

**Question 2: Focus on Sampling and Experimental Design****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                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>(a) Experimental units: 60 driveways.</p> <p>Treatments: Concrete with fibers and concrete without fibers.</p> <p>Response variable: Rating of the severity of the cracks after one year, on a scale of 0 to 10.</p> | <p><b>Essentially correct (E)</b> if the response satisfies the following three components:</p> <ol style="list-style-type: none"> <li>Identifies the experimental units as the driveways</li> <li>Identifies the treatments as concrete with fibers and concrete without fibers</li> <li>Identifies the response variable as the rating of the severity of the cracks</li> </ol> <p><b>Partially correct (P)</b> if the response satisfies only two of the three components.</p> <p><b>Incorrect (I)</b> if the response does not meet the criteria for E or P.</p> |

**Additional Notes:**

- The number of experimental units is not necessary to satisfy component 1.
- The experimental units can be identified as “new homes needing driveways” to satisfy component 1.
- The phrases “on a scale of 0 to 10” and “after one year” are not required for component 3.
- Reasonable synonyms for “rating of the severity” such as “rating,” “severity,” or “on a scale of 0 to 10” are acceptable towards satisfying component 3.
- Identification of the response variable as a mean, or average, does not satisfy component 3.

| Model Solution                                                                                                                                                                                                                                                                                                                                                        | Scoring                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>(b) Number the 60 driveways from 01 to 60. Using a random number generator, generate two-digit integers between 01 and 60. Ignore 00 and any number greater than 60 until 30 unique numbers are obtained. Assign the driveways with those 30 unique numbers to receive concrete with fibers and the remaining 30 driveways to receive concrete without fibers.</p> | <p><b>Essentially correct (E)</b> if the response satisfies the following two components:</p> <ol style="list-style-type: none"> <li>Describes how to correctly use a random number generator, or some other appropriate random process, to assign driveways that have been labeled 1 to 60 to concrete with fibers and concrete without fibers so that every possible random assignment is equally likely</li> <li>The random process results in an equal number of driveways assigned to the concrete with fibers and the concrete without fibers</li> </ol> <p><b>Partially correct (P)</b> if the response describes how to implement a random process that satisfies only one of the two components.</p> <p><b>Incorrect (I)</b> if the response does not meet the criteria for E or P.</p> |

**Additional Notes:**

- Examples of alternative random assignment processes that satisfy both component 1 and component 2 include:
  - Put 60 equally sized slips of paper labeled from 1 to 60 into a hat, mix well, and randomly select 30 slips of paper out of the hat, sampling without replacement. Assign the driveways with the numbers on the 30 selected slips of paper to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.
  - Put 30 white marbles and 30 black marbles in an urn and mix well. Randomly select one marble from the urn. If the marble is white, assign the driveway with label 1 to concrete with fibers, otherwise assign the driveway with label 1 to concrete without fibers. Randomly select a marble from the remaining 59 marbles in the urn. If the marble is white, assign the driveway with label 2 to concrete with fibers, otherwise assign the driveway with label 2 to concrete without fibers. Continue this process of randomly selecting marbles from the urn without replacement until 30 driveways are assigned to concrete with fibers. The other 30 driveways are assigned to concrete without fibers.
  - Spin a spinner with 60 equally sized sections numbered from 1 to 60, ignoring repeats, until 30 unique numbers are generated. Assign the driveways with those 30 numbers to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.
- The following random assignment processes are examples that would satisfy component 2 but would not satisfy component 1.
  - An example that does not specify without replacement: Put 60 equally sized slips of paper labeled from 1 to 60 into a bowl, mix well, and randomly select one slip out of the bowl and record the number. Continue this process of randomly selecting slips until 30 numbers are recorded. Assign the driveways with the 30 recorded numbers to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.

- An example of a stopping rule with equal probabilities: Put 30 white marbles and 30 black marbles into an urn and mix well. Randomly select one marble from the urn. If the marble is white, assign the driveway with label 1 to concrete with fibers, otherwise assign the driveway with label 1 to concrete without fibers. Place the selected marble back into the urn, mix well, and randomly select a marble from the urn. If the marble is white, assign the driveway with label 2 to concrete with fibers, otherwise assign the driveway with label 2 to concrete without fibers. Continue this process of randomly selecting marbles from the urn with replacement until 30 driveways are assigned to concrete with fibers. The other 30 driveways are assigned to concrete without fibers.
- An example where not all outcomes are equally likely: Toss 60 fair coins and record the number of heads. Continue tossing the 60 coins, and ignoring zero, until 30 unique numbers are recorded. Assign the driveways with the 30 recorded numbers to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.
- If there is not some type of labeling system (numbering the driveways), then component 1 is not satisfied.
- If it is not clear that the random assignment process allows every possible random assignment of driveways to type of concrete to be equally likely, then component 1 is not satisfied.
- If the response does not clearly indicate that random numbers are selected without using repeats, then component 1 is not satisfied.
- A response that only assigns driveways to groups and does not indicate how the groups correspond to concrete with fibers and concrete without fibers does not satisfy component 1.
- If a response describes two separate random assignment processes in detail (e.g., how to randomly assign number labels to driveways and how to randomly assign driveways to concrete type), score the combined random assignment process according to the two components.
- If the response describes two ways to make the same random assignment (e.g., how to randomly assign driveways to treatments), assign the score for the weaker random assignment process.

| Model Solution                                                                                                                                                                                                                                                                     | Scoring                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (c) The results were statistically significant, and because the driveways were randomly assigned to either the concrete with the fibers or the concrete without the fibers, there is evidence the treatment (type of concrete) caused the response (rating of severity of cracks). | <p><b>Essentially correct (E)</b> if the response satisfies the following two components:</p> <ol style="list-style-type: none"> <li>1. The response indicates that random assignment enables the conclusion that type of concrete caused the rating of severity of the cracks</li> <li>2. The explanation is in context of the problem</li> </ol> <p><b>Partially correct (P)</b> if the response satisfies only component 1,<br/> OR<br/> the response indicates random assignment limits the effect of confounding variables in context.</p> <p><b>Incorrect (I)</b> if the response does not meet the criteria for E or P.</p> |

**Additional Notes:**

- If the response also indicates an incorrect conclusion based on random assignment, such as generalization, representation of the population, reduction of bias, or reduction of variability, reduce the score from an E to a P or from a P to an I.
- Context includes either the treatment (concrete with fibers or concrete without fibers or “fibers”) or the severity rating.

| Scoring for Question 2                                                                                                                                                                                        | Score |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| <b>Complete Response</b><br>Three parts essentially correct                                                                                                                                                   | 4     |
| <b>Substantial Response</b><br>Two parts essentially correct and one part partially correct                                                                                                                   | 3     |
| <b>Developing Response</b><br>Two parts essentially correct and no part partially correct<br>OR<br>One part essentially correct and one or two parts partially correct<br>OR<br>Three parts partially correct | 2     |
| <b>Minimal Response</b><br>One part essentially correct and no part partially correct<br>OR<br>No part essentially correct and two parts partially correct                                                    | 1     |

**Question 2: Focus on Collecting 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                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>(a)</b> Treatments: New drug, placebo.<br><br>Experimental units: The 72 people who receive the new drug or placebo.<br><br>Response variable: Improvement in acne severity | <p><b>Essentially correct (E)</b> if the response satisfies the following three components:</p> <ol style="list-style-type: none"> <li>Identifies the treatments as new drug and placebo</li> <li>Identifies the experimental units as the 72 people (subjects, participants, twins) in the experiment</li> <li>Identifies the response variable as the improvement in acne severity</li> </ol> <p><b>Partially correct (P)</b> if the response satisfies only two of the three components.</p> <p><b>Incorrect (I)</b> if the response does not satisfy the criteria for E or P.</p> |

**Additional Notes:**

- To satisfy component 1, identification of the treatments must include both the placebo and the new drug.
- To satisfy component 2, the response must indicate that the experimental units are individual people. The response could refer to participants, subjects, twins, or members of the pairs of twins without explicitly mentioning the number 72. However, a response that states or implies that there are 36 experimental units (e.g., “the pairs of twins”) does not satisfy component 2.
- To satisfy component 3, the response must include the context of “acne” and “improvement” (e.g., “improvement in acne severity,” “acne improvement score”), but it does not need to include a reference to the scale, the dermatologist, two-week time periods, or treatments. Reasonable synonyms for improvement can be used, such as using “reduction” or “change” or by including the verbal descriptions of the scale (“no improvement” to “complete cure”). However, a description of a binary outcome (e.g., “whether or not the acne improves”) does not satisfy component 3.
- For responses that indicate the 36 pairs of twins are the experimental units, component 3 may be satisfied by indicating that the response variable is the improvement in acne severity or by indicating that the response variable is the difference in improvement in acne severity.
- If the response provides parallel solutions (i.e., two or more complete solutions without choosing or indicating which is to be scored), the response is scored based on the weaker of the two solutions. For example, if a response says that the experimental units are “the 72 participants and the scores from 0 to 100,” component 2 is not satisfied.

**Model Solution****Scoring**

- (b)** Improvement scores will vary due to many factors, including initial acne severity, what treatment is received, and other variables such as diet and genetics. Because the pairs of twins are similar in initial acne severity, pairing allows for the variation in improvement scores due to the treatment received to be distinguished from variation due to initial acne severity, unlike in a completely randomized design. Consequently, using the matched-pairs design will provide a more precise estimate of the mean difference in improvement in acne severity for the new drug compared to the placebo and make it easier to find convincing evidence that the new drug is better, if it really is better.
- Essentially correct (E)** if the response describes a statistical advantage of a matched-pairs design AND satisfies the following three components:
- The advantage pertains to an inference made after collecting the data (e.g., the ability to distinguish between the effects of the treatments or the precision of the estimate of the drug effect)
  - Indicates that the matched-pairs design is better by using a comparative word (e.g., easier, clearer, greater) or by making an explicit comparison to a completely randomized design
  - Includes context (e.g., “drug,” “improvement,” “acne,” or “twins”)
- Partially correct (P)** if the response describes a statistical advantage of a matched-pairs design AND satisfies one or two of the three components.
- Incorrect (I)** if the response does not satisfy the criteria for E or P.

**Additional Notes:**

- To be considered an advantage of a matched-pairs design, the advantage described must be true for a matched-pairs design and not be true for a completely randomized design. For example, saying that “random assignment allows us to conclude cause-and-effect” is true of both designs. Similarly, “this allows the dermatologist to make conclusions about people with differing acne severity” is true of both designs. Also, “reduces bias” and “reduces variability in the estimates of the individual treatment means” is true of neither design.
- Responses that describe only the set-up of a matched-pairs experiment do not satisfy the requirement to describe an advantage of a matched-pairs design. For example, the response “in a matched-pairs design, the members of each pair will be similar in terms of acne severity” does not describe an advantage. However, “in a matched-pairs design, we can compare two people with similar acne severity” does describe an advantage.
- Advantages of a matched-pairs design that satisfy component 1 include “makes it easier to determine if the drug is effective,” “gives a better estimate of the effect of the new drug,” “reduces variability in the estimate of the drug effect,” “makes the difference between the drug and the placebo more easily distinguishable,” and “gives a clearer picture of how well the drug works.”
- Advantages of a matched-pairs design that don’t satisfy component 1 include “accounts for a source of variability,” “controls for potentially confounding variables,” “allows you to distinguish variation due to severity from variation due to treatment,” “each person can be compared to someone similar,” “reduces variability,” “more balanced treatment groups,” and “more accurate results.”

- It is acceptable to provide a disadvantage of a completely randomized design rather than an advantage of the matched-pairs design (e.g., “The completely randomized design will make it harder to find convincing evidence that the new drug is better”).
- It is acceptable to use the term “blocking” as a synonym for “pairing.”
- A response that states that a matched-pairs design requires a smaller sample size to get power or precision equal to that in a completely randomized design and describes this advantage in context should be scored E.

| Model Solution                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Scoring                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>(c) For each pair of twins, label one person as twin A and label the other person as twin B. For each pair of twins, toss a coin. If the coin lands on heads, twin A gets the placebo and twin B gets the active drug. If the coin lands on tails, twin A gets the active drug and twin B gets the placebo.</p> <p>OR</p> <p>Label the members of each pair of twins as “Twin 1” and “Twin 2.” Using a random number generator, generate an integer from 1 to 2. Give the drug to the twin whose number is selected and the placebo to the twin whose number is not selected. Repeat for all pairs of twins.</p> <p>OR</p> <p>Label 1 notecard “A” and another notecard “B.” For each pair of twins, shuffle the cards and give one card to each twin. The twin who gets “A” receives the drug and the twin who gets “B” receives the placebo.</p> | <p><b>Essentially correct (E)</b> if the response randomly assigns the two treatments within pairs of twins AND satisfies the following three components:</p> <ol style="list-style-type: none"> <li>Uses a random process (e.g., flipping a coin, using a random number generator, shuffling cards) that gives each twin in a pair a 50% probability of getting the drug and a 50% probability of getting the placebo</li> <li>Describes how to use the random process to assign one specific twin in each pair to the drug and the other twin to the placebo</li> <li>Indicates that the random assignment process will be completed for each pair of twins</li> </ol> <p><b>Partially correct (P)</b> if the response randomly assigns the two treatments within pairs of twins AND satisfies only two of the three components for E.</p> <p><b>Incorrect (I)</b> if the response does not satisfy the criteria for E or P.</p> |

#### Additional Notes:

- A response that does not randomly assign both treatments within pairs of twins should be scored incorrect (I). Examples include a response that describes a completely randomized design, describes a crossover design where each person receives both treatments, uses pairs other than twins, does not use random assignment, or indicates that both twins in a pair receive the same treatment.
- For responses that use slips of paper or selecting items from a hat, the slips must be shuffled (or blindly drawn) or the hat mixed or shaken to have a random process and satisfy component 1.
- To satisfy component 2, the response must describe what to do for each possible outcome of the random process and specify which treatment each twin receives. For example, none of the following descriptions satisfy component 2:
  - “Roll a die. If it is 1–3, give the first twin the drug and the second twin the placebo.” (Response doesn’t describe what to do if the die is 4–6.)
  - “Have one member of each pair flip a coin. If it is heads, that twin gets the drug. If it is tails, that twin gets the placebo.” (Response doesn’t indicate what treatment the other twin will receive.)
  - “Flip a coin. If it is heads, give one twin the drug and the other twin the placebo. If it is tails, do the reverse.” (Response doesn’t specify which twin is getting the drug.)
  - “Label one slip of paper “A” and a second slip of paper “B.” Mix them in a hat and have each member of the pair choose one slip.” (Response doesn’t specify if A represents the new drug or the placebo.)
- Ignore any discussion about randomly selecting 36 pairs of twins to obtain subjects for the experiment. Likewise, ignore any discussion about how to perform the analysis for a paired design (e.g., “subtract the improvement scores for each pair of twins”).
- It is acceptable to refer to each pair of twins as a block.

| Scoring for Question 2                                                                                                                                                                                        | Score |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| <b>Complete Response</b><br>Three parts essentially correct                                                                                                                                                   | 4     |
| <b>Substantial Response</b><br>Two parts essentially correct and one part partially correct                                                                                                                   | 3     |
| <b>Developing Response</b><br>Two parts essentially correct and no part partially correct<br>OR<br>One part essentially correct and one or two parts partially correct<br>OR<br>Three parts partially correct | 2     |
| <b>Minimal Response</b><br>One part essentially correct and no part partially correct<br>OR<br>No part essentially correct and two parts partially correct                                                    | 1     |

2019

AP<sup>®</sup> CollegeBoard

# AP<sup>®</sup> Statistics

## Scoring Guidelines

## 2019 SCORING GUIDELINES

## Question 1

Intent of Question

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

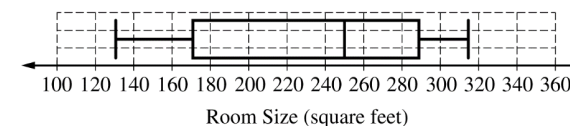
Solution**Part (a):**

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

**Part (b):**

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

DISTRIBUTION OF ROOM SIZES

**Part (c):**

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

## 2019 SCORING GUIDELINES

## Question 1 (continued)

Scoring

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

Section 1 is scored as follows:

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

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

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

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

Notes:

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

## 2019 SCORING GUIDELINES

## Question 1 (continued)

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

Section 2 is scored as follows:

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

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

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

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

Notes:

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

## 2019 SCORING GUIDELINES

## Question 1 (continued)

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

Section 3 is scored as follows:

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

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

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

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

Notes:

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

## 2019 SCORING GUIDELINES

## Question 1 (continued)

## 4 Complete Response

Three sections essentially correct

## 3 Substantial Response

Two sections essentially correct and one section partially correct

## 2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

## 1 Minimal Response

One section essentially correct

OR

No sections essentially correct and two sections partially correct

## 2019 SCORING GUIDELINES

## Question 2

**Intent of Question**

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

**Solution****Part (a):**

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

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

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

**Part (b):**

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

**Part (c):**

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

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

## 2019 SCORING GUIDELINES

## Question 2 (continued)

**Scoring**

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

**Part (a)** is scored as follows:

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

1. Identifies the 4 concentrations (or mixtures or sprays) as the treatments
2. Identifies the 20 containers as the experimental units
3. Identifies the number of insects that are still alive in each container as the response variable

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

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

*Notes:*

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

## 2019 SCORING GUIDELINES

## Question 2 (continued)

Part (b) is scored as follows:

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

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

OR

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

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

Notes:

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

Part (c) is scored as follows:

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

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

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

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

Notes:

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

## 2019 SCORING GUIDELINES

## Question 2 (continued)

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

## 2019 SCORING GUIDELINES

## Question 2 (continued)

## 4 Complete Response

Three parts essentially correct

## 3 Substantial Response

Two parts essentially correct and one part partially correct

## 2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

## 1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

## 2019 SCORING GUIDELINES

## Question 3

Intent of Question

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

Solution

## Part (a):

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

(ii)

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

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

## Part (b):

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

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

## Part (c):

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

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

Scoring

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

Part (a) is scored as follows:

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

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

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

Notes:

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

## 2019 SCORING GUIDELINES

## Question 3 (continued)

Part (b) is scored as follows:

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

Note: Examples of valid explanations with appropriate justifications include:

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

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

OR

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

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

Part (c) is scored as follows:

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

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

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

OR

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

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

## 2019 SCORING GUIDELINES

## Question 3 (continued)

Notes:

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

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

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

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

- For component 2, the boundary value and direction may be described in words, e.g.,  $P(\text{at least four people})$ .
- Component 2 may be satisfied by displaying a bar graph of a binomial distribution with the appropriate bars shaded.
- The response of  $1 - \text{binomcdf}(n = 5, p = 0.54, \text{upper bound} = 3) = 0.24$  is scored E since  $n$ ,  $p$ , and the boundary value are clearly identified.

The response of  $1 - \text{binomcdf}(n = 5, p = 0.54, 3) = 0.24$  is scored P since  $n$  and  $p$  are clearly identified and the boundary value is not identified.

The response of  $1 - \text{binomcdf}(5, 0.54, 3) = 0.24$  is scored I.

- A normal approximation to the binomial is not appropriate since  $np = 5 \times 0.54 = 2.7$  and  $2.7 < 5$ . A response using the normal approximation can score at most P. To score P, the response must include all of the following:
  - an indication that the probability calculated is a normal approximation for the binomial probability
  - a correct mean and standard deviation based on the binomial parameters
  - clear indication of boundary and direction with a z-score or diagram
  - the probability computed correctly.

An example of a response which meets these four criteria is

$$P\left(Z \geq \frac{4 - np}{\sqrt{np(1-p)}}\right) = P\left(Z \geq \frac{4 - (5)(0.54)}{\sqrt{(5)(0.54)(0.46)}}\right) \approx 0.1217, \text{ and the binomial distribution is mentioned.}$$

## 2019 SCORING GUIDELINES

## Question 3 (continued)

## 4 Complete Response

Three parts essentially correct

## 3 Substantial Response

Two parts essentially correct and one part partially correct

## 2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

## 1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

## 2019 SCORING GUIDELINES

## Question 4

Intent of Question:

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

Solution

Section 1:

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

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

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

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

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

Section 2:

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

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

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

## 2019 SCORING GUIDELINES

## Question 4 (continued)

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

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

The  $p$ -value is 0.0135.

Section 3:

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

**Scoring**

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

Section 1 is scored as follows:

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

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

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

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

Notes:

- Correct ways to state the null hypothesis that satisfy component 1:

$$H_0 : p_{17} = p_{14} \text{ or } H_0 : p_{17} - p_{14} = 0$$

$$H_0 : p_{17} \leq p_{14} \text{ or } H_0 : p_{17} - p_{14} \leq 0$$

$$H_0 : p_{14} \geq p_{17} \text{ or } H_0 : p_{14} - p_{17} \geq 0$$

Correct ways to state the alternative hypothesis that satisfy component 1:

$$H_a : p_{17} > p_{14} \text{ or } H_a : p_{17} - p_{14} > 0$$

$$H_a : p_{14} < p_{17} \text{ or } H_a : p_{14} - p_{17} < 0$$

Incorrect ways to state the null hypothesis that do not satisfy component 1:

$$H_0 : p_{17} < p_{14} \text{ or } H_0 : p_{17} - p_{14} < 0$$

$$H_0 : p_{14} > p_{17} \text{ or } H_0 : p_{14} - p_{17} > 0$$

Incorrect ways to state the alternative hypothesis that do not satisfy component 1:

$$H_a : p_{17} \neq p_{14} \text{ or } H_a : p_{17} - p_{14} \neq 0$$

$$H_a : p_{17} < p_{14} \text{ or } H_a : p_{17} - p_{14} < 0$$

## 2019 SCORING GUIDELINES

## Question 4 (continued)

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

Section 2 is scored as follows:

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

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

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

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

Notes:

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

## 2019 SCORING GUIDELINES

## Question 4 (continued)

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

Examples of acceptable quantities (comparison still must be made):

- 12, 49, 20, 32
- 12.017, 48.983, 20.02, 31.98
- $61(0.197)$ ,  $61(1 - 0.197)$ ,  $52(0.385)$ ,  $52(1 - 0.385)$

Examples of unacceptable quantities:

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

Section 3 is scored as follows:

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

- Provides justification of the conclusion based on a correct comparison between a stated  $p$ -value and an alpha value of 0.05.
- Provides a correct conclusion consistent with the alternative hypothesis.
- The conclusion is stated in context.

Partially correct (P)

if the response satisfies components 1 and 2

OR

if the response satisfies components 2 and 3

OR

if the response satisfies components 1 and 3 AND, based on the  $p$ -value from section 2, either

- the conclusion correctly rejects the null hypothesis but does not state that there is convincing evidence for the alternative hypothesis
- OR
- the conclusion correctly fails to reject the null hypothesis but does not state there is not convincing evidence for the alternative hypothesis.

## 2019 SCORING GUIDELINES

## Question 4 (continued)

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

Notes:

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

## 2019 SCORING GUIDELINES

## Question 4 (continued)

## Alternative Approach

## Two-Sided Confidence Interval for Difference in Two Population Proportions

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

*Notes:*

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

**Section 2** is scored as follows:

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

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

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

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

*Notes:*

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

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

*Notes:*

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

*Overall Notes:*

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

## 2019 SCORING GUIDELINES

## Question 4 (continued)

## Alternative Approach

## Chi-square test for homogeneity

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

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

*Notes:*

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

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

*Notes:*

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

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

*Notes:*

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

## 2019 SCORING GUIDELINES

## Question 4 (continued)

## 4 Complete Response

Three sections essentially correct

## 3 Substantial Response

Two sections essentially correct and one section partially correct

## 2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

## 1 Minimal Response

One section essentially correct

OR

No section essentially correct and two sections partially correct

## 2019 SCORING GUIDELINES

## Question 5

Intent of Question

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

Solution**Part (a):**

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

**Part (b):**

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

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

**Part (c):**

The company's expected gain for each warranty purchased is

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

Scoring

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

**Part (a)** is scored as follows:

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

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

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

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

*Notes:*

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

## 2019 SCORING GUIDELINES

## Question 5 (continued)

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

**Part (b)** is scored as follows:

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

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

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

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

*Notes:*

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

## 2019 SCORING GUIDELINES

## Question 5 (continued)

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

**Part (c)** is scored as follows:

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

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

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

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

*Notes:*

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

## 2019 SCORING GUIDELINES

## Question 5 (continued)

## 4 Complete Response

Three parts essentially correct

## 3 Substantial Response

Two parts essentially correct and one part partially correct

## 2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

## 1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

## 2019 SCORING GUIDELINES

## Question 6

Intent of Question

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

Solution**Part (a):**

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

**Part (b):**

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

**Part (c):**

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

**Part (d):**

- (i)  $(0.05)(15,000) = 750$  and  $(0.95)(15,000) = 14,250$ . The 5th percentile is a value, say  $x_{0.05}$ , such that at least 750 values in the table are less than or equal to  $x_{0.05}$  and at least 14,250 are greater than or equal to  $x_{0.05}$ . Cumulate frequencies starting with the smallest sample median listed in the table and going toward the largest (going down columns) until you first reach 750 values, to obtain  $x_{0.05} = \$2,500$ .
- (ii) Similarly,  $x_{0.95} = \$2,950$ .

**Part (e):**

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

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

## 2019 SCORING GUIDELINES

## Question 6 (continued)

## Part (f):

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

**Scoring**

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

**Section 1** is scored as follows:

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

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

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

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

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

**Section 2** is scored as follows:

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

1. Indicates that Emma would need to obtain every possible sample of 50 one-bedroom apartments.
2. Indicates that Emma would need to compute the median rental price for each sample.

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

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

**Section 3** is scored as follows:

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

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

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

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

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

## 2019 SCORING GUIDELINES

## Question 6 (continued)

**Section 4** is scored as follows:

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

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

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

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

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

Each essentially correct (E) section counts as 1 point, and a partially correct (P) section counts as  $\frac{1}{2}$  point.

**4 Complete Response**

**3 Substantial Response**

**2 Developing Response**

**1 Minimal Response**

If a response is between two scores (for example,  $2\frac{1}{2}$  points), use a holistic approach to decide whether to score up or down depending on the strength of the response and communication.

2017

AP<sup>®</sup> 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.

Part (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  $-16.46 + 35.02(1.4)$  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

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

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

$$= 0.2875 \pm 0.0992$$

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  $3,000 \times 0.3867$ , 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.

## 2017 SCORING GUIDELINES

## 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  $\pm$  (critical value)(SE/SD of proportion).

## 2017 SCORING GUIDELINES

## 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  $\frac{1}{2}$  point.

**4 Complete Response**

**3 Substantial Response**

**2 Developing Response**

**1 Minimal Response**

If a response is between two scores (for example,  $2\frac{1}{2}$  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

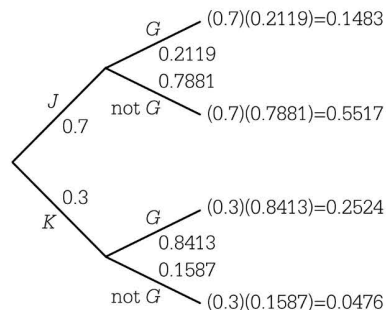
$K$ : melon is from Distributor K

$G$ : melon diameter is greater than 137 mm

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

For a randomly selected melon from the grocery store,

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

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

**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  $\sigma$  for the standard deviation or  $\bar{x}$  instead of  $\mu$  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

$$\begin{aligned}\chi^2 &= 2.093 + 0.395 + 0.817 + 1.322 \\ &\quad + 2.830 + 0.534 + 1.105 + 1.788 \\ &= 10.884.\end{aligned}$$

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

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

**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

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

## 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  $P(D)$  and says

$P(A) = \frac{1}{4}$  in (a-ii) and  $P(A) = \frac{1}{6}$  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

OR

One part essentially correct

No parts essentially correct and two parts partially correct