2025

AP® Statistics

Scoring Guidelines

Question 1: Focus on Exploring Data — 4 points

General Scoring Notes

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

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

Scoring Notes:

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

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

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

Scoring Notes:

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

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

Scoring Notes:

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

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

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

Question 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
A	Sampling method I is not an appropriate sampling method for the farmer to use to estimate the proportion of cabbage plants in the field that are affected by aphids. Sampling method I is a convenience sample where region 3 is not selected randomly. If the farmer's belief is correct, there may be fewer cabbage plants that are affected by aphids in region 3 than in most other regions of the cabbage field because region 3 is in the row farthest from the river. This may lead to an underestimate of the proportion of cabbage plants in the field that are damaged by aphids.	**Essentially correct (E)** if the response satisfies the following three components: 1. Indicates sampling method I is not an appropriate sampling method to obtain the estimate 2. Provides an explanation that refers to at least one of the following: • The region is likely not representative of the entire cabbage field. • The sampling process does not include random selection. • The sample is a convenience sample. 3. Provides sufficient context by including any two of the following: • Sampling unit (region) • Population (cabbage field or the 25 regions in the field) • Statistic or parameter (proportion of cabbage plants that are damaged by aphids) OR count (number of cabbage plants that are damaged by aphids) **Partially correct (P)** if the response satisfies component 1 AND either component 2 or component 3 *OR* if the response does not indicate whether or not sampling method I is an appropriate sampling method AND satisfies both components 2 and 3. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

References only to "a field" are sufficient to satisfy the population description in component 3.

	Model Solution	Scoring
B	The selection of row E is likely to provide an overestimate of the proportion of all cabbage plants in the field that are damaged by aphids. If the farmer's belief that the extent of aphid damage is greater for the regions in the cabbage field closer to the river is correct, then row E, which is the row of regions located closest to the river, is likely to have a greater proportion of cabbage plants damaged by aphids than regions farther from the river.	**Essentially correct (E)** if the response satisfies the following three components: 1. Indicates that the selection of row E is likely to produce an overestimate 2. Provides a justification that is based on the location of row E as the row located closest to the river 3. Links the location to why row E is likely to produce an overestimate (e.g., by referring to the farmer's belief that the extent of aphid damage is greater for the regions in the cabbage field closer to the river) **Partially correct (P)** if the response satisfies only two of the three components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

A response that indicates the selection of row E is likely to provide an underestimate may be scored P if the response provides a justification that ignores the farmer's belief (either by omission or by specific mention) but includes a valid reason for why the regions further from the river may have more aphid damage.

Model Solution	Scoring
C The farmer should write the region numbers from row A, 1 through 5, onto same-size slips of paper, then put the numbers into a hat, mix well, and select one of the numbers. The farmer should repeat this process for the region numbers of each of the other rows (i.e., row B, 6 through 10; row C, 11 through 15; row D, 16 through 20; row E, 21 through 25) and select one number from each row. This process will result in the selection of one region from each row. The farmer will examine every cabbage plant in each of the selected regions for aphid damage to determine the proportion of cabbage plants in the selected regions that are damaged by aphids. *Alternative Solution:* The farmer should use a random number generator to generate one two-digit integer from 01 to 05, one two-digit integer from 06 to 10, one two-digit integer from 11 to 15, one two-digit integer from 16 to 20, and one two-digit integer from 21 to 25. For each integer selected, the farmer should select the corresponding numbered region and examine every cabbage plant in each of the selected regions for aphid damage to determine the proportion of cabbage plants in the selected regions that are damaged by aphids.	**Essentially correct (E)** if the response satisfies the following four components: 1. Describes a random selection process that indicates the groupings used for the strata (by rows) 2. Describes how to correctly implement a random selection process for which the selection of regions within each stratum is equally likely 3. Describes a random selection process for which the selections across strata are independent 4. Describes a random selection process that results in the selection of one region from each stratum **Partially correct (P)** if the response satisfies only two or three of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- A response may satisfy component 1 by referring to the five regions in the rows instead of using row letters.
- For responses that use cards or slips of paper:
  - If the number of slips of paper (number of cards) does not equal 5 for each random selection, then component 1 is not satisfied. Slips of paper (cards) do not need to be specifically identified as equally sized.
  - If the response does not describe a thorough mixing (shuffling) of the slips of paper (cards), then component 2 is not satisfied.
- For responses that use a random number generator or table of random digits:
  - If it is not clear that a random selection process allows the selection of regions within each stratum to be equally likely, then component 2 is not satisfied.
  - If the response does not clearly indicate that a random number is generated from the region numbers within each of the five rows (e.g., only describes the generation of five random numbers from the two-digit integers from 01 to 25), then component 4 is not satisfied.

- For responses that use a fair die:
  - If a five-sided fair die is rolled for each of the five rows and the response clearly indicates the region numbers assigned to the values on the die for each roll, then component 4 is satisfied.
  - If a six-sided die is rolled for each of the five rows and the response clearly indicates which number is excluded (e.g., "if a 6 is rolled, roll again until a non-6 number is achieved") AND the response clearly indicates the region numbers (or columns) assigned to the values on the die for each roll, then component 4 is satisfied.
  - If a 25-sided fair die is rolled for each of the five selections, then component 4 is not satisfied without sufficient further justification because the selection of one region from each stratum is not guaranteed.
- If a response describes two separate random selection processes in detail (e.g., describes how to use a random number generator and slips of paper in a hat), score both descriptions according to the four components and use the lower score.
- If a response indicates that separate samples were taken from each of the strata, then component 3 is satisfied.
- A response that selects columns instead of regions within strata cannot satisfy component 3.

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

Question 3: Focus on Probability and Sampling Distributions — 4 points

General Scoring Notes

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

	Model Solution	Scoring
A	i. $P(\text{Rock Song}) = \frac{100}{1{,}000} = 0.10$ ii. $P(\text{Both Rock Songs}) = (0.10)(0.10) = 0.01$	**Essentially correct (E)** if the response satisfies the following four components: 1. In part A (i) the response calculates the correct probability. 2. In part A (i) the response provides supporting work for the correct probability. 3. In part A (ii) the response calculates the correct probability consistent with the answer in part A (i). 4. In part A (ii) the response provides supporting work consistent with the probability calculated in part A (i). **Partially correct (P)** if the response satisfies two or three of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- In part A (ii) a response that does not consider independence when solving (e.g., $\left(\frac{100}{1000}\right)\left(\frac{99}{999}\right) = 0.0099$) does not satisfy component 3 but may satisfy component 4.
- If probabilities are not labeled but are given in the order they are asked, then the response may earn an E. If the probabilities are presented in a different order, there must be a label to satisfy the components. In this case, sufficient labels include probability notation, context, work, or identification of the subparts (i) and (ii).

	Model Solution	Scoring
B	i. Let the random variable of interest, X, represent the number of the 20 songs played in one hour that are rock songs. It is stated that any song can be replayed at any time, which establishes that each rock song has probability $\frac{100}{1{,}000} = 0.10$ of being selected each hour and each song is independent from every other song. Therefore, X has a binomial distribution with $n = 20$ independent trials and probability of success $p = 0.10$ for each trial. ii. The expected value for the number of rock songs played in one hour is $np = 20(0.10) = 2$ songs.	**Essentially correct (E)** if the response satisfies at least three of the following four components: 1. In part B (i) the response defines the random variable as the number of rock songs played in one hour. 2. In part B (i) the response describes the distribution as binomial. 3. In part B (i) or B (ii) the response states that $n = 20$ and $p = 0.10$. 4. In part B (ii) the response correctly calculates the expected value AND provides supporting work for the calculation of the correct expected value. **Partially correct (P)** if the response satisfies only two of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- When defining the random variable, the response must include both "the number of rock songs" and "in one hour" or "out of 20 songs."
- A response that states $X \sim B(20,\ 0.1)$ satisfies components 2 and 3.
- If a response states the random variable has a distribution other than binomial (e.g., normal, left skewed, or uniform), part B cannot be scored E.
- Stating that songs are distributed randomly is not a distribution and should be considered extraneous.
- Examples that satisfy components 3 and 4 include:
  - $np = 20(0.10) = 2$
  - $np = 20\left(\frac{100}{1{,}000}\right) = 2$
  - $n = 20, p = 0.10, np = 2$
- An example that satisfies component 4 only:
  - $20(0.10) = 2$
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

Model Solution	Scoring
C i. The probability that in a particular hour 4 or more rock songs will be played is $P(X \geq 4) = 1 - P(X \leq 3)$ $P(X \geq 4) = 1 - \left[\binom{20}{0}(0.10)^0(0.90)^{20} + \binom{20}{1}(0.10)^1(0.90)^{19} + \binom{20}{2}(0.10)^2(0.90)^{18} + \binom{20}{3}(0.10)^3(0.90)^{17}\right]$ $P(X \geq 4) = 1 - 0.867 = 0.133$. ii. No, the probability that 4 or more rock songs would be played in an hour is 0.133, which is high enough to be reasonably attributed to chance alone. This probability is not small enough to provide evidence that the selection process was not truly random.	**Essentially correct (E)** if the response satisfies the following four components: 1. In part C (i) the response provides a correct probability. 2. In part C (i) the response shows work that supports the correct probability. 3. In part C (ii) the response indicates that there is not a strong reason to believe that the selection process was not truly random. 4. In part C (ii) the response provides an explanation that correctly links the probability to the decision. **Partially correct (P)** if the response satisfies only two or three of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- A response may satisfy component 2 by any of the following:
  - Graphical display: Displaying a bar graph of binomial probabilities including axes with scale with appropriate bars shaded.
  - Probability formula: For example,
    $1 - \binom{20}{0}(0.10)^0(0.90)^{20} - \binom{20}{1}(0.10)^1(0.90)^{19} - \binom{20}{2}(0.10)^2(0.90)^{18} - \binom{20}{3}(0.10)^3(0.90)^{17}$.
  - Calculator function notation: Using calculator function notation with clearly defined arguments. For example:
    - $1 - \text{binomcdf}(n = 20,\ p = 0.10, \text{upper bound} = 3)$ satisfies component 2 because the boundary value is clearly labeled.
    - $1 - \text{binomcdf}(n = 20,\ p = 0.10, 3)$ does not satisfy component 2 because the boundary value is not labeled.
    - $\text{Binomcdf}(n = 20, p = 0.10,$ lower bound =4, upper bound = 20) satisfies component 2 because the boundary value is clearly labeled.
  - Random Variable: $P(X \geq 4)$ or $1 - P(X \leq 3)$ with identification of the binomial distribution with correct parameters ($n = 20$ and $p = 0.10$) included in part C satisfies component 2.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.
- A response that indicates that the manager does have a strong argument that the selection process was not truly random (or responds "yes") that is adequately supported by an explanation based on an incorrectly calculated probability in part C (i) may satisfy components 3 and 4.
- A response that indicates that the manager does have a strong argument that the selection process was not truly random (or responds "yes") that is supported by a statement claiming the probability is low may satisfy components 3 and 4.

- A response that indicates that the manager does not have a strong argument that the selection process was not random supported by the calculation of the standard deviation of the binomial distribution, 1.342, and an explanation based on 4 being within two standard deviations of the expected value (mean) may earn credit for components 3 and 4.
- If a response gives two arguments, treat them as parallel solutions and score the weaker solution.
- A response that finds the probability of exactly 4 songs playing (e.g., $\text{binompdf}(20, 0.1, 4) = 0.089$) and explains that this is not strong evidence that the selection process was not truly random may still satisfy components 3 and 4.

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

Question 4: Focus on Inference — 4 points

General Scoring Notes

- This question is scored in three sections. Each section is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The first section includes statements of the null and alternative hypotheses and identification of the appropriate hypothesis test. The second section includes verifying the conditions for the test identified in the first section and calculating the value of the test statistic and the corresponding *p*-value. The third section includes the conclusion for the test identified in the first section. The response is then categorized based on the scores assigned to each section 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 section 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
Section 1	An appropriate inference procedure is a one-sample *z*-test for a population proportion. The null hypothesis is $H_0: p = 0.22$, and the alternative hypothesis is $H_a: p > 0.22$, where $p =$ the true proportion of students at Karen's high school that use the application at least once per week.	**Essentially correct (E)** if the response satisfies the following four components: 1. Identifies a one-sample *z*-test for a population proportion by name (e.g., "one-proportion *z*-test" but not merely "one-sample *z*-test") or by formula 2. States the correct equality for the null hypothesis with the value 0.22 3. States the correct direction for the one-sided alternative hypothesis consistent with the null hypothesis 4. Provides sufficient context for the parameter by including reference to the population proportion (true proportion of students at Karen's high school) AND the sampling units (students) AND the response variable (using the application) **Partially correct (P)** if the response satisfies three of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- If the response identifies the correct test by name but also states an unreasonable formula, then component 1 is not satisfied.
- If the response identifies the test using the correct formula but equating it with a *t* instead of a *z*, then component 1 is not satisfied.
- A response that states the null hypothesis as $H_0: p \leq 0.22$ satisfies component 2.

- Components 2 and 3 may be satisfied without regard to the symbol (or lack of symbol) used to represent the population parameter.
- Neither context nor the concept of the population is required to satisfy components 2 or 3.
- A response that states the hypotheses in words (e.g., "the null hypothesis is that the proportion is 0.22, and the alternative hypothesis is that the proportion is greater than 0.22") may satisfy components 2 and 3.
- A response that states the hypotheses in words and refers to the population in context (e.g., "the null hypothesis is that the population proportion of students at Karen's high school that use the application at least once per week is equal to 0.22 and the alternative hypothesis is that the population proportion is greater than 0.22") may satisfy components 2, 3, and 4.
- If the null hypothesis is incorrect, the response can satisfy component 3 with a correct alternative hypothesis.
- The elements of component 4 do not have to be satisfied with the statement of the hypotheses. They may be satisfied with the statement in the hypotheses, definition of the parameter, or the statement of the conclusion.
- If the statement of the hypotheses refers to population proportion and the conclusion refers to sample proportion (or vice versa), then the population aspect of component 4 is not satisfied.
- If the response clearly refers to the **sample** proportion instead of the **population** proportion using a $\hat{p}$, then component 4 is not satisfied unless the symbol used is defined as the **population** proportion.
- A response may satisfy the population aspect of component 4 by the following:
  - Referring to the population by using words such as "population," "all," or "true" when defining the parameter or in the statement of the conclusion of the inferential procedure.
  - Using notation such as p, p_0, or π when defining the hypothesis statements.

Confidence Interval Approach:

- If a one-sample z-interval for a population proportion is identified correctly by name (e.g., "one-proportion z-interval" but not merely "one-sample z-interval") or by formula, then component 1 is satisfied.
- If a response uses a one-sample z-interval for a population proportion, then component 4 is satisfied if the response indicates that it is a confidence interval for the true proportion of students at Karen's high school that use the application at least once per week.

	Model Solution	Scoring
Section 2	The independent observation condition for performing the one-sample z-test for a population proportion is satisfied. This is because the data were obtained from a random sample of 130 high school students from Karen's high school. Also, the sample of 130 students is less than 10% of the total number of students at this large high school, because $130 < 0.10(2{,}000)$ and the total number of students in Karen's high school is greater than 2,000. The 10% condition is required as sampling was conducted without replacement from a finite population. The number of expected successes and expected failures were both more than 10 because $130(0.22) = 28.6$ and $130(0.78) = 101.4$. Thus, the sample size is large enough to support the assumption that the sampling distribution of $\hat{p}$ is approximately normal. $\hat{p} = \frac{38}{130} \approx 0.2923$ Test statistic: $z = \frac{0.2923 - 0.22}{\sqrt{\frac{0.22(1-0.22)}{130}}} \approx 1.99$ $P(z > 1.99) \approx 0.023$	**Essentially correct (E)** if the response satisfies the following four components: 1. Checks the independence condition by referring to the random sample of 130 students AND indicates that 130 is less than or equal to 10% of 2,000 (i.e., $(130) \le (0.10)N$) 2. Checks that the sample size is large enough to support the assumption that the sampling distribution of $\hat{p}$ is approximately normal by verifying that the number of expected successes and expected failures were at least 10 by calculating the following values: $np_0 = 130(0.22) = 28.6$ and $n(1 - p_0) = 130(0.78) = 101.4$ 3. Correctly reports the value of the z-statistic 4. Reports the value for the correct p-value consistent with the test statistic and the procedure stated in Section 1 **Partially correct (P)** if the response satisfies only two or three of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- To satisfy the reference to the random selection of 130 students in component 1, it is minimally acceptable to state "random sample – check" or "SRS – check." However, component 1 is not satisfied if the response implies that random **assignment** was used or only states "random – check."
- To satisfy component 2, a direct comparison must be made against a standard criterion (5 or 10) using either actual values of the observed successes and failures (38 and 92) OR values for the expected successes and failures (28.6 and 101.4) OR formulas for the expected number of successes and failures with values inserted (or defined elsewhere), such as $130(0.22)$ and $130(1 - 0.22)$.
- If the response includes an inappropriate check of conditions, such as $n > 30$, then component 2 is not satisfied.
- A response that reports the correct value for the z-statistic but contains errors in supporting work satisfies component 3.
- A response that inputs correct values into the z-statistic formula but computes an incorrect value for the z-statistic, satisfies component 3.

- If the response incorrectly uses the sample proportion in calculating the standard error, the response does not satisfy component 3 but may satisfy component 4 ($z = 1.81$ and $p = 0.0351$).
  - The following combinations of z-statistics and p-values may satisfy component 4 but not component 3: $z = 1.76$ and $p = 0.039$ OR $z = 1.915$ and $p = 0.0276$.
- If the response satisfies component 4, any supporting work for the p-value may be treated as extraneous.
- To satisfy component 4, the p-value must be consistent with the alternative hypothesis and either the reported test statistic OR the correct test statistic (1.99).
- If the response compares the value of the test statistic to a critical value instead of reporting a p-value, then the critical value (1.645), or a critical value consistent with the stated alternative hypothesis, satisfies component 4.
- If the response omits identifying the hypotheses, the correct one-sided alternative hypothesis is assumed when scoring component 4.
- If the response omits a test statistic, the correct test statistic is assumed when scoring component 4.
- If an incorrect alternative hypothesis is stated, then the p-value must be consistent with the stated alternative hypothesis to satisfy component 4.

Confidence Interval Approach:

- If either a one-sided 95% confidence interval is correctly calculated as $(0.227, \infty)$ or a two-sided 90% confidence interval is correctly calculated as $(0.227, 0.358)$, then component 3 is satisfied.
- If the alternative hypothesis corresponds to $H_a: p > 0.22$ and only the lower end of a confidence interval is used to reach a conclusion, then component 4 is satisfied.
- Application of a confidence interval approach must be consistent with the stated alternative to satisfy component 4.
- A two-sided 95% confidence interval is $(0.214, 0.370)$ and does not satisfy component 3.

	Model Solution	Scoring
Section 3	Because this p-value is less than the $\alpha = 0.05$ significance level, the null hypothesis should be rejected. There is convincing statistical evidence that the population proportion of students at Karen's high school that use the application at least once per week is greater than Country W's proportion of 0.22.	**Essentially correct (E)** if the response satisfies the following two components: 1. Provides correct comparison of the p-value to alpha (p-value is less than/greater than alpha) AND provides a correct decision about the null and/or alternative hypothesis 2. States a conclusion in context, consistent with, and in terms of, the alternative hypothesis using nondefinitive language **Partially correct (P)** if the response satisfies only one of the two components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- To satisfy component 1, the response must clearly identify the number that is compared to alpha as a p-value (which can be identified anywhere in the response).
- If the comparison and decision are consistent with an incorrect p-value (or an incorrect value of the test statistic or an incorrect confidence interval), the response may satisfy component 1. This also applies to incorrect p-values that are unreasonable (e.g., p-value $= 3.2$).
- To satisfy the p-value comparison in component 1, the response can compare the value of the test statistic to an appropriate critical value (e.g., $z > 1.645$).
- If the response includes any statement equivalent to "accept the null hypothesis," component 1 is not satisfied.
- An explicit decision (Fail to Reject/Reject) is not required to satisfy component 1. The decision part of component 1 may be satisfied by implying the decision within the conclusion statement (insufficient evidence/sufficient evidence for the alternative hypothesis).
- If an explicit decision is stated and the conclusion is inconsistent with the decision, component 1 is not satisfied.
- To satisfy the context in component 2, the response must reference proportion, students (may be implied), and using the application OR answers the inference question.
- If the response omits hypotheses, assume the correct alternative hypothesis is provided when scoring component 2.
- If the response states incorrect hypotheses, component 2 may be satisfied by either stating a conclusion in terms of the incorrect hypothesis or by stating the conclusion by answering the inference question given in the stem (e.g., "the results from this study do not provide convincing statistical evidence that the proportion of students at Karen's high school that use the application at least once per week is greater than 0.22").
- Definitive language, such as the following, does not satisfy component 2: "proves the null," "proves the alternative," "accepts the alternative," and "there is no evidence for the alternative."
- Nondefinitive language, such as the following, is required to satisfy component 2: "evidence to accept the alternative," "there is evidence for the alternative," and "there is not sufficient evidence for the alternative."

- If components 1 and/or 2 are satisfied and the response provides an incorrect interpretation of the *p*-value, the score is lowered from E to P or P to I.

Confidence Interval Approach:

- If the alternative hypothesis is specified correctly as $H_a: p > 0.22$, then component 1 is satisfied if the justification is based on whether 0.22 is below the lower end of the confidence interval. If the alternative hypothesis is stated in the wrong direction, then component 1 is satisfied if the justification is based on whether 0.22 is above the upper end of the confidence interval.
- If an incorrect two-sided alternative hypothesis is specified, then component 1 is satisfied if the justification is based on whether 0.22 is included in the confidence interval.
- If no alternative hypothesis is specified in the response, then assume the correct alternative hypothesis is provided when scoring component 2.
- If the response includes an incorrect interpretation of the confidence interval, then the score for Section 3 is lowered from E to P or from P to I.

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

Question 5: Multi-Focus — 4 points

General Scoring Notes

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

	Model Solution	Scoring
A	i. Let random variable X represent the number of bedrooms in a randomly selected newly built house in the 2024 sample from Country B. The probability that a randomly selected house from the 2024 sample had fewer than 3 bedrooms is the probability that the house had either 1 or 2 bedrooms, which is $P(X < 3) = P(X = 1) + P(X = 2)$ $P(X < 3) = 0.12 + 0.22$ $P(X < 3) = 0.34$. ii. The average number of bedrooms per house for the sample of newly built houses in 2024 is $E(X) = 1(0.12) + 2(0.22) + 3(0.28) + 4(0.22) + 5(0.14) + 6(0.02)$ $E(X) = 3.10$ bedrooms.	**Essentially correct (E)** if the response satisfies at least three of the following four components: 1. In part A (i) the response correctly calculates the probability of 0.34. 2. In part A (i) the response provides supporting work or justification for component 1. 3. In part A (ii) the response correctly calculates the mean of 3.10. 4. In part A (ii) the response provides supporting work or justification for component 3. **Partially correct (P)** if the response satisfies only two of the components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- An arithmetic or transcription error in a response can be ignored if correct work is shown.
- Supporting work for finding the expected value must include at least two of the terms in the equation to show the pattern, such as $1(0.12) + 2(0.22) + \ldots$.
- A response to part A (ii) that indicates use of the appropriate formula for the mean of a discrete random variable in words (e.g., "the sum of x times p of x for all values of x") or describes an appropriate method (e.g., using the given distribution to create a hypothetical sample of 100 and then calculating the mean) may satisfy component 4.

Model Solution	Scoring
B i. Let $\mu =$ the population mean number of bedrooms in newly built houses in 2024 from Country B. The null hypothesis is $H_0: \mu = 2.9$ and the alternative hypothesis is $H_a: \mu \neq 2.9$. ii. A Type I error would be determining that the population mean number of bedrooms in newly built houses in 2024 from Country B is not equal to 2.9 when it is in fact 2.9.	**Essentially correct (E)** if the response satisfies the following four components: 1. In part B (i) the response states the correct equality for the null hypothesis with the value 2.9. 2. In part B (i) the response states the correct two-sided alternative hypothesis consistent with the null hypothesis. 3. In part B (i) the response provides sufficient context for the parameter by including a reference to the population mean AND the response variable (bedrooms) AND the sampling units (newly built houses). 4. In part B (ii) the response correctly defines a Type I error in context. **Partially correct (P)** if the response satisfies only three of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- Components 1 and 2 may be satisfied without regard to the symbol (or lack of symbol) used to represent the population parameter.
- Neither context nor the concept of the population is required to satisfy components 1 or 2.
- A response that states the hypotheses in words (e.g., "the null hypothesis is that the mean is 2.9, and the alternative hypothesis is that the mean is not equal to 2.9") may satisfy components 1 and 2.
- A response that states the hypotheses in words AND refers to the population in context (e.g., "the null hypothesis is that the population mean number of bedrooms in newly built houses in 2024 from Country B is equal to 2.9, and the alternative hypothesis is that the population mean number of bedrooms in newly built houses in 2024 from Country B is not equal to 2.9") may satisfy components 1, 2, and 3.
- If the response clearly refers to the **sample** mean instead of the **population** mean using words or a symbol (e.g., $\bar{x}$ or $\hat{\mu}$), then component 3 is not satisfied unless the symbol used is defined as the **population** mean.
- The phrase "mean number of bedrooms" or "mean number of bedrooms for houses" is not sufficient for the population aspect of component 3.
- A response may satisfy the population aspect of component 3 by the following:
  - Referring to the population by using words such as "population," "all," or "true" when defining the parameter.
  - Using notation such as μ when defining the hypothesis statements.
- A response that refers to "bedrooms" OR "rooms" may satisfy context for component 4.

	Model Solution	Scoring
C	Because the value 2.9 is not contained within the 97% confidence interval, the null hypothesis should be rejected. Therefore, there is convincing statistical evidence, at the $\alpha = 0.03$ level of significance, that the population mean number of bedrooms in newly built houses in 2024 from Country B is not equal to 2.9 (or is different than that in 2017).	**Essentially correct (E)** if the response satisfies the following two components: 1. States a conclusion consistent with and in terms of the alternative hypothesis using nondefinitive language 2. Provides a justification for the conclusion by indicating that 2.9 is not contained within the confidence interval **Partially correct (P)** if the response satisfies only one of the two components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- A response that only provides an interpretation of the given confidence interval is scored I.
- A response may satisfy the conclusion aspect of component 1 by using words such as "there is evidence to support the alternative," "there is statistical evidence that H_a is true," "I am 97% confident there is a difference in means," or "because 2.9 is not included in our interval, it is not a plausible value for the mean."
- If an explicit decision is stated and the conclusion is inconsistent with the decision, component 1 is not satisfied. A response that incorrectly indicates that 2.9 is contained in the confidence interval and then provides an otherwise correct conclusion based on 2.9 being contained in the interval satisfies component 1 but not component 2.
- A response that provides a conclusion that is consistent with an incorrect alternative hypothesis identified in part B satisfies component 1.
- A definitive response that states that the average number of bedrooms is 3.10 (the value of the sample mean or any other number) does not satisfy component 1, even if the response makes additional correct statements about the alternative hypothesis.
- A response that draws a conclusion that is clearly about the sample mean (e.g., a statement about "average number of bedrooms for newly built houses in the study") does not satisfy component 1.
- If the conclusion includes a definitive statement (e.g., "this proves that we have enough evidence that the mean number of bedrooms in newly built homes in 2024 is 2.9" or "Rodney is correct; the mean number of bedrooms in newly built homes in 2004 is 2.9"), then component 1 is not satisfied.
- A response that reverse engineers the given confidence interval to obtain a standard error using a reasonable critical value for a 97% confidence interval and then uses the standard error to compute a *t*-statistic and *p*-value to reach a correct conclusion satisfies component 1 but does not satisfy component 2.

Scoring for Question 5	Score
Complete Response Three parts essentially correct	**4**
Substantial Response Two parts essentially correct and one part partially correct	**3**
Developing Response Two parts essentially correct and no part partially correct *OR* One part essentially correct and one or two parts partially correct *OR* Three parts partially correct	**2**
Minimal Response One part essentially correct and no parts partially correct *OR* No part essentially correct and two parts partially correct	**1**

Question 6: Investigative Task — 4 points

General Scoring Notes

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

	Model Solution	Scoring
A	Because the *p*-value of 0.002 is less than the level of significance of 0.05, the null hypothesis should be rejected. There is convincing statistical evidence of a difference between the mean reading score for all children, similar to those who participated in the study, who would read the story at 9 a.m. and the mean reading score for all children, similar to those who participated in the study, who would read the story at 3 p.m.	**Essentially correct (E)** if the response satisfies the following two components: 1. Provides correct comparison of the *p*-value to alpha (*p*-value is less than α) AND provides a correct decision about the null and/or alternative hypothesis 2. States a conclusion in context, consistent with, and in terms of the stated alternative hypothesis using nondefinitive language **Partially correct (P)** if the response satisfies only one of the two components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- To satisfy the *p*-value comparison in component 1, the response can compare the value of the test statistic to an appropriate critical value; for example, $|t| > 1.985$ if $df = 97.489$, or $|t| > 2.01$ if $df = 49$.
- An explicit decision about the null hypothesis is not required to satisfy component 1.
- If an explicit decision is stated and the conclusion is inconsistent with the decision, component 1 is not satisfied.
- The decision part of component 1 may be satisfied by implying the decision within the conclusion statement (sufficient evidence for the alternative hypothesis).
- To satisfy component 2, the response must include reference to means, groups (e.g., 9 a.m. and 3 p.m.), the sampling units (e.g., children), and the variable of interest (reading score).
- Examples of nondefinitive language in component 2 include “evidence to accept the alternative,” “there is evidence for the alternative,” and “there is not sufficient evidence for the alternative.”
- Examples of definitive language in component 2 include “accepts the null,” “proves the null,” “proves the alternative,” “accepts the alternative,” “there is not evidence for the alternative,” and “no evidence for the alternative.”

- If components 1 and/or 2 are satisfied and the response provides an incorrect interpretation of the p-value, the score is lowered from E to P or P to I.
- The quality of communication for responses with score P should be considered if holistic scoring is required.

	Model Solution	Scoring
B	It was appropriate for Stefan to conduct a two-sample *t*-test instead of a paired *t*-test because the two groups are independent. Stefan used random assignment to place the 100 volunteer children into two groups, and there is no indication that the two groups of 50 children are paired in any meaningful way (e.g., age, reading comprehension level).	**Essentially correct (E)** if the response satisfies the following two components: 1. States that the two-sample *t*-test is appropriate because the groups of children are independent *OR* that the groups of children were *not* paired in a meaningful way 2. Explains why the groups are independent instead of paired, either by referencing the use of random assignment of the 100 children to the two groups *OR* by providing an explanation that clearly separates the two situations, such as an example of how the data could have been paired **Partially correct (P)** if the response satisfies only one of the two components required for E *OR* if the response states that a two-sample *t*-test is appropriate because different samples were used but does not explicitly state that the children in the two samples were not paired in a meaningful way. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- For component 2 it is not sufficient to state that the reason the groups are independent is that one group was in the morning and one group was in the afternoon or because they received different treatments.
- For component 2 a sufficient explanation is one that provides an example of how the data could have been paired instead, such as if twins had been used or each child read a story at both 9 a.m. and 3 p.m.
- The quality of communication for responses with score P should be considered if holistic scoring is required.

	Model Solution	Scoring
C	i. The value of the pooled standard deviation is $s_p = \sqrt{\frac{(4.12)^2 + (4.43)^2}{2}} \approx 4.28$. Therefore, the value of Cohen's d is $d = \frac{\lvert 15.2 - 17.9 \rvert}{4.28} \approx 0.63$. ii. Based on Table 2, a Cohen's d value of 0.63 would indicate that Stefan's results were somewhat practical or meaningful in real life.	**Essentially correct (E)** if the response satisfies the following two components: 1. In part C (i) the response correctly calculates the value of Cohen's d with work shown. 2. In part C (ii) the response indicates the correct level of practical importance that is consistent with component 1. **Partially correct (P)** if the response satisfies only one of the two components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- An arithmetic or transcription error in a response can be ignored if correct work is shown.
- To satisfy the work-shown aspect in component 1, the response must show correct work for Cohen's d. Showing work for the pooled standard deviation is not required to satisfy component 1.
- A response that reports a negative Cohen's d and indicates the practical importance is "not very meaningful in real life" or indicates the practical importance is less than "not very meaningful in real life" may satisfy component 2.

		Model Solution	Scoring
D	i.	The value of Cohen's *d* would decrease. If the standard deviation for the a.m. group and p.m. group were both greater than 4.43, the pooled standard deviation would be greater than 4.28. With a larger value in the denominator and the same value in the numerator, the value of Cohen's *d* would be smaller than 0.63.	**Essentially correct (E)** if the response satisfies the following three components: 1. In part D (i) the response states that the value of Cohen's *d* would be smaller. 2. In part D (i) the response provides a reason for the change in the value of Cohen's *d* that includes a correct reference to how Cohen's *d* will change based on an increase in the pooled standard deviation. 3. In part D (ii) the response states the decrease in Cohen's *d* would indicate there was less practical importance than the original study results. **Partially correct (P)** if the response satisfies only two of the three components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.
	ii.	The lower Cohen's *d* value would indicate less practical importance than that of the original results of Stefan's study.	

Scoring Notes:

- If only the pooled standard deviation was calculated and no value is computed for Cohen's *d* in part C, the response may satisfy all components for part D if it explains that the larger standard deviations would result in an increase in the pooled standard deviation, and therefore show a lower effect size or less practical importance.
- A response that uses Table 2 and indicates that it is unknown if Cohen's *d* would be in the same or lower category of practical importance may satisfy component 3.
- The quality of communication for responses with score P should be considered if holistic scoring is required.

Scoring for Question 6 Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as ½ point.	
	Score
Complete Response	**4**
Substantial Response	**3**
Developing Response	**2**
Minimal Response	**1**
If a response is between two scores (for example, 2 ½ points), use a holistic approach to decide whether to score up or down, depending on the strength of the response and quality of the communication.	