2.2 Statistics: Data Presentation and Interpretation: Test Pack

Total Marks

EDEXCEL 9MA0 · Calculator allowed · about 65 minutes

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

1 A company has 850 employees split across three departments: Admin (170 employees), Production (510 employees) and Sales (170 employees). The company wants to survey employees about a proposed new flexible working policy.

(a) State what is meant, in the context of sampling, by a 'sampling frame'. **(1)**

(b) The company decides to take a stratified sample of 40 employees, proportional to department size. Calculate the number of employees that should be sampled from each department. **(3)**

(c) Describe how systematic sampling could instead be used to select 40 employees from a numbered list of all 850 employees, and state one advantage this method has over simple random sampling in this context. **(3)**

(Total for Question 1 is 7 marks)

© Revision Library | revisionlibrary.co.uk

2 A cafe owner records, for 8 different days last winter, the outdoor temperature and the number of hot chocolates sold.

(a) State the type of correlation you would expect between outdoor temperature and the number of hot chocolates sold. **(1)**

(b) Describe the general shape of the scatter diagram you would expect to see for this data, referring to the direction of the trend. **(1)**

(c) The owner also wants to investigate the relationship between outdoor temperature and the number of cold drinks sold. State the type of correlation expected and explain your reasoning. **(2)**

(d) On one unusually cold day in December, a local event caused hot chocolate sales to be much higher than the temperature that day would otherwise suggest. Explain the effect this data point is likely to have on the product moment correlation coefficient (PMCC) calculated for temperature against hot chocolate sales across the whole data set. **(2)**

(Total for Question 2 is 6 marks)

3 The times, in minutes, taken by 9 runners to complete a 5 km training route are: 34, 36, 29, 41, 38, 33, 30, 45, 37.

(a) Calculate the mean time. **(2)**

(b) Calculate the standard deviation of these times. **(3)**

(c) A tenth runner joins the group with a time of 60 minutes, a clear outlier. Without recalculating, state and justify the effect that including this runner's time would have on the mean and on the standard deviation. **(2)**

(Total for Question 3 is 7 marks)

© Revision Library | revisionlibrary.co.uk

4 A scientist records data on height above sea level, h hundred metres, and the air temperature, T degrees Celsius, for several days. The regression line of T on h is T = 21 - 1.4h. The product moment correlation coefficient is r = -0.33.

(a) Interpret the value of the gradient of the regression line in context. **(1)**

(b) Use the regression line to estimate T when h = 10, and comment on the reliability of this estimate. **(3)**

(c) Comment on the strength and direction of the linear correlation indicated by r = -0.33. **(2)**

(Total for Question 4 is 6 marks)

© Revision Library | revisionlibrary.co.uk

5 The table summarises the times, in minutes, taken by 45 people to complete a puzzle.
Time: 0 - 5, 5 - 10, 10 - 15, 15 - 20, 20 - 25
Frequency: 7, 9, 3, 15, 11

Time (minutes)	Frequency
0 - 5	7
5 - 10	9
10 - 15	3
15 - 20	15
20 - 25	11

(a) Estimate the mean time. **(3)**

(b) Estimate the standard deviation of the times, to 3 significant figures. **(3)**

(c) Use linear interpolation to estimate the median time, to 3 significant figures. **(3)**

(Total for Question 5 is 9 marks)

---

6 Two classes, A and B, sat the same maths test, out of 50 marks. Box plots of their results gave the following five-figure summaries.
Class A: minimum = 10, Q1 = 18, median = 26, Q3 = 40, maximum = 48
Class B: minimum = 15, Q1 = 27, median = 30, Q3 = 33, maximum = 36

(a) Compare the medians and the interquartile ranges of the two classes, in the context of the test. **(3)**

(b) Using the quartiles, comment on the skewness of each class's distribution of marks. **(2)**

(c) A teacher claims 'Class B had less variation in test marks than Class A.' By considering both the range and the interquartile range, comment on whether this claim is fair. **(3)**

(Total for Question 6 is 8 marks)

© Revision Library | revisionlibrary.co.uk

7 A researcher recorded, for 10 UK towns, the number of ice cream shops, x, and the number of sunburn cases reported at local hospitals over one summer, y. The PMCC between x and y was found to be r = 0.91, based on values of x ranging from 2 to 15.

(a) The researcher concludes: 'Opening more ice cream shops causes more cases of sunburn.' Comment on this conclusion, suggesting a variable that could explain the high correlation observed. **(3)**

(b) State one condition that should be checked before using the regression line from this data to predict y for a town with x = 25 ice cream shops, and explain why this matters here. **(2)**

(c) A large data set recorded daily mean temperature and daily rainfall for a UK location across a whole year. Explain why it would not be appropriate to calculate a single PMCC using all of this data if you wanted to investigate the relationship between temperature and rainfall during summer months only. **(2)**

(Total for Question 7 is 7 marks)

© Revision Library | revisionlibrary.co.uk

Mark scheme · 2.2 Statistics: Data Presentation and Interpretation: Test Pack

Question 1

(a) a list of all the members of the population, from which the sample is taken, oe **B1**

(a) Answer: A list of all the members of the population, used to select the sample.

(b) correct method shown for at least one department, e.g. 40 x (170/850) **M1**

(b) two of the three values correct **A1**

(b) cao all three correct: Admin = 8, Production = 24, Sales = 8 **A1**

(b) Answer: Admin: 8, Production: 24, Sales: 8.

(c) sampling interval k = 850/40 = 21.25, so k = 21 is used (rounding down so at least 40 employees can be chosen) **B1**

(c) choose a random starting point between 1 and 21, then select every 21st employee after that from the numbered list **B1**

(c) valid advantage, e.g. quicker/simpler to carry out than generating 40 distinct random numbers, or spreads the sample evenly across the whole ordered list, oe **B1**

(c) Answer: k = 850/40 = 21.25, so k = 21; pick a random start between 1 and 21, then take every 21st employee after that. This is quicker to carry out than generating 40 separate random numbers.

Question 2

(a) negative correlation **B1**

(a) Answer: Negative correlation.

(b) points generally slope downwards from left to right, i.e. as temperature increases, hot chocolate sales decrease, oe **B1**

(b) Answer: The points would generally slope downwards from left to right: as temperature increases, the number of hot chocolates sold decreases.

(c) positive correlation **B1**

(c) valid reasoning in context, e.g. customers are more likely to buy cold drinks when it is warmer, so sales rise as temperature rises, oe **B1**

(c) Answer: Positive correlation, since customers are more likely to want a cold drink as the outdoor temperature increases.

(d) the effect on the PMCC cannot be determined without the data **B1**

(d) explains that a cold day with high sales can reinforce the negative association, but its position relative to the other points could make the PMCC more or less negative, oe **B1**

(d) Answer: The effect on the PMCC cannot be determined without the data. A cold day with high sales can reinforce the negative association, but the PMCC may become more or less negative depending on the position of this point relative to the other data.

Question 3

(a) attempt sum of the 9 values / 9 **M1**

(a) awrt 35.9 (minutes) **A1**

(a) Answer: awrt 35.9 minutes

(b) attempt sum of squares of the 9 values (or use calculator statistical mode) **M1**

(b) correct method: $\sqrt{\text{sum } x^2 / 9 - \text{mean}^2}$ **M1**

(b) awrt 4.82 (minutes) **A1**

(b) Answer: awrt 4.82 minutes

(c) mean would increase, since 60 is greater than the current mean (awrt 35.9), pulling the average up, oe **B1**

(c) standard deviation would increase, since 60 is a long way from the mean, adding a value with a large deviation increases the spread, oe **B1**

(c) Answer: The mean would increase (60 is well above the current mean, pulling it up); the standard deviation would increase (60 is far from the mean, so it adds a large deviation and increases the spread).

Question 4

(a) each extra hundred metre in h is associated with a change of -1.4 in T, so as h increases T decreases **B1**

(a) Answer: gradient -1.4: T decreases by 1.4 for each unit increase in h

(b) substitutes h = 10 **M1**

(b) T = 7.00 **A1**

(b) The weak correlation makes the prediction unreliable. The observed range is not given, so it is not possible to determine whether this is interpolation. Accept equivalent comments. **B1**

(b) Answer: 7 degrees Celsius. The weak correlation makes the prediction unreliable. The observed range is not given, so it is not possible to determine whether this is interpolation.

(c) weak **B1**

(c) negative correlation **B1**

(c) Answer: weak negative linear correlation

Question 5

(a) uses the midpoint of each class **M1**

(a) sum of f x midpoint = 632.5 **M1**

(a) mean = 632.5/45 = 14.06 minutes **A1**

(a) Answer: 14.06 minutes

(b) sum f x midpoint^2 = 11181.2 **M1**

(b) variance = 11181.2/45 - (14.0556)^2 **M1**

(b) standard deviation = 7.14 minutes **A1**

(b) Answer: 7.14 minutes

(c) identifies the median class 15 - 20 (cumulative frequency reaches 34 at its upper bound) **M1**

(c) interpolates: 15 + ((22.5 - 19)/15) x 5 **M1**

(c) median = 16.2 minutes **A1**

(c) Answer: 16.2 minutes

© Revision Library | revisionlibrary.co.uk

Question 6

(a) Class B has the higher median (30 compared with 26), so Class B typically scored higher on the test **B1**

(a) Class A has the larger IQR (22 compared with 6), so marks in Class A were far more spread out (variable) **B1**

(a) valid overall conclusion in context, e.g. Class B performed more consistently well, while Class A's results ranged from much weaker to much stronger performance, oe **B1**

(a) Answer: Class B has a higher median (30 vs 26), so its typical mark was higher; Class A has a much larger IQR (22 vs 6), so its marks were far more spread out. Overall, Class B performed more consistently, while Class A had a wider range of ability.

(b) Class A: Q3 - median (14) > median - Q1 (8), so Class A's marks are positively skewed **B1**

(b) Class B: Q3 - median (3) = median - Q1 (3), so Class B's marks are symmetrical (based on the quartiles) **B1**

(b) Answer: Class A: 40 - 26 = 14 and 26 - 18 = 8, so 14 > 8, meaning Class A's marks are positively skewed. Class B: 33 - 30 = 3 and 30 - 27 = 3, which are equal, so Class B's marks are symmetrical based on the quartiles.

(c) range of Class A = 38, range of Class B = 21, so the range is smaller for Class B **B1**

(c) IQR of Class A = 22, IQR of Class B = 6, so the IQR is also smaller for Class B **B1**

(c) since both measures of spread agree (both smaller for Class B), the claim is fair/well supported, oe **B1**

(c) Answer: Range: Class A = 48 - 10 = 38, Class B = 36 - 15 = 21. IQR: Class A = 22, Class B = 6. Both the range and the IQR are smaller for Class B, so the claim is a fair conclusion, supported by both measures of spread, not just one.

Question 7

(a) a strong correlation does not imply causation **B1**

(a) identifies a plausible confounding (third) variable, e.g. hours of sunshine or temperature **B1**

(a) explains that this variable is likely to affect both x and y directly, e.g. more sunshine leads to both more ice cream shop custom and more sunburn cases, rather than one causing the other, oe **B1**

(a) Answer: A strong correlation does not mean one variable causes the other. Here, the amount of sunshine (or temperature) is a likely confounding variable: warmer, sunnier weather leads to both more customers for ice cream shops and more sunburn, rather than the ice cream shops themselves causing sunburn.

(b) identifies that x = 25 lies outside the range of the given data (2 to 15), so this would be extrapolation **B1**

(b) explains that the linear relationship is not known to hold outside the observed range, so a prediction here would be unreliable, oe **B1**

(b) Answer: x = 25 is outside the observed range of x (2 to 15), so using the regression line here would involve extrapolation; there is no evidence that the linear relationship continues to hold beyond the data collected, so the prediction would be unreliable.

(c) the whole year's data includes other seasons (e.g. winter), where the relationship between temperature and rainfall may be different from summer, oe **B1**

(c) combining all seasons could obscure, distort or average out the pattern that actually exists specifically in summer, so the whole-year PMCC would not answer the question about summer months, oe **B1**

(c) Answer: The whole year's data includes other seasons, such as winter, where the relationship between temperature and rainfall may behave quite differently to summer. Combining all months together could mask or distort the true summer relationship, so a PMCC calculated across the whole year would not correctly represent the relationship in summer alone; only the summer months' data should be used.

© Revision Library | revisionlibrary.co.uk