Chief Reader Report on Student Responses:
2022 AP® Statistics Free-Response Questions

• Number of Students Scored	216,968			
• Number of Readers	1,114			
• Score Distribution	Exam Score	N	%At	
	5	32,163	14.8	
	4	48,182	22.2	
	3	50,818	23.4	
	2	35,785	16.5	
	1	50,020	23.1	
• Global Mean	2.89			

The following comments on the 2022 free-response questions for AP® Statistics were written by the Chief Reader, Dr. Ken Koehler, PhD. They give an overview of each free-response question and of how students performed on the question, including typical student errors. General comments regarding the skills and content that students frequently have the most problems with are included. Some suggestions for improving student preparation in these areas are also provided. Teachers are encouraged to attend a College Board workshop to learn strategies for improving student performance in specific areas.

Question 1

Task: Exploring Data
Max Score: 4
Mean Score: 1.92

What were the responses to the question expected to demonstrate?

The primary goals of this question were to access a student's ability to (1) use data presented as a scatterplot to describe a relationship between two variables within the context of a study; (2) identify and interpret the slope of a least-squares regression line; (3) interpret the coefficient of determination with respect to the proportion of variation in values of the response variable that can be explained by variation in the values of the explanatory variable; (4) identify the observation with the largest absolute residual using a scatterplot of the data with the least-squares regression line inserted; and (5) determine if the least-squares regression line overestimates or underestimates the value of the response for the identified observation and provide a justification based on a comparison of the identified observation to the least squares regression line.

This question primarily assesses skills in skill category 2: Data Analysis, and skill category 4: Statistical Argumentation. Skills required for responding to this question include (2.A) Describe data presented numerically or graphically, (2.C) Calculate summary statistics, relative positions of points within a distribution, correlation, and predicted response, and (4.B) Interpret statistical calculations and findings to assign meaning or assess a claim.

This question covers content from Unit 2: Exploring Two-Variable Data of the course framework in the AP Statistics Course and Exam Description. Refer to topics 2.4, 2.6, 2.7, and 2.8, and learning objectives DAT-1.A, DAT-1.D, DAT-1.F, DAT-1.G, and DAT-1.H.

How well did the responses address the course content related to this question? How well did the responses integrate the skills required on this question?

- In part (a) most responses identified at least two of the characteristics for a scatterplot, but many did comment on all five characteristics. The majority included direction, but some missed strength, form, or unusual features. Most of the responses included the context of mass and length of bullfrogs.
- In part (b) most responses were able to correctly identify the slope given a regression equation and include units in the context of their interpretation. However, many responses did not describe the change represented by the slope as a predicted, expected, or average change, to clearly distinguish it from an observed change in the response variable.
- In part (c) many responses were not able to correctly interpret r^2 in context. Difficulties included not only identifying the response variable in context, but also understanding the concept of r^2 as the percentage of variation in the response variable (mass) that can be explained by the variation in explanatory variable (length).
- In part (d-i) the majority of responses were able to identify the correct point with the largest absolute value residual. However, a number of responses chose the point that was the farthest away from the origin.
- In part (d-ii) most responses were able to correctly identify the prediction as an overestimate or underestimate (depending on the point identified in part (d-i)). However, some had difficulty in explaining why the point would be an overestimate or underestimate.

What common student misconceptions or gaps in knowledge were seen in the responses to this question?

Common Misconceptions/Knowledge Gaps	Responses that Demonstrate Understanding
• In part (a) failing to identify all characteristics of a scatterplot.	• The scatterplot reveals a strong, positive, roughly linear association between the mass and length of bullfrogs, with no unusual features.
• In part (b) communicating the concept of the slope of a least-squares regression line as representing an expected or predicted change using appropriate units of measurement.	• The expected mass of a bullfrog increases 6.086 grams for each additional millimeter of length.
• In part (c) interpretation of r^2.	• The proportion of the variation in the response variable (mass) that is accounted for by variation in the explanatory variable (length).
• In the interpretation of r^2 in part (c), not describing the response and explanatory variables in the context of the study.	• 81.9% of the variation in bullfrog mass can be explained by variation in bullfrog length.
• In part (d) identifying a point with the largest absolute residual.	• The response must select the point that has the largest vertical distance from the least-squares regression line.
• In part (d) explaining why the least-squares regression line will give an overestimate or an underestimate.	• Given the line is above the actual point, the least-squares regression line will be an overestimate. • A negative residual indicates that the predicted point will be higher than the actual point.

Based on your experience at the AP® Reading with student responses, what advice would you offer teachers to help them improve the student performance on the exam?

- Emphasize the importance of the characteristics of a scatterplot: direction of association, strength of association, form of association, and unusual features. Additionally, when writing about these characteristics, responses should be in the context of the study that provided the data.
- Remind students that a least-squares regression gives a *predicted* value. In their interpretation, it is important that they explain that the slope represents a "predicted," "estimated," or "expected" change.
- Discuss the importance of units in an interpretation of slope, and that for each additional unit change in the explanatory variable, the slope is the predicted change in the units of the response variable.
- Have students practice interpreting r^2, the proportion of variation in the response variable (y) that can be explained by the variation in explanatory variable (x).
    - **TIP:** It might be helpful to expose students to alternate language, such as:
        - r^2 represents the proportion of variation in *y* that is accounted for by the linear model.
        - r^2 represents the proportion of variation in *y* that is explained by its linear relation to *x*.
- Emphasize that when interpreting r^2 in context, it represents the proportion of the variability in the response variable.

- Discuss the relationship between the predicted values of a least-squares regression line and the actual values, and how that relates to positive or negative residuals.

What resources would you recommend to teachers to better prepare their students for the content and skill(s) required on this question?

- The *AP Statistics Course and Exam Description* (CED), effective Fall 2020, includes instructional resources for AP Statistics teachers to develop students' broader skills. Please see page 227 of the CED for examples of key questions and instructional strategies designed to develop skill 2.A, describe data presented numerically or graphically, and skill 2.C, calculate summary statistics, relative positions of points within a distribution, correlation, and predicted response, as well as page 232 for skill 4.B, interpret statistical calculations and findings to assign meaning or assess a claim. A table of representative instructional strategies, including definitions and explanations of each, is included on pages 213-223 of the CED. The strategy "Two Wrongs Make a Right," for example, may be helpful in developing students' abilities to identify all relevant features when describing the relationship between two variables in context based on a scatterplot.
- AP Classroom provides five videos focused on the content and skills to answer this question.
  - The daily video 2 for topic 2.4 discusses how to properly describe all the characteristics (direction, form, strength, unusual features, and context) of a scatterplot (see DAT-1.A.1).
  - The daily video 2 for topic 2.8 describes the precise interpretations of the slope of a linear regression model (see DAT-1.H.2). A key takeaway from this video that was relevant to this question is "The slope value tells you about the predicted change in y for every one-unit increase in x."
  - The daily video 3 for topic 2.8 demonstrates interpreting the coefficient of determination (r^2) (see DAT-1.G.4). A key takeaway from this video that was relevant to this question is "r^2 is the proportion of variation in the y variable that is explained by the x variable in the model."
  - The daily video 1 for topic 2.6 describes how the components of a linear regression model and demonstrates making predictions using that model (see DAT-1.D). A key takeaway from this video that was relevant to this question is using the linear regression model to make predictions about the response variable.
  - The daily video 1 for topic 2.7 discusses how to calculate and interpret residuals. Key takeaways of this video were especially relevant to this question: "Residual measure the difference between actual and predicted response values," and "Positive residual values indicate model under-prediction. Negatives indicate over-prediction."
- AP Classroom also provides topic questions for formative assessment of topics 2.4, 2.6, 2.7, and 2.8, as well as access to the question bank, which is a searchable database of past AP Questions on this topic.
- The Online Teacher Community features many resources shared by other AP Statistics teachers. For example, to locate resources to give your students practice discussing coefficient of determination, try entering the keywords "coefficient of determination" in the search bar, then selecting the drop-down menu for "Resource Library." When you filter for "Classroom-Ready Materials," you may find worksheets, data sets, practice questions, and guided notes, among other resources.

Question 2

Task: Collecting Data
Max Score: 4
Mean Score: 1.32

What were the responses to the question expected to demonstrate?

The primary goals of the question were to assess a student's ability to (1) identify the treatments, the experimental units, and the response variable from a description of an experiment; (2) identify a statistical advantage of a matched pairs design (such as increased ability to detect a treatment effect, reduced variability of the difference in treatment means, or more precise estimation of the treatment effect) relative to a completely randomized design; and (3) describe a correct procedure for randomly assigning two treatments to experimental units in a matched pairs experiment.

This question primarily assesses skills in skill category 1: Selecting Statistical Methods. Skills required for responding to this question include (1.B) Identify key and relevant information to answer a question or solve a problem, and (1.C) Describe an appropriate method for gathering and representing data.

This question covers content from Unit 3: Collecting Data of the course framework in the AP Statistics Course and Exam Description. Refer to topics 3.5 and 3.6 and learning objectives VAR-3.A, VAR-3.D, and VAR-3.B.

How well did the responses address the course content related to this question? How well did the responses integrate the skills required on this question?

- In part (a) most responses correctly identified the treatments and the response variable. However, many responses did not correctly identify the experimental units and some gave an incomplete identification of the response variable by failing to describe it in the context of the study as "improvement" in "acne" score.
- In part (b) most responses provided an advantage of a matched pairs design that was described in context and included an explicit or implicit comparison to a completely randomized design. However, many of these responses failed to identify a statistical advantage related to the ability to reach a conclusion from the experiment.
- In part (c) most responses described a method to randomly assign treatments so that one twin in each pair receives the placebo and the other twin receives the new drug. Most of these responses described the process in sufficient detail, but some did not.

What common student misconceptions or gaps in knowledge were seen in the responses to this question?

Common Misconceptions/Knowledge Gaps	Responses that Demonstrate Understanding
- In part (a) many responses incorrectly identified the experimental units as the "pairs of twins" instead of the individual twins. "Pairs of twins" is incorrect because specific treatments are assigned to each individual twin, not to each pair of twins.	- The experimental units are the 72 twins in the experiment.

• In part (b) very few responses clearly described a statistical advantage of a matched pairs design related to the conclusion that could be reached from the results of the study. Many responses focused on pre-conclusion advantages, such as accounting for a source of variability, reducing the potential of confounding, or being able to compare to someone very similar.	• A matched pairs design makes it easier to find convincing evidence that the new drug is better (more power) or gives a more precise estimate of the effectiveness of the drug (narrower confidence interval).
• In part (b) many responses used vocabulary incorrectly. For example, responses often used "accurate" in place of "precise" and misused terms like "bias," "confounding," and "skewed."	• 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.
• In part (b) many responses did not explicitly compare the matched pairs design to a completely randomized design.	• Unlike in a completely randomized design, 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.
• In part (c) some responses did not randomly assign treatments within pairs. That is, it was not clear that one member of the pair would receive the new drug and the other member of the pair would receive the placebo.	• In each pair, randomly assign one twin to the new drug and one twin to the placebo.
• In part (c) some responses did not sufficiently describe how to use the results of a random process (e.g., flipping a coin) for the assignment of treatments. For example, some responses described only what to do if the coin landed on heads or only what to do with the first twin. Other responses did not assign a treatment to a specific twin (e.g., if the coin flip is heads, give one twin the drug).	• 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.

• In part (c) some responses did not indicate that the random assignment procedure should be applied to each of the 36 pairs.	• For each pair of twins… OR Repeat for all 36 pairs of twins.

Based on your experience at the AP® Reading with student responses, what advice would you offer teachers to help them improve the student performance on the exam?

- Make sure to give students opportunities to practice identifying experimental units in varied contexts, such as in a matched pairs design. See also 2019 Free-Response Question 2.
- Make connections between different units of the course. When things are described as beneficial in Unit 3 (Collecting Data), the benefits are often pointing to greater power or precision. Because power and precision are not addressed until Units 6–9, spend time in those units revisiting concepts in Unit 3.
  - **TIP:** When discussing factors that affect the margin of error and factors that affect power, include study design (e.g., using a stratified random sample reduces the margin of error of an estimate, using pairing/blocking increases the power of a test).
- Encourage students to practice using statistical vocabulary in their responses and give detailed feedback to students as often as possible.
- When asked to compare two (or more) options, make sure students address both options. For example, by describing a positive aspect of one option and a negative aspect of the other option(s).
- Give students practice describing the process of random assignment for different designs (completely randomized, blocked, or matched pairs).
- Require students to be detailed in their descriptions of random assignment (and random selection for sampling items).
  - **TIP:** Have students imagine giving instructions to a computer that will only do exactly what it is told.
- Students should not assume that the reader will "fill in the rest" of a response. Readers can only award credit for what the student writes and cannot fill in any part of a response, no matter how obvious a student might think it is.

What resources would you recommend to teachers to better prepare their students for the content and skill(s) required on this question?

- The *AP Statistics Course and Exam Description* (CED), effective Fall 2020, includes instructional resources for AP Statistics teachers to develop students' broader skills. Please see page 225 of the CED for examples of key questions and instructional strategies designed to develop skills 1.B, identify key and relevant information to answer a question or solve a problem, and skill 1.C, describe an appropriate method for gathering and representing data. A table of representative instructional strategies, including definitions and explanations of each, is included on pages 213-223 of the CED. The strategy "Team Challenge," for example, may be helpful in developing students' abilities to identify the treatments, experimental units, and response variable for an experiment.
- AP Classroom provides three videos focused on the content and skills to answer this question.
  - The daily video 1 for topic 3.5 discusses the basic components of an experiment (see VAR-3.A).
  - The daily video 2 for topic 3.6 describes a matched pairs design (see VAR-3.D.1). A key takeaway from this video that was relevant to this question is "Matched pairs designs are a special form of randomized block design using blocks of two similar experimental units, one receiving each treatment. Another type of matched pairs design includes giving each experimental unit both treatments in a random order."
  - The daily video 2 for topic 3.5 discusses the elements of a well-designed experiment (see VAR-3.B.1). Key takeaways of this video were especially relevant to this question: "A well-designed experiment should include comparisons between at least two groups, random assignment of treatments to experimental units, replication of treatments to multiple experimental units, and control of possible confounding factors."
- AP Classroom also provides topic questions for formative assessment of topics 3.5 and 3.6, as well as access to the question bank, which is a searchable database of past AP Questions on this topic.
- The Online Teacher Community features many resources shared by other AP Statistics teachers. For example, to locate resources to give your students practice discussing experiments, try entering the keyword "experiments" in

the search bar, then selecting the drop-down menu for "Resource Library." When you filter for "Classroom-Ready Materials," you may find worksheets, data sets, practice questions, and guided notes, among other resources.

Question 3

Task: Probability and Sampling Distributions
Max Score: 4
Mean Score: 1.32

What were the responses to the question expected to demonstrate?

The primary goals of the question were to assess a student's ability to (1) calculate the probability that a bottle filling machine would underfill a bottle of shampoo using a specified normal distribution; (2) define a random variable as the number of underfilled bottles in a box of ten shampoo bottles; (3) describe the distribution of that random variable; (4) use the identified distribution to compute a probability, showing work; and (5) identify and compare relevant quantities, e.g., probabilities or z-scores, to justify a recommendation about whether a specific adjustment to the bottle filling machine should be made.

This question primarily assesses skills in skill category 3: Using Probability and Simulation, and skill category 4: Statistical Argumentation. Skills required for responding to this question include (3.A) Determine relative frequencies, proportion, or probabilities using simulation or calculations, and (4.B) Interpret statistical calculations and findings to assign meaning or assess a claim.

This question covers content from Unit 4: Probability, Random Variables, and Probability Distributions, and Unit 5: Sampling Distributions of the course framework in the AP Statistics Course and Exam Description. Refer to topics 4.3, 4.10, and 5.2, and learning objectives VAR-6.A, UNC-3.B, UNC-3.A, and VAR-4.B.

How well did the responses address the course content related to this question? How well did the responses integrate the skills required on this question?

- In part (a) most responses correctly calculated the probability and attempted to show some work to support their answer. However, many responses did not adequately identify parameters or clearly specify the correct event. A sketch of the normal distribution including a reasonably scaled horizontal axis with the bound for the event of interest clearly indicated and the corresponding probability shaded is often helpful.
- In part (b) a majority of responses did not correctly define the random variable as a numerical outcome of a random event, but instead defined it as a probability or as an event. When stating how the random variable is distributed, many responses listed the checks for the conditions of a binomial distribution rather than naming the binomial distribution and identifying the correct parameters. In addition, several responses identified the distribution as "normal," but then used the binomial formula to calculate a probability. Many responses did show the ability to calculate a binomial probability using the binomial formula or calculator notation. However, when calculating the cumulative probability, several responses did not use the correct boundary value to define the event of interest.
- In part (c) many responses correctly calculated the normal probability and made a comparison of the work in part (a) with the work in part (c). Some responses contained the correct probability but did not explicitly compare the probabilities or fully discuss the implications of a difference in z-scores. In addition, many responses began with "yes" or "no," which does not provide a conclusion about which programming method to use.

What common student misconceptions or gaps in knowledge were seen in the responses to this question?

Common Misconceptions/Knowledge Gaps	Responses that Demonstrate Understanding
• A number of responses used a *z*-score calculation without labeling the variable "*z*" and simply writing $\frac{0.5 - 0.6}{0.04} = -2.5$.	• $z = \frac{0.5 - 0.6}{0.04} = -2.5$
• Responses often used calculator notation without clearly labeling boundaries and parameters, such as normalcdf ($-\infty$, 0.5, 0.6, 0.04).	• normalcdf (lower bound $= -\infty$, upper bound $= 0.5, \mu = 0.6, \sigma = 0.04$) • binomcdf $(n = 10, p = 0.0062, x = 1)$
• Some responses made errors in statistical notation, such as referring to $\bar{x}$ rather than μ. The response should show that the parameters refer to the population.	• $\mu = 0.6$, $\sigma = 0.04$
• In this question, the random variable, A, is defined. Many responses to part (a) used another variable, such as $P(X) < 0.50$, but without defining the new variable.	• $P(A < 0.50)$
• Many responses to part (b) defined a random variable as a probability, which is not correct. ("random variable is the probability of getting at least two bottles underfilled.") A random variable is a quantitative outcome of a random event.	• The random variable, *X*, is the number of underfilled shampoo bottles in a box of 10 bottles.
• Several responses could not correctly determine the boundary value for a binomial calculation, incorrectly calculating $P(X = 2)$ or $1 - P(X \le 2)$.	• $P(X \ge 2) = 1 - P(X \le 1)$ $= 1 - binomcdf (n = 10, p = 0.0062, x = 1)$
• Many responses did not use the comparison of two correctly computed probabilities to support their choice of the original programming. (e.g., The probability of underfilling a bottle using the new programming is 0.023, whereas the probability for the original programming is 0.0062.)	• The probability of underfilling a bottle using the new programming is 0.023, which is greater than the probability for the original programming, which is 0.0062. This makes underfilling a shampoo bottle less likely with the original program.

• Several responses did not fully explain the implications of different *z*-score values and how this difference would justify their programming choice. (e.g., the *z*-score for the original programming is 2.5 standard deviations from the mean, which is less than the new programming, which is 2 standard deviations from the mean.)	• The *z*-score for the original programming is 2.5 standard deviations **below** the mean, which is less than the *z*-score for the new programming, which is 2 standard deviations below the mean. A lower *z*-score results in fewer underfilled bottles.

Based on your experience at the AP® Reading with student responses, what advice would you offer teachers to help them improve the student performance on the exam?

- Stress the importance of correct statistical notation. When making any calculation, first determine whether the values refer to a sample or the population.
- When using calculator function syntax, be sure to label all values.
- Have students practice defining random variables throughout the course. When doing so, ensure the response is in the context of the problem or data set.
  - **TIP:** A brief discussion of the scenario can often help students better understand the possible outcomes and better define the random variable.
- Have students practice identifying the type of distribution (e.g., normal, binomial, geometric) for a random variable, including identifying the parameters of the distribution.
- Have students practice comparing distributions, probabilities, and methods using explicit comparison words, such as "greater than," "more than," or "higher than." In addition, encourage them to discuss the implications of the comparison in the context of the problem.

What resources would you recommend to teachers to better prepare their students for the content and skill(s) required on this question?

- The *AP Statistics Course and Exam Description* (CED), effective Fall 2020, includes instructional resources for AP Statistics teachers to develop students' broader skills. Please see page 229 of the CED for examples of key questions and instructional strategies designed to develop skill 3.A, determine relative frequencies, proportions, or probabilities using simulations or calculations, and page 232 for skill 4.B, interpret statistical calculations and findings to assign meaning or assess a claim. A table of representative instructional strategies, including definitions and explanations of each, is included on pages 213-223 of the CED. The strategy "Odd One Out" for example, may be helpful in developing students' abilities to identify binomial distributions.
- AP Classroom provides two videos focused on the content and skills to answer this question.
  - The daily video 1 for topic 5.2 shows how to calculate probabilities for continuous random variables associated with a normal distribution and determine the interval associated with a given area in a normal distribution (see VAR-6.A).
  - The daily video for topic 4.10 discusses how to identify a binomial random variable and how to calculate binomial probabilities (see UNC-3.A and UNC-3.B). The key takeaways from this video that was relevant to this question are defining a random value, identifying the distribution and values of interest, determining probabilities using the binomial probability formula, and answering a question in context.
- AP Classroom also provides topic questions for formative assessment of topics 4.3, 4.10, and 5.2, as well as access to the question bank, which is a searchable database of past AP Questions on this topic.
- The Online Teacher Community features many resources shared by other AP Statistics teachers. For example, to locate resources to give your students practice discussing normal distribution, try entering the keywords "normal distribution" in the search bar, then selecting the drop-down menu for "Resource Library." When you filter for "Classroom-Ready Materials," you may find worksheets, data sets, practice questions, and guided notes, among other resources.

Question 4

Task: Inference
Max Score: 4
Mean Score: 2.28

What were the responses to the question expected to demonstrate?

The primary goals of the question were to assess a student's ability to (1) identify an appropriate procedure for constructing a confidence interval for a population proportion; (2) check conditions required for accurate application of the identified procedure; (3) calculate the confidence interval; (4) interpret the confidence interval in the context of the proportion of the relevant population that would give a specific response to a survey, and (5) use the confidence interval to determine if a statement about the population is justified.

This question primarily assesses skills in skill category 1: Selecting Statistical Methods, skill category 3: Using Probability and Simulation, and skill category 4: Statistical Argumentation. Skills required for responding to this question include (1.D) Identify an appropriate inference method for confidence intervals, (3.D) Construct a confidence interval, provided conditions for inference are met, (4.B) Interpret statistical calculations and findings to assign meaning or assess a claim, (4.C) Verify that inference procedures apply in a given situation, and (4.D) Justify a claim based on a confidence interval.

This question covers content from Unit 6: Inference for Categorical Data: Proportions of the course framework in the AP Statistics Course and Exam Description. Refer to topics 6.2 and 6.3 and learning objectives UNC-4.A, UNC-4.B, UNC-4.D, UNC-4.F, and UNC-4.G.

How well did the responses address the course content related to this question? How well did the responses integrate the skills required on this question?

This two-part question is scored in four sections. The response to part (a) is scored in three sections, the components of which may appear in any order in the response. The first section includes identification of the appropriate confidence interval in part (a). The second section includes verifying the conditions for inference in part (a) and calculating the values of the end points of the confidence interval. The third section includes the interpretation of the confidence interval in part (a). The fourth section includes the response to part (b).

Section 1:
- A substantial majority of responses failed to state the type of confidence interval to be constructed and only identified the procedure by using the calculator's name of "1 prop *z*-int." This notation does not indicate the confidence interval is for the *population* proportion.
- Many responses used nonstandard or inconsistent notation when referring to population and sample proportions.

Section 2:
- Most responses recognized that conditions must be checked before conducting a confidence interval, however, a substantial number of responses failed to properly check those conditions.
- The check of the independence condition was frequently incomplete. Most responses cited the condition of random sampling, but some failed to check the 10% condition. Further, some responses simply stated "random-stated" without indicating that data were obtained from a random sample.
- Some responses, when verifying the large sample condition, stated that the number was larger than 30, which indicates confusion between when to use a *t* or a *z* procedure.
- Most responses reported correct values of the confidence interval endpoints found from their calculator, with few attempting to directly calculate the interval using the formula.

- For responses that attempted to use the formula, some responses contained errors in using the formula, such as:
  - using the expected number of successes $(920)(0.59) = 542.8$ as the center of the interval instead of the sample proportion value of 0.59.
  - using a t^* value with $\text{df} = 919$.
  - writing the formula with incorrect notation.
- Some responses failed to recognize unrealistic values for proportions, i.e., their confidence interval for proportions was $(541, 620)$.

Section 3:
- Most responses did a very nice job of providing an interpretation of the confidence interval, which included the parameter of interest, in the context of the study.
- Some responses interpreted the confidence level and not the confidence interval.
- The most common error was the use of inconsistent terms by stating an interval was for a proportion, but instead providing an interval for a percentage.

Section 4:
- Most responses recognized how to use the provided interval to determine if a value was plausible.
- The most common issue was poor communication in describing how the interval showed there was convincing evidence.
- Some responses incorrectly used double negatives, i.e., "yes, the value is not in the interval so there is no evidence that the value is not .50," so contradicting the original "yes."
- Some responses overstated the finding that 0.50 is not in the interval and declared that the value of 0.50 is definitely not plausible.

What common student misconceptions or gaps in knowledge were seen in the responses to this question?

Most errors occurred in responses that poorly organized their work or poorly communicated their ideas. More specific errors are noted in the table below.

Common Misconceptions/Knowledge Gaps	Responses that Demonstrate Understanding
• Neither stating the parameter of interest nor the type of inference procedure to be implemented at the beginning of the response and/or using nonstandard notation for population proportion.	• One sample z-interval for a population proportion.
• Not stating the required conditions of a random sample and the 10% rule or listing these conditions without verifying them using information provided in the description of the study, i.e., only stating "SRS-check" and "$10n < N$."	• A random sample of 920 was selected, which is less than ten percent of all teenagers in the United States.

• Not verifying that the number of success and failures are more than 10, OR stating that these values should be greater than 30, which suggests that they are using a *t*-interval for a population mean, OR using incorrect/inconsistent notation when referring to the sample proportion.	• As $920(0.59) = 542.8$ and $920(1 - 0.59) = 377.2$, then the number of successes and failures are both larger than 10 so the sample is large enough to support a condition of normality of the sampling distribution.
• Incorrectly calculating values for the interval endpoints when using the formula by either 1. using 542.8 as the estimate, OR 2. using a t^* value with df = 919, instead of a z^* value of 1.96, OR 3. providing values that are unrealistic values for proportions.	• A 95% confidence interval for the population proportion is given by $\hat{p} \pm z^* \sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}} = 0.59 \pm 1.96 \sqrt{\dfrac{(0.59)(0.41)}{920}}$, which is 0.59 ± 0.032, and the interval is $(0.5573, 0.6219)$.
• Using inconsistent terms when interpreting the interval by providing values that are percents, but stating that these are proportions.	• We can be 95% confident that the proportion of all teenagers in the United States who would respond that they use a streaming service every day is between 0.558 and 0.622.
• Implying that we have proven that the value of 0.50 is not possible, e.g., "as 0.50 is not in the interval, the value is NOT 0.50."	• The sample data provide convincing statistical evidence that the proportion of all teenagers in the United States who would use a streaming service every day is not 0.5.
• Not using the interval as directed to answer the question and instead using a hypothesis test.	• The value of 0.50 is not contained in the 95% confidence interval.

Based on your experience at the AP® Reading with student responses, what advice would you offer teachers to help them improve the student performance on the exam?

- Provide opportunities for students to practice writing skills from the beginning of the course.
    - **TIP:** Assign previously released AP problems as assessments.
    - **TIP:** Teach students organizational strategies (e.g., state/plan/do/conclude).
- Encourage students to classify variables as either categorical or quantitative when looking at all data sets.
    - **TIP:** Provide students with a multivariable data set and a research question and have them decide which variable(s) could be used to answer that question and explain why.
- Encourage students to define the population parameter of interest in context.
- Emphasize the importance of checking conditions before conducting inference procedures.
- Assess student's ability to quickly decide on appropriate inference procedures.
- Assess students' knowledge of standard statistical notation.
    - **TIP:** Assign a notation quiz.
- Encourage students to name inference procedures in words instead of by formula or by using a calculator name.
- Emphasize the differences when writing the large enough sample size conditions for hypothesis tests and confidence intervals for proportions.
- Emphasize that the independence condition is checked by 1) random sampling AND 2) the sample size is less than 10% of the population.

What resources would you recommend to teachers to better prepare their students for the content and skill(s) required on this question?

- The *AP Statistics Course and Exam Description* (CED), effective Fall 2020, includes instructional resources for AP Statistics teachers to develop students' broader skills.
    - Section 1: Please see pages 226, 230, and 232 of the CED for examples of key questions and instructional strategies designed to develop skill 1.D, identify an appropriate inference method for confidence intervals, skill 3.D, construct a confidence interval, provided conditions for inference are met, skill 4.B, interpret statistical calculations and findings to assign meaning or assess a claim, and skill 4.C, verify that inference procedures apply in a given situation.
    - Section 2: Please see page 232 of the CED for examples of key questions and instructional strategies designed to develop skill 4.D, justify a claim based on a confidence interval.
- A table of representative instructional strategies, including definitions and explanations of each, is included on pages 213-223 of the CED. The strategy "Build the Model Solution," for example, may be helpful in developing students' abilities to use precise language when justifying a claim based on a confidence interval.
- AP Classroom provides three videos focused on the content and skills to answer this question.
    - The daily video 1 for topic 6.2 discusses how to identify the procedure and check the conditions for constructing a confidence interval for a population proportion (see UNC-4.A.1 and UNC-4.B.2). The key takeaways from this video that were relevant to this question are identifying an appropriate confidence interval procedure for a population proportion and verifying the conditions when calculating a confidence interval for a population proportion.
    - The daily video 2 for topic 6.2 describes how to calculate a confidence interval for a population proportion (see UNC-4.D.1).
    - The daily video 3 for topic 6.3 demonstrates how to interpret a confidence interval for a population proportion and use the interval to justify a claim (see UNC-4.F.2 and UNC-4.G.1). A key takeaway from this video that was relevant to this question is "If all the values in the confidence interval are consistent with the claim, there is convincing evidence for the claim. If one or more of the values in the confidence interval are inconsistent with the claim, there is not convincing evidence for the claim."
- AP Classroom also provides topic questions for formative assessment of topics 6.2 and 6.3, as well as access to the question bank, which is a searchable database of past AP Questions on this topic.
- The Online Teacher Community features many resources shared by other AP Statistics teachers. For example, to locate resources to give your students practice discussing confidence intervals, try entering the keywords "confidence intervals" in the search bar, then selecting the drop-down menu for "Resource Library." When you filter for "Classroom-Ready Materials," you may find worksheets, data sets, practice questions, and guided notes, among other resources.

Question 5

Task: Multi-Focus
Max Score: 4
Mean Score: 0.87

What were the responses to the question expected to demonstrate?

The primary goals of the question were to assess a student's ability to (1) compute and compare sample medians for data presented in two dotplots; (2) explain why a conclusion based only on the difference between two sample means may not necessarily be true; and (3) use results of a simulation, presented in a dotplot, to justify a conclusion about whether a study provides convincing statistical evidence that one treatment is better than another treatment.

This question primarily assesses skills in skill category 1: Selecting Statistical Methods, skill category 2: Data Analysis, and skill category 4: Statistical Argumentation. Skills required for responding to this question include (1.E) Identify an appropriate inference method for significance tests, (2.A) Describe data presented numerically or graphically, and (4.B) Interpret statistical calculations and findings to assign meaning or assess a claim.

This question covers content from Unit 1: Exploring One-Variable Data and Unit 7: Inference for Quantitative Data: Means of the course framework in the AP Statistics Course and Exam Description. Refer to topics 1.6, 7.8, and 7.9, and learning objectives UNC-1.H, VAR-7.F, and DAT-3.H.

How well did the responses address the course content related to this question? How well did the responses integrate the skills required on this question?

- In part (a) most responses correctly identified the medians for both treatment groups, made a direct comparison, and included complete context. A few responses failed to properly identify the median of a data set with an even or odd sample size. Furthermore, a few responses failed to make a clear directional comparison. Additionally, some responses did not contain all criteria required for complete context. Often the word "reduction" was omitted when referring to the response variable "reduction in blood pressure."
- In part (b) many responses struggled to use statistical language correctly when describing the variability of the distribution of differences of the sample means. In addition, many responses recognized the need for an inference procedure, but struggled to explain why.
- In part (c) many responses did not demonstrate the correct use of the simulation results to find a *p*-value or used the data from the dotplots from the problem stem to perform an inference procedure.

What common student misconceptions or gaps in knowledge were seen in the responses to this question?

Common Misconceptions/Knowledge Gaps	Responses that Demonstrate Understanding
- Failing to make a direct comparison of the medians. - Failing to include full context for the response variable.	- The median reduction in blood pressure for those who consumed dark chocolate is 7 mmHg, which is greater than the median reduction in blood pressure for those who consumed white chocolate, which is 0 mmHg.
- Failing to use correct terminology to describe the variability in the sampling distribution of differences in sample means.	- The researcher's conclusion may not be true because the difference of 5.66 mmHg may be caused by random chance if the standard error of the sampling distribution of the difference in sample means is large.

• Failing to focus on the variability of the difference of sample means and instead focused on the variability within the samples.	
• Failing to demonstrate how to use the results of a simulation to find a *p*-value.	• Using the simulated distribution above, only 3 of the 120 trials resulted in a difference in reduction of blood pressure of 5.66 mmHg or greater. This means the *p*-value equals 0.025.

Based on your experience at the AP® Reading with student responses, what advice would you offer teachers to help them improve the student performance on the exam?

- Provide students the opportunity to use correct statistical language to describe and explain statistical procedures and concepts.
- Require students to use complete context in all responses.
- Allow students to use simulation techniques to perform hypothesis tests.
  - **TIP:** Have students perform small simulations with manipulatives early in the year before moving to larger simulations using technology.

What resources would you recommend to teachers to better prepare their students for the content and skill(s) required on this question?

- The *AP Statistics Course and Exam Description* (CED), effective Fall 2020, includes instructional resources for AP Statistics teachers to develop students' broader skills.
  - Section 1: Please see page 227 of the CED for examples of key questions and instructional strategies designed to develop skill 2.A, describe data presented numerically or graphically.
  - Section 2: Please see page 226 of the CED for examples of key questions and instructional strategies designed to develop skill 1.E, identify an appropriate inference method for significance tests.
  - Section 3: Please see page 232 of the CED for examples of key questions and instructional strategies designed to develop skill 4.B, interpret statistical calculations and findings to assign meaning or assess a claim.
- A table of representative instructional strategies, including definitions and explanations of each, is included on pages 213-223 of the CED. The strategy "Simulation," for example, may be helpful in reinforcing the role of probability in inference.
- AP Classroom provides three videos focused on the content and skills to answer this question.
  - The daily video for topic 1.6 describes the characteristics of quantitative data distributions (see UNC-1.H.1). The key takeaway from this video that is relevant to this question was the best ways to discuss the important characteristics when describing a distribution of quantitative data.
  - The daily video 1 for topic 7.8 demonstrates how to state the null and alternative hypotheses for a significance test of a difference in means (see VAR-7.F.1).
  - The daily video 2 for topic 7.9 discusses how to interpret the *p*-value and state a conclusion for a significance test of a difference in means (see DAT-3.H). The key takeaways from this video that was relevant to this question are strategies to interpret the *p*-value and state a conclusion in a significance test for the difference of two population means.
- AP Classroom also provides topic questions for formative assessment of topics 1.6, 7.8, and 7.9, as well as access to the question bank, which is a searchable database of past AP Questions on this topic.
- The Online Teacher Community features many resources shared by other AP Statistics teachers. For example, to locate resources to give your students practice discussing simulations, try entering the keyword "simulations" in the search bar, then selecting the drop-down menu for "Resource Library." When you filter for "Classroom-Ready Materials," you may find worksheets, data sets, practice questions, and guided notes, among other resources.

Question 6

Task: Investigative Task
Max Score: 4
Mean Score: 1.55

What were the responses to the question expected to demonstrate?

The primary goals of the question were to assess a student's ability to (1) use information presented as a table of counts to compute relative frequencies of successful and unsuccessful treatment of allergies at each of two clinics; (2) use the computed relative frequencies to determine which of two clinics is more successful in treating allergy sufferers; (3) recognize that the data were obtained from an observational study; (4) justify a decision about whether a causal inference may be made based on the type of study that produced the data; (5) use information presented in two mosaic plots to compute and compare relative frequencies of successfully treating mild and severe allergy sufferers for each of two clinics; (6) use information presented in two mosaic plots to determine whether mild or severe allergy sufferers are more likely to be treated for each clinic; and (7) use previous answers to explain why two conclusions about which clinic is more successful may be different (or the same).

This question primarily assesses skills in skill category 2: Selecting Statistical Methods, and skill category 4: Statistical Argumentation. Skills required for responding to this question include (2.B) Construct numerical or graphical representations of distributions, (2.D) Compare distributions or relative positions of points within a distribution, and (4.A) Make an appropriate claim or draw an appropriate conclusion.

This question covers content from Unit 1: Exploring One-Variable Data, Unit 2: Exploring Two-Variable Data, and Unit 3: Introductions to Planning a Study of the course framework in the AP Statistics Course and Exam Description. Refer to topics 1.3, 2.2, and 3.2 and learning objectives UNC-1.A, DAT-2.B, and UNC-1.P.

How well did the responses address the course content related to this question? How well did the responses integrate the skills required on this question?

- In part (a) most responses reported the correct conditional relative frequencies of successful and unsuccessful treatments at each clinic. Most responses also included a correct comparison of the relevant relative frequencies to make a correct decision. However, some responses confused frequencies and relative frequencies. In addition, many responses included relative frequencies of the overall total instead of the conditional relative frequencies for each clinic. Lastly, many students did not make it clear what was being compared, for example, stating, "Clinic A has a higher relative frequency," without indicating relative frequency of successes, or indicating it was making a comparison with Clinic B.
- In part (b) many responses correctly indicated a causal inference cannot be made from an observational study. In addition, many responses indicated this was not an experiment or that there was no random assignment of clinics to the patients. Many responses argued that confounding variables were not controlled for but failed to identify a plausible confounding variable that could affect success rates for treating allergies at the two clinics. Furthermore, some responses stated that a claim of a cause-and-effect relationship would be valid based on the result of a significance test or the fact that conditions for inference were satisfied. Finally, several responses indicated that no conclusion could be reached or that results could not be generalized, instead of focusing on causal inference.
- In part (c) most responses clearly indicated the correct severity level for both clinics in both sub-parts. Many responses included a correct comparison of relative frequencies to justify responses. However, several responses justified an answer by stating, for example, "There is a higher proportion of successful treatments for mild cases than severe cases," without providing sufficient justification, using quantitative information provided by the mosaic plots. Very few responses referred to a relevant characteristic of the mosaic plots, which was not required in this question, but is something students should be comfortable doing.

- In part (d) very few responses provided a clear explanation for the apparent discrepancy between the more successful clinic in part (a-ii) and the physician's conclusion. While several responses showed a general understanding of why the paradox exists, communication tended to be weak. In addition, several responses failed to use answers from part (c) and simply provided numerical justification for the more successful clinic identified in part (a-ii) and the physician's conclusion.

What common student misconceptions or gaps in knowledge were seen in the responses to this question?

Common Misconceptions/Knowledge Gaps	Responses that Demonstrate Understanding
• Several responses used frequencies or relative frequencies based on the overall total instead of conditional relative frequencies.	• The correct conditional relative frequencies for each clinic are in the following table: <table><tr><td></td><td>Clinic A</td><td>Clinic B</td></tr><tr><td>Unsuccessful Treatment</td><td>$\frac{51}{139} = 0.367$</td><td>$\frac{33}{68} = 0.485$</td></tr><tr><td>Successful Treatment</td><td>$\frac{88}{139} = 0.633$</td><td>$\frac{35}{68} = 0.515$</td></tr></table>
• Several responses did not include a clear comparison of the relevant relative frequencies.	• Clinic A appeared to be more successful than Clinic B. In Clinic A, 63.3% of patients were treated successfully. This is higher than the percentage of patients successfully treated at Clinic B, which was 51.5%.
• Some responses indicated that a statistically significant result would allow the researchers to make a causal conclusion, due simply to a significance test or because conditions for inference are satisfied.	• No, it would not be appropriate for researchers to conclude Clinic A causes a higher proportion of successful treatments than Clinic B. This was an observational study, not a randomized experiment. Without random assignment of clinics to patients, it is possible a confounding variable, such as severity of allergy, could lead to the association between clinic and success rate. For example, perhaps Clinic A sees a higher proportion of mild cases than Clinic B does, and Clinic B sees a higher proportion of severe cases than Clinic A does. If, in general, mild cases are treated more successfully than severe cases, we will not be able to determine if Clinic A's higher success rate is due to the clinic or to the severity of the allergy they tend to treat.
• Some responses stated that no conlusion could be made or that results could not be generalized to the population.	• No, because this is not an experiment, a cause-and-effect relationship between clinic and success rate cannot be established.
• Many responses indicated that a conclusion (or cause-and-effect conclusion) cannot be made because the sample sizes are different.	• No, because this is an observational study and not a randomized experiment, a causal inference cannot be made.

• Several responses refered to the correct proportions, but did not specifically report them.	• Clinic A treated mild cases more successfully than severe cases. Clinic A's success rate for mild cases was 75%, higher than that of severe cases at 28.6%.
• Very few responses referred to a characteristic of the mosaic plot.	• Clinic A: Mild allergy sufferers are more likely to be treated than severe allergy sufferers. In the mosaic plot for Clinic A, the width of the bar for "Mild" cases is almost three times wider than the bar for "Severe" cases. • Clinic B: Severe allergy sufferers are more likely to be treated than mild allergy sufferers. In the mosaic plot for Clinic B, the width of the bar for "Mild" cases is not nearly as wide as the bar for "Severe" cases.
• Many responses did not focus on the reasons for the paradox, and instead, simply justified the physician's conclusion.	• Both clinics treated mild allergy sufferers more successfully than severe allergy sufferers. For both clinics, their proportion of successful cases for mild allergy sufferers was higher than their proportion of successful cases for severe allergy sufferers. Furthermore, Clinic A treated a much higher proportion of mild cases than Clinic B did, and Clinic B treated a much higher proportion of severe cases than Clinic A did.
• Many responses did not use answers from part (c) as the question prompted.	• As was shown in part (c-i), mild allergies were treated more successfully than severe allergies. As shown in part (c-ii), Clinic A treated more mild cases than severe cases and Clinic B treated more severe cases than mild cases.

Based on your experience at the AP® Reading with student responses, what advice would you offer teachers to help them improve the student performance on the exam?

- Students need to understand what they are expected to do when presented with a specific question.
  - **TIP:** In the AP Statistics Course and Exam Description (CED), page 239 contains a list of "Task Verbs in Free-Response Questions." Make sure students are familiar with this list, as well as the descriptions for what a response should include when each of these verbs is used.
- Students need to be encouraged to read questions carefully and pay attention to specific information they are given. For example, if the question reads, "If there is a statistically significant result," it is unnecessary to conduct a hypothesis test—the result has been provided.
- Teachers should help students understand why each element of the design of a proper study is important. Students should be able to address which feature of a design allows for causal inference, a generalization to the population, a reduction in bias, control for possible confounding variables, and a decrease in variability in a sampling distribution.
- If a question states what an answer should be based on, it is important that responses include justification based on what is prompted.
  - **TIP: Example 1:** If a question starts, "Based on the design of this study," the response should be focused on elements of the design of the study (observational study vs. experiment, random assignment, whether or not there is a control group, replication, etc.).
  - **Example 2:** If a question starts, "Using your answers to parts (a) and (b)," it is expected that the response makes explicit reference to the answers presented for those parts.

- When making a decision based on a comparison, responses need to be explicit about what is being compared. Comparative phrases, such as "more than," "larger than," should be included in responses, along with a clear indication of what is being compared.
  - **TIP:** Make sure students have plenty of practice making comparisons and basing specific decisions on those comparisons.
- This is the first free-response question on the AP Statistics exam that asked students to interpret a mosaic plot since mosaic plots appeared in the CED. Most students addressed this question by using numerical values provided in the plot, which was fine. However, students should also be comfortable interpreting characteristics of a mosaic plot in a case where numbers are not provided.
  - **TIP:** Make sure students have practice interpreting mosaic plots in which different amounts of specific information is provided.
    - Students should know which feature of a mosaic plot is important in making a specific conclusion (heights of bars, widths of bars, areas of bars, ratios of areas of bars).
    - Students should understand what additional information a mosaic plot provides that a segmented bar chart does not provide.
    - Students should be able to draw a mosaic plot if presented with data and interpret it.

What resources would you recommend to teachers to better prepare their students for the content and skill(s) required on this question?

- The *AP Statistics Course and Exam Description* (CED), effective Fall 2020, includes instructional resources for AP Statistics teachers to develop students' broader skills. Please see page 227 of the CED for examples of key questions and instructional strategies designed to develop skill 2.B, construct numerical or graphical representations of distributions, page 228 for examples of key questions and instructional strategies designed to develop skill 2.D, compare distributions or relative positions of points within a distribution, and page 231 for examples of key questions and instructional strategies designed to develop skill 4.A, make an appropriate claim or draw an appropriate conclusion.
- A table of representative instructional strategies, including definitions and explanations of each, is included on pages 213-223 of the CED. The strategy "Create Representations," for example, may allow students to have a deeper understanding of the differences in quantitative graphs.
- AP Classroom provides three videos on focused on the content and skills to answer this question.
  - The daily video for topic 1.3 describes how the information provided in a table can be used to describe a categorical variable (see UNC-1.A.1). The key takeaway from this video that are relevant to this question was that "counts and relative frequencies (percentages or proportions) of categorical data reveal information that can be used to justify claims about the data in content."
  - The daily video for topic 2.2 demonstrates how to utilize two-way tables and graphical displays to determine if two categorical variables are associated. (see UNC-1.P).
  - The daily video for topic 3.2 explores different types of studies and the scope of conclusions that can be drawn from them (see DAT-2.B).
- AP Classroom also provides topic questions for formative assessment of topics 1.3, 2.2, and 3.2, as well as access to the question bank, which is a searchable database of past AP Questions on this topic.
- The Online Teacher Community features many resources shared by other AP Statistics teachers. For example, to locate resources to give your students practice discussing relative frequency, try entering the keywords "relative frequency" in the search bar, then selecting the drop-down menu for "Resource Library." When you filter for "Classroom-Ready Materials," you may find worksheets, data sets, practice questions, and guided notes, among other resources.