AP® CollegeBoard

AP® Statistics
Sample Student Responses and Scoring Commentary

Inside:

Free Response Question 6

☑ **Scoring Guideline**

☑ **Student Samples**

☑ **Scoring Commentary**

Question 6: Investigative Task   4 points

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

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

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

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

Model Solution	Scoring
(b) During the years in the new stadium, the average per-game attendance increases linearly, from about 16,000 people in 2000 to about 27,000 people in 2016. However, during the years in the old stadium, there is no obvious increasing or decreasing trend over time for the average per-game attendance. The average attendance appears to vary about an average of approximately 16,000 attendees per game from 1970 to 1999.	**Essentially correct (E)** if the response satisfies the following three components: 1. Describes the direction of the trend in average per-game attendance in the new stadium as increasing (positive) 2. Describes the direction of the trend in average per-game attendance in the old stadium as relatively constant (e.g., "no association," "flat") *OR* Describes the direction of the trend in the old stadium as positive but less steep (e.g., "less positive," "flatter") than the trend in the new stadium 3. Provides sufficient context by including the two groups (old stadium, new stadium) AND the explanatory variable (time or year) AND the response variable (average attendance or attendance) or the units of the response variable (thousands of people or people) **Partially correct (P)** if the response satisfies only two of the three components **Incorrect (I)** if the response does not meet the criteria for E or P.

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

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

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

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

Additional Notes:
- The response can use any combination of 2 of the 3 explanatory variables (stadium and wins, stadium and year, year and wins). The response cannot introduce a new variable (e.g., weather, having popular players) to satisfy any component.
- An incorrectly described association cannot be used to satisfy components 1 through 3.
- To satisfy component 4 the response must discuss all three variables: the response variable (attendance) and the two explanatory variables from components 1 and 2.

- The strength of the response to part (d), especially communication in component 4, can be considered in holistic scoring. A well-communicated response to part (d) should be considered a positive. A poorly-communicated response to part (d), should be considered a negative.

Scoring for Question 6
Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as ½ point.

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

If a response is between two scores (for example, 2 ½ points), use a holistic approach to decide whether to score up or down, depending on the strength of the response and quality of the communication.
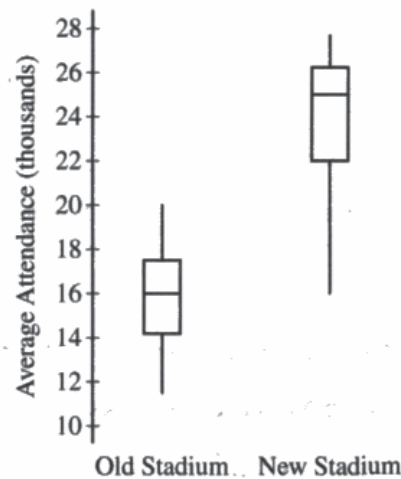
Begin your response to **QUESTION 6** on this page.

Part B

Question 6

Spend about 25 minutes on this part of the exam.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

6. Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.

Old Stadium   New Stadium

(a) Compare the distributions of average attendance between the old and new stadiums.

Shape: The distribution of avg attendance at the old stadium is roughly uniform while at the new stadium is skewed to the left.

Center: The median avg attendance at the old stadium is 16,000 attendees while much higher at 25,000 attendees at the new stadium.

Spread: The range of avg attendance at the new stadium (about 12,000) is greater than at the old stadium (about 8,000).

Outliers: There is at least one potential outlier of avg attendence at the new stadium at approximately 16,000 attendees.

Continue your response to **QUESTION 6** on this page.

The following scatterplot shows average attendance versus year.
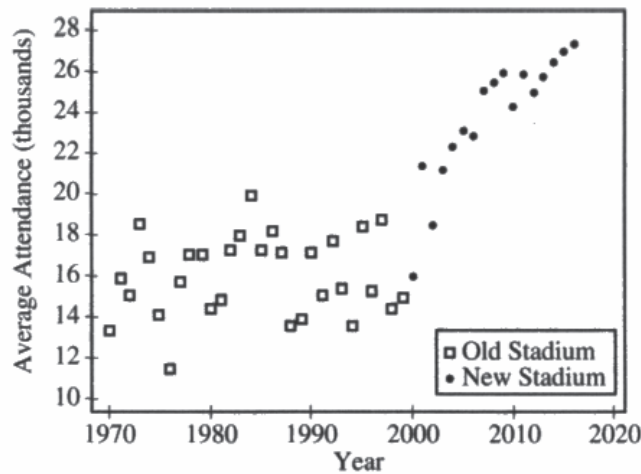
(b) Compare the trends in average attendance over time between the old and new stadium.

There has been no trend of increasing or decreasing average attendance over time at the old stadium, but average attendance has rapidly increased over time at the new stadium, although the rate of increase is slowing down. The new stadium has seen rapid growth in attendance over the years while the old stadium has seen no significant change in trends over time.

Continue your response to **QUESTION 6** on this page.
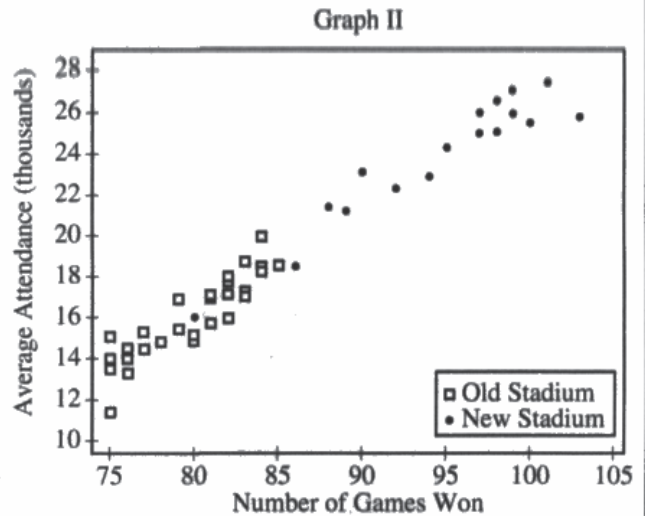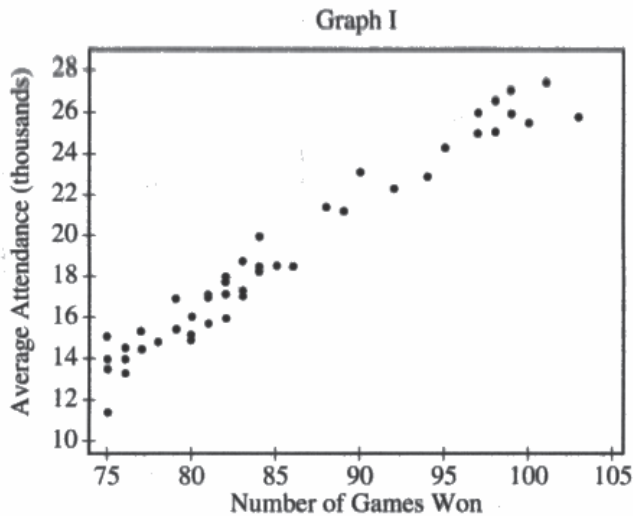
(c) Consider the following scatterplots.

Graph I

Graph II

□ Old Stadium
• New Stadium

(i) Graph I shows the average attendance versus number of games won for each year. Describe the relationship between the variables.

There is a strong positive linear association between the number of games won and average attendance for each year.

(ii) Graph II shows the same information as Graph I, but also indicates the old and new stadiums. Does Graph II suggest that the rate at which attendance changes as number of games won increases is different in the new stadium compared to the old stadium? Explain your reasoning.

Graph II does not suggest a change in rate for games in the new stadium compared to the old stadium because there is only a very minor shift in rate between the two clusters of old vs new stadium years and the same regression line could reasonably represent the whole graph.

Page 14   **GO ON TO THE NEXT PAGE.**

Continue your response to **QUESTION 6** on this page.

(d) Consider the three variables: number of games won, year, and stadium. Based on the graphs, explain how one of those variables could be a confounding variable in the relationship between average attendance and the other variables.

The number of games won could be a confounding variable in the relationship between average attendance, and year or stadium. While the graphs have shown a clear association between year and average attendance (average attendance rapidly increases in 2000 and on) and between stadium and average attendance, (far more people on average attended games at the new stadium than the old stadium), the number of games won in a year is a lurking variable for both of these relationships. In the new stadium, the baseball team won way more games per year than they did at the old stadium, but the rate at which average attendance increased with the number of games won didn't change when the stadium changed, suggesting the new stadium didn't cause the increase in average attendance. The average attendance <u>did</u> increase as the years went on, but for most of the years with high attendance (at the new stadium), the team also won more games than before. Therefore, the number of games won per year confounded the relationships between average attendance and year or stadium and explained the ^variation in average attendance.

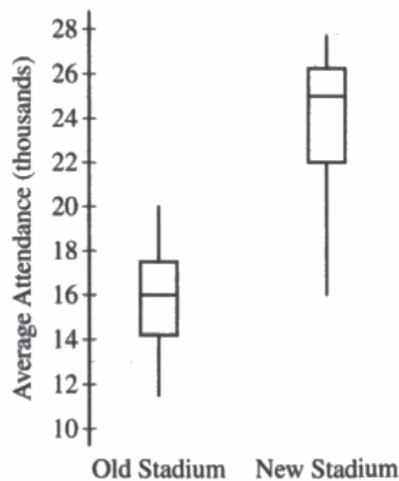Begin your response to **QUESTION 6** on this page.

Part B

Question 6

Spend about 25 minutes on this part of the exam.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

6. Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.


(a) Compare the distributions of average attendance between the old and new stadiums.

The distributions of average attendance between old and new stadiums ⊕ have a similar range. The mean (median) for the old stadium is ≈ 16 versus the new being ≈ 25. There is a drastic difference in these two numbers. Additionaly, it looks as if the boxplut for the new stadium is skewed left or downwards.

Continue your response to **QUESTION 6** on this page.

The following scatterplot shows average attendance versus year.
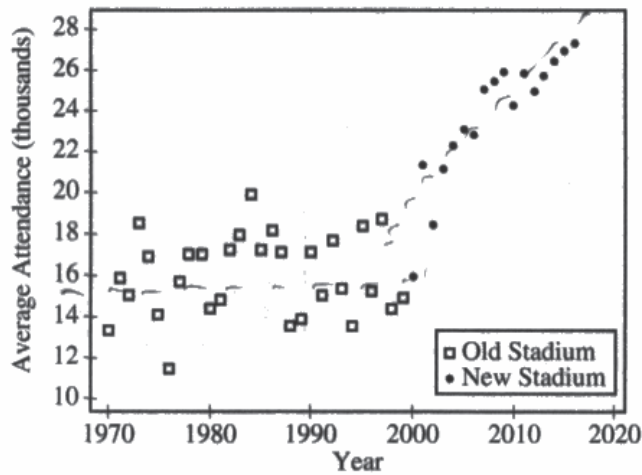
(b) Compare the trends in average attendance over time between the old and new stadium.

The average attendance of the old stadium remained fairly constant between 1970 and 2000, when attendance suddenly stopped. After 2000, the new stadiums attendance grew rapidly and continues to grow it seems, up to 2020. There is an obvious difference between the ~~gradual~~ constant level of attendance in the old stadium and positive, rapid growth of attendance in the new stadium.

Continue your response to **QUESTION 6** on this page.
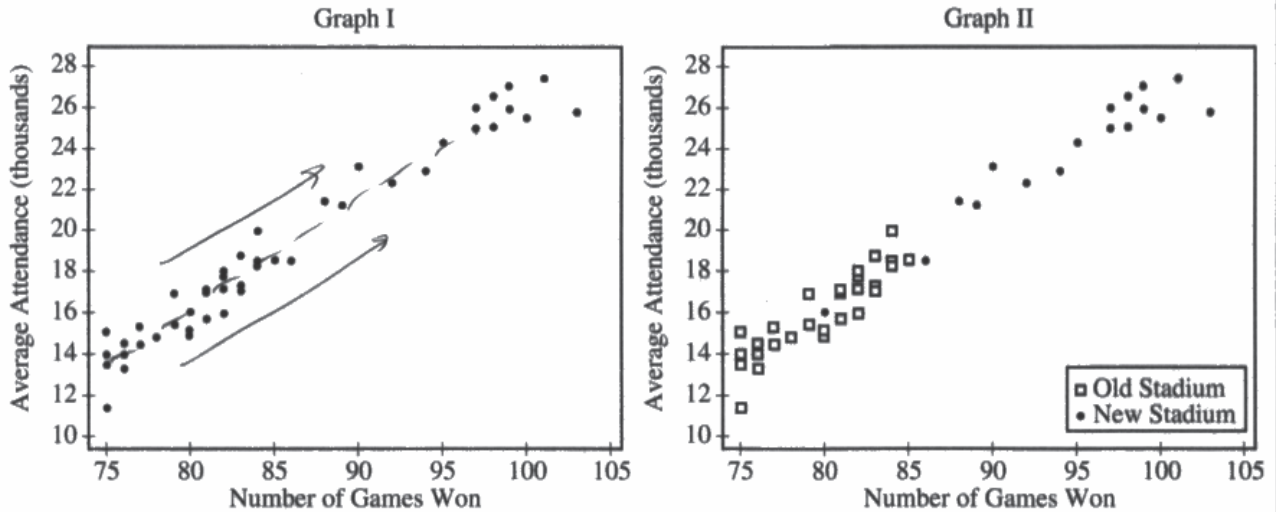
(c) Consider the following scatterplots.

Graph I — Graph II

(i) Graph I shows the average attendance versus number of games won for each year. Describe the relationship between the variables.

There is a strong, positive linear association between games won and average attendence in graph 1. As the number of games won increases, the average attendence grows as well.

(ii) Graph II shows the same information as Graph I, but also indicates the old and new stadiums. Does Graph II suggest that the rate at which attendance changes as number of games won increases is different in the new stadium compared to the old stadium? Explain your reasoning.

No, both graphs are identical in their growth and dot placement. The only difference is in graph two the dots are characterized by which stadium the game was played/won at.

Continue your response to **QUESTION 6** on this page.

(d) Consider the three variables: number of games won, year, and stadium. Based on the graphs, explain how one of those variables could be a confounding variable in the relationship between average attendance and the other variables.

The number of games won and stadium could each be a confounding variable in two different scenerios. For one, the more games won by a team, the more likely it is that fans will attend games. In the previous question there is an obvious relationship between this variable and average games won.

Additionally, a new stadium may attract more fans, which may be the reason the average attendance has increased. This is not as convincing since the attendance continues to rise even though the stadium has been new for nearly 20 years. The more likely confounding variable is games won.

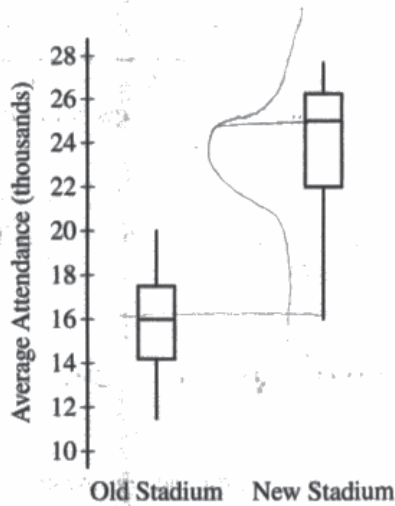Begin your response to **QUESTION 6** on this page.

Part B

Question 6

Spend about 25 minutes on this part of the exam.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

6. Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.
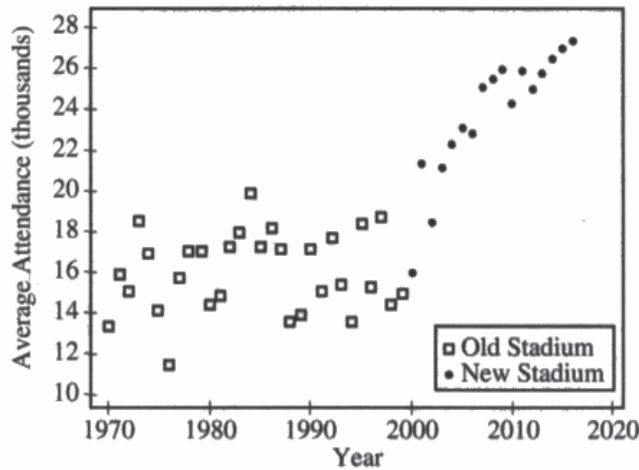

Old Stadium   New Stadium

(a) Compare the distributions of average attendance between the old and new stadiums.

Old Stadiums: - Approximately normal distribution
              - Med: 16,000
              - Range: 9 (smaller spread)

New Stadiums: - Strongly skewed left   - Range: 10
              - Med: 25,000
              - Possible outliers

Continue your response to **QUESTION 6** on this page.

The following scatterplot shows average attendance versus year.
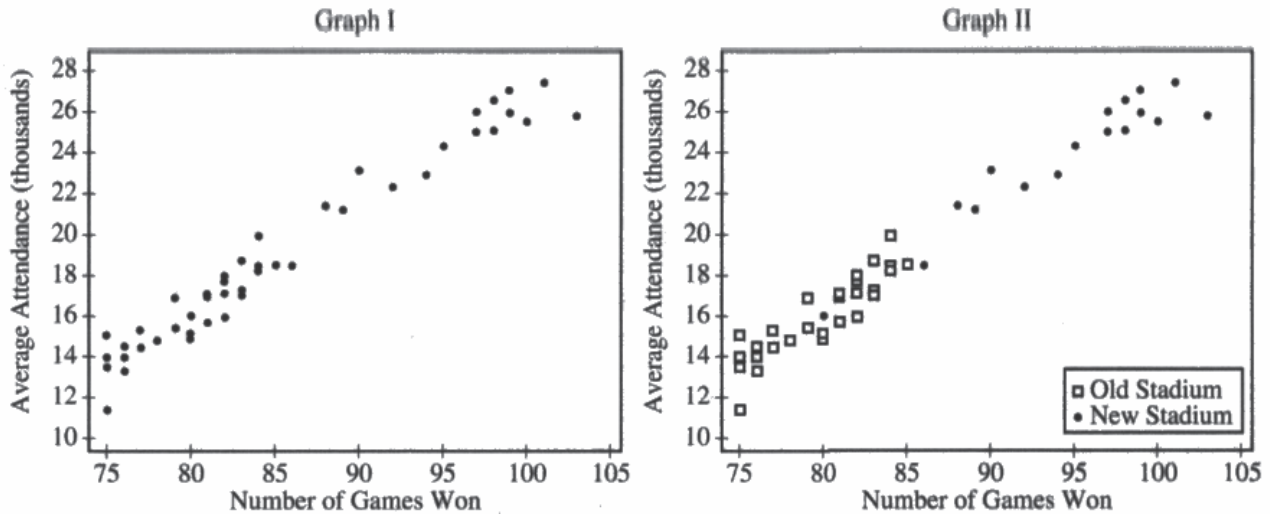

(b) Compare the trends in average attendance over time between the old and new stadium.

Old Stadium: The old studium does not have a strong correlation of attendance every year.

New Stadium: The new stadium has a stronger correlation with a positive slope of attendance.

Continue your response to **QUESTION 6** on this page.

(c) Consider the following scatterplots.

Graph I

Graph II

(i) Graph I shows the average attendance versus number of games won for each year. Describe the relationship between the variables.

The variables have a positive relationship. As the number of games won goes up, the average attendance also goes up. The correlation is strong and demonstrates a positive relationship between the 2 variables.

(ii) Graph II shows the same information as Graph I, but also indicates the old and new stadiums. Does Graph II suggest that the rate at which attendance changes as number of games won increases is different in the new stadium compared to the old stadium? Explain your reasoning.

The graph indicates that the average attendance in the new stadium increases much faster than in the old stadium. Both stadiums continue to have a positive correlation, One is just a faster incline

Page 14 **GO ON TO THE NEXT PAGE.**

Continue your response to **QUESTION 6** on this page.

(d) Consider the three variables: number of games won, year, and stadium. Based on the graphs, explain how one of those variables could be a confounding variable in the relationship between average attendance and the other variables.

Stadium may be a cofounding variable in the relationship between attendance. Attendance may fluctuate based on what stadium games are held out. Individuals may not make it to a game if the stadium is far away. This could decrease attendance in the stadiums. Some people may not like a certain stadium and choose not to go also.

Question 6

Note: Student samples are quoted verbatim and may contain spelling and grammatical errors.

Overview

The primary goals of this question were to assess a student's ability to (1) compare two distributions using information provided by side-by-side boxplots; (2) use information in a scatterplot to compare trends across time for two sets of data; (3) use information in a scatterplot to describe the relationship between two variables; (4) compare the rates of change for one variable as another variable changes for two data sets displayed in a scatterplot; (5) use information presented in several graphs to explain how a variable could be a confounding variable with respect to the relationship between two other variables.

This question assesses skills in multiple categories, including 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.C) Describe an appropriate method for gathering and representing data, (2.A) Describe data presented numerically or graphically, (2.D) Compare distributions or relative positions of points within a distribution, (4.A) Make an appropriate claim or draw an appropriate conclusion, and (4.B) Interpret statistical calculations and findings to assign meaning or assess a claim.

This question covers content from multiple units, including Unit 1: Exploring One-Variable Data, Unit 2: Exploring Two-Variable Data, and Unit 3: Collecting Data of the course framework in the AP Statistics Course and Exam Description. Refer to topics 1.9, 2.4, 2.8, and 3.5, and learning objectives DAT-1.A, DAT-1.H, UNC-1.N, and VAR-3.A.

Sample: 6A
Score: 4

The response earned the following: part (a) – E; part (b) – E; part (c) – E; part (d) – E.

In part (a) the response indicates that the median average attendance is greater in the new stadium using a comparison phrase ("much higher"), satisfying component 1. The response indicates that the range of average attendance is greater in the new stadium using a comparison phrase ("is greater than"), satisfying component 2. The response also includes numerical estimates for the medians and ranges, although these are not required to satisfy components 1 or 2. The response provides sufficient context by including the name of the response variable ("avg attendance"), satisfying component 3. The response also discusses outliers and shape, which are considered extraneous. Because the response includes all three components, part (a) was scored as essentially correct (E).

In part (b) the response states that "there has been no trend of increasing or decreasing average attendance over time at the old stadium," satisfying component 2. The response states that "average attendance has rapidly increased over time at the new stadium," satisfying component 1. The response provides sufficient context by including the two groups (old, new), the explanatory variable (time), and the response variable (average attendance), satisfying component 3. Because the response includes all three components, part (b) was scored as essentially correct (E).

In part (c-i) the response describes the direction as "positive," satisfying component 1. The response describes the form as "linear," satisfying component 2. The response describes the strength as "strong," satisfying component 3. In part (c-ii) the response indicates that the rates are about the same ("does not suggest a change in rate"), satisfying component 4. Note that the response also states there is "only a very minor shift in rate," which also satisfies component 4. The response provides an explanation by stating "the same regression line could

Question 6 (continued)

reasonably represent the whole graph," satisfying component 5. Because the response includes at least four of the five components, part (c) was scored as essentially correct (E).

In part (d) the response states that there is an association "between stadium and average attendance," satisfying component 1. The response states that there is an association between wins and attendance by referencing the "rate at which average attendance increased with the number of games won," satisfying component 2. The response states that there is an association between stadium (from component 1) and wins (from component 2) by saying "in the new stadium, the baseball team won way more games per year than they did in the old stadium," satisfying component 3. The response explains the idea of confounding by describing that the potential confounding variable (wins) could be the cause of the association between stadium and attendance ("suggesting the new stadium didn't cause the increase in average attendance" and "number of games won … explained the variation in average attendance"). Component 4 is satisfied. Note that the response also satisfies components 1–4 using years and wins as the explanatory variable. Because the response includes all four components, part d was scored essentially correct (E).

Sample: 6B
Score: 2

The response earned the following: part (a) – I; part (b) – E; part (c) – P part (d) – P.

In part (a) the response provides estimated values for the centers of both distributions, states that there is a "drastic difference" between them, but does not specify which distribution has a greater center. Component 1 is not satisfied. The response incorrectly states that the distributions have a similar range. Component 2 is not satisfied. Note that a response that indicates that the variability is similar can satisfy component 2 if the response specifically refers to the IQRs. The response provides sufficient context by including the name of the response variable ("average attendance"), satisfying component 3. The correct statement about the shape of the distribution in the new stadium is considered a positive for holistic scoring. Because the response includes only component 3, part (a) was scored as incorrect (I).

In part (b) the response describes the trend at the old stadium as "fairly constant" and later as "constant." Either statement satisfies component 2. The response states that the "attendance grew rapidly and continues to grow it seems, up to 2020" and later describes the trend in the new stadium as "positive." Either statement satisfies component 1. The response provides sufficient context by including the two groups (old, new), the explanatory variable (e.g., "between 1970 and 2000"), and the response variable (attendance). Component 3 is satisfied. Because the response includes all three components, part (b) was scored as essentially correct (E).

In part (c-i) the response describes the direction as "positive," satisfying component 1. The response describes the form as "linear," satisfying component 2. The response describes the strength as "strong," satisfying component 3. In part (c-ii) the response compares Graph I and Graph II and does not answer the question by comparing the associations for the old and new stadiums in Graph II. Components 4 and 5 are not satisfied. Because the response includes three of the five components, part (c) was scored as partially correct (P).

In part (d) the response states, "the more games won by a team, the more likely it is that fans will attend games," satisfying component 1. The response states that "a new stadium may attract more fans," satisfying component 2. The response does not describe the association between wins (from component 1) and stadium (from component 2), so component 3 is not satisfied. The response suggests that either more wins or a new stadium could be the cause and states that games won is the "more likely confounding variable." However, the response doesn't explain the idea of confounding—that it is impossible to know which of the variables is the cause. Instead, the response seems to focus on which variable would be a better predictor of attendance. Component 4 is not satisfied. Because two of the four components are satisfied, and the response makes a clear reference to a graph from earlier parts ("In the previous question"), part (d) was scored as partially correct (P).

Question 6 (continued)

Sample: 6C
Score: 1

The response earned the following: part (a) – P; part (b) – P; part (c) – I; part (d) - I.

In part (a) the response provides estimated values for the median of both distributions but does not compare them using an explicit comparison phrase. Component 1 is not satisfied. The response also provides estimated values for the ranges of both distributions and compares them using a comparison phrase ("smaller spread"), satisfying component 2. The response does not provide sufficient context by including the name of the response variable ("attendance") or the units of the response variable ("people"). Component 3 is not satisfied. The response also discusses possible outliers and shape, which are considered extraneous. However, the statement that the distribution in the old stadium is "approximately normal" is not supported by the boxplot and is considered a negative for holistic scoring. Because the response includes only component 2, part (a) was scored as partially correct (P).

In part (b) the response describes the trend of the new stadium as "positive," satisfying component 1. The response does not describe the direction of the trend at the old stadium. Note that descriptions of strength ("not a strong correlation") do not address direction. Component 2 is not satisfied. The response provides sufficient context by including the two groups (old, new), the explanatory variable (years), and the response variable (attendance). Component 3 is satisfied. Because the response includes two of the three components, part (b) was scored as partially correct (P).

In part (c-i) the response describes the direction as "positive," satisfying component 1. The response does not address form, so component 2 is not satisfied. The response describes the strength as "strong," satisfying component 3. In part (c-ii) the response incorrectly states that the rate is greater at the new stadium. Components 4 and 5 are not satisfied. Because the response includes two of the five components, part (c) was scored as incorrect (I).

In part (d) the response states, "Attendance may fluctuate based on what stadium games" are played, satisfying component 1. The response does not identify an association between attendance and a different explanatory variable, so component 2 is not satisfied. The response does not state that stadium (from component 1) is related to a different explanatory variable, so component 3 is not satisfied. Other than mentioning the existence of a confounding variable, the response does not explain the idea of confounding. Component 4 is not satisfied. Because the response includes one of the four components, part (d) was scored incorrect (I).