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 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 histogram of dissolved oxygen concentration in Alaskan streams with water temperatures colder than 8°C is unimodal and skewed left with a median between 11 and 12 mg/l. The first quartile is in the bin from 10-11 mg/l and the third quartile is in the bin from 12-13 mg/l, so the IQR is approximately 2 mg/l. There do not appear to be any high outliers, but there are several potential low outliers because the values in the 2-3, 4-5, and 5-6 bins are all certainly more than 1.5 IQR below the first quartile.	**Essentially correct (E)** if the description of the distribution satisfies component 1 *AND* at least three of components 2-5: 1. Includes context of dissolved oxygen concentration 2. Shape: The distribution is skewed left 3. Center: The center of the distribution is between 11 mg/l and 12 mg/l 4. Spread: Refers to at least one measure of variation (i.e., the range is between 14 – 2 = 12 mg/l and 13 – 3 = 10 mg/l; all values are between 2 mg/l and 14 mg/l or the IQR is approximately 2 mg/l) 5. Unusual features: potential outliers or a gap between 3 mg/l and 4 mg/l **Partially correct (P)** if the response satisfies component 1 and two components out of components 2-5 *OR* if the response satisfies at least three out of components 2-5 but does not satisfy component 1. **Incorrect (I)** if the response does not meet the criteria for E or P.

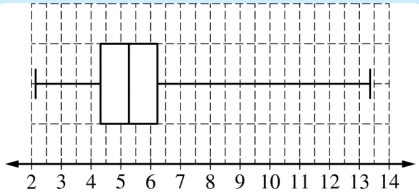
Additional Notes:

- **Context**
  - Component 1 can be satisfied with a reference to dissolved oxygen concentration, dissolved oxygen, amount of oxygen, or mg/l.

- **Shape:**
  - Component 2 cannot be satisfied if a response describes the histogram as normal or approximately normal.
- **Center:**
  - A response that addresses center using definitive language such as "the mean (median) of the distribution is 11.5" cannot satisfy component 3.
  - A response that addresses center using approximate language such as "the median of the distribution is approximately 11.5" must, for any single measure of center, specify a numeric value that is between 11 mg/l and 12 mg/l, inclusive, to satisfy component 3.
- **Spread:**
  - A response recognizing all values in the sample fall between 2 mg/l and 14 mg/l, satisfies component 4 only for these exact endpoints and need not appeal to a specific measure of spread such as the range.
  - A response that uses interval language must use it correctly. For example, "the observations range from 2 mg/l and 14 mg/l," satisfies component 4 because it correctly indicates that all observations are between 2 mg/l and 14 mg/l, inclusive. However, a statement such as "the range is between 2 mg/l and 14 mg/l," is incorrect because the range is a single number, i.e., 14 – 2 = 12 mg/l.
  - A response that appeals to a specific measure of spread using approximate language, such as "the IQR is approximately 2," must specify a numeric value within the bounds appropriate to that measure of spread shown in the following table.

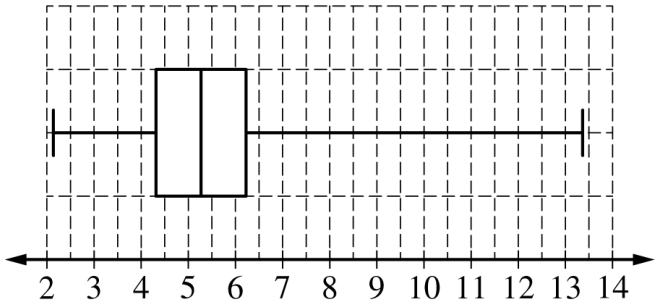
Statistic	Min (mg/l)	Max (mg/l)
Range	10	12
IQR	1.5	2.5
Standard Deviation	1.7	1.8

- **Unusual Features:**
  - Component 5 cannot be satisfied if a response indicates that the distribution has an unusual feature other than potential outliers or a gap.
  - Definitive language such as "there is an outlier" does not satisfy component 5.

Model Solution	Scoring
(b)  Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Warmer than 8˚C	**Essentially correct (E)** if the response satisfies four or five of the following five components: 1. Constructs the box beginning at Q1 2. Constructs the box ending at Q3 3. Locates the median within the box 4. Extends the lines on the boxplot to the minimum value 5. Extends the lines on the boxplot to the maximum value **Partially correct (P)** if the response includes a boxplot with only three of the five components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response that shows a graph or plot other than a boxplot should be scored incorrect (I).
- To be correctly placed:
  - The minimum must be between the tick marks for 2 and 2.5
  - The first quartile must be between the tick marks for 4 and 4.5
  - The median must be between the tick marks for 5 and 5.5
  - The third quartile must be between the tick marks for 6 and 6.5
  - The maximum must be between the tick marks for 13 and 13.5



Model Solution	Scoring
(c) If the researchers' belief is correct, then streams with water temperature colder than 8°C are healthier for wildlife. The distribution of dissolved oxygen concentration for colder streams has a higher center because its median (between 11 mg/l and 12 mg/l) is larger than the median for warmer streams (5.43 mg/l). The shape of the distribution of dissolved oxygen concentration for colder streams is different from the shape of the distribution for warmer streams. The distribution of values of dissolved oxygen concentration for colder streams is skewed to the left but the distribution of values for warmer streams is skewed to the right. Both distributions have a similar spread because they both have similar IQR values — approximately 2 mg/l for the colder streams and 1.73 mg/l for the warmer streams.	**Essentially correct (E)** if the response satisfies component 1 *AND* at least two of components 2-4: 1. States colder streams are healthier for wildlife 2. Directly compares the centers of the two distributions 3. Indicates warmer streams are skewed right and colder streams are skewed left 4. Directly compares the spreads **Partially correct (P)** if the response satisfies component 1 *AND* one of components 2-4. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- While both distributions have potential outliers, it is not necessary to compare them for this part of the question.
- A response based on an incorrect part (a) or part (b) may be scored E or P if it justifies the choice consistent with their answers to part (a) or part (b).
- In order to satisfy component 4, a response may compare at least one of the values of the approximate range, IQR, or standard deviation with the value of the same statistic from the other distribution.

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**

Question 5: Multi-Focus — 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 sample median reduction in blood pressure for those eating dark chocolate, 7 mmHg, is greater than the sample median reduction in blood pressure for those eating white chocolate, about 0 mmHg.	**Essentially correct (E)** if the response satisfies the following three components: 1. The medians are correctly determined with the median for dark chocolate equal to 7 mmHg and the median for white chocolate between −2 mmHg and 2 mmHg, inclusive 2. The computed medians are compared correctly 3. The context of the response variable (reduction in blood pressure) and the treatment groups (dark chocolate and white chocolate) is included **Partially correct (P)** if the response satisfies two of the three components and, if stated, the reported values of the medians are reasonable *OR* if the response satisfies one of components 2 or 3 and the reported values of the medians are reasonable. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- To be scored as reasonable, the reported value of the median for the dark chocolate group must be between 5 mmHg and 8 mmHg, inclusive, and the reported value for the median of the white chocolate group must be between −2 mmHg and 3 mmHg, inclusive.

Model Solution	Scoring
(b) The researcher's conclusion may not necessarily be true because looking at the difference in sample means alone does not consider the variability in the sampling distribution of the differences in sample means. Another random assignment of subjects to dark and white chocolate may result in a sample mean reduction in blood pressure for subjects assigned to dark chocolate that is smaller than the mean reduction for subjects assigned to white chocolate. The variability in the sampling distribution of potential differences in sample means must be considered in making a conclusion about convincing statistical evidence. Further, an inference procedure can assess the likelihood that the observed difference in sample means occurred by random chance if the population means are equal.	**Essentially correct (E)** if the response satisfies at least two of the following three components: 1. Indicates that the researchers failed to consider the variability in the sampling distribution of differences in sample means (or that this should be considered) OR states that a different random assignment could produce different sample means 2. Includes an explanation that a difference of 5.66 mmHg could have occurred by random chance, even if the population means are equal 3. Indicates that an inference procedure is needed **Partially correct (P)** if the response satisfies only one of the three components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response may be scored E if it states that no conclusion can be made about the relationship between the population mean blood pressure reduction for those who consume dark chocolate and those who consume white chocolate because the sample was drawn only from the subpopulation of *healthy* adults.

Model Solution	Scoring
(c) The observed value of the sample statistic $\bar{x}_{dark} - \bar{x}_{white}$ is 5.66 mmHg. The graph of simulation results reveals that a difference of 5.66 mmHg or larger occurred in only 3 of the 120 trials. Thus, the p-value is approximately equal to $\frac{3}{120} = 0.025$. Thus, assuming that there is no difference in mean blood pressure reduction, there is an approximate probability of 0.025 of getting a difference of 5.66 mmHg or larger by chance alone. Because this approximate p-value is less than 5 percent, there is convincing evidence that adding dark chocolate to the diet will result in a greater mean reduction in blood pressure than adding white chocolate to the diet for people similar to those who participated in the study.	**Essentially correct (E)** if the response satisfies the following four components: 1. Calculates the correct p-value of 0.025 2. Provides supporting work from the simulation for the calculation of the p-value 3. Provides a correct conclusion in the context of whether there is convincing evidence that adding dark chocolate to the diet will result in greater average reduction in blood pressure than adding white chocolate to the diet 4. Justifies the conclusion by directly comparing the calculated p-value to 0.05 **Partially correct (P)** if the response satisfies only two or three of the four components *OR* if the conclusion (component 3) is correct but is not in context and at least one additional component is satisfied *OR* if the conclusion (component 3) is in context and correct, but the justification is based on an analysis of the data shown in the dotplots (e.g., a two-sample t-confidence interval, a two-sample t-test, an appropriate nonparametric test, or an "exact" randomization test). **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- Component 2 is satisfied by simply noting that only three of the simulated trials resulted in a difference of at least 5.66 mmHg, the observed difference.
- A response that incorrectly evaluates the p-value may still satisfy components 3 and 4 if the response is consistent with the use of the incorrect p-value.
- The hypotheses are not required because they are given in the stem.

Scoring for Question 5	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**

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)	The boxplots reveal that the team tended to have a higher average per-game attendance during years in the new stadium than during years in the old stadium because the median value of about 25,000 attendees per game during years in the new stadium is greater than the median value of about 16,000 attendees per game during years in the old stadium. The interquartile ranges (IQRs) are similar, which indicates that variability in average per-game attendance is similar during the years in the two stadiums, but the range of average per-game attendance is slightly larger during the years in the new stadium. There are no years with average per-game attendance identified as an outlier for either stadium.	**Essentially correct (E)** if the response satisfies the following three components: 1. Indicates that the median average per-game attendance is greater in the new stadium *OR* Indicates that the average per-game attendance is usually (typically) greater in the new stadium 2. Indicates that the variability in average per-game attendance is greater in the new stadium *OR* Indicates that the variability in the two stadiums is roughly the same according to the IQRs 3. Provides sufficient context by including the response variable (average attendance or attendance) or the units of the response variable (thousands of people or people) **Partially correct (P)** if the response satisfies only two of the three components *OR* if the response satisfies only component 1 *OR* if the response satisfies only component 2. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- To satisfy component 1, the response may refer only to "center" and does not need to specify which measure of center (e.g., median, mean) is being used.

- To satisfy component 2, a response may refer only to "variability" or "spread" and does not need to specify which measure of variability (e.g., range or IQR) is being used. However, if the response states that the variability is about the same, the response must explicitly refer to the IQRs.
- To satisfy components 1 and 2, the stadiums must be identified (e.g., "old," "new"), and an explicit comparison phrase (e.g., "greater than," "about the same as") must be used. Separate lists of characteristics alone or summary statistics alone do not count as a comparison.
- To satisfy components 1 and 2, numerical values are not required. However, if they are included, they should be reasonably correct. Numerical values can be reported in units of people (e.g., median = 16,000) or in thousands of people (e.g., median = 16).
- Any mention of shape is ignored in the scoring of part (a) because complete shape information cannot be obtained from a boxplot. Statements about shape not clearly supported by the boxplots (e.g., "the old stadium distribution is approximately normal") should be considered a negative in terms of holistic scoring. However, statements about shape that are supported by the boxplots (e.g., "the old stadium distribution is roughly symmetric," "the new stadium distribution is skewed to the left") should be considered a positive.

	Model Solution	Scoring
(b)	During the years in the new stadium, the average per-game attendance increases linearly, from about 16,000 people in 2000 to about 27,000 people in 2016. However, during the years in the old stadium, there is no obvious increasing or decreasing trend over time for the average per-game attendance. The average attendance appears to vary about an average of approximately 16,000 attendees per game from 1970 to 1999.	**Essentially correct (E)** if the response satisfies the following three components: 1. Describes the direction of the trend in average per-game attendance in the new stadium as increasing (positive) 2. Describes the direction of the trend in average per-game attendance in the old stadium as relatively constant (e.g., "no association," "flat") *OR* Describes the direction of the trend in the old stadium as positive but less steep (e.g., "less positive," "flatter") than the trend in the new stadium 3. Provides sufficient context by including the two groups (old stadium, new stadium) AND the explanatory variable (time or year) AND the response variable (average attendance or attendance) or the units of the response variable (thousands of people or people) **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.

Additional Notes:

- Only describing an association or correlation as "strong" or "weak" addresses strength and not the direction of the trend and does not satisfy components 1 or 2.
- Only describing an association as "linear" or "non-linear" addresses form and not the direction of the trend and does not satisfy components 1 or 2.
- Numerical values, including years (e.g., "from 2000 to 2016"), are not required for any component. However, a response that includes years in numerical form (e.g., "1970") satisfies the context requirement for the explanatory variable in component 3.
- Component 1 can also be satisfied if the response provides an estimated value for the correlation for the new stadium that is positive.
- Component 2 can also be satisfied if the response provides an estimated value for the correlation for the old stadium of 0 (or approximately 0). Providing positive correlations for both stadiums does not satisfy component 2 because it is impossible to compare the steepness of the trends using their correlations.

Model Solution	Scoring
(c) **(i)** Graph I indicates a strong, positive, linear relationship between average per-game attendance and the number of games won during the 47 years of the team's existence. Average per-game attendance increases linearly, with an average increase of about 500 attendees per game for each additional game won. Variation about the linear trend in attendance is relatively small and about the same for any number of games won. **(ii)** No. Graph II suggests that the rates at which average per-game attendance increases as the number of games won increases are about the same for the two stadiums. A line drawn through the points for the old stadium has about the same slope as (or may have a slightly larger slope than) a line drawn through the points for the new stadium.	**Essentially correct (E)** if the response satisfies four or five of the following five components: 1. In part (c-i) describes the direction of the relationship as positive 2. In part (c-i) describes the form of the relationship as linear or nearly linear 3. In part (c-i) describes the strength of the relationship as very strong, strong, or moderately strong 4. In part (c-ii) indicates that the rates are about the same for the two stadiums (or slightly larger for the old stadium) 5. In part (c-ii) provides an explanation that indicates that if a line were drawn through the points for the old stadium, the slope would be roughly the same (or slightly greater than) the slope of a line through the points for the new stadium **Partially correct (P)** if the response satisfies only three of the five components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response that provides an estimated value of the correlation satisfies component 1 if the estimated correlation is positive, but an estimated correlation cannot satisfy components 2 or 3.
- A response need not include the word "positive" to satisfy component 1. For example, "the average attendance is higher when the team has more wins" satisfies component 1. Likewise, a response need not include the word "strong" to satisfy component 3. For example, "variation about the linear trend in attendance is relatively small" satisfies component 3.
- Correct comments on homogeneous variation in part (c-i) (i.e., the variation about the linear trend in attendance is about the same for any number of games won) should be considered a positive in terms of holistic scoring.
- Responses that satisfy all 5 components should be considered a positive in terms of holistic scoring.
- Context is not required in part (c) because it has already been assessed in parts (a) and (b).

Model Solution	Scoring
(d) The number of games won could be a confounding variable for assessing the potential effect of opening the new stadium on average per- game attendance. The boxplots in part (a) show that average per-game attendance tended to be higher for games in the new stadium than for games in the old stadium, but the cause of the increase may actually be that attendees were more excited about attending games for teams that were better at winning. The scatterplots in part (c) show that average per-game attendance has a strong positive correlation with games won, and the team tended to win more games playing in the new stadium than in the old stadium.	**Essentially correct (E)** if the response provides an explanation that satisfies the following four components: 1. States that there is an association between attendance and one of the explanatory variables (stadium, year, wins) 2. States that there is an association between attendance and a different one of the explanatory variables (stadium, year, wins) 3. States that there is an association between the two explanatory variables (stadium, year, wins) identified in components 1 and 2 4. Explains the idea of confounding by describing that the variable identified as a potential confounding variable could be the cause of the association between attendance and the other explanatory variable identified in components 1 and 2 *OR* Explains the idea of confounding by stating that it is impossible to know which of the two explanatory variables identified in components 1 and 2 may be the cause of the increase in attendance **Partially correct (P)** if the response satisfies only three of the four components *OR* if the response satisfies only two of the four components and justifies at least one of the statements in components 1 through 3 by referring to the appropriate graph from parts (a) through (c) (e.g., "based on the boxplots," "in part (b)"). **Incorrect (I)** if the response does not otherwise meet the criteria for E or P.

Additional Notes:

- The response can use any combination of 2 of the 3 explanatory variables (stadium and wins, stadium and year, year and wins). The response cannot introduce a new variable (e.g., weather, having popular players) to satisfy any component.
- An incorrectly described association cannot be used to satisfy components 1 through 3.
- To satisfy component 4 the response must discuss all three variables: the response variable (attendance)
- The strength of the response to part (d), especially communication in component 4, can be considered in holistic scoring. A well-communicated response to part (d) should be considered a positive. A poorly-communicated response to part (d), should be considered a negative.

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.	

2018

AP CollegeBoard

AP Statistics

Scoring Guidelines

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 $\bar{y}$.
- 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 <u>why</u> 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 <u>how</u> 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 $P(\text{yes})$ 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.

AP® STATISTICS
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)

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.

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

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 $\frac{189}{221} \approx 0.8552$.

AP STATISTICS
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.

AP STATISTICS
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 $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} > 1.645$. With $\mu = 122$, $\sigma = 15$, and $n = 100$, we get $\dfrac{\bar{x} - 122}{1.5} > 1.645$. Therefore, $\bar{x} > 124.4675$.

Part (c):

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

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

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 $\bar{x}$ 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.

AP STATISTICS
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.

AP STATISTICS
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

2017

AP® CollegeBoard

AP Statistics
Scoring Guidelines

2017 SCORING GUIDELINES

Question 1

Intent of Question

The primary goals of this question were to assess a student's ability to (1) explain statistical terms used when describing the relationship between two variables; (2) interpret the slope of a linear regression equation; and (3) calculate a value of y when given a regression equation, a value of x, and a residual.

Solution

Part (a):

In the context of a scatterplot in which y represents weight and x represents length, the following are defined.

A positive relationship means that wolves with higher values of length also tend to have higher weights.

A linear relationship means that as length increases by one meter, weight tends to change by a constant amount, on average.

A strong relationship means that the data points fall close to a line (or curve).

Part (b):

The slope of 35.02 indicates that two wolves that differ by one meter in length are predicted to differ by 35.02 kilograms in weight, with the longer wolf having the greater weight.

Part (c):

In general, a residual is equal to actual weight minus predicted weight, or equivalently,

$$\text{actual weight} = \text{predicted weight} + \text{residual}.$$

For the wolf with length 1.4 meters and residual of -9.67, the predicted weight is

$$-16.46 + 35.02(1.4) = 32.568 \text{ kilograms}.$$

Therefore, the actual weight of the wolf is $32.568 + (-9.67) = 22.898$ kilograms.

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 includes the following four components:
1. A reasonable definition of positive
2. A reasonable definition of linear
3. A reasonable definition of strong
4. At least one definition in context

2017 SCORING GUIDELINES

Question 1 (continued)

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

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

Notes:

- The description of a positive relationship should clearly indicate that relatively low values of one variable tend to appear with relatively low values of the other variable, and relatively high values of the first variable tend to appear with relatively high values of the other variable.
  - Examples of acceptable responses:
    - As length increases, so does weight.
    - Longer wolves weigh more.
    - The points on the graph go up as you move from left to right.
  - Examples of unacceptable responses:
    - As length goes up, weight changes.
    - Both length and weight get bigger.
    - The correlation is greater than 0.
- The description of a linear relationship can take one of two approaches: the data pattern (data points exhibit the pattern of a line in the graph) or the constant rate of change (as the explanatory variable changes, the response variable exhibits a constant rate of change).
  - Examples of acceptable responses:
    - The points generally follow a straight line.
    - The relationship between x and y is straight.
    - Length and width have a constant slope.
  - Examples of unacceptable responses:
    - The points all line up.
    - You can draw a straight line through the points.
    - There is a positive correlation.
    - Every increase in x yields a 35.02 increase in y.
- The description of strong should indicate how close points are to a line.
  - Examples of acceptable responses:
    - Observed values are close to predicted values.
    - Deviations from the least-squares regression line are small.
    - The correlation coefficient is close to 1.
  - Examples of unacceptable responses:
    - All the points are close together.
    - The scatterplots are clustered together.
    - There is a high positive correlation.
- Context can be shown by referring to length and weight or by using meters and kilograms.
- Sketches and graphs can be used to help clarify definitions, but a sketch alone cannot satisfy a definition component.

Part (b) is scored as follows:

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

1. The correct value of 35.02 for the slope.
2. An interpretation that includes an increase of a specified amount of weight for each unit increase in length.
3. An indication that the relationship is not exact by using words such as "on average" or "predicted weight."

2017 SCORING GUIDELINES

Question 1 (continued)

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

Note: If the response identifies the slope as -16.46 (the intercept value), the second component is satisfied only if the response states that for each one-meter increase in length there is a *decrease* in predicted or average weight of 16.46 kilograms.

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

art (c) is scored as follows:

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

1. A correct computation for the predicted value 32.568 kilograms.
2. A correct computation for the actual weight 22.9 kilograms using the given residual and the predicted value.

Partially correct (P) if the response provides a correct computation for the predicted value but is not able to complete the correct calculation of the actual weight, including if the residual is defined in the wrong direction as (predicted weight) – (actual weight) to give an answer of 42.24 kilograms;

OR

if the response provides an incorrect value for the predicted weight, but then uses that value correctly to determine the actual weight as (predicted weight) + residual = (predicted weight) $+(-9.67)$;

OR

if the response provides a correct answer for the actual weight but does not give sufficient information to determine how it was calculated.

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

Notes:

- The expression $-16.46 \quad 35.02(1.4)$ is enough to satisfy the first component.
- The equation $-16.46 + 35.02(1.4) - 9.67 = 22.9$ satisfies both components.
- Arithmetic mistakes are overlooked if they do not lead to an unreasonable answer (such as a negative value). For example, $32.568 + (-9.67) = 21.9$ satisfies the second component.

2017 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

2017 SCORING GUIDELINES

Question 2

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 and (2) use a confidence interval for a proportion to find a confidence interval for a dollar amount that can be calculated using that proportion.

Solution

Part (a):

Step 1: Identify the appropriate confidence interval by name or formula and check appropriate conditions.

The appropriate procedure is a one-sample z-interval for a population proportion p. In this case, the population is all customers of the restaurant who ask for a water cup, and p is the proportion of that population who will fill the cup with a soft drink. The appropriate formula is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Conditions:

1. Random sample
2. Large sample (number of successes $n\hat{p} \geq 10$ and number of failures $n(1-\hat{p}) \geq 10$)

For condition 1 the stem of the problem states that a random sample of customers who asked for a water cup was used.

For condition 2 the number of successes (filled cup with soft drink) is 23 and the number of failures is 57, both of which are greater than 10.

Step 2: Correct mechanics

The sample proportion is $\hat{p} = \frac{23}{80} = 0.2875$. The confidence interval is

$$0.2875 \pm 1.96\sqrt{\frac{0.2875(1 \quad 0.2875)}{80}}$$

$$= 0.2875 \pm 1.96(0.0506) \qquad \text{or } 0.1883 \text{ to } 0.3867.$$

$$= 0.2875 \pm 0.0992$$

Step 3: Interpretation

We can be 95 percent confident that in the population of all customers of the restaurant who ask for a water cup, the proportion who will fill it with a soft drink is between 0.1883 and 0.3867.

Part (b):

Using the confidence interval in part (a), a 95 percent interval estimate for the number of customers in June who asked for a water cup but then filled it with a soft drink is $3{,}000 \times 0.1883$ to $3{,}000 \times 0.3857$, or 565 to 1,160. At a cost of \$0.25 per customer, a 95 percent interval estimate for the cost to the restaurant in June is \$141.25 to \$290.00.

Question 2 (continued)

Scoring

This question is scored in four sections. Section 1 consists of step 1 in part (a), section 2 consists of step 2 in part (a), section 3 consists of step 3 in part (a), and section 4 consists of part (b). 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 one-sample *z*-interval for a proportion is identified (either by name or formula) *AND* both conditions (random sampling and large sample) are adequately addressed.

Partially correct (P) if the response identifies the correct procedure *BUT* adequately addresses only one of the two required conditions;
OR
if the response does not identify the correct procedure *BUT* adequately addresses both required conditions.

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

Notes:

- Stating the large sample condition without verifying it is not sufficient. The response must use specific numerical values to adequately address the condition.
- If the response includes additional inappropriate conditions, such as $n \geq 30$ or requiring a normal population, then the response earns at most P for section 1.
- Stating and checking a condition about the size of the sample relative to the size of the population is appropriate but not required.
- Any statement of hypotheses, description of the population, or definition of the parameter should be considered extraneous. However, if such statements are included and incorrect, they are considered as poor communication in terms of holistic scoring.

Section 2 is scored as follows:

Essentially correct (E) if the response gives the correct 95 percent confidence interval. Supporting work is not required, but if included, it must be correct.

Partially correct (P) if the response gives a correct confidence interval with incorrect (but appropriate) supporting work shown;
OR
if the response gives an incorrect but reasonable confidence interval with appropriate supporting work shown — for instance, if a value other than 1.96 is used for the critical value or a value other than $\frac{23}{80}$ is used for $\hat{p}$.

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

Notes:

- Appropriate supporting working must have the form:
  proportion ± (critical value)(SE/SD of proportion).

Question 2 (continued)

- A confidence interval that has both endpoints outside the interval from 0 to 1 is considered unreasonable.

Section 3 is scored as follows:

Essentially correct (E) if the response provides an appropriate interpretation of the interval that includes the following three components:
1. Conveys inference about a population proportion
2. Demonstrates a clear understanding that the parameter is the proportion of the water cup population that fills the cup with a soft drink
3. Mentions 95 percent confidence and interprets it correctly using words such as "We can be 95 percent confident" or "With 95 percent confidence"

Partially correct (P) if the response provides an appropriate interpretation of the interval that includes the first component *AND* only one of the other two components;
OR
if the response provides 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:

- Clear indication of an inference to the random sample of 80 customers rather than to the population does not satisfy the first component and is scored I.
- Stating values that are unrealistic as proportions or percentages (including blanks) in the interpretation lowers the score one level (from E to P, or from P to I).
- When both the interpretation and the level of the interval are given, only the interpretation is scored. If the interpretation of the confidence level is incorrect, it is considered as poor communication in terms of holistic scoring.
- Any interpretation that implies the interval has a 95 percent chance (or possibility or probability) of capturing the population proportion is scored I.

Section 4 is scored as follows:

Essentially correct (E) if the response gives a correct interval estimate for the cost to the restaurant *AND* shows enough work to indicate how the interval was found.

Partially correct (P) if the response gives the correct interval estimate for the *number* of customers (565 to 1,160) who would fill the water cup with soda, including showing work, but does not multiply by \$0.25 to find the interval for the cost;
OR
if the response gives the correct interval estimate for the *expected cost* to the restaurant for an individual who asked for a water cup (\$0.05 to \$0.10), including showing work, but does not multiply by 3,000 to find the interval for the total cost;
OR
if the response makes a reasonable attempt to find the endpoints of the interval as (0.1883)(3,000)(0.25) and (0.3867)(3,000)(0.25), but makes an error such as using the sample size of 80 instead of the population size of 3,000;

2017 SCORING GUIDELINES

Question 2 (continued)

OR

if the response gives the correct interval without showing how it was found.

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

Notes:

- The response in section 4 earns an E if it follows correctly from an incorrect interval found in part (a), even if the interval is unreasonable.
- Units ($) are not required to earn an E.

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.

2017 SCORING GUIDELINES

Question 3

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate a probability from a normal distribution; (2) calculate a weighted probability from two individual probabilities; and (3) calculate a conditional probability for dependent events when individual and joint probabilities are provided.

Solution

Part (a):

Let X denote the diameter of a randomly selected melon from Distributor J. X has an approximately normal distribution with mean 133 mm and standard deviation 5 mm.

The z-score for a diameter of 137 mm is $z = \frac{137 - 133}{5} = \frac{4}{5} = 0.8$.

Therefore, $P(X > 137) = P(Z > 0.8) = 1 - 0.7881 = 0.2119$.

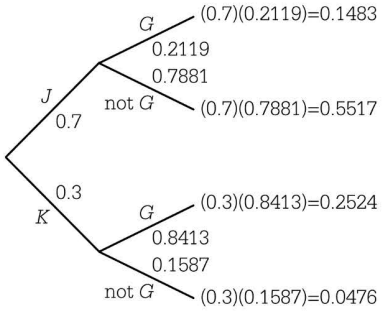
Part (b):

Define events:
J: melon is from Distributor J
K: melon is from Distributor K
G: melon diameter is greater than 137 mm

For a randomly selected melon from the grocery store,

$$\begin{aligned} P(G) &= P(G \mid J) \times P(J) + P(G \mid K) \times P(K) \\ &= (0.2119)(0.7) + (0.8413)(0.3) \\ &= 0.1483 + 0.2524 \\ &= 0.4007 \end{aligned}$$

OR



From the tree diagram, $P(G) = P(G \text{ and } J) + P(G \text{ and } K) = 0.1483 + 0.2524 = 0.4007$.

2017 SCORING GUIDELINES

Question 3 (continued)

Part (c):

Using the events defined in part (b), the requested probability is

$$P(J \mid G) = \frac{P(J \text{ and } G)}{P(G)} = \frac{P(G \mid J)P(J)}{P(G)} = \frac{(0.2119)(0.7)}{0.4007} = \frac{0.1483}{0.4007} = 0.3701.$$

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 provides the following three components:

1. Normality and parameters: Indicates use of a normal (or approximately normal) distribution and clearly identifies the correct parameter values. (Showing correct components in a z-score, labeling the mean and standard deviation in a normalcdf calculator statement, or drawing a normal curve with sufficient values marked to indicate the mean and standard deviation are all sufficient methods for identifying the correct parameter values.)
2. Boundary and direction: Uses the correct boundary value of $x = 137$ or $z = 0.8$ and the correct direction. (Showing correct boundary and direction in a probability statement using "greater than," labeling the lower limit and upper limit in a normalcdf calculator statement, drawing a normal curve with 137 labeled and an indication that the area of interest is to the right of 137, or a conclusion in words using "greater than.")
3. Probability: Reports the correct normal probability consistent with the response's setup described in components 1 and 2.

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

Incorrect (I) if the response gives the answer with no work shown or otherwise does not satisfy the criteria for E or P.

Notes:

- An inconsistency in calculations lowers the score for part (a) by one level (from E to P, or from P to I). For instance, if the response states the requested probability as $P(X > 137)$, but actually computes $P(X < 137)$.
- An error in statistical notation, such as using s instead of σ for the standard deviation or $\overline{x}$ instead of μ for the mean, does not satisfy component 1.
- A correct direction can be obtained by using a left direction and an upper bound of 137 and then subtracting the value of the cumulative probability from 1. However, simply showing the calculation $1 - 0.7881 = 0.2119$ does NOT give a correct direction.
- Consider any steps of a hypothesis test as extraneous work.
- Standard notations such as N(133,5) or N(133,25) satisfy the first component.
- A sketch of the normal curve with 133 and 138 in the appropriate positions satisfies the first component.
- Use of a value other than 137 does not satisfy the second component.
- If the only error in part (a) is the reversal of the numerator for the z-score $(133 - 137)$, the response is scored as P.

2017 SCORING GUIDELINES

Question 3 (continued)

- For calculator notation:
  - ncdf is sufficient to indicate use of the normal distribution.
  - Stating the lower and upper bounds is sufficient for identifying boundary and direction.
  - Minimum notation, such as "u" for upper, is sufficient for identifying boundary.
  - The minimum value for an upper bound must be at least 4 standard deviations above the mean.

Part (b) 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 method.

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

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

Notes:

- A response that uses 0.3 for J and 0.7 for K earns a P if no other mistakes are made.
- Tree diagrams:
  - If used, the multiplication is implied, but the addition must be indicated.
  - If not used, both the multiplication and the addition must be indicated.
- The shown work might be found in part (c).
- Probabilities can be shown in a table. Addition is implied if "Total" is labeled, as shown.

Intersection	G	Not G	Total
J	0.1483	0.5517	0.7
K	0.2524	0.0576	0.3
Total	0.4007	0.5993	

Part (c) is scored as follows:

Essentially correct (E) if the probability is computed correctly *AND* work is shown that illustrates how the probability was found.

Partially correct (P) if the response provides a strategy for finding the probability for dependent variables, such as the appropriate formula or end results from the correct tree diagram, but uses one or more incorrect values or does not use the end result from the tree;
OR
if the response gives the correct probability but not enough work is shown to determine how it was found.

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

2017 SCORING GUIDELINES

Question 3 (continued)

Notes:

- Part (c) earns an E if a correct answer follows from incorrect values used from part (a) and/or part (b), provided all values are between 0 and 1, inclusive.
- A transcription error in a response can be ignored if correct work is shown.
- For any part, 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 is lowered by one level (from E to P, or from P to I).
- In any part, unsupported answers are scored as I.

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

2017 SCORING GUIDELINES

Question 4

Intent of Question

The primary goals of this question were to assess a student's ability to use boxplots to (1) compare multiple sets of data; (2) identify which set of data is most likely to have produced a particular summary value; and (3) determine which variable is most useful for classifying a new observation.

Solution

Part (a):

The median value for the percent of chemical Z in the pottery pieces is similar for all three sites, at about 7 percent. The ranges for the percent of chemical Z are much different for the three sites, with the smallest range of about 2 percent (from 6 percent to 8 percent) at Site II, a range of about 6 percent (from about 4 percent to 10 percent) at Site I, and the largest range of about 8 percent (from about 3 percent to 11 percent) at Site III.

Part (b):

(i) The piece most likely originated at Site III. Although values outside of the range of data observed in the samples would be possible, using the available data results in approximate minimum and maximum sums of the percents for the three chemicals as shown in the table below. Site III is the only site in which 20.5 falls between the sums of the minimum and maximum values.

	Site I		Site II		Site III	
Chemical	Min	Max	Min	Max	Min	Max
X	6	8	5	7	5	7.5
Y	11	15	1.9	4	6	8
Z	4	10	6	8	3	11
Sum	21	33	12.9	19	14	26.5

(ii) Chemical Y would be most useful, because the distribution of the percentages of total weights at the three sites do not overlap. The distributions of chemicals X and Z have substantial overlap.

Scoring

This question is scored in three sections. Section 1 consists of part (a), section 2 consists of part (b-i), and section 3 consists of part (b-ii). 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 response includes the following three components:

1. Recognition that the medians or centers are almost the same for the three sites
2. Recognition that the variability (ranges, IQRs, spread) is different across the three sites
3. Context is included

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.

2017 SCORING GUIDELINES

Question 4 (continued)

Notes:

- In all sections, comments about shape should be ignored because complete shape information is not obtainable from boxplots.
- Responses are not required to give numerical values. If responses provide numerical values, any reasonable approximation from the boxplots is acceptable.
- Because the boxplots are all symmetric, it is acceptable if the response discusses means instead of medians.
- Any discussion of chemical X and chemical Y is considered extraneous.
- Context is satisfied by any of the following references: site, chemical, weight, total weight, X, Y, or Z.

Section 2 is scored as follows:

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

1. Site III is chosen.
2. Sums of the minimum and maximum are computed for the three chemicals at each site.
3. A reasonable numerical justification is given involving sums of a statistical measure across the three chemicals to choose Site III.

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.

Notes:

- If the response computes only the sum of the minimums for Site I and the sum of the maximums for Site II and recognizes that this is sufficient, the response is scored E.
- If an alternative measure is used that involves sums of the three chemicals, such as the sum of the medians or the sums of the first quartiles and sums of the third quartiles, instead of the minimum and maximum sums, the second component is not satisfied, but the third component might be satisfied.
    - If the response explicitly or implicitly compares the alternate sum to the other two sites (for example, by indicating that the sum is the closest to 20.5 percent or by listing the sums for all three sites) the response is scored P.
    - If the response does not have an implicit or explicit comparison, the response is scored I.
- If either Site I or Site II is identified as the correct choice, no matter how that choice is justified, the response is scored I.
- The approximate sums of the medians are 27.5 for Site I, 16 for Site II, and 20 for Site III.

Section 3 is scored as follows:

Essentially correct (E) if the response chooses chemical Y *AND* gives a reasonable justification based on the fact that the distributions of chemical Y are distinctive across sites.

Partially correct (P) if the response chooses chemical Y *AND* provides justification based on the boxplots, but does not clearly explain that the distributions of chemical Y are distinctive across sites;
OR
if the response correctly discusses that the distributions of chemical Y are distinctive across sites, but never explicitly chooses chemical Y as the best choice, for instance, by stating only that there is substantial overlap across sites for chemicals X and Z but no overlap for chemical Y.

2017 SCORING GUIDELINES

Question 4 (continued)

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

Notes:

- To justify that the distributions of chemical Y are distinctive across sites, the justification must address both location and variability of the boxplots; for example, by stating that the boxplots do not overlap for chemical Y.
- If the response chooses chemical X or Z *OR* chooses chemical Y with no reasonable justification, the response is scored I.
- The justification that the distributions of chemical Y are distinctive across sites:
    - The following are acceptable because both location and variability are addressed. Such responses are scored E.
        - The boxplots for chemical Y do not overlap, or the boxplots for chemicals X and Z overlap.
        - All values of Site I are high, all values of Site II are low, and all values of Site III are in the middle.
        - The ranges never intersect.
        - The boxplots share no data.
        - Has completely different percentages at each site.
    - The following are incomplete justifications and are scored P.
        - The boxplots vary.
        - Chemical Y varies the most.
        - Chemical Y has the greatest variation.
        - The variation between/among sites is the largest.
        - The boxplots are different.
        - The medians/means differ.
        - The medians/means are most variable.
        - There is a difference in the percentages of chemical Y for each site.
        - The distribution of percents differs the most among the sites.

2017 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 sections essentially correct and two sections partially correct

2017 SCORING GUIDELINES

Question 5

Intent of Question

The primary goal of this question was to assess a student's 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 a student's 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: State a correct pair of hypotheses.

The null hypothesis is that age group at diagnosis and gender are independent (that is, they are not associated) for the population of people currently being treated for schizophrenia.

The alternative hypothesis is that age group at diagnosis and gender are not independent for the population of people currently being treated for schizophrenia.

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

The appropriate test is a chi-square test of independence.

The conditions for this test are satisfied because:

1. The question states that the sample was randomly selected.
2. The expected counts for the eight cells of the table are at least 5, as seen in the following table, with expected counts shown below observed counts.

	Age at Diagnosis				
	20 to 29	30 to 39	40 to 49	50 to 59	Total
Women	46	40	21	12	119
	56.91	36.22	17.25	8.62	
Men	53	23	9	3	88
	42.09	26.78	12.75	6.38	

Step 3: Find the value of the test statistic and the *p*-value.

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

$$\chi^2 = 2.093 + 0.395 + 0.817 + 1.322 + 2.830 + 0.534 + 1.105 + 1.788 = 10.884.$$

The *p*-value is $P(\chi^2 \geq 10.884) = 0.012$, based on $(4-1) \times (2-1) = 3$ degrees of freedom.

2017 SCORING GUIDELINES

Question 5 (continued)

Step 4: State the conclusion in context, with linkage to the *p*-value.

Because the *p*-value is very small (for instance much smaller than $\alpha = 0.05$), we reject the null hypothesis and conclude that the sample data provide strong evidence that there is an association between age group at diagnosis and gender for the population currently being treated for schizophrenia.

Scoring

This question is scored in three sections. Section 1 consists of steps 1 and 2 (stating the correct hypotheses, identifying the appropriate test procedure, and checking the technical conditions); section 2 consists of step 3 (performing the correct mechanics); and section 3 consists of step 4 (stating a correct conclusion with justification). 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 correctly includes the following three components:
1. States BOTH hypotheses correctly with context included in at least one of them
2. Identifies a chi-square test of independence by name or formula
3. Verifies appropriate conditions that minimally include the condition for the expected counts and do not include any incorrect conditions (such as normality)

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.

Notes:
- Stating the expected count condition is not sufficient; the condition must be checked by reporting the expected counts, or minimally by showing that the smallest expected count is at least 5.
- The random sample condition was stated in the stem so need not be explicitly checked.
- If the null and alternative hypothesis are correctly stated in terms of population proportions, component 1 is satisfied. For example:
  $H_0 : p_1 = p_2 = p_3 = p_4$, where p_i is the population proportion of women at each indicated age group, 1, 2, 3, or 4, who are currently being treated for schizophrenia.
  H_a : At least one of the population proportions, p_1, p_2, p_3, p_4, differs from the other three.
  OR
  H_a : The population proportions for the four age groups are not all the same.

Section 2 is scored as follows.

Essentially correct (E) if the response correctly calculates the following two values:
1. The value of the chi-square test statistic
2. The *p*-value, critical value, or *p*-value range from chi-square table

Partially correct (P) if the response correctly calculates only one of the two values.

2017 SCORING GUIDELINES

Question 5 (continued)

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

Notes:
- If the response makes an error in one calculation, subsequent calculations are considered correct if they follow correctly from the initial miscalculation.
- With 3 degrees of freedom, the correct critical value is 7.81 for a significance level of 0.05 and 11.34 for a significance level of 0.01.
- Work does not have to be shown for calculations of test statistic or *p*-value. However, if incorrect work (other than minor arithmetic/transcription errors) is shown it is considered to be an incorrect calculation of the respective component, even if the correct value is given.
- If a response provides a test statistic that is not a chi-square test statistic, section 2 is scored I.

Section 3 is scored as follows.

Essentially correct (E) if the response includes the following three components:
1. A correct conclusion about the alternative hypothesis.
2. Justification of the conclusion based on linkage between the *p*-value and a reasonable alpha (or linkage between test statistic and critical value).
3. The conclusion is stated in context.

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.

Notes:
- If the response provides a correct decision, in context, with linkage to the *p*-value, but the decision is stated in terms of the null hypothesis with no conclusion about the alternative hypothesis, component 1 is not satisfied.
- Incorrect statistical statements are considered incorrect conclusions for the hypothesis test and do not satisfy component 1.
- 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 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 explaining how the conclusion follows from the *p*-value. For example, stating that because the *p*-value is small, we reject the null hypothesis or stating that because the *p*-value is large, we do not reject the null hypothesis.
- A decision about the null hypothesis (reject H_0 or fail to reject H_0) is not required, but if such a statement is given the scoring of the decision is considered in component 2.

2017 SCORING GUIDELINES

Question 5 (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

Section 1 partially correct and the other two sections incorrect

2017 SCORING GUIDELINES

Question 6

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate probabilities associated with treatment and control group memberships for two different methods of random assignment and (2) justify which method of random assignment is more appropriate in a given situation.

Solution

Part (a):

(i) Let T (tail) represent being assigned to the treatment group and H (head) represent being assigned to the control group for each coin flip. The process stops when either the treatment group or the control group has two members. The outcomes and their probabilities are as follows.

Arrangement	A	B	C	D	E	F
Chip outcomes	TT	THT	THH	HH	HTH	HTT
Calculation	$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$	$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$	$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$	$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$	$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$	$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$
Probability	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

(ii) Man 1 and Man 2 are assigned to the same group for arrangements A and D, so the probability is

$$P(\text{A}) + P(\text{D}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Part (b):

(i) Let T represent being assigned to the treatment group and C represent being assigned to the control group for each chip drawn. The process stops when either the treatment group or the control group has two members. The probabilities differ from the coin flip method because chips are drawn *without* replacement. The outcomes and their probabilities are as follows.

Arrangement	A	B	C	D	E	F
Chip outcomes	TT	TCT	TCC	CC	CTC	CTT
Calculation	$\left(\frac{2}{4}\right)\left(\frac{1}{3}\right)$	$\left(\frac{2}{4}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$	$\left(\frac{2}{4}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$	$\left(\frac{2}{4}\right)\left(\frac{1}{3}\right)$	$\left(\frac{2}{4}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$	$\left(\frac{2}{4}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

(ii) Man 1 and Man 2 are assigned to the same group for arrangements A and D, so the probability is

$$P(\text{A}) + P(\text{D}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

2017 SCORING GUIDELINES

Question 6 (continued)

Part (c):

Use the chip method. The chip method gives equal probability to all possible arrangements, but the coin method does not, as shown in the tables from parts (a-i) and (b-i). Furthermore, the coin method is more likely to result in imbalanced treatment groups with regard to students and teachers, based on the probabilities in parts (a-ii) and (b-ii). If food preferences for teachers are different than for students, the imbalance is a problem. For example, if one treatment group consists entirely of students, it would be impossible to know if a difference in the response variable is due to the treatment (type of meal) or the role of the person at the school (teacher or student).

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 includes the following three components:
1. The correct probability for each arrangement in the table of (i).
2. Appropriate justification is shown in (i).
3. The correct probability is given in (ii).

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

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

Notes:
- For component 2, examples of appropriate justification include the following:
  - Shows calculation of a correct probability for at least one of arrangements A and D *AND* shows calculation of a correct probability for at least one of arrangements B, C, E, and F. Calculations must be linked to a specific arrangement with a label, by proximity to an arrangement in the table, or by following the structure of the table or display of arrangements.
  - Lists correct coin outcomes for at least one of arrangements A and D *AND* lists correct coin outcomes for at least one of arrangements B, C, E, and F (for example, A: TT and B: THT). Lists of outcomes must be linked to a specific arrangement with a label, by proximity to an arrangement in the table, or by following the structure of the table or display of arrangements.
  - Shows a tree diagram that ends with six branches corresponding to the six arrangements. Probabilities or coin outcomes must be listed on the branches, but arrangements do not need to be identified.
  - Shows a tree diagram that ends with eight (or 16) branches where the branches leading to at least one of arrangements A and D are identified, and the branches leading to at least one of arrangements B, C, E, and F are identified. Probabilities or coin outcomes must be listed on the branches.
- If an incorrect probability is given in part (ii), component 3 is satisfied if the probability is between 0 and 1 and:
  - is the result of a minor arithmetic error; *OR*
  - is the result of adding the probabilities of arrangements A and D from the table in (i).

2017 SCORING GUIDELINES

Question 6 (continued)

Part (b) is scored as follows:

Essentially correct (E) if the response includes the following two components:
1. The correct probability for each arrangement in the table of (i).
2. The correct probability is given in (ii).

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

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

Note: If an incorrect probability is given in (ii), component 2 is satisfied if the probability is between 0 and 1 and
- is the result of a minor arithmetic error; *OR*
- is the result of adding the probabilities of arrangements A and D from the table in (i); *OR*
- work is shown that demonstrates the same incorrect approach is used in both parts (a-ii) and (b-ii). For example, component 2 is satisfied if the response forgets to include $P(\mathbf{D})$ and says $P(\text{A}) = \frac{1}{4}$ in (a-ii) and $P(\text{A}) = \frac{1}{6}$ in part (b-ii). However, unlabeled answers of $\frac{1}{4}$ in (a-ii) and $\frac{1}{6}$ in (b-ii) would not satisfy component 2 because it is unclear if the same approach is being used.

Part (c) is scored as follows:

Essentially correct (E) if the response chooses the chip method AND includes the following three components:
1. Provides a statistical benefit to the chip method. For example, stating that all arrangements are equally likely.
2. Provides a statistical drawback to the coin method. For example, stating that the coin method is more likely to result in imbalanced treatment groups.
3. States that the responses (opinions, food preferences) of teachers and students might be different.

Partially correct if the response chooses the chip method and includes only one or two of the three components.

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

Notes:
- A correct comparative statement can be used to satisfy both components 1 and 2. For example, stating that the coin method is *more likely* to result in a disproportionate number of students in one treatment group, or that the chip method has a *greater chance* of providing roughly equivalent treatment groups.
- Benefits and drawbacks must be about the probabilities of the arrangements (from the tables in (i)), or the probabilities of imbalanced (balanced) groups (from the questions in (ii)), or both.
- If the response chooses the coin method, look back to see if the coin method should be the preferred method based on incorrect work from parts (a) and (b). If so, score part (c) using the corresponding three components.

2017 SCORING GUIDELINES

Question 6 (continued)

- If the response says that it does not matter which method to use, look back to see if the tables are the same in parts (a-i) and (b-i) or if the probabilities are the same in parts (a-ii) and (b-ii). If so, part (c) is partially correct if the response justifies the decision by stating that the two tables or the two probabilities are the same. To be essentially correct, the response also needs to satisfy component 3.
- If the response does not choose a method, deduct one component from the number of correct components.

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
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):

- 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.

- 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 percent 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 Choc 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 percent 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 d 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 mial 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 nterval 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
2015 Scoring Guidelines

2015 SCORING GUIDELINES

Question 1

Intent of Question

The primary goals of this question were to assess a student's ability to (1) compare features of two distributions of data displayed in boxplots and (2) identify statistical measures that are important in making decisions based on data sets.

Solution

Part (a):

The median salary is approximately the same for both corporations. The range and interquartile range of the salaries are greater for Corporation A than for Corporation B. The two highest salaries at Corporation A are outliers while Corporation B has no outliers.

Part (b):

(i) Five years after starting, at least 3 out of 30 (10%) of the salaries at Corporation A are greater than the maximum salary at Corporation B. If I accept the offer from Corporation A, I might be able to make a higher salary at Corporation A than at Corporation B.

(ii) Five years after starting, the minimum salary at Corporation B is greater than at Corporation A. In fact, at Corporation A it looks like some people are still making the starting salary of $36,000 and never received a raise in the five years since they were hired. So if I work at Corporation A, I might never receive a raise in salary.

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 the following four components:
1. A correct comparison of center.
2. A correct comparison of spread.
3. A discussion of the outliers for Corporation A.
4. The response is in context.

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

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

Note: Any mention of shape should be ignored because complete shape information cannot be determined from a boxplot.

2015 SCORING GUIDELINES

Question 1 (continued)

Part (b) is scored as follows:

Essentially correct (E) if the response includes the following four components:
1. In part (b-i) a relevant statistical measure is identified (or described) or a relevant statistical comparison is provided that supports the choice of Corporation A.
2. In part (b-i) an explanation is provided for why the measure or comparison is relevant.
3. In part (b-ii) a relevant statistical measure is identified (or described) or a relevant statistical comparison is provided that supports the choice of Corporation B.
4. In part (b-ii) an explanation is provided for why the measure or comparison is relevant.

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

Incorrect (I) if the response includes none or one of the four components.

Note: If a response does not provide a statistical measure or comparison in part (b-i) or (b-ii), the second and fourth components can still be satisfied if an acceptable explanation is provided that would follow from a relevant statistical measure or comparison. For example, if the response in part (b-i) only states "At Corporation A, I have the potential to earn a higher salary," the second component is satisfied.

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

2015 SCORING GUIDELINES

Question 2

Intent of Question

The primary goals of this question were to assess a student's ability to (1) use confidence intervals to test a question about a proportion and (2) understand the relationship between sample size and margin of error in a confidence interval for a proportion.

Solution

Part (a):

(i) No. The confidence interval is $(0.09, 0.21)$, which includes the value of 0.20. Therefore, it is plausible that the computer program is generating discounts with a probability of 0.20, and the confidence interval does not provide convincing statistical evidence that the program is not working as intended.

(ii) No. The confidence interval includes values from 0.09 to 0.21, so any value in that interval is a plausible value for the probability that the computer is using to generate discounts.

Part (b):

The formula for computing the margin of error for a proportion includes the square root of the sample size in the denominator. For a random sample that is four times the size of the original sample, the margin of error can be determined by dividing the margin of error of the original sample by two. Therefore, the new margin of error is 0.03.

Part (c):

Using the margin of error of 0.03 obtained from the second sample, the confidence interval for p is 0.15 ± 0.03 or $(0.12, 0.18)$. The interval does not include 0.20, and therefore, there is convincing evidence that the computer program is not working as intended and is not generating discounts with a probability of 0.20.

Scoring

This question is scored in four sections. Section 1 consists of part (a-i); section 2 consists of part (a-ii); section 3 consists of part (b); and section 4 consists of part (c). 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 states that because the interval contains 0.20, it does not provide convincing statistical evidence that the computer program is not working as intended.

Partially correct (P) if the response indicates that it is necessary to check whether the value of 0.20 is in the computed interval, but there are errors in implementation. Examples of errors include:

- The response notes that 0.20 is within the interval but does not draw a conclusion.
- The response has an arithmetic error in the computation of the endpoints of the interval but provides a correct conclusion with justification that is consistent with the computed interval.

2015 SCORING GUIDELINES

Question 2 (continued)

Incorrect (I) if the response does not recognize how to use the confidence interval to check whether the computer is working correctly;
OR
if the response states that the interval shows that the proportion is equal to 0.20;
OR
if the response notes that 0.20 is within the interval and concludes that the program is not working as intended;
OR
if the response otherwise does not meet the criteria for E or P.

Section 2 is scored as follows:

Essentially correct (E) if the response concludes that there is not convincing statistical evidence that the computer program generates the discount with a probability of 0.20 *AND* justifies the conclusion by noting that there are values other than 0.20 in the interval.

Partially correct (P) if the response correctly concludes that there is not convincing evidence that the computer program generates the discount with a probability of 0.20, but provides incomplete reasoning to justify the conclusion.

Examples of incomplete reasoning include:

- stating that 0.20 is a plausible value for the proportion of discounts without giving further explanation;
- indicating that having 0.20 in the interval does not prove that 0.20 is the true proportion; and
- providing a generic statistical argument, such as reference to the fact that the null hypothesis should never be accepted, or stating that not rejecting the null hypothesis is not proof that 0.20 is the true proportion of bills discounted.

Note: If an incorrect interval is computed in part (a-i) that does not contain 0.20, and in part (a-ii) the response concludes that the program is not generating discounts with a probability of 0.20 because 0.20 is not in the interval, section 2 is scored as P.

Incorrect (I) if the response states that there is evidence that the computer program generates the discount with a probability of 0.20;
OR
if the response correctly concludes that there is not convincing evidence that the computer program generates the discount with a probability of 0.20 *AND* provides incorrect or no justification;
OR
if the response otherwise does not meet the criteria for E or P.

Section 3 is scored as follows:

Essentially correct (E) if the response gives the correct value of 0.03 as the new margin of error by using the correct formula or by providing a correct explanation that recognizes that quadrupling the sample size divides the margin of error from the original sample by two.

Note: If the response relies on the margin of error formula and calculates a value that would round to 0.030, the response is scored as E.

2015 SCORING GUIDELINES

Question 2 (continued)

Partially correct (P) if the response uses the correct margin of error formula or provides a correct explanation that recognizes that quadrupling the sample size divides the margin of error from the original sample by two, but calculates an incorrect margin of error that is less than 0.06;
OR
if the response does not calculate a margin of error but provides a correct explanation that recognizes that quadrupling the sample size divides the margin of error from the original sample by two;
OR
if the response gives the correct margin of error without correct justification.

Incorrect (I) if the response does not recognize in any way that the new margin of error depends on the square root of the sample size;
OR
if the response calculates a margin of error that is greater than or equal to 0.06;
OR
if the response otherwise does not meet the criteria for E or P.

Section 4 is scored as follows:

Essentially correct (E) if the conclusion states that there is convincing evidence that the computer program is not working as intended because, based on the new margin of error, the interval does not contain 0.20;
OR
if the conclusion states the intended value of 0.20 is greater than the upper boundary of 0.18;
OR
if the conclusion states the margin of error is smaller than the difference between the sample proportion and the intended long-run proportion of 0.20.

Notes:
- If the margin of error was computed incorrectly in part (b), but a correct answer to part (c) is consistent with this incorrect margin of error, section 4 is scored as E.
- If no specific margin of error or confidence interval was given in parts (b) or (c), but the response provides a complete and correct conclusion with reference to a smaller margin of error (or narrower confidence interval), section 4 is scored as E.

Partially correct (P) if an interval is incorrectly constructed using the margin of error from part (b), but a correct conclusion is given and justified for the computed interval;
OR
if an interval is computed correctly using the margin of error from part (b) and an argument is made based on whether or not 0.20 is in the interval, but the conclusion is incorrect;
OR
if no specific margin of error or confidence interval has been given in parts (b) or (c), but the response concludes that because the margin of error has decreased (or the confidence interval is narrower), 0.20 is not in the interval.

Incorrect (I) if an interval is computed correctly using the margin of error from part (b) and a correct conclusion is given, but no argument is made based on whether or not 0.20 is in the interval;
OR
if the response otherwise does not meet the criteria for E or P.

2015 SCORING GUIDELINES

Question 2 (continued)

Notes:
- If a response includes a confidence level, the level can be ignored because no confidence level was provided.
- If the response bases a conclusion on the relative location of 0.20 within the interval (for example, 0.20 is near the edge of the interval), the response is scored as I.
- A response that provides additional incorrect explanation lowers the score in section 4 by one level (that is, from E to P, or from P to I).

Each essentially correct (E) section counts as 1 point. Each 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 overall 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) perform a probability calculation from a discrete random variable; (2) calculate the expected value of a discrete random variable; (3) perform a conditional probability calculation from a discrete random variable; and (4) use probabilistic thinking to make a prediction about how an expected value will change given a condition about the random variable.

Solution

Part (a):

The probability that at least one ATM is working when the mall opens is:

$$P(X \geq 1) = 0.21 + 0.40 + 0.24 = 0.85.$$

Part (b):

The expected value of the number of ATMs that are working when the mall opens is:

$$E(X) = 0(0.15) + 1(0.21) + 2(0.40) + 3(0.24) = 1.73 \text{ machines.}$$

Part (c):

probability that all three ATMs are working when the mall opens, given that at least one ATM is working is:

$$P(X = 3 \mid X \geq 1) = \frac{P(X = 3 \text{ and } X \geq 1)}{P(X \geq 1)} = \frac{P(X = 3)}{P(X \geq 1)} = \frac{0.24}{0.85} \approx 0.282$$

Part (d):

Given that at least one ATM is working when the mall opens, the expected value of the number of working ATMs would be greater than the expected value calculated in part (b). By eliminating the possibility of 0 working ATMs, the probabilities for 1, 2, and 3 working ATMs all increase proportionally, so the expected value must increase.

Scoring

(b), (c), and (d) are 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 with work shown.

Partially correct (P) if the correct answer is given, but no work is shown;
OR
ppropriate work is shown but the answer is incorrect or missing;
OR

SCORING GUIDELINES

Question 3 (continued)

ne of the incorrect cumulative probabilities $P(X < 1)$, $P(X \leq 1)$, or $P(X > 1)$ is stated *AND* computed correctly.

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

Note: The probability can be calculated as $1 - P(X = 0) = 1 - 0.15 = 0.85$.

Part (b) is scored as follows:

Essentially correct (E) if the expected value is computed correctly with work shown.

Partially correct (P) if the correct answer is given, but no work is shown;
OR
if appropriate work is shown but the answer is incorrect or missing.

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

Part (c s scored as follows:

Essentially correct (E) if the probability is computed correctly, with work shown that includes correct 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 (numerator or denominator) correct in numerical value and the other incorrect;
OR
if the correct answer is given, but no work is shown.

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

Part (d) is scored as follows:

Essentially correct (E) if the response provides the correct answer (greater than) with a reasonable explanation based on the fact that with $X = 0$ eliminated, the probabilities for $X = 1$, $X = 2$, $X = 3$ *all* increase;
OR
if the response provides the correct answer (greater than) with a reasonable explanation based on fact that with $X = 0$ eliminated, the balance point of the distribution increases;
OR
if the response provides the correct answer (greater than) and the conditional expected value is computed correctly with work shown.

Partially correct (P) if the response provides the correct answer (greater than) with a weak explanation, such as "Yes, because 0 is eliminated";
OR

SCORING GUIDELINES

Question 3 (continued)

if the response provides the correct answer (greater than) with a reasonable but incorrect attempt to calculate the conditional expected value using a revised probability distribution with only the values $X = 1$, $X = 2$, and $X = 3$, and corresponding probabilities greater than or equal to those in the original probability distribution.

Incorrect (I) if the response does not provide the correct answer (greater than);
OR
if the response provides the correct answer (greater than) with an incorrect explanation or no explanation.

Note: The conditional expected value is:

$$1\left(\frac{0.21}{0.85}\right) + 2\left(\frac{0.40}{0.85}\right) + 3\left(\frac{0.24}{0.85}\right) \approx 1(0.247) + 2(0.471) + 3(0.282) = 2.04 \text{ machines.}$$

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 score up or down, depending on the overall strength of the response and communication.

2015 SCORING GUIDELINES

Question 4

Intent of Question

The primary goal of this question was to assess a student's 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 a student's 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 p_{asp} represent the population proportion of adults similar to those in the study who would have developed colon cancer within the six years of the study if they had taken a low-dose aspirin each day. Similarly, let p_{plac} represent the population proportion of adults similar to those in the study who would have developed colon cancer within the six years of the study if they had taken a placebo each day.

The hypotheses to be tested are $H_0 : p_{asp} = p_{plac}$ versus $H_a : p_{asp} < p_{plac}$ or equivalently, $H_0 : p_{asp} - p_{plac} = 0$ versus $H_a : p_{asp} - p_{plac} < 0$.

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

The appropriate procedure is a two-sample *z*-test for comparing proportions.

Because this is a randomized experiment, the first condition is that the volunteers were randomly assigned to one treatment group or the other. The condition is satisfied because we are told that the volunteers were randomly assigned to take a low-dose aspirin or a placebo.

The second condition is that the sample sizes are large, relative to the proportions involved. The condition is satisfied because all sample counts are large enough; that is, 15 with colon cancer in aspirin group, 26 with colon cancer in placebo group, $500 - 15 = 485$ cancer-free in aspirin group, and $500 - 26 = 474$ cancer-free in placebo group.

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

The sample proportions who developed colon cancer are $\hat{p}_{asp} = \frac{15}{500} = 0.030$ and $\hat{p}_{plac} = \frac{26}{500} = 0.052$.

The combined sample proportion who developed colon cancer is $\hat{p}_{combined} = \frac{15 + 26}{500 + 500} = 0.041$.

The test statistic is $z = \frac{0.030 - 0.052}{\sqrt{0.041(1 - 0.041)\left(\frac{1}{500} + \frac{1}{500}\right)}} \approx -1.75$ (-1.7542 from calculator).

The *p*-value is $P(Z \le -1.75) = 0.0401$ (0.0397 from calculator), where Z has a standard normal distribution.

2015 SCORING GUIDELINES

Question 4 (continued)

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

Because the *p*-value is less than the given significance level of $\alpha = 0.05$, we reject the null hypothesis. The data provide convincing statistical evidence that the proportion of all adults similar to the volunteers who would develop colon cancer if they had taken a low-dose aspirin every day is less than the proportion of all adults similar to the volunteers who would develop colon cancer if they had not taken a low-dose aspirin every day.

Scoring

Steps 1, 2, 3, and 4 are scored as essentially correct (E), partially correct (P), or incorrect (I).

Step 1 is scored as follows:

Essentially correct (E) if the response identifies correct parameters *AND* both hypotheses are labeled and state the correct relationship between the parameters.

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

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

Note: Either defining the parameters in context, or simply using common parameter notation with subscripts clearly relevant to the context, such as p_{asp} and p_{plac}, is sufficient.

Step 2 is scored as follows:

Essentially correct (E) if the response correctly includes the following three components:
1. Identifies the correct test procedure (by name or by formula).
2. Notes that the use of random assignment satisfies the randomness condition.
3. Checks for approximate normality of the test statistic by citing that all four counts are larger than some standard criterion such as 5 or 10.

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

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

Notes:
- For the randomness component, it is (minimally) acceptable to say "random assignment — check" but not acceptable to say "random — check" or "SRS — check." The important concept here is that it is random assignment, and not random sampling, that is required. If the response implies that the study used a random sample, the randomness component is not satisfied, regardless of whether random assignment is correctly addressed.
- The normality check may use the expected counts under the null hypothesis in place of observed counts.

2015 SCORING GUIDELINES

Question 4 (continued)

Step 3 is scored as follows:

Essentially correct (E) if the response correctly calculates both the test statistic and a *p*-value that is consistent with the stated alternative hypothesis.

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;
OR
if the response reports the correct *p*-value but no calculations or test statistic are shown.

Incorrect (I) if the response fails to meet the criteria for E or P.

Note: The *p*-value is considered correct if it is consistent with the alternative stated in the response and the calculated test statistic, even if those are incorrect.

Step 4 is scored as follows:

Essentially correct (E) if the response provides a correct conclusion in context, with justification based on linkage between the *p*-value and the given $\alpha = 0.05$.

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:
- The conclusion must be related to the alternative hypothesis.
- If the *p*-value is incorrect, then step 4 is scored as E if the response includes proper linkage and a conclusion in context consistent with that *p*-value.
- If the *p*-value is less than 0.05, wording that states or implies that the alternative hypothesis is *proven* lowers the score one level (that is, from E to P or P to I) in step 4.
- If the *p*-value is incorrect and greater than 0.05, wording that states or implies that the null hypothesis is *accepted* lowers the score one level (that is, from E to P or P to I) in step 4.

2015 SCORING GUIDELINES

Question 4 (continued)

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.

SCORING GUIDELINES

Question

Intent of Question

The primary goals of this question were to assess a student's ability to (1) use the information provided by a scatterplot to describe the relationship between two quantitative variables; (2) interpret and use the information given by lines displayed on a scatterplot; and (3) use a regression equation to estimate a predicted value of y for a given x value.

Solution

Part

There is a moderately strong, positive, linear relationship between height and arm span so that taller students tend to have longer arm spans.

Part (b):

The line in Graph 2 is the one that is helpful. For each student, the graph illustrates whether arm is equal to height (points on the line), arm span is greater than height (points above the line), or arm span is less than height (points below the line).

Classification	Square	Tall Rectangle	Short Rectangle
Frequency	3	4	5

Part (c):

The predicted arm span is $\hat{y} = 11.74 + 0.8247x = 11.74 + 0.8247(61) = 62.05$ inches.

Scoring

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 includes the following four components:
1. The relationship is approximately linear.
2. The relationship is positive.
3. There is a moderate to strong association (or relationship).
4. The response is given in context.

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.

SCORING GUIDELINES

Question 5 (continued)

Notes:

- Correlation alone is not sufficient to indicate a linear relationship. However, linear correlation sufficient to indicate a linear relationship.
- Reporting that the correlation coefficient is 0.81, the correlation coefficient is close to 1.0, or some other value or range of values for the correlation coefficient is not sufficient to satisfy component 3.

Part (b) is scored as follows:

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

1. The line in Graph 2 is selected.
2. A reasonable explanation for selecting the line in Graph 2 that links the body shapes to the regions defined by the line is given. The explanation should indicate that square shapes (arm span is equal to height) are represented by points on the line, tall rectangle shapes (arm spa less than height) are represented by points below the line, and short rectangle shapes (arm span is greater than height) are represented by points above the line.
3. Correct counting of body types demonstrated by reporting correct frequencies (3, 4, and 5) reporting correct proportions $\left(\frac{1}{4} = \frac{3}{12}, \frac{1}{3} = \frac{4}{12}, \text{ and } \frac{5}{12}\right)$ for the square, tall rectangle, and short rectangle body types, respectively.
4. The correct frequencies (3, 4, and 5) are reported in the table.

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.

Notes:

- Selecting the regression line on Graph 1 cannot satisfy either component 1 or component 2.
- To satisfy component 1, it is sufficient to refer to the $y = x$ line without explicitly mentioning Graph 2.
- Use of incorrect labels, such as regression line or least-squares regression line, in referring to the $y = x$ line in Graph 2 is an incorrect use of terminology that should be strongly discouraged, but it is ignored in this scoring rubric.
- The explanation for selecting Graph 2 is acceptable if it explicitly includes at least two of the following: (i) square body shapes (arm span equal to height) are represented by points line, (ii) tall rectangle body shapes (arm span less than height) are represented by points *below* the line, or (iii) short rectangle body shapes (arm span greater than height) are represented by points *above* the line.
- Incorrectly counting the points on the $y = x$ line in Graph 2 is considered a minor error. Frequencies (2, 4, and 5) are accepted for both component 3 and component 4. Reporting corresponding proportions $\left(\frac{2}{11}, \frac{4}{11}, \frac{5}{11}\right)$ or $\left(\frac{2}{12}, \frac{4}{12}, \frac{5}{12}\right)$ in the table satisfies component does not satisfy component 4.
- Frequencies reported in the table as (2, 5, 4) or (3, 5, 4), interchanging the counts for tall and short rectangle body shapes, satisfies component 3, but does not satisfy component 4. Reporting corresponding proportions, $\left(\frac{2}{11}, \frac{5}{11}, \frac{4}{11}\right)$ $\left(\frac{3}{12}, \frac{5}{12}, \frac{4}{12}\right)$, in the table does not satisfy either component 3 or component 4.

SCORING GUIDELINES

Question 5 (continued)

Part (c) is scored as follows:

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

1. Correct formula for predicting arm span with 61 inserted for x.
2. Correct value for the predicted arm span.
3. Units for the predicted arm span given as inches.

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.

Notes:

- Any value for the predicted arm span between 61.5 and 62.5 is acceptable; values outside th interval do not satisfy component 2. This allows for inaccuracy in obtaining a value for the predicted arm span directly from the line displayed on Graph 1 and for rounding in applying the prediction formula; but it excludes predictions based on the $y = x$ line and other unreasonable predictions.
- Reporting a prediction equation that is similar to the equation given in the stem, with 61 inserted for x, satisfies components 1 and 2 if the value of the predicted arm span is between 61.5 and 62.5 inches. Otherwise, neither of those two components is satisfied. This could occur if a student enters the data from the graph into a calculator to compute the least squares regression line, instead of using the formula provided in the stem.

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 two parts partially correct

OR

One part essentially correct, one part partially correct, and one part incorrect

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

2015 SCORING GUIDELINES

Question 6

Intent of Question

The primary goals of this question were to assess a student's ability to (1) describe how sample data would differ using two different sampling methods; (2) describe the sampling distribution of the sample mean for two different sampling methods; and (3) choose the sampling method that will result in the best estimate of the population mean.

Solution

Part (a):

No, a sample obtained using Method 2 will not be representative of all tortillas made that day. The sample obtained using Method 2 will only represent the tortillas from one production line, not from the entire population because the distributions of diameters for the two production lines are different.

Part (b):

Method 1 was most likely used to select this sample. The bimodal shape in the histogram of sample data indicates that tortillas were selected from both production lines, which is what would happen using Method 1. Method 2 would be likely to produce a unimodal distribution of diameters centered at either 5.9 inches or 6.1 inches.

Part (c):

Method 2 would result in less variability in the sample of 200 tortillas on a given day because the sample comes from only one production line. Because the distributions of diameters are not the same for the two production lines, selecting tortillas from both lines as in Method 1 would result in more variable sample data.

Part (d):

The sampling distribution of the sample mean diameter for samples obtained using Method 1 would be approximately normal with mean 6 inches and standard deviation $\frac{0.11}{\sqrt{200}} \approx 0.0078$ inch.

Part (e):

Method 1 would result in less variability in the sample means over the 365 days, because with Method 2, roughly half of the sample means will be around 5.9 inches and the other half will be around 6.1 inches. With Method 1, however, the sample means will all be very close to 6 inches, as indicated by the standard deviation in part (d).

Part (f):

Method 1 is more likely to produce a sample mean close to 6 inches. Because the sample mean is an unbiased estimator for both methods, the manager should pick the method that would result in less variability in the distribution of the sample mean. Based on the answer to part (e), Method 1 results in less variability in the distribution of the sample mean.

2015 SCORING GUIDELINES

Question 6 (continued)

Scoring

This question is scored in three sections. Section 1 consists of parts (a), (b), and (c), section 2 consists of part (d), and section 3 consists of parts (e) and (f). 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 three components:
1. In part (a) the response says no *AND* either argues that the sample will only be selected from one production line and not the entire population (that is, the sample will only represent one production line) *OR* argues that the tortillas from the two productions lines are different.
2. In part (b) the response chooses Method 1 *AND* refers to a relevant characteristic of the histogram (shape, center, or variability) that matches what would be expected when using Method 1 (or that does not match what would be expected when using Method 2).
3. In part (c) the response chooses Method 2 *AND* either justifies by stating that the sample comes from only one production line (does not come from both production lines) *OR* justifies by comparing the possible range of diameters for the two methods.

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.

Note: If a response includes more than one justification in an individual part, score the weaker of the two justifications. For example, a response for part (a) that says "No, because only one line was selected and because the sample size was too small" does not satisfy the first component because the sample size argument is incorrect.

Section 2 is scored as follows:

Essentially correct (E) if the response includes the following three components:
1. The shape is approximately normal.
2. The mean is 6 inches.
3. The standard deviation is $\frac{0.11}{\sqrt{200}}$ inch.

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.

Note: It is not necessary to include units (inches) for the mean or standard deviation.

Section 3 is scored as follows:

Essentially correct (E) if the response includes the following two components:
1. In part (e) the response chooses Method 1 *AND* describes the sampling distribution of the sample mean for Method 2 as having some *sample means* close to the mean of production line A (5.9 inches) and the other *sample means* close to the mean of production line B (6.1 inches).

2015 SCORING GUIDELINES

Question 6 (continued)

2. In part (f) the response chooses Method 1 *AND*
   - refers to a correct answer in part (e);

   OR
   - describes the sampling distribution of the sample mean for Method 2 as having some *sample means* close to the mean of production line A (5.9, less than 6) and the other *sample means* close to the mean of production line B (6.1, greater than 6);

   OR
   - argues that on a single day, it would be preferable to get a sample with a mean around 6 rather than getting a sample that would have a mean around 5.9 (less than 6) or a mean around 6.1 (greater than 6).

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

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

Notes:

- In part (e), the response must be clear that there is *more than one* mean being described (365 sample means).
  - Correct: The sample means will be around 5.9 or 6.1.
  - Not correct: The sample mean will be around 5.9 or 6.1.

- Parts (e) and (f) are not satisfied if the response does not imply the sample means *vary* from the population means in both methods (or does not imply that the sample mean varies from the population mean when making the single day argument in part (f)). However, if the mistake is made in part (e), do not penalize the response in part (f) for the same mistake.
  - Correct: The sample means will be around 5.9 or 6.1 instead of close to 6.
  - Not correct: The sample means will be 5.9 or 6.1 instead of close to 6.
  - Not correct: The sample means will be 6 instead of around 5.9 or 6.1.

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 either one or two sections 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