H01-30min Data Cleaning and Outliers (Higher): 30-minute Pack

Total Marks

EDEXCEL 1ST0 · Calculator allowed · about 30 minutes

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

1 A delivery company records the time, in minutes, taken for 10 parcels to be delivered from the depot to customers' homes.

Parcel	1	2	3	4	5	6	7	8	9	10
Time (min)	18	22	(not recorded)	19	145	21	20	17	23	19

(a) Parcel 3's delivery time was not recorded. Give one advantage and one disadvantage of imputing (estimating) this missing value using the mean of the other delivery times, rather than deleting parcel 3's record entirely. **(2)**

(b) Parcel 5's delivery time is recorded as 145 minutes, far higher than every other parcel. Explain why the company should investigate this value rather than immediately deleting it or replacing it with the mean. **(2)**

(c) The company decides to impute parcel 3's missing value using the mean of the other 9 recorded times (treating parcel 5's 145 minutes as genuine for now). Work out this imputed value, correct to 1 decimal place. **(2)**

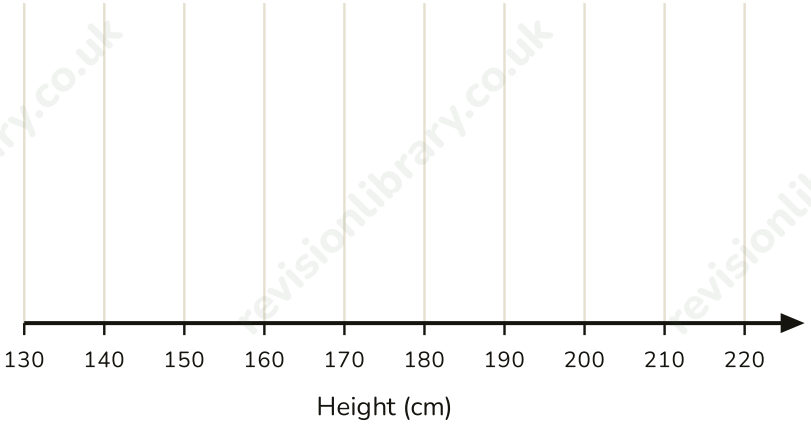
.....

(d) Comment on whether this imputed value from part (c) is a sensible estimate for parcel 3's delivery time. **(1)**

(Total for Question 1 is 7 marks)

2 A gardener records the height, in cm, of 15 sunflower plants at the end of the growing season, listed here in ascending order.

142, 145, 148, 150, 151, 153, 154, 156, 158, 159, 161, 163, 165, 168, 210



(a) Find the median, the lower quartile (Q1) and the upper quartile (Q3) of the 15 heights. **(3)**

.....

(b) Show that the tallest sunflower (210 cm) is an outlier, using the boundary Q3 + 1.5 x IQR. **(2)**

(c) On the grid provided, draw a box plot for the 14 non-outlier heights (that is, excluding the 210 cm sunflower). Plot the 210 cm sunflower separately, as an isolated point beyond the whisker. **(3)**

(Total for Question 2 is 8 marks)

3 A machine operator records the diameter, in mm, of 15 ball bearings and calculates two summary totals from her spreadsheet: the sum of the diameters, sum of x = 309.6, and the sum of the squared diameters, sum of x^2 = 6476.86.

(a) Using sum of x = 309.6, work out the mean diameter as originally recorded. **(1)**

.....

(b) A check of the machine's calibration log shows that one ball bearing's diameter was recorded as 29.6 mm, but should have been 19.6 mm (a digit was mistyped). Work out the corrected value of sum of x. **(2)**

.....

(c) Work out the corrected value of sum of x^2. **(2)**

.....

(d) Hence work out the corrected mean and standard deviation of the 15 diameters. Give the standard deviation correct to 2 decimal places. You may use the formula standard deviation = square root of [(sum of x^2)/n - mean^2]. **(3)**

.....

(e) Given that the standard deviation calculated from the original (uncorrected) sum of x and sum of x^2 was 2.40 mm (2 dp), comment on the effect that correcting the single misrecorded value had on the standard deviation. **(1)**

(Total for Question 3 is 9 marks)

Mark scheme · H01-30min Data Cleaning and Outliers (Higher): 30-minute Pack

Question 1

- (a) **B1** advantage: keeps the sample size at 10 / keeps some information about parcel 3, so later calculations can still use all the deliveries oe
- (a) **B1** disadvantage: the imputed value is not a genuine measurement and may not reflect parcel 3's true delivery time, especially if it was unusual oe
- **(a) Answer: Advantage: keeps the full sample of 10 parcels for later calculations. Disadvantage: the imputed value is not a real measurement and may not reflect what actually happened to parcel 3.**
- (b) **B1** 145 minutes might reveal a genuine problem, e.g. a lost or delayed parcel, which is valuable information for the business oe
- (b) **B1** deleting or replacing it straight away would hide this problem and could make future delivery-time estimates unrealistically low/optimistic oe
- **(b) Answer: 145 minutes could reveal a genuine delivery problem (e.g. a lost parcel); deleting or replacing it immediately would hide this and could make future delivery-time estimates too optimistic.**
- (c) **M1** 18+22+19+145+21+20+17+23+19, oe (=304), divided by 9
- (c) **A1** 33.8 (awrt 33.8) cao
- **(c) Answer: 33.8 minutes (awrt)**
- (d) **B1** no - because parcel 5's extreme 145 minutes was included in the calculation, it pulls the mean far higher than a typical delivery (all other times are between 17 and 23 minutes), so 33.8 minutes is not realistic; parcel 5 should be resolved before imputing parcel 3 oe
- **(d) Answer: No; including the extreme 145-minute delivery pulls the imputed value up to 33.8 minutes, far higher than a typical delivery time (17 to 23 minutes), so it is not a sensible estimate.**

Question 2

- (a) **B1** median = 156 cao
- (a) **B1** Q1 = 150 cao
- (a) **B1** Q3 = 163 cao
- **(a) Answer: Median = 156 cm, Q1 = 150 cm, Q3 = 163 cm**
- (b) **M1** IQR = 13 and 163 + 1.5x13, oe
- (b) **A1** 182.5, and 210 > 182.5 so it is an outlier, cao
- **(b) Answer: 182.5 cm; since 210 > 182.5, it is an outlier.**
- (c) **B1** box drawn correctly from Q1 = 150 to Q3 = 163 with the median line at 156
- (c) **B1** whiskers drawn correctly to the minimum (142) and to the maximum of the remaining values (168)
- (c) **B1** the 210 cm sunflower plotted as a separate point, clearly apart from the whisker
- **(c) Answer: Box from 150 to 163 with median line at 156; whiskers to 142 and 168; the 210 cm sunflower plotted as a separate point.**

Question 3

- (a) **B1** 20.64 mm cao
- **(a) Answer: 20.64 mm.**
- (b) **M1** 309.6 - 29.6 + 19.6, oe
- (b) **A1** 299.6 cao
- **(b) Answer: sum of x = 299.6.**
- (c) **M1** 6476.86 - 29.6^2 + 19.6^2, oe
- (c) **A1** 5984.86 cao
- **(c) Answer: sum of x^2 = 5984.86.**
- (d) **M1** mean = 299.6/15, oe
- (d) **A1** 19.97 (awrt 19.97) cao
- (d) **A1** SD = 0.24 (awrt 0.24) cao
- **(d) Answer: Mean = 19.97 mm (awrt), standard deviation = 0.24 mm (awrt).**
- (e) **B1** correcting the value reduced the standard deviation hugely, from about 2.40 mm to about 0.24 mm (roughly one tenth of its original size), showing how much a single incorrect entry can inflate a standard deviation oe
- **(e) Answer: The standard deviation fell dramatically, from about 2.40 mm (using the recorded value) to about 0.24 mm (corrected), about one tenth the size, showing how much one wrong entry can inflate the spread.**