Past-paper walk-through

Graphical Representations of Summary Statistics ▪ 1.8

eXercise

Questions in this set

- 2026 Q1
- 2023 Q1
- 2021 Q6
- 2019 Q1
- 2015 Q1

Reading the mark scheme

- **B1** a mark that stands on its own – the point is either made or not.
- **M1** a *method* mark: the right approach, even if the number is wrong.
- **A1** an *answer* mark, and it usually needs its M mark first.
- **C1** a working mark on the way to the answer.
- // separates answers the examiner accepts equally.
- **ecf** = error carried forward: a later part is marked on your own earlier answer.
- **owtte** = "or words to that effect" – the wording need not match.

Question 1

2026 ▪ Q1

A goat farmer raises two breeds of goats, Breed H and Breed J. He wants to compare the weights of the two breeds and takes independent random samples of 14 goats from each breed. The weight, in pounds, of each goat is measured. The weights of the 14 Breed H goats are recorded.

Weights of Breed H Goats (pounds)

48, 48, 55, 56, 56, 57, 62, 66, 72, 72, 72, 73, 80, 80

(a) Use the data in the list to determine the five-number summary for the sample of Breed H goats.

Question 1

2026 ▪ Q1

A goat farmer raises two breeds of goats, Breed H and Breed J. He wants to compare the weights of the two breeds and takes independent random samples of 14 goats from each breed. The weight, in pounds, of each goat is measured. The weights of the 14 Breed H goats are recorded.

Weights of Breed H Goats (pounds)

48, 48, 55, 56, 56, 57, 62, 66, 72, 72, 72, 73, 80, 80

(b) The distribution of weight for Breed J goats is displayed in the boxplot.
[Boxplot titled “Weights of Breed J Goats (pounds)”; horizontal axis “Weight (pounds)” from 50 to 90, gridlines at 50, 55, 60, 65, 70, 75, 80, 85, 90; left whisker starts at about 50, box left edge (Q1) at about 57, median line at about 64, box right edge (Q3) at about 80, right whisker ends at about 86.]

Question 1

Use the five-number summary from part A and the boxplot of Breed J to compare the center and variability for the distribution of weight for the Breed H goats and the distribution of weight for the Breed J goats, in context.

Question 1

2026 ▪ Q1

A goat farmer raises two breeds of goats, Breed H and Breed J. He wants to compare the weights of the two breeds and takes independent random samples of 14 goats from each breed. The weight, in pounds, of each goat is measured. The weights of the 14 Breed H goats are recorded.

Weights of Breed H Goats (pounds)

48, 48, 55, 56, 56, 57, 62, 66, 72, 72, 72, 73, 80, 80

(c)(i) A stem-and-leaf plot is another way to display the weights of goats. The distribution of weight for the Breed H goats is displayed in the given stem-and-leaf plot.

Weights of Breed H Goats (pounds)

Question 1

```
4 | 8 8
5 | 5 6 6 7
6 | 2 6
7 | 2 2 2 3
8 | 0 0
```

Question 1

Key: $4|8 = 48$ pounds

Consider the five-number summary from part A and the stem-and-leaf plot.

If a boxplot were created from the five-number summary found in part A, what characteristic of the shape of the distribution of weight for the Breed H goats would be apparent from the stem-and-leaf plot but not from the boxplot?

(c)(ii) Explain why a boxplot would not display the characteristic of the shape of the distribution identified in part C (i).

Question 1

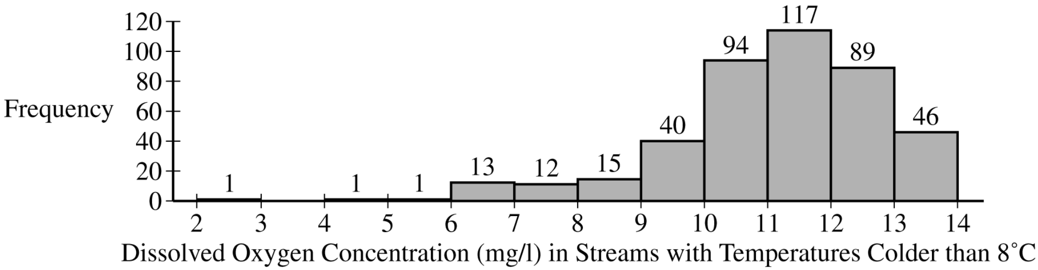
2023 ▪ Q1

As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than $8°$C and from many streams with temperatures warmer than $8°$C. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l). [Histogram titled "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than $8°$C"; x-axis "Dissolved Oxygen Concentration (mg/l)" from 2 to 14 (bin width 1); y-axis "Frequency" from 0 to 120. Bar heights (frequency) by bin: [2,3): 1, [3,4): 0, [4,5): 1, [5,6): 1, [6,7): 13, [7,8): 12, [8,9): 15, [9,10): 40, [10,11): 94, [11,12): 117, [12,13): 89, [13,14): 46.]
(a) The researchers constructed the histogram shown for the dissolved oxygen concentration in streams from the sample with water temperatures <u>colder</u>

Question 1

than 8°C. Based on the histogram, describe the distribution of dissolved oxygen concentration in streams with water temperatures <u>colder</u> than 8°C.

Question 1



Frequency
120
100
80
60
40
20
0
1
1
1
13
12
15
40
94
117
89
46
2
3
4
5
6
7
8
9
10
11
12
13
14
Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than 8°C

Question 1

2023 ▪ Q1

As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than $8°$C and from many streams with temperatures warmer than $8°$C. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l).

[Histogram titled "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than $8°$C"; x-axis "Dissolved Oxygen Concentration (mg/l)" from 2 to 14 (bin width 1); y-axis "Frequency" from 0 to 120. Bar heights (frequency) by bin: [2,3): 1, [3,4): 0, [4,5): 1, [5,6): 1, [6,7): 13, [7,8): 12, [8,9): 15, [9,10): 40, [10,11): 94, [11,12): 117, [12,13): 89, [13,14): 46.]

(b) The researchers computed the summary statistics shown in the table for the dissolved oxygen concentration in streams from the sample with water temperatures <u>warmer</u> than $8°$C. Use the summary statistics to construct a box plot for

Question 1

the dissolved oxygen concentration in streams with water temperatures <u>warmer</u> than 8°C. Do not indicate outliers.

Min	Q1	Median	Q3	Max	Mean	Std. Dev.
2.10	4.39	5.43	6.12	13.45	5.54	1.64

[Blank grid provided for the box plot; x-axis "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Warmer than 8°C" numbered 2 to 14.]

Question 1

2023 ▪ Q1

As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than 8°C and from many streams with temperatures warmer than 8°C. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l).

[Histogram titled "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than 8°C"; x-axis "Dissolved Oxygen Concentration (mg/l)" from 2 to 14 (bin width 1); y-axis "Frequency" from 0 to 120. Bar heights (frequency) by bin: [2,3): 1, [3,4): 0, [4,5): 1, [5,6): 1, [6,7): 13, [7,8): 12, [8,9): 15, [9,10): 40, [10,11): 94, [11,12): 117, [12,13): 89, [13,14): 46.]

(c) The researchers believe that streams with higher dissolved oxygen concentration are generally healthier for wildlife. Which streams are generally healthier for wildlife, those with water temperature <u>colder</u> than 8°C or those with water

Question 1

temperature <u>warmer</u> than 8°C? Using characteristics of the distribution of dissolved oxygen concentration for temperatures <u>colder</u> than 8°C and characteristics of the distribution of dissolved oxygen concentration for temperatures <u>warmer</u> than 8°C, justify your answer.

Solution 1

(a) Mark scheme

- The distribution of dissolved oxygen concentration (mg/l) in Alaskan streams colder than 8°C is unimodal and skewed left, with median between 11 and 12 mg/l. Q1 falls in the 10-11 mg/l bin and Q3 in the 12-13 mg/l bin, so IQR $\approx$ 2 mg/l (acceptable IQR range 1.5-2.5; acceptable range-of-data 10-12 mg/l; acceptable SD 1.7-1.8 mg/l). There are no apparent high outliers, but the 2-3, 4-5, and 5-6 mg/l bins are all more than 1.5$\times$IQR below Q1, so there are several potential low outliers (a comment on a gap between 3 and 4 mg/l is also acceptable for this component). Full credit requires component 1 (context: dissolved oxygen concentration/amount of oxygen/mg per liter) AND at least 3 of: (2) shape - skewed left (describing it as normal/approximately normal fails this

Solution 1

component), (3) center - a numeric value for median/mean between 11 and 12 mg/l (definitive language like 'the median is 11.5' does NOT satisfy this component; it must be phrased approximately, e.g. 'approximately 11.5'), (4) spread - any correctly used measure of variation (IQR, range, or SD) within the bounds above, (5) unusual features - potential outliers or a gap (stating a *different* unusual feature, or definitive language like 'there is an outlier', does not satisfy this component). Partial credit: component 1 + only 2 of components 2-5, OR at least 3 of components 2-5 without component 1.

Solution 1

(b) Mark scheme

- Construct a boxplot (not any other graph type) for dissolved oxygen concentration in streams warmer than 8°C using: Min = 2.10, Q1 = 4.39, Median = 5.43, Q3 = 6.12, Max $\approx$ 13 mg/l (do not flag outliers - draw the whiskers all the way to the min and max). The box spans from Q1 to Q3 with the median line inside; whiskers extend from the box ends to the minimum and maximum values. Full credit (4 or 5 of 5 components): (1) box begins at Q1 (between the 4 and 4.5 tick marks), (2) box ends at Q3 (between 6 and 6.5), (3) median located within the box (between 5 and 5.5), (4) line extended to the minimum value (between 2 and 2.5), (5) line extended to the maximum value (between 13 and 13.5). Partial

Solution 1

credit: only 3 of the 5 components. A response that draws a graph other than a boxplot (e.g. a histogram) should be scored incorrect regardless of accuracy.

Solution 1

(c) Mark scheme

- Streams with water temperature colder than 8°C are generally healthier for wildlife (assuming the researchers' belief that higher dissolved oxygen means healthier streams). The colder-stream distribution has a higher center - its median (between 11 and 12 mg/l) is larger than the warmer streams' median (5.43 mg/l). The two distributions differ in shape: colder streams are skewed left while warmer streams are skewed right. Both distributions have similar spread - their IQRs are comparable ($\approx$2 mg/l for colder streams vs. $\approx$1.73 mg/l for warmer streams). Full credit requires component 1 (states colder streams are healthier for wildlife) AND at least 2 of: (2) directly compares the centers of the two distributions, (3) indicates warmer streams are skewed right and colder streams are skewed left, (4)

Solution 1

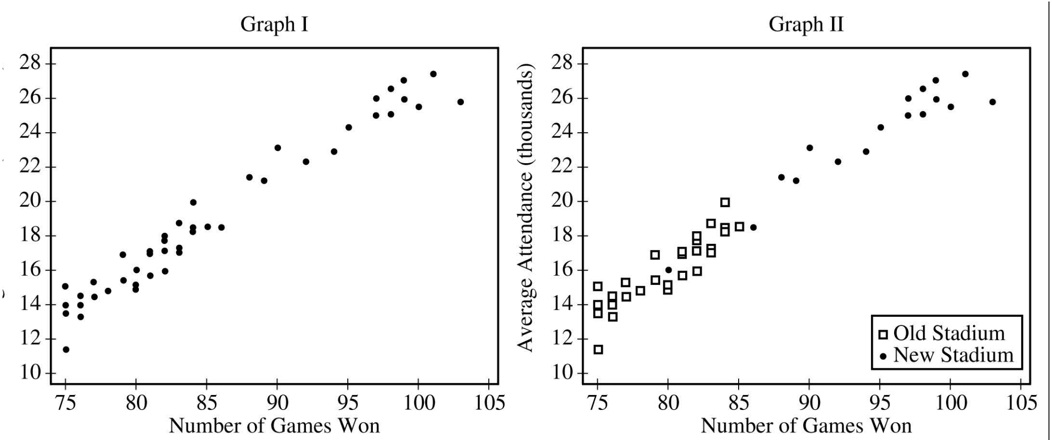
directly compares the spreads (comparing at least one of range, IQR, or SD between the two distributions). Partial credit: component 1 + only 1 of components 2-4. A response based on an incorrect part (a) or (b) may still earn credit here if it is consistent with those earlier answers. Comparing outliers between the two distributions is not necessary for this part.

Question 6

2021 ▪ Q6

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.
[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]
(a) Compare the distributions of average attendance between the old and new stadiums.

Question 6

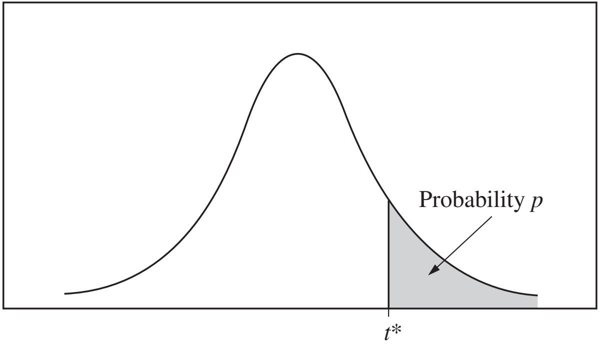


Graph I
Graph II
28
26
24
22
20
18
16
14
12
10
75
80
85
90
95
100
105
Number of Games Won
Average Attendance (thousands)
Old Stadium
New Stadium

Question 6

3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.



Question 6

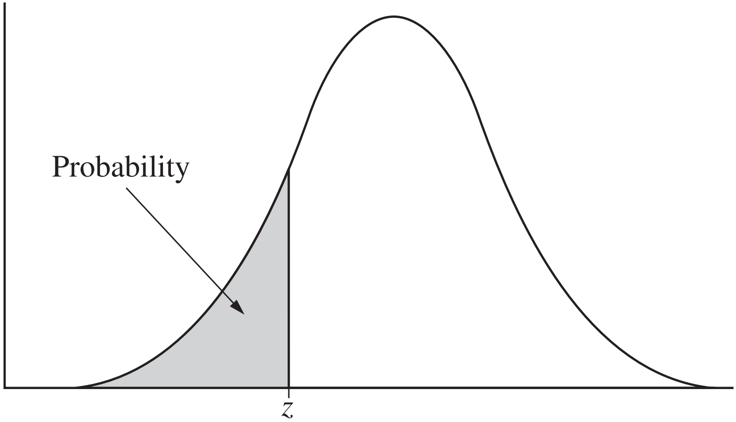


Table entry for z is the probability lying below z.

Question 6

2021 ▪ Q6

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.

[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]

(b) The following scatterplot shows average attendance versus year.

[Scatterplot titled with y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2, x-axis "Year" from 1970 to 2020 in increments of 10 (gridlines at 1970, 1980, 1990, 2000, 2010, 2020). Legend distinguishes open squares = Old

Question 6

Stadium and filled circles = New Stadium. Old Stadium square points span roughly 1970-2000, scattered between about 11 and 20 with no strong trend (values fluctuate between roughly 13-20 throughout). New Stadium circle points span roughly 2000-2018, showing a clear upward trend from about 16-21 in the early 2000s rising steadily to about 26-27.5 by the late 2010s.]

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

Question 6

2021 ▪ Q6

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.

[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]

(c) Consider the following scatterplots.

[Graph I: scatterplot titled "Graph I", y-axis "Average Attendance (thousands)" (label shown on Graph II) from 10 to 28 in increments of 2, x-axis "Number of Games Won" from 75 to 105 in increments of 5. All points plotted as plain dots (stadium not

Question 6

distinguished) showing a positive, fairly linear association: points cluster from about (75,11-15) at the low end rising to about (100-103,26-28) at the high end.]
[Graph II: scatterplot titled “Graph II”, same axes as Graph I (“Average Attendance (thousands)” from 10 to 28, “Number of Games Won” from 75 to 105). Same data as Graph I but now marked with open squares = Old Stadium and filled circles = New Stadium (legend shown). Old Stadium (square) points cluster in the lower-left region, roughly games won 75-86 and attendance 11-20. New Stadium (circle) points cluster in the upper-right region, roughly games won 75-103 and attendance 16-27.5, appearing to lie along a steeper positive trend than the Old Stadium points.]
(c)(i) Graph I shows the average attendance versus number of games won for each year. Describe the relationship between the variables.
(c)(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

Question 6

number of games won increases is different in the new stadium compared to the old stadium? Explain your reasoning.

Question 6

2021 ▪ Q6

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.

[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from about 22 to 26.5, median line at about 25.]

(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.

Solution 6

(a) Mark scheme

- The boxplots show the new stadium tended to have higher average per-game attendance (median $\approx 25,000) thantheoldstadium(median \approx$
$16,000). TheIQRs(spread/variability)aresimilarbetweenthetwostadiums, thoughtheoverallrange$
$gameattendance. Creditrequires(1)acorrectcentercomparison(new >$
$old, withstadiumsidentifiedandanexplicitcomparisonphrase), (2)acorrectvariabilitycomparison(s$

Solution 6

(b) Mark scheme

- In the new stadium (years 2000-2016), average per-game attendance shows an increasing (positive) linear trend, rising from about 16,000 in 2000 to about 27,000 in 2016. In the old stadium (years 1970-1999), there is no clear increasing or decreasing trend – attendance fluctuates around an average of about 16,000 attendees per game with no obvious pattern over time. Credit requires (1) describing the new-stadium trend as increasing/positive, (2) describing the old-stadium trend as flat/no association (or positive but much less steep than the new stadium), AND (3) context identifying both stadiums, the explanatory variable (time/year), and the response variable (average attendance) or its units.

Solution 6

(c)

Mark scheme

- This leaf is the shared stem introducing Graph I (a plain scatterplot of average attendance vs. number of games won, stadium not distinguished) and Graph II (the same data, but marked by stadium: old vs. new) that parts (c)(i) and (c)(ii) below are based on. It poses no separate question of its own, so there is no independent scorable answer here – the graded reasoning about these two graphs is in leaves 6ci and 6cii.

Solution 6

(c)(i) Mark scheme

- Graph I shows a strong, positive, linear (or nearly linear) relationship between average per-game attendance and number of games won across the team's history: attendance increases roughly linearly with wins, by about 500 additional attendees per game for each additional game won, and the scatter about this linear trend is relatively small and fairly consistent (homogeneous) across the range of wins. Credit (part of a combined 5-component part-(c) score with 6cii) for describing direction (positive), form (linear), and strength (strong/moderately strong) of the relationship, ideally also noting the roughly constant variation about the trend.

Solution 6

(c)(ii) Mark scheme

- No – Graph II suggests the rate at which average attendance increases with games won is about the same for the two stadiums (or slightly larger for the old stadium): a line drawn through the old-stadium points would have roughly the same slope as (or a slightly steeper slope than) a line drawn through the new-stadium points, so there is no evidence the new stadium has a meaningfully different attendance-vs-wins rate than the old stadium. Credit (part of the same combined part-(c) score as 6ci) for correctly comparing the two implied slopes as about equal (or old $\geq new)withanexplanationreferencingfittingalinethrougheachstadium'spoints.$

Solution 6

(d) Mark scheme

- Number of games won could be a confounding variable for the effect of the new stadium on attendance. There is an association between attendance and stadium (part (a): new stadium had higher average attendance) and an association between attendance and games won (part (c): strong positive relationship); additionally, the team tended to win more games while playing in the new stadium than in the old stadium, so games won is also associated with stadium. Because these three variables are entangled, it is impossible to tell whether the higher attendance in the new stadium was actually caused by the stadium itself or simply by the team winning more games during that period – games won confounds the relationship between stadium and attendance. Credit requires (1)-(2) stating an

Solution 6

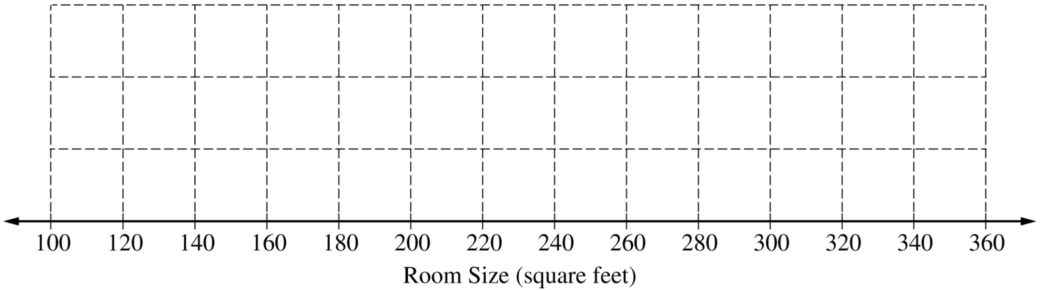
association between attendance and each of two explanatory variables (stadium, year, or wins), (3) stating an association between those same two explanatory variables, AND (4) explaining the confounding logic (the identified variable could be the true cause, or it's impossible to tell which variable is responsible).

Question 1

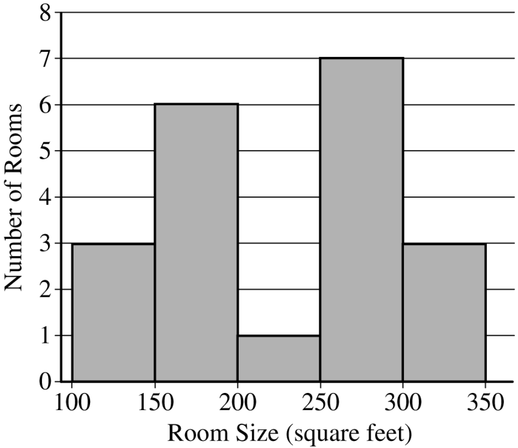
2019 ▪ Q1

The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram.
[Histogram titled “Room Size (square feet)”; y-axis “Number of Rooms” from 0 to 8; x-axis “Room Size (square feet)” with tick marks at 100, 150, 200, 250, 300, 350. Bars: 100-150 has height 3; 150-200 has height 6; 200-250 has height 1; 250-300 has height 7; 300-350 has height 3.]
(a) Based on the histogram, write a few sentences describing the distribution of room size in the residence hall.

Question 1


100
120
140
160
180
200
220
240
260
280
300
320
340
360
Room Size (square feet)

Question 1


Number of Rooms
8
7
6
5
4
3
2
1
0
100
150
200
250
300
350
Room Size (square feet)

Question 1

Mean	Standard Deviation	Min	Q1	Median	Q3	Max
231.4	68.12	134	174	253.5	292	315

Question 1

2019 ▪ Q1

The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram.
[Histogram titled "Room Size (square feet)"; y-axis "Number of Rooms" from 0 to 8; x-axis "Room Size (square feet)" with tick marks at 100, 150, 200, 250, 300, 350. Bars: 100-150 has height 3; 150-200 has height 6; 200-250 has height 1; 250-300 has height 7; 300-350 has height 3.]

(b) Summary statistics for the sizes are given in the following table.

Mean	Standard Deviation	Min	Q1	Median	Q3	Max
231.4	68.12	134	174	253.5	292	315

Question 1

Determine whether there are potential outliers in the data. Then use the following grid to sketch a boxplot of room size.
[Number line grid titled "Room Size (square feet)" with tick marks from 100 to 360 in increments of 20 (100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300, 320, 340, 360), for sketching a boxplot.]

Question 1

2019 ▪ Q1

The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram.

[Histogram titled “Room Size (square feet)”; y-axis “Number of Rooms” from 0 to 8; x-axis “Room Size (square feet)” with tick marks at 100, 150, 200, 250, 300, 350. Bars: 100-150 has height 3; 150-200 has height 6; 200-250 has height 1; 250-300 has height 7; 300-350 has height 3.]

(c) What characteristic of the shape of the distribution of room size is apparent from the histogram but not from the boxplot?

Solution 1

(a) Mark scheme

- The distribution of room sizes is bimodal and roughly symmetric, with two clusters: about 100-200 square feet and about 250-350 square feet. The center of the distribution is between 200 and 300 square feet (e.g., median ≈ 253.5, mean ≈ 231.4). The spread can be stated as: the range is a value between 150 and 250 square feet, OR the IQR is a value between 50 and 150 square feet, OR all room sizes are between 100 and 350 square feet. The response must include context (room size in square feet). Full credit requires all four components: (1) bimodal shape / two peaks / two clusters (NOT unimodal or normal), (2) center between 200 and 300 sq ft, (3) a valid spread statement, (4) context.

Solution 1

(b) Mark scheme

- Outlier check: IQR = Q3 - Q1 = 292 - 174 = 118 square feet. Fences: $Q_1 - 1.5(IQR) = 174 - 1.5(118) = -3$ and $Q_3 + 1.5(IQR) = 292 + 1.5(118) = 469$. There are no potential outliers because the minimum (134 sq ft) does not fall below -3, and the maximum (315 sq ft) does not exceed 469. Boxplot: sketch a standard box-and-whisker (no mean shown) using the five-number summary min=134, Q1=174, median=253.5, Q3=292, max=315 – on the grid the minimum should fall between 120-140, Q1 between 160-180, the median between 240-260, Q3 between 280-300, and the maximum between 300-320 (square feet).

Solution 1

(c)

Mark scheme

- The histogram clearly shows the bimodal nature of the distribution of room sizes (two distinct peaks/clusters), but this bimodal shape is not apparent from the boxplot – a boxplot only displays the five-number summary and cannot reveal multiple modes. (A response based only on symmetry/skewness, rather than the bimodal/unimodal shape, does not earn credit here.)

Question 1

2015 ▪ Q1

STATISTICS

SECTION II

Part A

Questions 1-5

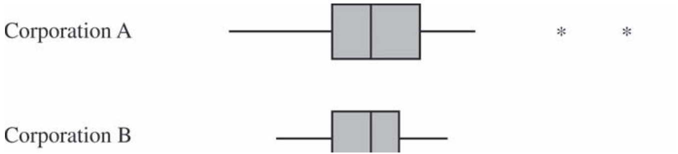
Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

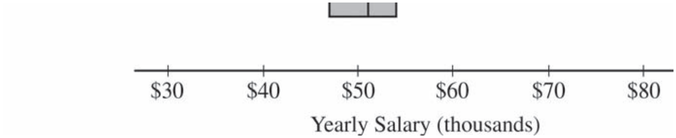
Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Question 1

1. Two large corporations, A and B, hire many new college graduates as accountants at entry-level positions. In 2009 the starting salary for an entry-level accountant position was \$36,000 a year at both corporations. At each corporation, data were collected from 30 employees who were hired in 2009 as entry-level accountants and were still employed at the corporation five years later. The yearly salaries of the 60 employees in 2014 are summarized in the boxplots below.



Question 1



(a) Write a few sentences comparing the distributions of the yearly salaries at the two corporations.

(b) Suppose both corporations offered you a job for $36,000 a year as an entry-level accountant.

(i) Based on the boxplots, give one reason why you might choose to accept the job at corporation A.

(ii) Based on the boxplots, give one reason why you might choose to accept the job at corporation B.

Solution 1

(a) Mark scheme

- The median salary is about the same for both corporations. The range and interquartile range of salaries are greater for Corporation A than for Corporation B. The two highest salaries at Corporation A are outliers, while Corporation B has no outliers. (Ignore any mention of shape - complete shape information cannot be determined from a boxplot.) Full credit needs all four: (1) a correct comparison of center, (2) a correct comparison of spread, (3) a discussion of Corporation A's outliers, and (4) the response given in context (referring to salaries/corporations, not just numbers).

Solution 1

(b)(i)

Mark scheme

- Sample reason to choose Corporation A: five years after starting, at least 3 of 30 (10%) of the salaries at Corporation A are greater than the maximum salary at Corporation B, so accepting the offer at Corporation A gives the potential to earn a higher salary than is possible at Corporation B. Credit requires identifying a relevant statistical measure or comparison that supports choosing Corporation A, with an explanation of why it is relevant.

Solution 1

(b)(ii)

Mark scheme

- Sample reason to choose Corporation B: five years after starting, the minimum salary at Corporation B is greater than the minimum at Corporation A; at Corporation A some employees are still earning the starting salary of 36,000 dollars and appear never to have received a raise in five years, so working at Corporation A risks never getting a raise, making Corporation B the safer choice. Credit requires identifying a relevant statistical measure or comparison that supports choosing Corporation B, with an explanation of why it is relevant.