Question 1

(Edexcel 6683, Jan 2010, Q2)

Worked Solution

19 employees; stem-and-leaf: 0|7, 1|88, 2|4468, 3|2333459, 4|00000.

(a) Median score

$n = 19$; median is the 10th value. Ordered data: 7, 18, 18, 24, 24, 26, 28, 32, 33, **33**, 33, 34, 35, 39, 40, 40, 40, 40, 40.

Median = 33.

(b) Interquartile range

Q_1: average of 5th and 6th values $= (24 + 26)/2 = 25$; or using $\lceil n/4 \rceil$ = 5th value $= 24$.

Q_3: average of 14th and 15th values $= (35 + 40)/2 = 37.5$; or 15th value $= 40$.

Using the mark-scheme standard: $Q_1 = 24$, $Q_3 = 40$, IQR $= 16$.

$Q_1 = 24$, $Q_3 = 40$, IQR $= 16$.

(c) Only one employee will undergo retraining

Outlier threshold: $Q_1 - 1.0 \times \text{IQR} = 24 - 16 = 8$.

Only the value 7 is below 8. No other value is below 8 (the next lowest is 18).

The only value below the outlier threshold of $Q_1 - \text{IQR} = 24 - 16 = 8$ is the score of 7, so only one employee will undergo retraining.

(d) Box plot

Box from $Q_1 = 24$ to $Q_3 = 40$, median line at 33. Outlier at 7 marked with $\times$. Lower whisker extends to 18 (the lowest non-outlier value). Upper whisker to 40 (maximum $= Q_3$, no upper outliers).

Question 2

(Edexcel 6683, Jun 2007, Q5a–e)

Worked Solution

Histogram of swimming times t (minutes). Bars: $[5,10)$: fd=2; $[10,14)$: fd=4; $[14,18)$: fd=6; $[18,25)$: fd=5; $[25,40)$: fd=1.

(a) Complete the frequency table

Frequencies = fd × class width:

$$[5,10): 2\times 5 = 10\checkmark, \quad [10,14): 4\times 4 = 16\checkmark, \quad [14,18): 6\times 4 = 24\checkmark$$

$$[18,25): 5\times 7 = 35, \quad [25,40): 1\times 15 = 15$$

$[18,25)$: frequency = 35; $[25,40)$: frequency = 15.

(b) Number who took longer than 20 minutes

From $[18,25)$: the portion $[20,25)$ has width 5 out of 7, so frequency $= \frac{5}{7}\times 35 = 25$. All of $[25,40)$: 15.

$25 + 15 = 40$ people took longer than 20 minutes.

(c) Estimated mean

Total $n = 10+16+24+35+15 = 100$. Midpoints: 7.5, 12, 16, 21.5, 32.5.

$$\bar{t} = \frac{10(7.5)+16(12)+24(16)+35(21.5)+15(32.5)}{100} = \frac{75+192+384+752.5+487.5}{100} = \frac{1891}{100} = 18.9$$

Estimated mean = 18.91 minutes.

(d) Standard deviation

$$\sigma = \sqrt{\frac{\sum ft^2}{n} - \bar{t}^2}$$

$$\sum ft^2 = 10(7.5)^2 + 16(12)^2 + 24(16)^2 + 35(21.5)^2 + 15(32.5)^2$$
$$= 562.5 + 2304 + 6144 + 16168.75 + 15843.75 = 41023$$

$$\sigma = \sqrt{\frac{41023}{100} - 18.91^2} = \sqrt{410.23 - 357.59} = \sqrt{52.64} \approx 7.25$$

Standard deviation ≈ 7.26 minutes.

(e) Median and quartiles

$n = 100$. Using $n/2 = 50$ for median.

Cumulative: < 10: 10; < 14: 26; < 18: 50; < 25: 85; < 40: 100.

Q_2 lies at the boundary of $[14, 18)$:

$$Q_2 = 14 + \frac{50 - 26}{24} \times 4 = 14 + 4 = 18 \text{ (or } 18.1 \text{ using } n + 1)$$

Q_1: $n/4 = 25$. Lies in $[10, 14)$:

$$Q_1 = 10 + \frac{25 - 10}{16} \times 4 = 10 + 3.75 = 13.75$$

Q_3: $3n/4 = 75$. Lies in $[18, 25)$:

$$Q_3 = 18 + \frac{75 - 50}{35} \times 7 = 18 + 5 = 23$$

$Q_1 \approx 13.75$ min, $Q_2 = 18$ min, $Q_3 = 23$ min.

Question 3

(Edexcel 6683, Jun 2012, Q5)

Worked Solution

450 cars were recorded on a road with a 30 mph speed limit. A histogram of speed s (mph) is given.

Key fact from the histogram

$$\boxed{\begin{array}{c}\text{Total area} = 22.5 \text{ large squares} \\ \Rightarrow 1 \text{ large square} = \dfrac{450}{22.5} = 20 \text{ cars}\end{array}}$$

(a) Cars exceeding the speed limit by at least 5 mph

We require:

$$s > 35$$

So we use the classes:

$$[35, 40), \quad [40, 45), \quad [45, 50)$$

Class	Area (large squares)	Frequency (cars)
$[35, 40)$	4.5	$4.5 \times 20 = 90$
$[40, 45)$	2	$2 \times 20 = 40$
$[45, 50)$	small	negligible

$\boxed{90}$ cars were exceeding the speed limit by at least 5 mph.

(b) Estimated mean speed

Use:

$$\bar{s} = \frac{\sum fx}{\sum f}$$

From the histogram (using midpoints and frequencies):

$$\sum fx = 12975, \qquad \sum f = 450$$

$$\bar{s} = \frac{12975}{450} = 28.8\overline{3}$$

Estimated mean speed $\approx$ **28.8** mph.

(c) Estimated median speed

$$n = 450 \quad \Rightarrow \quad \frac{n}{2} = 225$$

From cumulative frequency, the median lies in:

$$[25, 30)$$

Using linear interpolation:

$$Q_2 = 25 + \frac{225 - 180}{240} \times 5$$

$$Q_2 \approx 28.1 \text{ mph}$$

Estimated median $\approx$ **28.1** mph.

Question 4

(OCR 4732, Jun 2005, Q5)

Worked Solution

Cumulative frequency graph for 1200 candidates' marks.

(i) Interquartile range

Reading from the graph at cumulative frequencies 300 (Q_1) and 900 (Q_3): $Q_1 \approx 45$, $Q_3 \approx 69$ (reading to nearest mark).

IQR $= Q_3 - Q_1 \approx 69 - 45 = 24$ marks (accept 23–25).

(ii) x if 40% scored more than x marks

40% scored more, so 60% scored $\leq x$: cumulative frequency $= 0.6 \times 1200 = 720$.

Reading from the graph at cf $= 720$: $x \approx 63$.

$x \approx 63$ marks.

(iii) Number who scored more than 68 marks

From the graph at mark $= 68$: cf ≈ 860, so number above $= 1200 - 860 = 340$.

Approximately 340 candidates scored more than 68 marks.

(iv) P(all five scored more than 68 marks)

$p = 340/1200 \approx 0.2833$.

$$P(\text{all five}) = \left(\frac{340}{1200}\right)^5 \approx 0.2833^5 \approx 0.00183$$

$P \approx 0.00183$.

(v) Effect of the additional information on the IQR estimate

If marks in 35–55 are evenly distributed, the cumulative frequency curve in that region is a straight line, which is how it was assumed when reading off quartiles. The estimate of Q_1 from the graph was already based on linear interpolation in that range, so the IQR estimate is likely to be accurate (or slightly too low/high depending on exact reading).

The IQR estimate should be approximately correct (or may be slightly too low), since the marks in the range 35–55 are evenly spread, which is consistent with the linear interpolation assumption used when reading from the cumulative frequency curve.

Question 5

(OCR 4732, Jun 2008, Q6)

Worked Solution

Year 12: 128 males (sector 120°), remaining females.

(i)(a) Number of females in Year 12

Males occupy 120° out of 360°, so $\frac{120}{360} = \frac{1}{3}$ are male.

$\frac{1}{3}$ of total $= 128 \Rightarrow$ total $= 384$. Females $= 384 - 128 = 256$.

Number of females in Year 12 = 256.

(i)(b) Why Year 13 may not have more males than Year 12

A pie chart shows proportions, not totals. Year 13 may have a smaller total number of students, so although males represent a larger proportion (150° out of 360°), the actual number of males could still be less than 128.

(ii)(a) Comment on the statement about the pie chart and box plots

The student's statement is incorrect. The pie chart shows that females make up the larger proportion of Year 12 (since their sector is larger than the males' 120° sector). The box plots show the distribution of marks, not the number of students — the box plots are consistent with the pie chart. There is no contradiction.

(ii)(b) Two comparisons between female and male performance

(1) Females generally achieved higher marks: the median for females is higher than the median for males.
(2) The female marks are less spread out: the IQR for females is smaller than for males.

(ii)(c) Advantage and disadvantage of box-and-whisker plots vs histograms

Advantage: Box plots clearly show the median, quartiles, and range, and are easy to compare two datasets side-by-side.
Disadvantage: Box plots do not show individual data values or the exact shape/distribution of the data (e.g. multimodality); histograms convey more information about the distribution.

(iii) Combined mean for all 128 males

$\bar{x}_{\text{all}} = \frac{102\times51+26\times59}{128} = \frac{5202+1534}{128} = \frac{6736}{128} = 52.625$.

Mean mark for all 128 males $= 52.625 \approx 52.6$.

Question 6

(OCR 4732, Jan 2009, Q5i–iii)

Worked Solution

Stem-and-leaf of 23 plum masses (g): stems 5,6,7,8,9.

(i) Median and IQR

$n = 23$; median is the 12th value. Reading from the diagram in order: the data contains values in the 50s (6 values: 55,56,57,58,58,59), 60s (8 values), 70s (9 values), 80s (1 value), 90s (1 value).

Ordered data count: 12th value is in the 60s. Listing: 55,56,57,58,58,59,61,62,63,65,66,68,69, ... The 12th value = 68.

Q_1: 6th value = 59; Q_3: 18th value = 75.

Median = 68 g; IQR = $Q_3 - Q_1 = 75 - 59 = 16$ g.

(ii) One advantage of IQR over standard deviation

The IQR is not affected by outliers or extreme values, whereas the standard deviation is sensitive to extreme values.

(iii) Advantage and disadvantage of stem-and-leaf vs box-and-whisker plot

Advantage: A stem-and-leaf diagram shows every individual data value and retains the original data; the mode and exact frequencies can be identified.
Disadvantage: Harder to read the median and quartiles directly compared to a box plot; two datasets are harder to compare side-by-side.

Question 7

(OCR 4732, Jun 2014, Q1)

Worked Solution

Back-to-back stem-and-leaf for tree heights (metres, nearest 0.1 m). Species A: $n = 21$.

(i) Median and IQR for species A

Ordered data (from stem-and-leaf):

5.9, 6.1, 6.4, 6.5, 6.5, 6.9, 7.2, 7.3, 7.3, 7.3, **7.4**, 7.5, 7.6, 7.6, 7.6, 7.7, 7.8, 8.0, 8.3, 8.4, 8.5

Median

$$n = 21 \quad \Rightarrow \quad \text{median is the 11th value}$$

$$\text{Median} = 7.4 \text{ m}$$

Quartiles

$$Q_1 = 6\text{th value} = 6.9 \qquad Q_3 = 16\text{th value} = 7.7$$

Alternatively (using interpolation):

$$Q_3 = \frac{7.7 + 7.8}{2} = 7.75$$

Interquartile Range

$$\text{IQR} = Q_3 - Q_1 = 7.75 - 6.9 = 0.85 \text{ m}$$

Median = 7.4 m (or 7.45 m if interpolated)
IQR $\approx$ 0.85 m (accept answers in the range 0.85–1.18 m)

(ii) Back-to-back stem-and-leaf for species B

Species B heights: 7.6, 5.2, 8.5, 5.2, 6.3, 6.3, 6.8, 7.2, 6.7, 7.3, 5.4, 7.5, 7.4, 6.0, 6.7.

Arranged on the left side of the same stems 5–8:

B	stem	A
4 2 2	5	9
8 7 7 3 3 0	6	1 4 5 5 9
6 5 4 3 2	7	2 3 3 3 4 5 6 6 6 7 8
5	8	0 3 4
		5

Key: 8|6|4 means 6.8 (B) and 6.4 (A).

(iii) Two comparisons

(1) Species A has a higher median height than species B, so species A trees tend to be taller on average.
(2) Species A has a larger IQR (more spread), suggesting greater variation in height; species B heights are more consistent.

Question 8

(OCR 4732, Jan 2011, Q1)

Worked Solution

200 candidates, two papers. Cumulative frequency graphs given.

(i) Median for each paper

Read off at cumulative frequency $= 100$.

From the graph: Paper 1 median ≈ 38; Paper 2 median ≈ 61.

Paper 1 median ≈ 38 marks; Paper 2 median ≈ 61 marks.

(ii) Which paper was easier?

Paper 2 was easier, since its median (≈ 61) is higher than Paper 1's median (≈ 38), indicating that students scored higher marks on Paper 2 on average.

(iii) IQRs and comment on variation

Paper 1: $Q_1 \approx 25$, $Q_3 \approx 55$, IQR ≈ 30.

Paper 2: $Q_1 \approx 46$, $Q_3 \approx 73$, IQR ≈ 27.

Paper 2 has a smaller IQR, so Paper 2 marks are slightly less varied. The suggestion is supported (just).

Paper 1 IQR ≈ 30; Paper 2 IQR ≈ 27. Paper 2 marks are slightly less varied, so the suggestion is (just) supported.

(iv) Number gaining grade A on Paper 2

Grade A on Paper 1 starts at (median of P1) minus 10: Paper 2 grade A threshold = Paper 1 grade A threshold $+10$.

25 candidates gained grade A on Paper 1, so Paper 1 grade A threshold is at cf $= 175$ (top 25 out of 200). Reading from the graph at cf $= 175$: mark ≈ 68. Paper 2 grade A threshold $= 68 + 10 = 78$.

Reading Paper 2 cf at 78: cf ≈ 163, so number above $= 200 - 163 = 37$.

Approximately 37 candidates gained grade A on Paper 2.

(v) New mean and standard deviation after adding 1 mark to all Paper 1 scores

New mean $= 36.5 + 1 = 37.5$. Standard deviation remains 28.2 (adding a constant to all values does not change the spread).

Question 9

(OCR 4732, Jan 2012, Q5)

Worked Solution

Hours of sunshine at a resort over 21 days. Table: 0: 0 days; 1–3: 6 days; 4–6: 9 days; 7–9: 4 days; 10–15: 2 days.

(i)(a) Frequency density of the 1–3 class

Class width for 1–3 measured to nearest hour: boundaries 0.5 to 3.5, width = 3.

$$\text{fd} = \frac{6}{3} = 2$$

Frequency density of the 1–3 class = 2.

(i)(b) Height of the 10–15 bar (scale: 2 cm to 1 unit of fd)

Class boundaries 9.5 to 15.5, width = 6.

$$\text{fd} = \frac{2}{6} = \frac{1}{3}$$

Height in cm $= \frac{1}{3} \times 2 = \frac{2}{3}$ cm.

Height $= \frac{2}{3}$ cm ≈ 0.667 cm.

(ii) First two points of cumulative frequency graph

The graph is plotted at the upper class boundaries.

(0.5, 0) and (3.5, 6).

(iii)(a) Estimated mean and standard deviation

Midpoints: 0 (unused, 0 days), 2, 5, 8, 12.5.

$$\bar{x} = \frac{0 + 6(2) + 9(5) + 4(8) + 2(12.5)}{21} = \frac{0 + 12 + 45 + 32 + 25}{21} = \frac{114}{21} \approx 5.43$$

$$\sigma = \sqrt{\frac{\sum fx^2}{n} - \bar{x}^2} = \sqrt{\frac{6(4) + 9(25) + 4(64) + 2(156.25)}{21} - 5.43^2}$$

$$= \sqrt{\frac{24 + 225 + 256 + 312.5}{21} - 29.48} = \sqrt{\frac{817.5}{21} - 29.48} = \sqrt{38.93 - 29.48} = \sqrt{9.45} \approx 3.07$$

Estimated mean ≈ 5.43 hours; estimated standard deviation ≈ 3.08 hours.

(iii)(b) Why these are only estimates

Because the actual values within each class are unknown; we use class midpoints to represent all values in a class, which is only an approximation (the data are given in grouped form).

End of Worked Solutions