $\chi^2 = 5.99$.

Section 3 is scored as follows:

Essentially correct (E) if the response provides a correct conclusion about the alternative hypothesis context, *AND* provides justification based on linkage between the *p*-value and conclusion.

Partially correct (P) if the response provides a correct conclusion in context, but without justification based on linkage to the *p*-value;
OR
if the response provides a correct conclusion, with linkage to the *p*-value, but not in context;
OR
if the response provides a correct decision stated in terms of the null hypothesis in context, with linkage to the *p*-value, but no conclusion is made about the alternative hypothesis.

Incorrect (I) if the response does not meet the criteria for E or P.

SCORING GUIDELINES

Question 2 (continued)

Notes:

- If the conclusion is consistent with the *p*-value from section 2, and also in context with justification based on linkage to the *p*-value, section 3 is scored as E.
- If no alpha level is given, the solution must be explicit about the linkage by giving a correct interpretation of the *p*-value or by explaining how the conclusion follows from the *p*-value such as saying: "Because the *p*-value is small, we reject the null hypothesis." or "Because the alue is large, we do not reject the null hypothesis."
- If a conclusion contains language that suggests that the response refers to the sample data, conclusion component is not correct, unless the same error occurred in the statement of hypotheses in section 1.

Section 4 is scored as follows:

Essentially correct (E) if the response, in context, concludes that ad type A had little effect id type B had an effect, both supported by the observed proportions (or counts) from the table.

Partially correct (P) if the response correctly concludes that ad type A had little effect, and ad type B had an effect but does not provide correct numerical justification;
OR
if the response compares all of the proportions (or counts) as required, but without correctly concluding ad effectiveness;
OR
if the response correctly describes the effect of one of the ad types A or B (in context, with correct numerical justification) but not the other;
OR
if the response compares all of the proportions (or counts) as required, but not in context.

Incorrect (I) if the response does not meet the criteria for E or P.

Each essentially correct (E) section counts as 1 point, and a partially correct (P) section counts as ½ point.

4 Complete Response

3 Substantial Response

2 Developing Response

1 Minimal Response

If a response is between two scores (for example, 2½ points), use a holistic approach to decide whether score up or down, depending on the strength of the response and communication.

SCORING GUIDELINES

Question

Intent of Question

The primary goals of this question were to assess a student's ability to (1) identify explanatory and response variables from the description of a research study; (2) indicate and justify whether a study is observational or experimental; and (3) explain what confounding means in the context of a particular study with a specific confounding variable.

Solution

Part (a):

The explanatory variable is the person's degree of cigarette smoking. The response variable is whether the person develops Alzheimer's disease during the course of the study.

Part (b):

This is an observational study because the people in the study were not assigned to a certain degree of cigarette smoking. Rather, the degree of cigarette smoking for each person was passively observed and recorded, not manipulated by the researchers.

Part (c):

nfounding variable is one that is related to the explanatory variable and possibly influences response variable. In this case it seems plausible that people who exercise more regularly might be more health conscious, therefore, less likely to smoke cigarettes than people who do not exercise regularly. Similarly, it's possible that people who exercise more regularly are less likely to develop Alzheimer's disease than people who do not exercise regularly. If both of these relationships turn out to be true, then smoking cigarettes would be associated with developing Alzheimer's disease due to the association of both variables with exercise, even if there were no cause-and-effect relationship between smoking cigarettes and developing Alzheimer's disease.

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if both variables are described correctly. A correct description includes some degree of status of the variables, such as smoking versus not smoking and developing Alzeimer' versus not developing Alzheimer's.

Partially correct (P) if one variable is described correctly and one is not described correctly;
OR
ıeither variable is described correctly but smoking is mentioned as the explanatory variable, and Alzheimer's is mentioned as the response variable;
OR
the explanatory and response variables are correctly described but are interchanged.

Incorrect if the response does not meet the criteria for E or P.

SCORING GUIDELINES

Question 3 (continued)

Note: Describing the variables only as smoking and Alzheimer's without a reference to levels/status not sufficient for E. But making the connection between smoking and explanatory and between Alzheimer's and response is sufficient for P.

Part (b) is scored as follows:

Essentially correct (E) if the response selects the correct type of study (observational) and provides the correct explanation that smoking status was not assigned *OR* that smoking status was only observed.

Partially correct (P) if the response selects the correct type of study (observational) and provides correct explanation but does not refer to smoking status as the variable that would have been assigned had it been an experiment; for instance, by simply stating that treatment was not assigned;
OR
if the response selects the correct type of study (observational) and provides an explanation that says that smoking is observed (or that smoking status and Alzheimer's are observed) without a modifier for observed (such as, only, just, merely, simply) *AND* without indicating that treatments were not assigned.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- A response that states the explanatory variable was not assigned without naming smoking status is sufficient for E if the explanatory variable is correctly defined in part (a).
- A response that states that smoking status is ONLY observed (or that smoking status and Alzheimer's are ONLY observed) is sufficient for E.
- A response that says that Alzheimer's is observed without mentioning smoking status is scored I.
- A response that provides an incorrect statistical explanation (such as, "the study is observational because an experiment must have a control group") lowers the score in part (b) by one level (from E to P or from P to I).
- If Alzheimer's is given as the explanatory variable and smoking is given as the response variable in part (a), then part (b) should be scored accordingly with the two variables interchanged.
- Any phrase that refers to the "effect of smoking on Alzheimer's" or "the association between smoking and Alzheimer's" (rather than smoking status) should be ignored.

Part (c) is scored as follows:

Essentially correct (E) if the response includes the following two components:
1. Provides a reasonable explanation that exercise status is related to smoking status.
2. States that exercise status might influence whether the person develops Alzheimer's disease.

Partially correct (P) if the response describes only one of the two components;
OR
if the response only describes how smoking and exercise jointly influence whether the person develops Alzheimer's;

SCORING GUIDELINES

Question 3 (continued)

if the response mentions both of the associations in each component, but also includes an incorrect statistical statement. For instance, part (a) is correct, but in part (c), the explanatory and response variables are interchanged.

Incorrect (I) if the response does not meet the criteria for E or P.

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

SCORING GUIDELINES

Question 4

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate a probability using basic probability rules or the geometric distribution; (2) recognize that a probability calculation for independent events does not depend on the previous outcomes of those events; and (3) assess whether a claim about the probability of a single event is reasonable based on a calculated probability of a series of those events.

Solution

Part (a):

If the failure rate for the super igniters is 15 percent, then the probability that each igniter fails is 0.15, and the probability that it does not fail is 0.85. Therefore the probability that the first 30 igniters tested do not fail is $(0.85)^{30} \approx 0.0076$. The solution can also be written as $(1 - 0.15)^{30} \approx 0.0076$.

Part (b):

Given that there are no failures in the first 30 trials, the probability that the first failure occurs on the trial is 0.15, and the probability that it does not occur on the 31st but occurs on the 32nd trial $(0.85)(0.15) = 0.1275$. Therefore the probability that the first failure occurs on the 31st or 32nd super igniter tested is $0.15 + 0.1275 = 0.2775$.

Note that this is equivalent to asking for the probability that the first failure occurs on the first or second trial, which is $0.15 + (0.85)(0.15) = 0.2775$.

Part (c):

The result of the probability calculation in part (a) provides a reason to believe that the failure rate of the super igniters is less than 15 percent. The calculated probability of 0.0076 shows that there is less than a 1 percent chance that 30 or more igniters in a row would not fail if the failure rate was ercent. This probability is smaller than conventional significance levels such as $\alpha = 0.05$ $\alpha = 0.01$, and thus is small enough to make it reasonable to believe that the failure rate of the super igniters is less than 15 percent.

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response gives the correct probability *AND* correct justification.

Partially correct (P) if the response correctly notes that the answer is the probability that there will be 30 successes in 30 attempts, but does not carry out a correct probability calculation;
OR

SCORING GUIDELINES

Question 4 (continued)

if the response defines the random variable X as the trial with the first failure, identifies X as having a geometric distribution with $p = 0.15$, and writes the desired probability as $P(X > 30)$, but does not carry out a correct probability calculation;

OR

if the response defines the random variable X as the number of failures in the first 30 attempts, identifies X as a binomial random variable with $p = 0.15$ and $n = 30$, and writes the desired probability as $P(X = 0)$, but does not carry out a correct probability calculation;

OR

if the response gives the correct probability but, in specifying a geometric or binomial distribution, has an incorrect or incomplete definition of parameters or value(s) of the random variable.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: Justification can be given using the multiplication rule; *OR* by defining X to be the trial with the first failure, recognizing that X has a geometric distribution, and using that information to find $P(X > 30)$; *OR* by defining X to be the number of failures in the first 30 attempts, and then finding $P(X = 0)$ using either probability rules or the binomial distribution with $n = 30$ and $p = 0.15$.

Part (b) is scored as follows:

Essentially correct (E) if the response gives the correct probability AND correct justification.

Partially correct (P) if the response makes a reasonable attempt to calculate a geometric, binomial, or conditional probability, but does not successfully carry out the calculation;

OR

if the response gives the correct probability but, in specifying a geometric or binomial distribution, has an incorrect or incomplete definition of parameters or value(s) of the random variable.

Incorrect (I) if the response finds an incorrect probability resulting from an unreasonable attempt to calculate a geometric, binomial, or conditional probability or otherwise does not meet the criteria for E or P.

Note: Similar to part (a) justification can be given using probability rules; *OR* by stating that X geometric where X is the trial with the first failure, then finding $P(X = 1 \text{ or } X = 2)$; *OR* by stating that s the number of failures in two trials and finding $1 - P(X = 0)$ or $P(X = 1 \text{ or } X = 2)$ using the binomial distribution.

Part (c) is scored as follows:

Essentially correct (E) if the response states that it is reasonable to believe that the failure rate is less than 15 percent *AND* bases this decision on the fact that the probability of 30 consecutive successful launches with a failure rate of 15 percent (that is, answer from part (a)) is small *AND* does so in the context of the situation.

Partially correct (P) if the response otherwise satisfies the criteria for an (E) but does so without any context;

SCORING GUIDELINES

Question 4 (continued)

if the response states a significance level and makes a decision in a context that is appropriate to the given probability in part (a) and the stated significance level but does not explicitly compare the probability and the significance level (no linkage).

Incorrect (I) if the response does not explicitly make a decision about whether it is reasonable to conclude that the failure rate is less than 15 percent (For example: "As seen in Part (a), if the failure rate is 15 percent then the probability of 30 successful launches in a row is very small.");
OR
if the response otherwise does not meet the criteria for E or P.

Notes:

- Justification based on the probability can come by stating a significance level and noting that the probability is smaller than the significance level *OR* by simply stating that the probability of 0.0076 is small *OR* by referring to the expected number of failures (4.5) as being very unlikely because zero failures is more than two standard deviations below 4.5.
- If the response bases the decision on the expected number of failures (4.5) for $n = 30$ and $p = 0.15$ without referencing why zero failures would be considered to be too far below 4.5 to give reason to doubt the stated 15 percent failure rate, the response is scored P.
- If the calculation in part (a) is incorrect, the answer in part (c) needs to be consistent with the answer in part (a), unless the value is recalculated in part (c).

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct
OR
One part essentially correct and one or two parts partially correct
OR
Three parts partially correct

1 Minimal Response

One part essentially correct
OR
No parts essentially correct and two parts partially correct

SCORING GUIDELINES

Question 5

Intent of Question

The primary goals of this question were to assess a student's ability to (1) construct and interpret a confidence interval for a population proportion; (2) explain why one of the conditions for inference necessary; and (3) explain why a suggested procedure for constructing a confidence interval is incorrect.

Solution

Part (a):

The appropriate procedure is a one-sample z-interval for a population proportion. The problem stated the conditions for inference have been met, so they do not need to be checked. A 95 percent confidence interval for the population proportion is given as $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which is

$$0.37 \pm 1.96\sqrt{\frac{(0.37)(0.63)}{1{,}048}} \approx 0.37 \pm 0.03 = (0.34, 0.40).$$

We are 95 percent confident that the population proportion of all adults in the U.S. who would have chosen the economy statement is between 0.34 and 0.40.

Part (b):

The condition is necessary because the formula for the confidence interval relies on the fact that the binomial distribution can be approximated by a normal distribution which then results in the sampling distribution of $\hat{p}$ being approximately normal. The approximation does not work well unless both $n\hat{p}$ and $n(1-\hat{p})$ are at least 10.

Part (c):

The suggested procedure is not appropriate because one of the requirements for using a two-sample interval for a difference between proportions is that the two proportions are based on two independent samples. In the situation described the two proportions come from a single sample and thus are not independent.

Scoring

This question is scored in three sections. Section 1 consists of the mechanics of the confidence interval in part (a), section 2 consists of interpreting the confidence interval in part (a), and section 3 consists of parts (b) and (c).

Section 1 is scored as follows:

Essentially correct (E) if the response includes the following two components:
1. States the correct procedure by name or formula
2. Calculates the confidence interval

Partially correct (P) if the response includes only one of the two components.

Incorrect (I) if the response does not meet the criteria for E or P.

SCORING GUIDELINES

Question 5 (continued)

Notes:

- A formula with correct values is sufficient for component 2.
- Component 2 can never be satisfied if the response contains an incorrect formula.

Section 2 is scored as follows:

Essentially correct (E) if the response includes the following four components for the interval interpretation:

1. Estimates a proportion
2. Infers about the population
3. States 95 percent confidence
4. Includes context

Partially correct (P) if the response satisfies components 1 and 2 *AND* satisfies only one of components 3 or 4;
OR
if the response gives a correct interpretation of the confidence *level* in context without interpreting the specific interval.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- Any indication of an inference to the adults sampled rather than the population does not satisfy component 2 and is scored as I.
- Stating values that are unrealistic as proportions (including blanks) in the interpretation lower the score by one level (from E to P or from P to I).
- When both the interpretation of the interval and the level are given, only the interpretation of the interval is scored.

Section 3 is scored as follows:

Essentially correct (E) if the response includes the following two components:

1. Part (b) states that the condition implies the sampling distribution of $\hat{p}$ is approximately normal *OR* the normal approximation to the binomial distribution is appropriate.
2. Part (c) indicates the procedure is not appropriate because the two proportions come from single sample (dependent) rather than two (independent) samples.

Partially correct (P) if the response includes only one of the two components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- Referring to the sampling distribution needing to be normal without explicitly stating $\hat{p}$ satisfies component 1.
- Referring to normal approximation must specify "to binomial" to satisfy component 1.
- Discussing the two ample nterval versus the two roportion nterval is extraneous information.

SCORING GUIDELINES

Question 5 (continued)

- In part (c) component 2 can never be satisfied if a procedure that is not possible for one sample of categorical data is suggested.

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

SCORING GUIDELINES

Question

Intent of Question

The primary goals of this question were to assess a student's ability to (1) use a scatterplot to comment on eport about the relationship between two variables and interpret the slope for the least-squares regression line summarizing this relationship; (2) describe the relationship between two variables scatterplot when a categorical variable is introduced and compare a characteristic of the distribution variable for different categories of individuals in a scatterplot; and (3) describe how the associations between two variables for each category of individuals in a scatterplot differ from the overall association i the same scatterplot.

Solution

Part (a):

The scatterplot supports the newspaper report about number of semesters needed to complete an academic program and starting salary because it shows a positive association between these two variables.

Part

The slope is 1.1594. For each additional semester needed to complete an academic program, predicted starting salary increases by €1,159.40.

Part (c):

For the business majors alone, there is a strong, negative, linear association between number of semesters and starting salary. Business majors who need a greater number of semesters to complete an academic program tend to have lower starting salaries.

Part (d):

Business majors have the lowest median starting salary at around €38,000, followed by physics majors at around €48,000, and then chemistry majors with the highest median starting salary at around €55,000.

Part (e):

The newspaper report should be modified to account for major. Overall, majors that take longer to complete tend to have higher starting salaries, with chemistry the highest, physics the next highest, and business the lowest. However, within a major, students who take a greater number of semesters tend to have lower starting salaries.

Scoring

This question is scored in three sections. Section 1 consists of parts (a) and (b), section 2 consists of parts (c) and (d), and section 3 consists of part (e). Sections 1, 2, and 3 are scored as essentially correct (E), partially correct (P), or incorrect (I).

SCORING GUIDELINES

Question 6 (continued)

Section 1 is scored as follows:

Essentially correct (E) if the response includes the following five components:

1. In part (a) the response addresses the positive association.
2. In part (a) the response uses the positive association to justify that the scatterplot support newspaper report.
3. In part (b) the response correctly identifies the numerical value of the slope from the computer output.
4. In part (b) the response interprets the slope as the change in starting salary for each additional semester, in context.
5. In part (b) the interpretation of slope includes nondeterministic language (e.g., "predicted starting salary" or equivalent) when interpreting the slope.

Partially correct (P) if the response includes three or four of the five components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In part (a) the response can use phrases such as "positive association (correlation, relationship)," "increasing relationship," or describe a positive association (e.g., "the starting salaries are higher when there is a greater number of semesters") to satisfy component 1 However, describing the relationship between only two points does not satisfy this component.
- In part (a) comments about linearity and strength should be ignored, even if the response implies these are required (e.g., "Yes, because the relationship is strong, positive, and linear.").
- In part (a) responses that answer "no," "maybe," "kind of," "somewhat," or equivalent do not satisfy component 2.
- In part (a) a response that says "no, because the three clusters of points each have a negative association" does not satisfy component 1 or component 2.
- In part (a) no context is required, but variable names are required in part (b) to satisfy component 4.
- In part (b) a response that states the equation $\hat{y} = 34.018 + 1.1594x$ does not satisfy component 3 unless the slope is specifically identified or used in the interpretation.
- In part (b) a response that incorrectly identifies the numerical value of the slope can still satisfy components 4 and 5 using the incorrect value.
- In part (b) the 1-unit increase in number of semesters must be stated or implied to satisfy component 4 (e.g., for each semester, for every semester). A response that states or implies an unspecified number of semesters does not satisfy this component (for example, as semesters increase).
- In part (b) examples of nondeterministic language include "predicted starting salary," "expected starting salary," "estimated starting salary," "typical starting salary," "average starting salary," "starting salary, on average," "our model predicts," and so on. However, "about," "approximately," and "according to the model" do not satisfy component 5.
- In part (b) no units are required for the change in predicted starting salary (which means OK to say 1.1594 euros, 1159.40, 1.1594, 1,159.40 dollars, and so on).
- In all parts it is acceptable if a response refers to salary rather than starting salary.

SCORING GUIDELINES

Question 6 (continued)

Section 2 is scored as follows:

Essentially correct (E) if the response includes the following five components:

1. In part (c) the response states that the association is negative.
2. In part (c) the response states that the association is strong *OR* linear *OR* both.
3. In part (c) the response refers to both variables (semesters, salary) in context.
4. In part (d) the response correctly compares the three majors.
5. In part (d) the response provides reasonable values for the median salaries or refers to "median starting salaries" when describing a characteristic of the graph.

Partially correct (P) if the response includes three or four of the five components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In part (c) the response can use phrases such as "negative association (correlation, relationship)," "decreasing relationship," "inversely related," or describe a negative association (for example, "the starting salaries are smaller when there is a greater number of semesters") satisfy component 1. However, describing the relationship between only two points does not satisfy this component.
- In part (c) responses that describe the relationship incorrectly as weak or nonlinear do satisfy component 2, even if the other characteristic is described correctly.
- Only drawing a line on the scatterplot does not satisfy component 2.
- In part (d) a response that says "chemistry has the highest median starting salary and business has the lowest median starting salary" (or the equivalent) implies that physics is in the middle, and satisfies component 4.
- In part (d) if no values are provided for the medians, the response must use the phrase "median starting salary" at least once to satisfy component 5.
- In all parts it is acceptable if a response refers to salary rather than starting salary.

Section 3 is scored as follows:

Essentially correct (E) if the response states that there is a negative association for each of the majors *AND* the response notes the overall positive association.

Partially correct (P) if the response states that there is a negative association for each of the majors BUT does not note the overall positive association;
OR
if the response states that there is a negative association for one or two specific majors (for example, for business majors) *AND* the response notes the overall positive association.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- Additional ways to note the overall positive association include stating that the original report was correct, stating that majors that take longer to complete tend to have higher starting salaries, or the equivalent.

SCORING GUIDELINES

Question 6 (continued)

- A response that does not explicitly name the majors or refer to the three majors but makes a general statement such as "However, for an academic major, there is a negative association" satisfies the requirement to state the negative association for each of the majors.
- A response that states the original report was wrong, or incorrect, or should be retracted, cannot satisfy the requirement to note the overall positive association. However, a response that says the original report might be misleading (or the equivalent) but still notes the overall positive association satisfies the requirement to note the overall positive association.
- In all parts it is acceptable if a response refers to salary rather than starting salary.

4 Complete Response

All three sections essentially correct

3 Substantial Response

Two sections essentially correct and one section partially correct

OR

Sections 1 and 2 partially correct and section 3 essentially correct

OR

Sections 1 and 3 essentially correct and section 2 incorrect

OR

Sections 2 and 3 essentially correct and section 1 incorrect

2 Developing Response

Sections 1 and 2 essentially correct and section 3 incorrect

OR

Section 1 essentially correct and sections 2 and 3 partially correct

OR

Sections 1 and 3 partially correct and section 2 essentially correct

OR

One section essentially correct and one section partially correct

OR

All three sections partially correct

1 Minimal Response

One section essentially correct and two sections incorrect

OR

Two sections partially correct and one section incorrect

OR

One section partially correct and two sections incorrect

AP® Statistics
2014 Scoring Guidelines

2014 SCORING GUIDELINES

Question 1

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate conditional proportions from a two-way table; (2) comment on association between two categorical variables as displayed in a graph; and (3) draw an appropriate conclusion from the *p*-value of a chi-square test.

Solution

Part (a):

The proportion of on campus residents who participate in at least one extracurricular activity is $\frac{17+7}{33} = \frac{24}{33} \approx 0.727$. The proportion of off campus residents who participate in at least one extracurricular activity is $\frac{25+12}{67} = \frac{37}{67} \approx 0.552$.

Part (b):

The graph reveals that on campus residents in this sample are more likely to participate in extra-curricular activities than off campus residents. The proportions who participate in two or more extra-curricular activities are similar between the two groups but slightly greater for on campus residents (on campus: 0.212, off campus: 0.179). On campus residents have a greater proportion who participate in one activity (on campus: 0.515, off campus: 0.373) and a smaller proportion who participate in no extracurricular activities (on campus: 0.273, off campus: 0.448) than off campus residents.

Part (c):

The *p*-value of 0.23 is greater than conventional significance levels such as $\alpha = 0.10$ or $\alpha = 0.05$ or $\alpha = 0.01$. Therefore, the *p*-value indicates that the sample data do not provide strong enough evidence to conclude that participation in extracurricular activities differs between on and off campus residents in the population of all students at the university.

2014 SCORING GUIDELINES

Question 1 (continued)

Scoring

Parts (a), (b), and (c) were scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response correctly performs *both* calculations with work shown.

Partially correct (P) if the response correctly performs one of the two calculations with work shown;
OR
if the response provides both correct answers with no work shown;
OR
if the response calculates the proportion of students involved in exactly one extracurricular activity rather than the proportion of students involved in at least one extracurricular activity for both groups, with work shown.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: Answers reported as fractions rather than decimals are acceptable.

Part (b) is scored as follows:

Essentially correct (E) if the response correctly compares proportions between the two groups of students for at least two of the three categories.

Partially correct (P) if the response correctly lists proportions for at least two categories for the two groups but does not make an explicit comparison between the two groups;
OR
if the response correctly compares the relative values of the proportions between the two groups of students for only one of the three categories.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- A response without any reference to percentages or proportions is scored as at most P, (for example, a response that attempts to compare counts).
- A response that treats bar graphs as distributions of a quantitative variable lowers the score one level (that is, from E to P, or from P to I) in part (b).

2014 SCORING GUIDELINES

Question 1 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response states a correct conclusion in the context of the study *AND* provides correct justification/decision of that conclusion based on linkage to the *p*-value.

Partially correct (P) if the response provides no conclusion in context but does provide correct justification/decision based on linkage to the *p*-value;
OR
if the response provides a correct conclusion in context but with incorrect or missing justification/decision based on linkage to the *p*-value;
OR
if the response provides the conclusion in context and correct justification/decision based on linkage to the *p*-value but states a conclusion equivalent to accepting the null hypothesis.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:
- Justification based on the *p*-value can be given by stating a significance level and noting that the *p*-value is larger than the significance level *OR* by simply stating that the *p*-value is large.
- A conclusion that is equivalent to "accept the null hypothesis", either as a stated decision or as a conclusion in context, cannot be scored as E. Such a response is scored as P if it includes both content and correct justification based on linkage to *p*-value. If such a response lacks either context or linkage, it is scored as I.

4 Complete Response

All three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and one part incorrect
OR
One part essentially correct and one or two parts partially correct
OR
Three parts partially correct

1 Minimal Response

One part essentially correct and two parts incorrect
OR
Two parts partially correct and one part incorrect

2014 SCORING GUIDELINES

Question 2

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate a probability; (2) assess whether a claim about randomness is questionable in light of a calculated probability; and (3) determine whether a description of a simulation method achieves a correct simulation of a random process.

Solution

Part (a):

The probability that all 3 people selected are women can be calculated using the multiplication rule, as follows:

P(all three selected are women)

$= P$(first is a woman) $\times P$(second is a woman|first is a woman) $\times P$(third is a woman|first two are women)

$= \frac{3}{9} \times \frac{2}{8} \times \frac{1}{7} \approx 0.012$

Part (b):

The probability calculated in part (a) does provide a reason to doubt the manager's claim that the selections were made at random. The calculation shows that there is only about a 1.2% chance that random selection would have resulted in three women being selected. The probability is small enough that it may cast doubt on the manager's claim that the selections were made at random.

Part (c):

No, the process does not correctly simulate the random selection of three women from a group of nine people of whom six are men and three are women. The random selection of three people among nine is done *without* replacement. However, in the simulation with the dice, the three dice rolls in any given trial are independent of one another, indicating a selection process that is done *with* replacement.

2014 SCORING GUIDELINES

Question 2 (continued)

Scoring

Parts (a), (b), and (c) were scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response correctly computes the probability of selecting the three women, and shows how the probability was computed.

Partially correct (P) if the response shows only one of the following:

Gives the correct probability of $\frac{1}{84}$ (0.012 or 0.011 is acceptable) but does not show how it was computed;
OR
Correctly shows how the probability should be computed, but does not carry the computation through correctly;
OR
Correctly computes (showing work) only the numerator, or only the denominator of the correct answer. (For example, $\frac{1}{9}\times\frac{1}{8}\times\frac{1}{7}\approx 0.002$, or $\frac{3}{9}\times\frac{2}{9}\times\frac{1}{9}\approx 0.008$, or $\frac{3}{9}\times\frac{3}{8}\times\frac{3}{7}\approx 0.054$);
OR
Mistakenly assumes independence and calculates (showing work) the binomial probability $\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}=\frac{1}{27}\approx 0.037$.

Incorrect (I) if the response does not meet the criteria for E or P.

Part (b) is scored as follows:

Essentially correct (E) if the response states that the probability from part (a) is small (or insufficiently small), makes an appropriate decision consistent with the probability being small (or insufficiently small), and does so in the context of this situation.

Partially correct (P) if the response shows only one of the following:

Otherwise satisfies the criteria for an E but does so without any context;
OR
States a significance level and makes a decision in context that is appropriate to the given probability in part (a) and the stated significance level, but does not explicitly compare the probability and the significance level;
OR
Otherwise satisfies the criteria for an E but does not explicitly make a decision about whether there is reason to doubt the manager's claim. (For example: "The probability of selecting the three women from among the nine employees is very small so it is unlikely to occur by chance.")

Incorrect (I) if the response does not meet the criteria for E or P.

2014 SCORING GUIDELINES

Question 2 (continued)

Notes:

- Each of the following situations is one in which a response that otherwise would be scored as E should be scored as P, and a response that otherwise would be scored as P should be scored as I:
  - The response includes a statement that the small probability *proves* that the manager did not make the selection at random (or any equivalent wording).
  - The response includes a statement that clearly interprets the probability from part (a) to be the probability that the manager selected the three people at random.
- Each of the following situations is one in which the response is scored as I:
  - The decision is inconsistent with the justification (*e.g.*, "The probability is very small, so there is no reason to doubt the manager's claim").
  - The response states or implies that because the selection of three women was *not impossible,* there is *no reason to doubt* the manager's claim.

Part (c) is scored as follows:

Essentially correct (E) if the response answers no *AND* states that the dice outcomes in the proposed simulation are independent *AND* states that the genders of the selected convention attendees are dependent. The table below shows statements that should be considered equivalent to the required statements of independence and dependence.

Independence of dice outcomes	Dependence of genders
• The three dice outcomes are independent. • The probability of rolling a 5 or a 6 is the same on all three dice. • The dice simulation actually simulates sampling with replacement.	• The genders of the three people are dependent (or not independent). • The probability of selecting a woman changes after each selection. • The people are sampled without replacement.

OR

Essentially correct if the response answers no *AND* computes the correct probability that a trial of the simulation will indicate the selection of three women $\left(\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}\approx 0.037\right)$ *AND* states that the probability is different from the probability found in part (a).

Partially correct (P) if the response correctly answers no and either:

States only that the dice outcomes are independent or states only that the genders of the selected convention attendees are dependent, but not both;

OR

Otherwise meets the criteria for E but has poor communication. An example of poor communication is: "No, because it selects with replacement. It isn't possible for the same person to be selected twice." (There is an apparent shift between the two sentences from describing the simulation to describing the actual selection of people, but that is not made clear.)

2014 SCORING GUIDELINES

Question 3

Intent of Question

The primary goals of this question were to assess a student's ability to (1) perform a probability calculation from a normal distribution; (2) explain an implication of examining the distribution of a sample mean rather than the distribution of a single measurement; and (3) perform a probability calculation involving independent events using the multiplication rule.

Solution

Part (a):

Because the distribution of the daily number of absences is approximately normal with mean 120 students and standard deviation 10.5 students, the *z*-score for an absence total of 140 students is $z = \frac{140 - 120}{10.5} \approx 1.90$. The table of standard normal probabilities or a calculator reveals that the probability that 140 or fewer students are absent is 0.9713. So the probability that more than 140 students are absent (and that the school will lose some state funding) is $1 - 0.9713 = 0.0287$.

Part (b):

High School A would be *less* likely to lose state funding. With a random sample of 3 days, the distribution of the sample mean number of students absent would have less variability than that of a single day. With less variability, the distribution of the sample mean would concentrate more narrowly around the mean of 120 students, resulting in a smaller probability that the mean number of students absent would exceed 140.

In particular, the standard deviation of the sample mean number of absences, $\overline{x}$, is $\frac{\sigma}{\sqrt{n}} = \frac{10.5}{\sqrt{3}} = 6.062$. So the *z*-score for a sample mean of 140 is $\frac{140 - 120}{6.062} \approx 3.30$. The probability that High School A loses funding using the suggested plan would be $1 - 0.9995 = 0.0005$, as determined from the table of standard normal probabilities or from a calculator, which is less than a probability of 0.0287 obtained for the plan described in part (a).

Part (c):

For any one typical school week, the probability is $\frac{2}{5} = 0.4$ that the day selected is not Tuesday, not Wednesday, or not Thursday. Therefore, because the days are selected independently across the three weeks, the probability that none of the three days selected would be a Tuesday or Wednesday or Thursday is $(0.4)^3 = 0.064$.

2014 SCORING GUIDELINES

Question 3 (continued)

Scoring

Parts (a), (b), and (c) were scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response provides the following three components:
1. Indicates use of a normal distribution and clearly identifies the correct parameter values (showing correct components of a *z*-score calculation is sufficient).
2. Uses the correct boundary value (140, 140.5, or 141 is acceptable).
3. Reports the correct normal probability consistent with components 1 and 2

OR

if the response reports a probability of 0.025 with justification based on the empirical rule for an acceptable boundary value (140, 140.5, or 141 is acceptable).

Partially correct (P) if the response correctly provides only two of the three components listed above.

OR

if the response provides an incorrect probability of 0.05 with justification based on the empirical rule for an acceptable boundary value (140, 140.5, or 141 is acceptable).

Incorrect (I) if the response does not satisfy the criteria for E or P.

Note: An inconsistency in calculations lowers the score for part (a) by one level (that is, from E to P or from P to I).

Part (b) is scored as follows:

Essentially correct (E) if the response provides the correct answer of less likely *AND* the following three components:
1. Clearly references the distribution of the sample mean.
2. Indicates that the variability of the distribution is smaller.
3. Indicates that the distribution is centered at 120.

OR

if the response provides the correct answer of less likely *AND* the following two components:
1. Correctly calculates the probability that the sample mean would exceed 140 (arithmetic errors are not penalized).
2. Correctly compares this probability to the probability in part (a).

Partially correct (P) if the response provides the correct answer of less likely *AND* only two of the following three components:
1. Clearly references the distribution of the sample mean.
2. Indicates that the variability of the distribution is smaller.
3. Indicates that the distribution is centered at 120.

OR

if the response provides the correct answer of less likely *AND* correctly calculates the probability that the sample mean would exceed 140 (arithmetic errors are not penalized) *BUT* does not correctly compare this probability with the probability in part (a).

2014 SCORING GUIDELINES

Question 3 (continued)

Incorrect (I) if the response does not meet the criteria for E or P, including if the response provides the incorrect answer or provides the correct answer of less likely with no explanation or an incorrect explanation.

Note: An equivalent approach is to use the total number of absences for 3 days. The sampling distribution of the total number of absences for the 3 days is approximately normal, with mean $3(120) = 360$ absences and standard deviation $3(6.026) \approx 18.187$ absences. The *z*-score for a total of $3(140) = 420$ absences is: $\frac{420 - 360}{18.187} \approx 3.30$. Such a response is scored E if the response provides the correct answer of less likely and references the distribution of the sample total, and includes the correct mean and standard deviation.

Part (c) is scored as follows:

Essentially correct (E) if the response correctly calculates the probability *AND* shows sufficient work.

Partially correct (P) if the response reports the correct probability but shows no work or does not show sufficient work;
OR
if the response uses the multiplication rule involving three events but does so incorrectly and/or with an incorrect probability of not selecting a Tuesday, Wednesday, or Thursday.

Incorrect (I) if the response does not meet the criteria for E or P.

4 Complete Response

All three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and one part incorrect
OR
One part essentially correct and one or two parts partially correct
OR
Three parts partially correct

1 Minimal Response

One part essentially correct and two parts incorrect
OR
Two parts partially correct and one part incorrect

2014 SCORING GUIDELINES

Question 4

Intent of Question

The primary goals of this question were to assess a student's ability to (1) describe why the median might be preferred to the mean in a particular context; (2) compare the relative merits of two sampling plans; and (3) describe a consequence of nonresponse in a particular study.

Solution

Part (a):

The median is less affected by skewness and outliers than the mean. With a variable such as income, a small number of very large incomes could dramatically increase the mean but not the median. Therefore, the median would provide a better estimate of a typical income value.

Part (b):

Method 2 is better than Method 1. A sample obtained from Method 1 could be biased because of the voluntary nature of the response. It is plausible that class members with larger incomes might be more likely to return the form than class members with smaller incomes. The mean income for such a sample would overestimate the mean income of all class members. With Method 2, despite the smaller sample size, the random selection is likely to result in a sample that is more representative of the entire class and produce an unbiased estimate of mean yearly income of all class members.

2014 SCORING GUIDELINES

Question 4 (continued)

Scoring

This question is scored in three sections. Part (b) has two components: (1) identifying a relevant characteristic for each sampling method; (2) indicating the effect of the biased method on the estimate of the mean income. Section 1 consists of Part (a); section 2 consists of part (b), component 1; and section 3 consists of part (b), component 2. Sections 1, 2, and 3 are scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response includes the following two components:
1. Describes how skewness or outliers affect the mean or do not affect the median.
2. Makes a conjecture about a relevant characteristic of the distribution of incomes, such as skewness or an outlier.

Partially correct (P) if the response includes only one of the two components listed above.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:
- For Component 1, examples of responses that are acceptable include:
  - The mean is affected by skewness (outliers).
  - The median is not affected by skewness (outliers).
  - The mean is greater (less) than the median when there is right (left) skewness or outliers.
- For Component 1, examples of responses that are not acceptable include:
  - Don't use the mean for skewed distributions or distributions with outliers.
  - Use the median for skewed distributions or distributions with outliers.
  - Responses that include an incorrect statement about means and/or medians, such as for right skewed distributions, the median will be higher than the mean.
- It is possible to satisfy *both* components with a single sentence, such as, "If there was a billionaire in the sample, the mean would be higher than the median."
- If a response argues that using the *mean* is a more appropriate way to estimate the typical income, then reduce the score in section 1 by one level (that is, from E to P or from P to I).

Section 2 is scored as follows:

Essentially correct (E) if the response chooses Method 2 *AND* includes the following two components:
1. Identifies a relevant characteristic of Method 1.
2. Identifies a relevant characteristic of Method 2.

Partially correct (P) if the response chooses Method 2 *AND* includes only one of the two components listed above

OR

if the response includes both components but does not choose a method.

Incorrect (I) if the response chooses Method 1 *OR* otherwise does not meet the criteria for E or P.

2014 SCORING GUIDELINES

Question 4 (continued)

Notes:

- Responses that do not explicitly choose Method 2 can still earn an E for section 2 if the choice is clearly implied. The choice of Method 2 is clearly implied if the response only discusses negative characteristics of Method 1 and only discusses positive characteristics of Method 2, such as, Method 1 is biased but Method 2 uses a random sample.
- Responses that *compare* the two methods can satisfy *both* components, such as, saying that Method 1 is more biased or that nonresponse will be less of a problem with Method 2.
- Responses that refer to the nonresponse bias as *voluntary response bias*, *response bias, undercoverage* can still earn an E.
- Discussions of conditions for inference should be considered extraneous and ignored.

Section 3 is scored as follows:

Essentially correct (E) if the response includes the following two components:

1. Indicates the incomes of responders may be different from the incomes of nonresponders.
2. Indicates the biased sampling method may produce a misleading estimate/conclusion about the mean income, including direction, for example, "The sample mean is likely to be higher than the mean of the population."

Partially correct (P) if the response provides only one of the two components listed above.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- A single sentence can satisfy the first component of section 2 and the first component of section 3. (For example, "In method 1, rich people are more likely to respond.")
- For component 2, either direction is acceptable but the direction must be consistent with the identified bias. Saying only that Method 2 will be more accurate or more representative does not satisfy component 2.
- If a response addresses possible nonresponse bias in Method 2, the response can still satisfy both components of section 3.
- Responses that focus on the larger sample size in Method 1 can satisfy component 2 if such responses describe the effect as reducing the variability of the estimate. (For example, "I would use Method 1 since the larger sample size would give less variability of the mean.")
- Responses that focus on untruthful survey answers can satisfy component 2 if the effect on the estimate is appropriate. (For example, "People contacted in Method 2 might say they make more money than they actually do. This would make the estimated mean too high.")

2014 SCORING GUIDELINES

Question 4 (continued)

4 Complete Response

All three sections essentially correct

3 Substantial Response

Two sections essentially correct and one section partially correct

2 Developing Response

Two sections essentially correct and one section incorrect

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

1 Minimal Response

One section essentially correct and two sections incorrect

OR

Two sections partially correct and one section incorrect

2014 SCORING GUIDELINES

Question 5

Intent of Question

The primary goal of this question was to assess students' ability to identify, set up, perform, and interpret the results of an appropriate hypothesis test to address a particular question. More specific goals were to assess students' ability to (1) state appropriate hypotheses; (2) identify the appropriate statistical test procedure and check appropriate conditions for inference; (3) calculate the appropriate test statistic and *p*-value; and (4) draw an appropriate conclusion, with justification, in the context of the study.

Solution

Step 1: States a correct pair of hypotheses.

Let μ_{diff} represent the population mean difference in purchase price (woman – man) for identically equipped cars of the same model, sold to both men and women by the same dealer, in the county.

The hypotheses to be tested are $H_0 : \mu_{\text{diff}} = 0$ versus $H_a : \mu_{\text{diff}} > 0$.

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a paired *t*-test.

The conditions for the paired *t*-test are:

1. The sample is randomly selected from the population.
2. The population of price differences (woman – man) is normally distributed, or the sample size is large.

The first condition is met because the car models and the individuals were randomly selected. The sample size ($n = 8$) is not large, so we need to investigate whether it is reasonable to assume that the population of price differences is normally distributed. The dotplot of sample price differences reveals a fairly symmetric distribution, so we will consider the second condition to be met.

Step 3: Correct mechanics, including the value of the test statistic and *p*-value (or rejection region).

The test statistic is $t = \dfrac{585 - 0}{\dfrac{530.71}{\sqrt{8}}} \approx 3.12$.

The *p*-value, based on a *t*-distribution with $8 - 1 = 7$ degrees of freedom, is 0.008.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the *p*-value is very small (for instance, smaller than $\alpha = 0.05$), we reject the null hypothesis. The data provide convincing evidence that, on average, women pay more than men in the county for the same car model.

2014 SCORING GUIDELINES

Question 5 (continued)

Scoring

Each of steps 1, 2, 3, and 4 were scored as essentially correct (E), partially correct (P), or incorrect (I).

Step 1 is scored as follows:

Essentially correct (E) if the response identifies the correct parameter *AND* states correct hypotheses.

Partially correct (P) if the response identifies the correct parameter *OR* states correct hypotheses, but not both.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: Defining the parameter symbol in context or simply using common parameter notation is sufficient.

Step 2 is scored as follows:

Essentially correct (E) if the response identifies the correct test procedure (by name or by formula) *AND* checks *both* conditions correctly.

Partially correct (P) if the response correctly completes two of the three components (identification of procedure, check of randomness condition, check of normality condition).

Incorrect (I) if the response does not meet the criteria for E or P.

Note: The random sampling condition can be verified by referring to the random selection of car models or to the random selection of male and female car buyers.

Step 3 is scored as follows:

Essentially correct (E) if the response correctly calculates both the test statistic and the *p*-value.

Partially correct (P) if the response correctly calculates the test statistic but not the *p*-value;
OR
if the response calculates the test statistic incorrectly but then calculates the correct *p*-value for the computed test statistic.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: If the response identifies a *z*-test for a mean as the correct procedure in step 2, then the response can earn a P in step 3 if both the test statistic and the *p*-value are calculated correctly.

2014 SCORING GUIDELINES

Question 5 (continued)

Step 4 is scored as follows:

Essentially correct (E) if the response provides a correct conclusion in context, also providing justification based on linkage between the *p*-value and the conclusion.

Partially correct (P) if the response provides a correct conclusion with linkage to the *p*-value, but not in context;
OR
if the response provides a correct conclusion in context, but without justification based on linkage to the *p*-value.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- If the conclusion is consistent with an incorrect *p*-value from step 3 and also in context with justification based on linkage to the *p*-value, step 4 is scored as E.
- A response that performs a two-sample *t*-test with correct calculations should fail to reject H_0. A conclusion that is equivalent to "accept H_0" (such as "we conclude that women pay the same amount as men, on average"), either as a stated decision or as a conclusion in context, cannot be scored as E. Such a response will be scored as P provided that the conclusion is in context with linkage. Such a response will be scored as I if it lacks either context or linkage.

Each essentially correct (E) step counts as 1 point. Each partially correct (P) step counts as ½ point.

4 Complete Response

3 Substantial Response

2 Developing Response

1 Minimal Response

If a response is between two scores (for example, 2½ points), use a holistic approach to decide whether to score up or down, depending on the overall strength of the response and communication.

2014 SCORING GUIDELINES

Question 6

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate and interpret a residual value; (2) answer questions about residual plots; (3) compare associations between two scatterplots; and (4) identify an appropriate explanatory variable to include in a regression model based on residuals from simpler regression models.

Solution

Part (a):

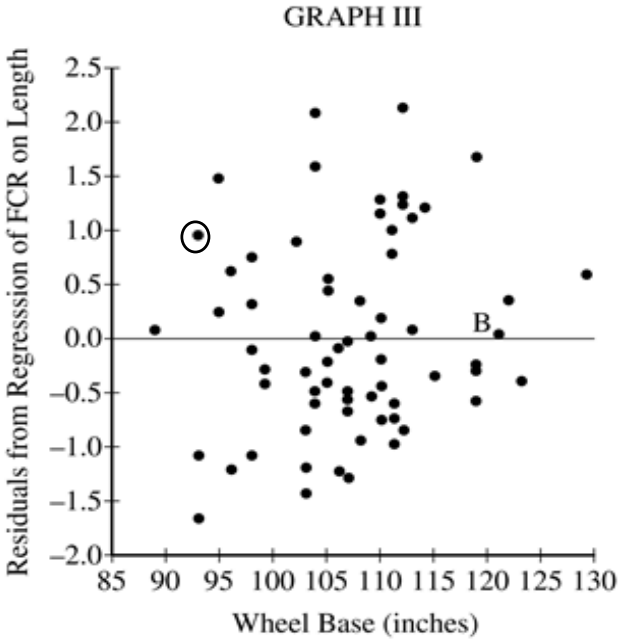
For a car with length 175 inches, the predicted value for the car's FCR, based on the least squares regression line, is

$$\text{predicted FCR} = -1.595789 + 0.0372614(175) \approx 4.92 \text{ gallons per 100 miles.}$$

The actual FCR for the car is 5.88, so the residual is $5.88 - 4.92 = 0.96$. The residual value means that the car's FCR is 0.96 gallons per 100 miles greater than would be predicted for a car of its length.

Part (b):

(i) The point with a wheel base of 93 inches and a residual of 0.96 gallons per 100 miles is circled in graph III below.



(ii) Point B corresponds to a car with an actual FCR that is very close to the FCR that would be predicted for a car with its length by the regression model which predicts FCR using the explanatory variable length.

2014 SCORING GUIDELINES

Question 6 (continued)

Part (c):

Graph II reveals a moderate association that is positive and linear. In contrast, there is a weak association that is positive and linear in graph III. The association between engine size and residual (from predicting FCR based on length) is stronger than the association between wheel base and residual (from predicting FCR based on length).

Part (d):

Engine size is a better choice than wheel base for including with length in a regression model for predicting FCR. The stronger association between engine size and residual (from predicting FCR based on length) indicates that engine size is more useful than wheel base for reducing the variability in FCR values that remains unexplained (as indicated by residuals) after predicting FCR based on length.

Scoring

Parts (a), (b), (c), and (d) were scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response provides the following two components:
1. A correct residual value with supporting calculation.
2. A correct interpretation of the residual value, in context.

Partially correct (P) if the response includes only one of the two components listed above.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:
- If the residual value is incorrect, the interpretation should be considered correct if it follows from the incorrect residual value.
- Correct interpretation of the residual must include the correct direction and magnitude of the FCR value away from the predicted FCR value.
- A calculated residual value which is slightly different from 0.96 due to the number of significant digits is acceptable.

Part (b) is scored as follows:

Essentially correct (E) if the response provides the following two components:
1. Circles the correct point in graph III.
2. Provides a reasonable interpretation of the car associated with point B having a residual near 0 that refers to predicting FCR based on length.

Partially correct (P) if the response correctly provides only one of the two components listed above.

Incorrect (I) if the response does not meet the criteria for E or P.

2014 SCORING GUIDELINES

Question 6 (continued)

Note: A correct response for the second component must include reference to the observed FCR value of the car represented by point B, not the point B itself.

Part (c) is scored as follows:

Essentially correct (E) if the response correctly provides the following three components:
1. A description of form *AND* direction for both graphs.
2. A description of the strength of association for both graphs.
3. A comparison between the two graphs.

Partially correct (P) if the response correctly provides only two of the three components listed above.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:
- Part (c) is focused on the comparison of graph II and graph III. Inferences drawn from patterns in these graphs are considered in part (d).
- Linear is needed for form in graph II.
- Graph III may be described as having no association between wheel base and the residuals of FCR based on length, which is sufficient for describing the form, direction and strength of association of graph III.

Part (d) is scored as follows:

Essentially correct (E) if the response indicates the correct choice with a sound justification based on the following two components:
1. The strong(er) association.
2. Reducing the variability that remains unexplained in the model which predicts FCR based on length.

Partially correct (P) if the response indicates the correct choice and provides a justification based on only one of the two components which are listed above.

Incorrect (I) if the response indicates the incorrect choice;
OR
if the response indicates the correct choice but does not mention either of the two components which are listed above.

Note: Describing the variables in graph II and graph III as residuals is not required but can be used positively in holistic scoring. Incorrect descriptions of graph II or graph III or the variables in graphs are not acceptable.

2014 SCORING GUIDELINES

Question 6 (continued)

Each essentially correct (E) part counts as 1 point. Each partially correct (P) part counts as ½ point.

4 **Complete Response**

3 **Substantial Response**

2 **Developing Response**

1 **Minimal Response**

If a response is between two scores (for example, 2½ points), use a holistic approach to decide whether to score up or down, depending on the overall strength of the response and communication.