Past-paper walk-through

Comparing Distributions of a Quantitative Variable ▪ 1.9

eXercise

Questions in this set

- 2026 Q1
- 2025 Q1
- 2023 Q1
- 2022 Q5
- 2021 Q6
- 2018 Q5
- 2017 Q4
- 2016 Q6
- 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

2025 ▪ Q1

The manager of an automotive company is interested in comparing the gas mileages for cars manufactured in Country A and cars manufactured in Country B. The manager selected a random sample of 100 cars manufactured in Country A and a random sample of 100 cars manufactured in Country B. The gas mileages for each sample, in miles per gallon (mpg), are summarized in the boxplots.

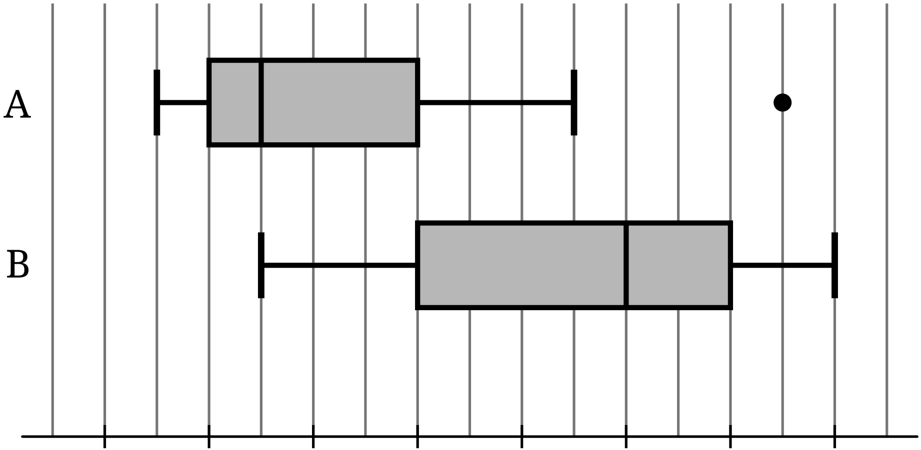
Boxplots of Gas Mileage for Each Country

[Boxplots comparing gas mileage (mpg) for Country A and Country B, drawn above a shared horizontal axis "Gas Mileage (mpg)" numbered 12 to 40 in increments of 4. Country A: left whisker to about 14, box (Q1) starts at about 16, median line at about 17, box (Q3) ends at about 24, right whisker to about 30, and an outlier point at about 38. Country B: left whisker to about 18, box (Q1) starts at about 24, median line at about 32, box (Q3) ends at about 36, right whisker to about 40.]

Question 1

(a) Compare the distributions of gas mileage for the sample of cars manufactured in Country A and the sample of cars manufactured in Country B.
(b) For the distribution of gas mileage for the sample of cars manufactured in Country A, would you expect the mean to be greater than 18 mpg, less than 18 mpg, or equal to 18 mpg? Justify your answer.
(c)(i) The manager will create a new boxplot with the combined data from the sample of cars manufactured in Country A and the sample of cars manufactured in Country B. What is the range of the combined data set? Justify your answer.
(c)(ii) What is a possible value of the median of the combined data set? Justify your answer by referencing the boxplots shown.

Question 1



Solution 1

(a) Mark scheme

- Country A's gas mileage distribution has a lower center than Country B's: median A $=$ 18 mpg is less than median B $=$ 32 mpg. Spread: the IQR of A (8 mpg) is less than the IQR of B (12 mpg), even though the range of A (24 mpg) is slightly greater than the range of B (22 mpg). The car in Country A with 38 mpg (the maximum of A) is an outlier; Country B's distribution has no outliers. Full credit (1 mark) requires at least 3 of these 4 components: (1) directly compares center for the two distributions, (2) directly compares spread (IQR or range) for the two distributions, (3) indicates Country A has an outlier, (4) sufficient context (both country names AND 'gas mileage'/'mpg'). Do not credit shape descriptions like

Solution 1

'skewed'/'symmetric' toward these components (boxplots can't fully establish shape).

Solution 1

(b) Mark scheme

- The mean is expected to be greater than 18 mpg. The median of Country A's gas-mileage distribution is 18 mpg. Because the distribution has an outlier to the right (is right-skewed, due to the 38 mpg car), the mean — which is not resistant to outliers — is pulled above the resistant median toward the higher values. Full credit needs: (1) states the mean is expected to be greater than 18 mpg, (2) states 18 as the median value, (3) a reasonable justification connecting the right-skew/outlier to mean $>$ median (a bare 'because the distribution is skewed' is too weak; an appropriate numerical argument also qualifies, but averaging portions of the five-number summary does not).

Solution 1

(c)(i) Mark scheme

- Range of the combined data set = 26 mpg. The maximum of the combined data is 40 mpg (Country B's maximum, since Country A's boxplot shows all its values below 40). The minimum of the combined data is 14 mpg (Country A's minimum, since Country B's boxplot shows all its values above 14). Range = 40 − 14 = 26 mpg. Full credit needs: (1) correctly calculates 26 mpg as the range, (2) justifies it by identifying 40 as the combined maximum and 14 as the combined minimum (a response that only restates 'ranges from 14 to 40' without computing 26 does not satisfy component 1 but may satisfy component 2).

Solution 1

(c)(ii) Mark scheme

- A possible median of the combined data set is 24 mpg. With 200 total values, the median is a value where at least 100 of the combined gas mileages are $\leq$ it and at least 100 are $\geq$ it. From Country A's boxplot, Q3 $=$ 24 mpg, so at least 75 of A's 100 values are ≤ 24. From Country B's boxplot, Q1 $=$ 24 mpg, so at least 75 of B's 100 values are ≥ 24 (and thus at least 25 of B's values are ≤ 24, and at least 25 of A's values are ≥ 24). Combining, at least 100 of the 200 combined values are ≤ 24 and at least 100 are ≥ 24, so 24 is a valid possible median. Full credit needs: (1) calculates 24 mpg as a possible median, (2) justifies it by showing at least half the combined values are ≤ 24 and at least half are ≥ 24 (referencing

Solution 1

the boxplots' quartiles, a near-100th-value argument, or counting boxplot segments are all acceptable justifications).

Question 1

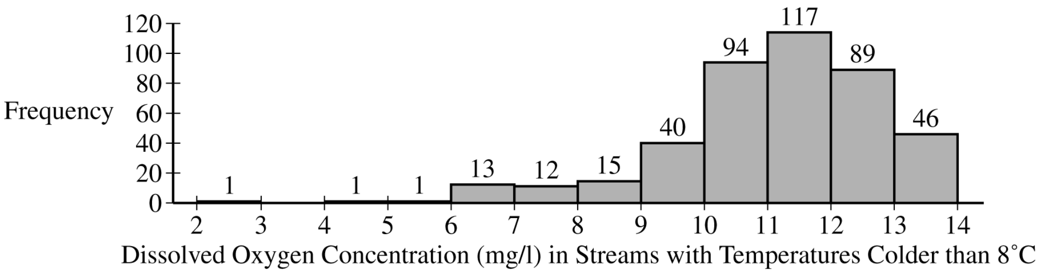
2023 ▪ Q1

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 ≈ 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 ≈ 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 5

2022 ▪ Q5

Studies have shown that foods rich in compounds known as flavonoids help lower blood pressure. Researchers conducted a study to investigate whether there was a greater reduction in blood pressure for people who consumed dark chocolate, which contains flavonoids, than people who consumed white chocolate, which does not contain flavonoids. Twenty-five healthy adults agreed to participate in the study and add 3.5 ounces of chocolate to their daily diets. Of the 25 participants, 13 were randomly assigned to the dark chocolate group and the rest were assigned to the white chocolate group. All participants had their blood pressure recorded, in millimeters of mercury (mmHg), before adding chocolate to their daily diets and again 30 days after adding chocolate to their daily diets.

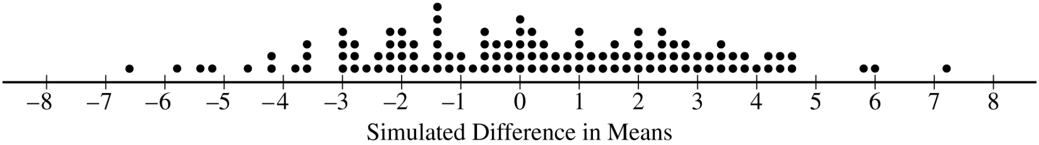
The reduction in blood pressure (before minus after) for each of the participants in the two groups is shown in the dotplots below.

Question 5

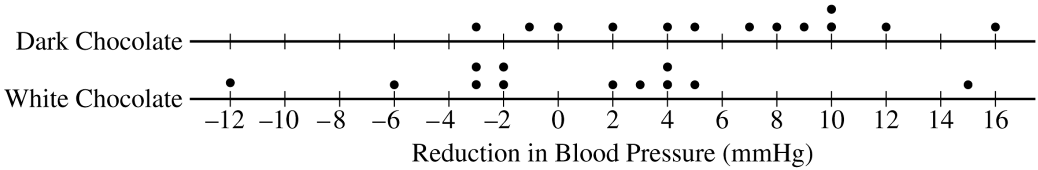
[Dotplots titled with x-axis “Reduction in Blood Pressure (mmHg)” from -12 to 16. “Dark Chocolate” row: approximate dot values at -2, 1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 10 (highest stack), 16. “White Chocolate” row: approximate dot values at -12, -6, -2, -1, 3, 4, 4, 5, 16.]

(a) Determine and compare the medians of the reduction in blood pressure for the two groups.

Question 5



Question 5


Dark Chocolate
White Chocolate
–12 –10 –8 –6 –4 –2 0 2 4 6 8 10 12 14 16
Reduction in Blood Pressure (mmHg)

Question 5

2022 ▪ Q5

Studies have shown that foods rich in compounds known as flavonoids help lower blood pressure. Researchers conducted a study to investigate whether there was a greater reduction in blood pressure for people who consumed dark chocolate, which contains flavonoids, than people who consumed white chocolate, which does not contain flavonoids. Twenty-five healthy adults agreed to participate in the study and add 3.5 ounces of chocolate to their daily diets. Of the 25 participants, 13 were randomly assigned to the dark chocolate group and the rest were assigned to the white chocolate group. All participants had their blood pressure recorded, in millimeters of mercury (mmHg), before adding chocolate to their daily diets and again 30 days after adding chocolate to their daily diets.

The reduction in blood pressure (before minus after) for each of the participants in the two groups is shown in the dotplots below.

Question 5

[Dotplots titled with x-axis "Reduction in Blood Pressure (mmHg)" from -12 to 16. "Dark Chocolate" row: approximate dot values at -2, 1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 10 (highest stack), 16. "White Chocolate" row: approximate dot values at -12, -6, -2, -1, 3, 4, 4, 5, 16.]

(b) The researchers found the mean reduction in blood pressure for those who consumed dark chocolate is $\bar{x}_{dark} = 6.08$ mmHg and the mean reduction in blood pressure for those who consumed white chocolate is $\bar{x}_{white} = 0.42$ mmHg.

One researcher indicated that because the difference in sample means of 5.66 mmHg is greater than 0 there is convincing statistical evidence to conclude that the population mean blood pressure reduction for those who consume dark chocolate is greater than for those who consume white chocolate. Why might the researcher's conclusion, based only on the difference in sample means of 5.66 mmHg, not necessarily be true?

Question 5

2022 ▪ Q5

Studies have shown that foods rich in compounds known as flavonoids help lower blood pressure. Researchers conducted a study to investigate whether there was a greater reduction in blood pressure for people who consumed dark chocolate, which contains flavonoids, than people who consumed white chocolate, which does not contain flavonoids. Twenty-five healthy adults agreed to participate in the study and add 3.5 ounces of chocolate to their daily diets. Of the 25 participants, 13 were randomly assigned to the dark chocolate group and the rest were assigned to the white chocolate group. All participants had their blood pressure recorded, in millimeters of mercury (mmHg), before adding chocolate to their daily diets and again 30 days after adding chocolate to their daily diets.

The reduction in blood pressure (before minus after) for each of the participants in the two groups is shown in the dotplots below.

Question 5

[Dotplots titled with x-axis "Reduction in Blood Pressure (mmHg)" from -12 to 16. "Dark Chocolate" row: approximate dot values at -2, 1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 10 (highest stack), 16. "White Chocolate" row: approximate dot values at -12, -6, -2, -1, 3, 4, 4, 5, 16.]

(c) A simulation was conducted to investigate whether there is a greater reduction of blood pressure for those who consume dark chocolate than for those who consume white chocolate. The simulation was conducted under the assumption that no difference exists. The results of 120 trials of the simulation are shown in the following dotplot.

[Dotplot titled with x-axis "Simulated Difference in Means" from -8 to 8; roughly symmetric, mound-shaped distribution of 120 dots centered near 0, tallest stacks between -2 and 3, tapering to isolated dots near -8, -7, -6, -5, 6, 7.]

Use the results of the simulation to determine whether the results from the 25 participants in the study provide convincing statistical evidence, at a 5 percent level of

Question 5

significance, that adding dark chocolate to a daily diet will result in a greater reduction in blood pressure, on average, than adding white chocolate to a daily diet. Justify your answer.

Solution 5

(a)

Mark scheme

- The sample median reduction in blood pressure for those eating dark chocolate is 7 mmHg, which is greater than the sample median reduction for those eating white chocolate, about 0 mmHg. Full credit needs all three: (1) the medians are correctly determined (dark chocolate median $= 7$ mmHg; white chocolate median between -2 and 2 mmHg, inclusive), (2) the computed medians are correctly compared, (3) context is included (the response variable, reduction in blood pressure, and the two treatment groups, dark and white chocolate).

Solution 5

(b) Mark scheme

- The researcher's conclusion may not necessarily be true because looking at the difference in sample means alone does not account for the variability in the sampling distribution of the difference in sample means. A different random assignment of subjects to dark and white chocolate could result in a sample mean reduction for dark chocolate that is smaller than for white chocolate. The variability in the sampling distribution of potential differences in sample means must be considered when concluding convincing statistical evidence; an inference procedure can assess the likelihood that the observed difference in sample means (5.66 mmHg) occurred by random chance if the population means are actually equal. Full credit needs at least two of the following three: (1) indicates the

Solution 5

researchers failed to consider the variability in the sampling distribution of differences in sample means (or that this should be considered), or states a different random assignment could produce different sample means, (2) explains that a difference of 5.66 mmHg could have occurred by random chance even if the population means are equal, (3) indicates that an inference procedure is needed. (A response may also earn full credit by instead correctly noting that no population conclusion can be drawn at all, because the sample was drawn only from the subpopulation of healthy adults.)

Solution 5

(c) Mark scheme

- The observed value of the sample statistic $\bar{x}_{dark} - \bar{x}_{white}$ is 5.66 mmHg. The simulation (under the assumption of no true difference) shows that a difference of 5.66 mmHg or larger occurred in only 3 of the 120 trials, so the (simulation-based) p-value is approximately $\frac{3}{120} = 0.025$. Because this p-value is less than 0.05, there is convincing statistical evidence that adding dark chocolate to the diet results in a greater mean reduction in blood pressure than adding white chocolate, for people similar to those in the study. (The hypotheses themselves are not required since they are given in the stem.) Full credit needs all four components: (1) calculates the correct p-value of 0.025, (2) provides

Solution 5

supporting work from the simulation for the calculation (e.g. noting only 3 of the 120 trials met or exceeded 5.66 mmHg), (3) provides a correct conclusion in context about whether there is convincing evidence that dark chocolate gives a greater mean blood-pressure reduction than white chocolate, (4) justifies the conclusion by directly comparing the calculated p-value to 0.05.

Question 6

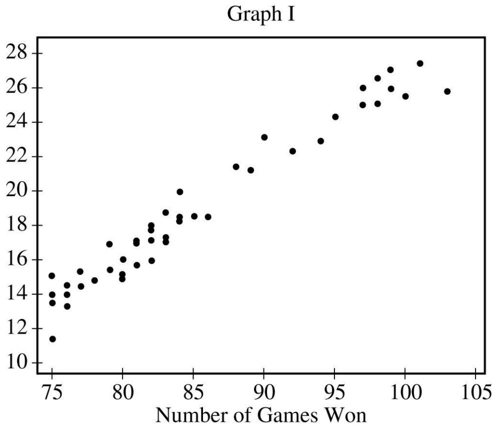
2021 ▪ Q6

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.

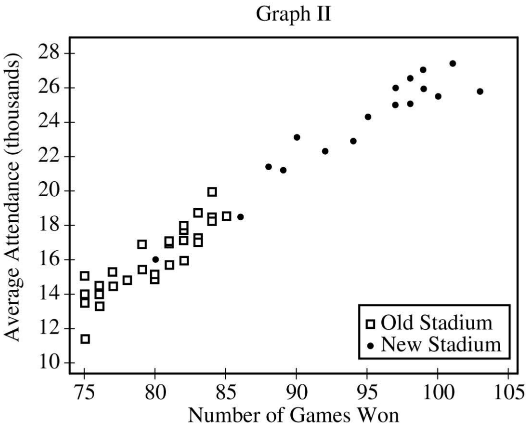
[Two vertical boxplots side by side, y-axis "Average Attendance (thousands)" from 10 to 28 in increments of 2. "Old Stadium" boxplot: whisker from about 11.5 to 20, box (Q1 to Q3) from about 14 to 17.5, median line at about 16. "New Stadium" boxplot: whisker from about 16 to 27.5, box (Q1 to Q3) from 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
28
26
24
22
20
18
16
14
12
10
75
80
85
90
95
100
105
Number of Games Won

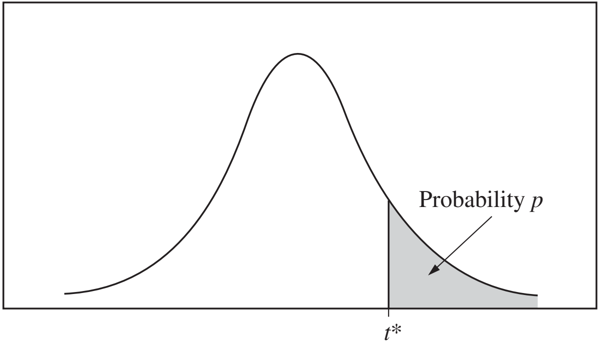


Graph II
Average Attendance (thousands)
28
26
24
22
20
18
16
14
12
10
75
80
85
90
95
100
105
Number of Games Won
Old Stadium
New Stadium

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

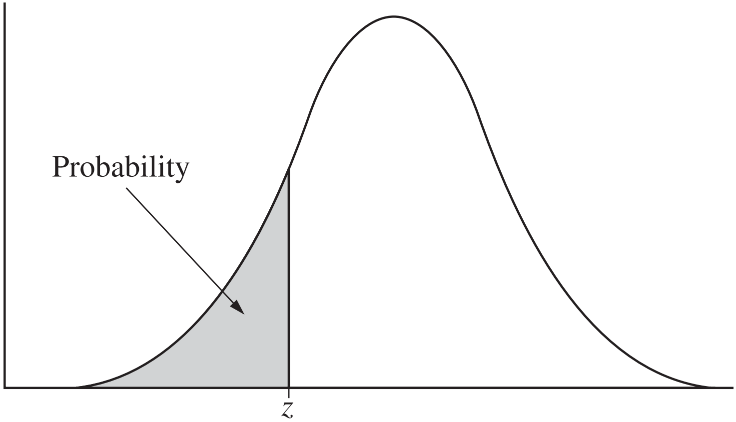


Table entry for z is the probability lying below z.

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 5

2018 ▪ Q5

The following histograms summarize the teaching year for the teachers at two high schools, A and B.

[Histogram “High School A”: x-axis “Teaching Year” labeled at 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, 34 (each bar spans a 3-year interval, left endpoint inclusive); y-axis “Frequency” from 0 to 80 in increments of 10. Approximate bar heights by interval start: 1→46, 4→48, 7→24, 10→45, 13→29, 16→4, 19→1, 22→0, 25→1, 28→1, 31→1, 34→0.]

[Histogram “High School B”: x-axis “Teaching Year” labeled at 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, 34 (each bar spans a 3-year interval, left endpoint inclusive); y-axis “Frequency” from 0 to 80 in increments of 10. Approximate bar heights by interval start: 1→79, 4→34, 7→28, 10→29, 13→19, 16→12, 19→8, 22→3, 25→4, 28→2, 31→3, 34→0.]

Question 5

Teaching year is recorded as an integer, with first-year teachers recorded as 1, second-year teachers recorded as 2, and so on. Both sets of data have a mean teaching year of 8.2, with data recorded from 200 teachers at High School A and 221 teachers at High School B. On the histograms, each interval represents possible integer values from the left endpoint up to but not including the right endpoint.

(a) The median teaching year for one high school is 6, and the median teaching year for the other high school is 7. Identify which high school has each median and justify your answer.

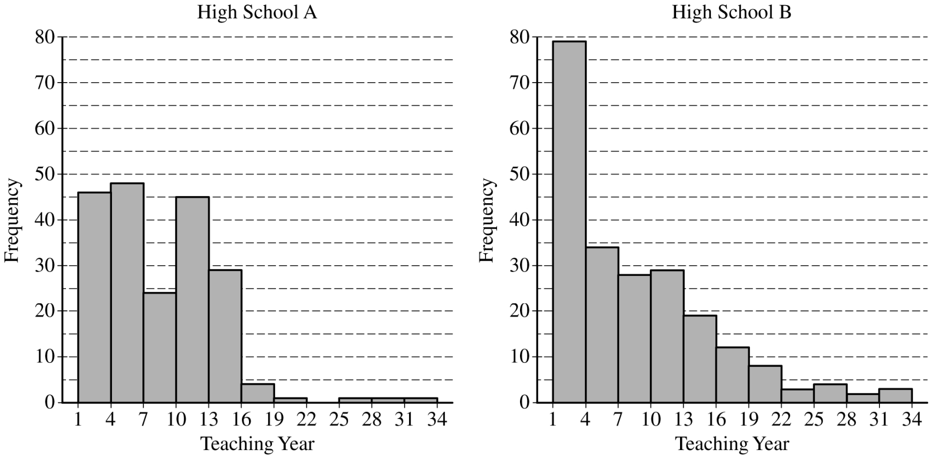
(b) An additional 18 teachers were not included with the data recorded from the 200 teachers at High School A. The mean teaching year of the 18 teachers is 2.5. What is the mean teaching year for all 218 teachers at High School A?

(c) The standard deviation of the teaching year for the 221 teachers at High School B is 7.2. If one teacher is selected at random from High School B, what is the probability

Question 5

that the teaching year for the selected teacher will be within 1 standard deviation of the mean of 8.2 ? Justify your answer.

Question 5


High School A
High School B
Frequency
Teaching Year
80
70
60
50
40
30
20
10
0
1
4
7
10
13
16
19
22
25
28
31
34

Solution 5

(a) Mark scheme

- High School A has median teaching year 7; High School B has median teaching year 6. Reasoning: High School A's median needs 100 of its 200 values at or below it and 100 at or above; only 94 of A's values are below 7 while 106 are at least 7, so A's median lies in the interval [7,10) and cannot be less than 7. High School B's median is the 111th of 221 ordered values; more than half of B's values (113) are less than 7, so B's median lies in [4,7) and must be less than 7 — hence 6. (Equivalent argument: High School B's distribution is strongly right-skewed while A's is bimodal with a few high outliers; a strongly right-skewed distribution tends to have a mean substantially larger than its median, and since both schools have the same mean (8.2), the more right-skewed distribution, B, must be the one

Solution 5

with the larger mean-to-median gap — i.e., the smaller median, 6.) Full credit requires: (1) stating median $= 7$ for School A and median $= 6$ for School B, (2) a reasonable explanation of how the decision was made, and (3) either explicitly using/applying the definition of median as the criterion, OR stating that School B is more skewed and explaining why the more-skewed distribution (B) has the larger mean-median separation. Describing either distribution as left-skewed, normal, or approximately normal lowers an otherwise-complete response by one level.

Solution 5

(b) Mark scheme

- Combined (weighted-average) mean for all 218 teachers at High School A: $\dfrac{(200)(8.2)+(18)(2.5)}{200+18}=\dfrac{1640+45}{218}=\dfrac{1685}{218}\approx 7.73$ years. Full credit requires both: (1) the correct answer, 7.73, and (2) enough work shown that the answer was obtained as a weighted average of the two individual group means (200 teachers at 8.2 years and 18 teachers at 2.5 years), not a simple unweighted average of 8.2 and 2.5.

Solution 5

(c) Mark scheme

- The interval mean $\pm$ 1 standard deviation is 8.2 ± 7.2, i.e. from 1.0 to 15.4 years. Since teaching year is recorded as an integer, this interval covers teaching years 1 through 15. Summing the heights of the five histogram bars covering that range (bins 1–4, 4–7, 7–10, 10–13, 13–16): $79 + 34 + 28 + 29 + 19 = 189$. Probability $= 189/221 \approx 0.8552$. Full credit requires: (1) correctly determining the appropriate interval is 1 to 15.4 (or 1 to 15 teaching years, or recognizing that all histogram bins up to 16 fall within one SD of the mean), (2) correctly summing the counts of data values in that interval based on the histogram (a count that is slightly off due to difficulty reading exact bar heights is still acceptable), and (3) computing the probability using 221 as the denominator. Using the Empirical Rule

Solution 5

or a normal-distribution approximation instead of the actual finite population/histogram counts is NOT acceptable and earns no credit for this part.

Question 4

2017 ▪ Q4

The chemicals in clay used to make pottery can differ depending on the geographical region where the clay originated. Sometimes, archaeologists use a chemical analysis of clay to help identify where a piece of pottery originated. Such an analysis measures the amount of a chemical in the clay as a percent of the total weight of the piece of pottery. The boxplots below summarize analyses done for three chemicals—X, Y, and Z—on pieces of pottery that originated at one of three sites: I, II, or III.
[Three side-by-side sets of boxplots grouped by Site (Site I, Site II, Site III), each set showing boxplots for chemicals X, Y, Z on the x-axis. Shared y-axis "Percent of Total Weight" from 0 to 16 (gridlines every 2). A legend distinguishes Site I (open box), Site II (gray box), Site III (hatched box).

Question 4

Site I: Chemical X — min 6, Q1 $\approx$ 6.5, median $\approx$ 7, Q3 $\approx$ 7.5, max 8. Chemical Y — min 11, Q1 $\approx$ 12, median $\approx$ 13, Q3 $\approx$ 14, max 15. Chemical Z — min 4, Q1 $\approx$ 5, median $\approx$ 7, Q3 $\approx$ 9, max 10.

Site II: Chemical X — min 3.5, Q1 $\approx$ 5, median $\approx$ 6, Q3 $\approx$ 6.5, max 7. Chemical Y — min 1.8, Q1 $\approx$ 2.5, median $\approx$ 3, Q3 $\approx$ 3.5, max 4.5. Chemical Z — min 6, Q1 $\approx$ 6.5, median $\approx$ 7, Q3 $\approx$ 7.5, max 8.

Site III: Chemical X — min 4.8, Q1 $\approx$ 5.5, median $\approx$ 6, Q3 $\approx$ 7, max 7.5. Chemical Y — min 6, Q1 $\approx$ 6.5, median $\approx$ 7, Q3 $\approx$ 7.5, max 8. Chemical Z — min 2.8, Q1 $\approx$ 5, median $\approx$ 7, Q3 $\approx$ 9, max 11.2.]

(a) For chemical Z, describe how the percents found in the pieces of pottery are similar and how they differ among the three sites.

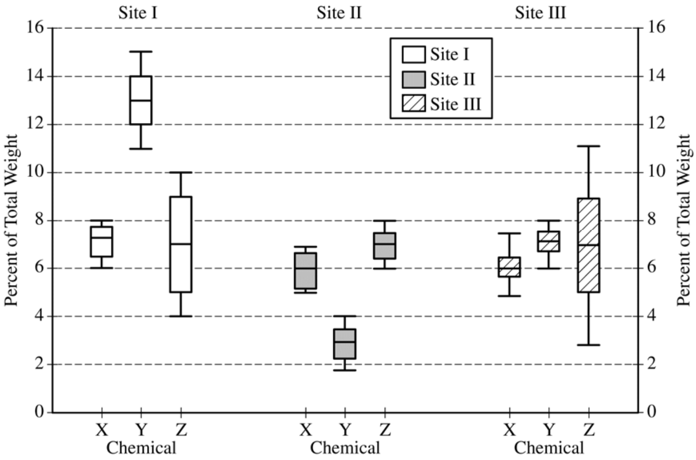
(b)(i) Consider a piece of pottery known to have originated at one of the three sites, but the actual site is not known.

Question 4

Suppose an analysis of the clay reveals that the sum of the percents of the three chemicals X, Y, and Z is 20.5%. Based on the boxplots, which site—I, II, or III—is the most likely site where the piece of pottery originated? Justify your choice.

(b)(ii) Suppose only one chemical could be analyzed in the piece of pottery. Which chemical—X, Y, or Z—would be the most useful in identifying the site where the piece of pottery originated? Justify your choice.

Question 4


Site I
Site II
Site III
Site I
Site II
Site III
Percent of Total Weight
16
14
12
10
8
6
4
2
0
X
Y
Z
Chemical
X
Y
Z
Chemical
X
Y
Z
Chemical

Solution 4

(a) Mark scheme

- The median percent of chemical Z is similar across all three sites (about 7 percent at each site). The ranges (variability), however, differ substantially: Site II has the smallest range, about 2 percent (from about 6 percent to 8 percent); Site I has a range of about 6 percent (from about 4 percent to 10 percent); Site III has the largest range, about 8 percent (from about 3 percent to 11 percent).

Solution 4

- Full credit needs three components: (1) recognition that the medians/centers are almost the same for the three sites; (2) recognition that the variability (ranges, IQRs, spread) is different across the three sites; (3) context is included (reference to site, chemical, weight, total weight, X, Y, or Z). Comments about shape should be ignored/are not required (complete shape information isn't obtainable from boxplots); numerical values are not required, but if given, any reasonable approximation from the boxplots is acceptable.

Solution 4

(b)(i) Mark scheme

- The piece most likely originated at Site III. Using the boxplots, approximate minimum and maximum percents for the three chemicals (X, Y, Z) give the following approximate min/max sums at each site: Site I sums range from about 21 to 33; Site II sums range from about 12.9 to 19; Site III sums range from about 14 to 26.5. Since the piece of pottery's chemical sum is 20.5 percent, Site III is the only site in which 20.5 falls between the sums of the minimum and maximum values.

Solution 4

- Full credit needs three components: (1) Site III is chosen; (2) sums of the minimum and maximum are computed for the three chemicals at each site; (3) a reasonable numerical justification involving those sums across the three chemicals is given to choose Site III (e.g. noting 20.5 falls between Site III's min-sum and max-sum but not between the other two sites').

Solution 4

(b)(ii) Mark scheme

- Chemical Y would be most useful, because the distributions of the percentages of chemical Y at the three sites do not overlap (e.g. all values at Site I are high, all values at Site II are low, and all values at Site III are in the middle – the boxplots share no data / the ranges never intersect). The distributions of chemicals X and Z, by contrast, have substantial overlap across the three sites, making them less useful for identifying the site of origin.

Solution 4

- Full credit needs chemical Y chosen AND a reasonable justification based on the fact that chemical Y's distributions are distinctive across sites – the justification must address BOTH location and variability (e.g. "the boxplots for chemical Y do not overlap" is acceptable). Justifications that address only one aspect (e.g. "chemical Y varies the most," "the medians differ," "the boxplots are different") are incomplete and earn only partial credit.

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Question 6

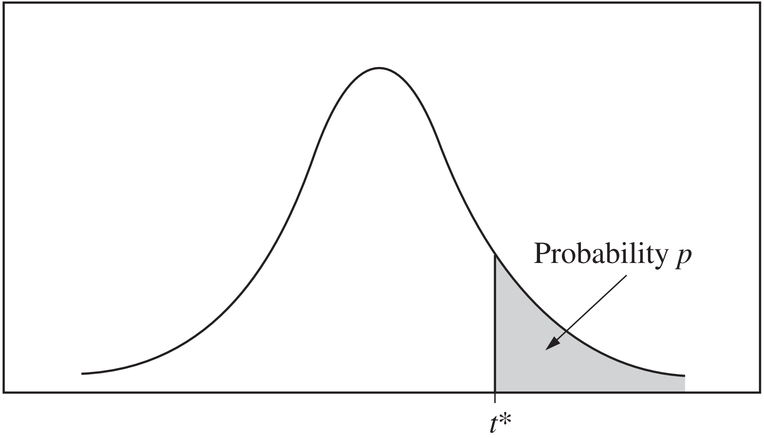
Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

(a) Does the scatterplot support the newspaper's report about number of semesters and starting salary? Justify your answer.

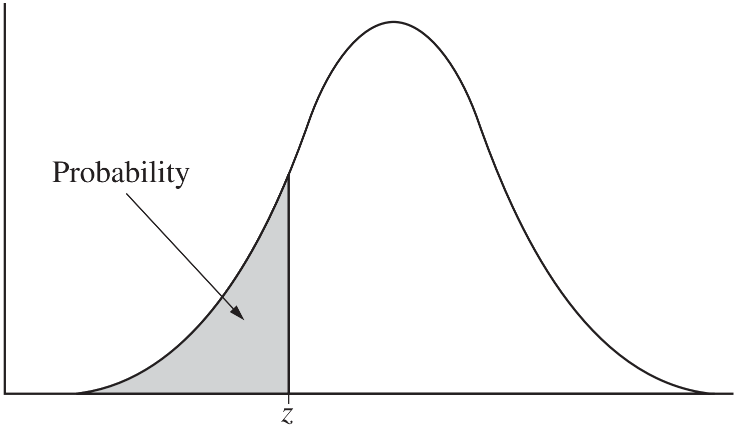
Question 6

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

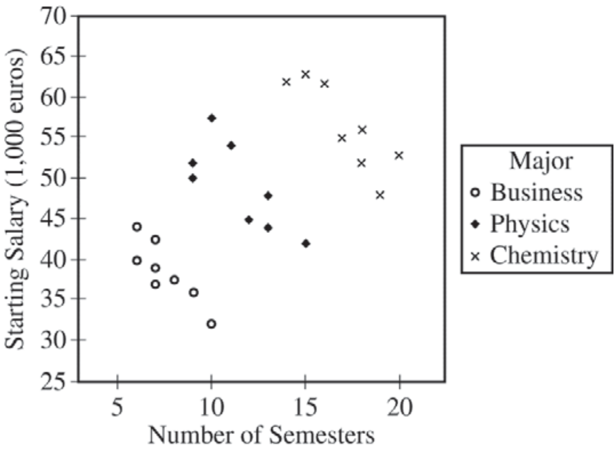


Question 6

Table entry for z is the probability lying below z.



Question 6



Starting Salary (1,000 euros)
70
65
60
55
50
45
40
35
30
25
5
10
15
20
Number of Semesters
Major
Business
Physics
Chemistry

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(b) Identify the slope of the least-squares regression line, and interpret the slope in context.

An independent researcher received the data from the newspaper and conducted a new analysis by separating the data into three groups based on the major of each person. A revised scatterplot identifying the major of each person is shown below.

[Revised scatterplot, same axes as before (y-axis "Starting Salary (1,000 euros)" approximately 25 to 65, x-axis "Number of Semesters" approximately 5 to 20), with points marked by three different symbols per the legend "Major: Business (square), Physics (circle), Chemistry (diamond)". Business-major points cluster at lower semesters with lower-to-moderate salaries; Physics- and Chemistry-major points spread across higher semesters with a range of salaries, generally higher for more semesters.]

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(c) Based on the people in the sample, describe the association between starting salary and number of semesters for the business majors.

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(d) Based on the people in the sample, compare the median starting salaries for the three majors.

Question 6

2016 ▪ Q6

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program and were selected to represent the population of that country, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
[Scatterplot titled with y-axis "Starting Salary (1,000 euros)" ranging approximately 25 to 65, x-axis "Number of Semesters" ranging approximately 0 to 20. Points scattered showing a general upward trend from lower-left (fewer semesters, lower salary, around 30) to upper-right (more semesters, higher salary, up to about 65), with considerable scatter throughout.]
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

$S = 7.37702$ nbsp;nbsp; R-Sq $= 33.5\%$ nbsp;nbsp; R-Sq(adj) $= 30.5\%$

Question 6

(e) Based on the analysis conducted by the independent researcher, how should the newspaper report be modified to give a better description of the relationship between the number of semesters and the starting salary for the people in the sample?

Solution 6

(a)

Mark scheme

- Yes – the scatterplot supports the newspaper's report, because it shows a positive association between number of semesters needed to complete an academic program and starting salary (in general, needing more semesters is associated with a higher starting salary). Full credit requires both addressing the positive association AND using it to justify that the scatterplot supports the report.

Solution 6

(b)

Mark scheme

- The slope of the least-squares regression line is 1.1594 (from $\hat{y} = 34.018 + 1.1594x$). Interpretation: for each additional semester needed to complete an academic program, the predicted (model-estimated / average) starting salary increases by about €1,159.40 (equivalently 1.1594 thousand euros). Full credit requires the correct numerical slope value AND an in-context interpretation using nondeterministic language (e.g. "predicted"/"estimated"/"on average" – not a deterministic claim that salary WILL increase).

Solution 6

(c)

Mark scheme

- For the business majors alone, there is a strong, negative, linear association between number of semesters and starting salary: business majors who need a greater number of semesters to complete their academic program tend to have LOWER starting salaries. Full credit requires stating the association is negative, describing it as strong and/or linear, and referring to both variables (semesters, salary) in context.

Solution 6

(d)

Mark scheme

- Comparing the three majors by median starting salary: business majors have the lowest median starting salary (around €38,000), followed by physics majors (around €48,000), with chemistry majors having the highest median starting salary (around €55,000). Full credit requires a correct comparison of all three majors, with reasonable median-salary values (or explicit reference to "median starting salary").

Solution 6

(e) Mark scheme

- The newspaper's report should be modified to account for major: overall, majors that take longer to complete DO tend to have higher starting salaries (chemistry highest, physics next, business lowest – the original positive association holds across majors). However, WITHIN each individual major, students who take a greater number of semesters tend to have LOWER starting salaries (a negative association) – the opposite of the overall pattern. Full credit requires noting the negative within-major association for each major AND still acknowledging the overall positive association across majors (a Simpson's-paradox-type reversal).

Question 1

2015 ▪ Q1

STATISTICS

SECTION II

Part A

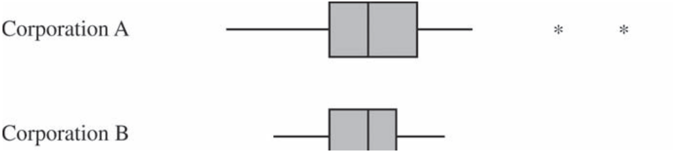
Questions 1-5

Spend about 65 minutes on this part of the exam.

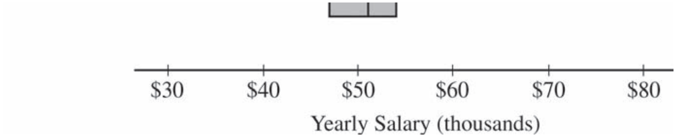
Percent of Section II score—75

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.

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.