Begin your response to **QUESTION 1** on this page.

STATISTICS

SECTION II

Total Time—1 hour and 30 minutes

6 Questions

Part A

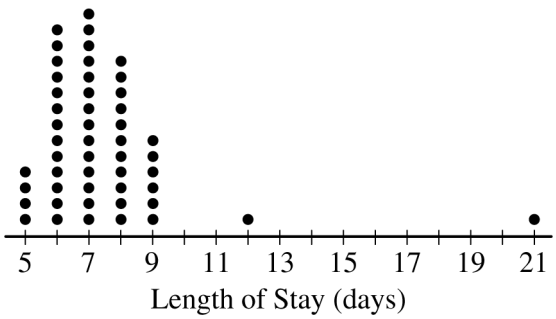
Questions 1-5

Spend about 1 hour and 5 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.

1. The length of stay in a hospital after receiving a particular treatment is of interest to the patient, the hospital, and insurance providers. Of particular interest are unusually short or long lengths of stay. A random sample of 50 patients who received the treatment was selected, and the length of stay, in number of days, was recorded for each patient. The results are summarized in the following table and are shown in the dotplot.

Length of stay (days)	5	6	7	8	9	12	21
Number of patients	4	13	14	11	6	1	1



(a) Determine the five-number summary of the distribution of length of stay.

GO ON TO THE NEXT PAGE.

Use a pencil or pen with black or dark blue ink only. Do NOT write your name. Do NOT write outside the box.

Continue your response to **QUESTION 1** on this page.

(b) Consider two rules for identifying outliers, method A and method B. Let method A represent the $1.5 \times \text{IQR}$ rule, and let method B represent the 2 standard deviations rule.

(i) Using method A, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

(ii) The mean length of stay for the sample is 7.42 days with a standard deviation of 2.37 days. Using method B, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

(c) Explain why method A might identify more data points as potential outliers than method B for a distribution that is strongly skewed to the right.

GO ON TO THE NEXT PAGE.

Begin your response to **QUESTION 2** on this page.

2. Researchers will conduct a year-long investigation of walking and cholesterol levels in adults. They will select a random sample of 100 adults from the target population to participate as subjects in the study.

   (a) One aspect of the study is to record the number of miles each subject walks per day. The researchers are deciding whether to have subjects wear an activity tracker to record the data or to have subjects keep a daily journal of the miles they walk each day. Describe what bias could be introduced by keeping the daily journal instead of wearing the activity tracker.

GO ON TO THE NEXT PAGE.

Use a pencil or pen with black or dark blue ink only. Do NOT write your name. Do NOT write outside the box.

Question 1: Focus on Exploring Data — 4 points

General Scoring Notes

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

	Model Solution	Scoring
(a)	The five-number summary of the distribution of length of stay is: Minimum = 5 days Lower quartile (Q_1) = 6 days Median = 7 days Upper quartile (Q_3) = 8 days Maximum = 21 days	**Essentially correct (E)** if the response provides correct values for ALL FIVE of the summary statistics with labels (minimum, lower quartile, median, upper quartile, and maximum). **Partially correct (P)** if the response provides correct values for only THREE or FOUR of the summary statistics with labels. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- Any discussion of the mean, IQR, or the standard deviation of length of stay should be ignored in scoring.
- Inclusion or omission of units of measurement (days) has no bearing on scoring.
- If the response includes exactly 5 unlabeled numbers expressed together as a vertical or horizontal list, interpret the numbers as being labeled as the minimum, lower quartile, median, upper quartile, and maximum, respectively.
- A response that includes only five numbers that are correct values for the five-number summary without providing a complete set of labels or not putting them in an ordered list may be scored P.

Model Solution	Scoring
(b) **(i)** The patients who stayed for 12 days and 21 days are considered outliers using method A. An outlier using method A is a value greater than $1.5 \times \text{IQR}$ above the third quartile (Q_3) or more than $1.5 \times \text{IQR}$ below the first quartile (Q_1). Because $Q_1 - 1.5 \times \text{IQR} = 6 - 1.5(8 - 6) = 3,$ then any values below 3 are considered outliers. There are no such values. Because $Q_3 + 1.5 \times \text{IQR} = 8 + 1.5(8 - 6) = 11,$ then any values above 11 are considered outliers. **(ii)** The patient who stayed for 21 days is the only outlier using method B. An outlier using method B is a value located 2 or more standard deviations above, or below, the mean. Because $\text{Mean} \pm 2 \times \text{SD} = 7.42 \pm 2(2.37),$ then any value that is outside of the interval $(2.68, 12.16)$ is considered an outlier.	**Essentially correct (E)** if the response satisfies the following four components: 1. Correctly identifies the two outliers in part (b-i) as the patients who stayed for 12 days and 21 days 2. Provides a justification for part (b-i) by calculating the lower and upper outlier criteria for the $1.5 \times \text{IQR}$ rule (e.g., "using method A, an outlier is any value below 3 days or above 11 days") 3. Correctly identifies the one outlier in part (b-ii) as the patient who stayed for 21 days 4. Provides a justification for part (b-ii) by calculating the lower and upper outlier criteria for the 2 standard deviations rule (e.g., "using method B, an outlier is any value below 2.68 days or above 12.16 days") **Partially correct (P)** if the response satisfies only two or three of the four components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response for part (b-ii) that manually computes the standard deviation as 2.374 and then uses it to construct an interval of $(2.672, 12.168)$ satisfies component 4.
- Component 1 and component 2 are satisfied if the response to part (b-i) uses correct calculations with incorrect values of summary statistics reported in the response to part (a).

Model Solution	Scoring
(c) Quartiles and the IQR are less sensitive to extreme values in strongly skewed distributions than the mean and standard deviation. Relative to the quartiles, the mean is pulled more toward the extreme values in the longer tail of a strongly skewed distribution. For a distribution that is strongly skewed to the right, the sample mean will be pulled more toward the extreme values in the longer right tail of the distribution than the sample median, and the ratio of the standard deviation to the IQR will tend to be larger than that for more nearly symmetric distributions. As a result, this pulls the value of the outlier criterion for method B, $\text{Mean} + 2 \times \text{SD}$, more toward the extreme values in the right tail of the distribution than the outlier criterion for method A, $Q_3 + 1.5 \times \text{IQR}$. This decreases the ability of method B to identify outliers relative to method A, which means that method A may identify more outliers than method B for a distribution that is strongly skewed to the right.	**Essentially correct (E)** if the response satisfies the following two components: 1. Indicates that the mean is pulled more toward the extreme values in the longer right tail for a strongly right-skewed distribution than the quartiles (or median) OR indicates that the ratio of the standard deviation to the IQR tends to be larger for strongly skewed distributions than for more nearly symmetric distributions 2. Provides an explanation that links effects of skewness on an increased ability of method A to detect outliers relative to method B (e.g., "the larger shift in the mean relative to the shift in the median (or quartiles) has a greater effect on decreasing the ability of method B to detect outliers compared to method A" OR "the larger increase in the standard deviation, relative to the IQR, results in a greater increase in the range of non-outlier values for method B compared to method A") **Partially correct (P)** if the response satisfies only one of the two components. **Incorrect (I)** if the response does not meet the criteria for E or P.

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

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

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

Additional Notes:

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

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

Additional Notes:

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

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

Additional Notes:

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