

## Past-paper walk-through

Introducing Statistics: Do the Data We Collected Tell the Truth? ■ 3.1

eXercise

## Questions in this set

- 2021 Q6

## 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 6

2021 ■ Q6

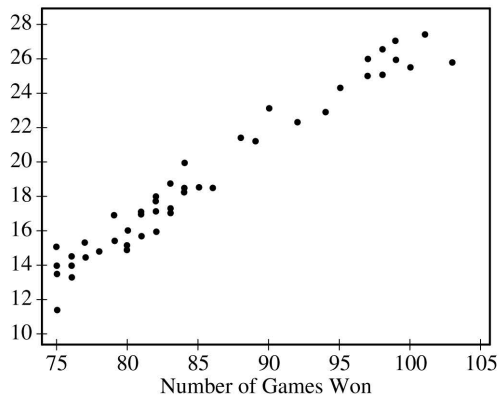
Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the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

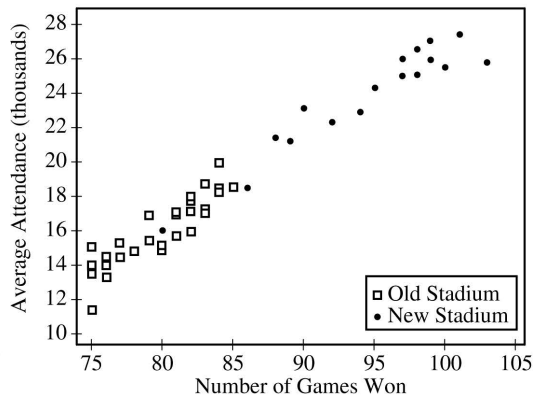
**(a)** Compare the distributions of average attendance between the old and new stadiums.

## Question 6

Graph I



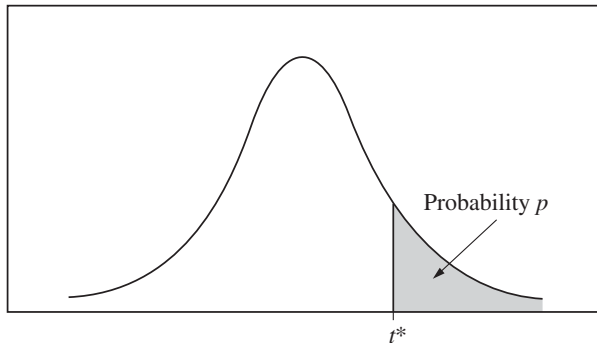
Graph II



# Question 6

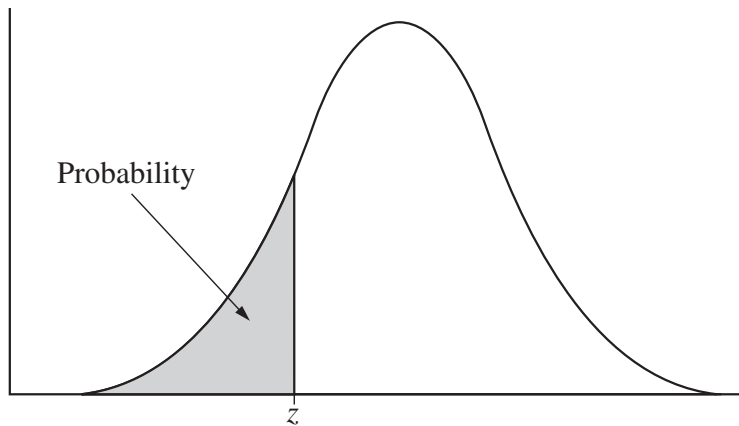
|     |       |       |       |       |       |       |       |       |       |       |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 3.0 | .9981 | .9981 | .9981 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| 3.1 | .9990 | .9991 | .9991 | .9991 | .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  $\approx 25,000$ ) *than the old stadium* (median  $\approx 16,000$ ). *The IQRs (spread/variability) are similar between the two stadiums, though the overall range of game attendance. Credit requires (1) a correct center comparison (new > old, with stadiums identified and an explicit comparison phrase), (2) a correct variability comparison (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) with an explanation referencing fitting a line through each stadium's points.

## 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).