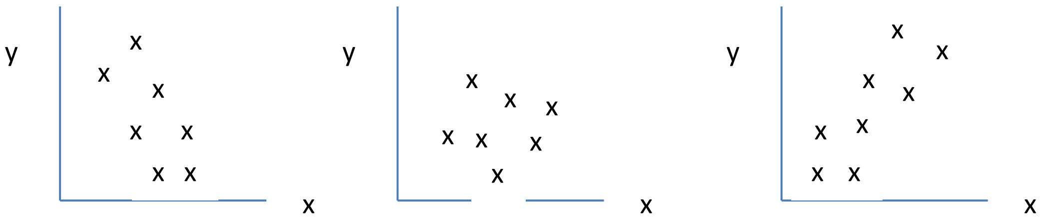
1. The scatter diagrams below were drawn by a student.



The student calculated the value of the product moment correlation coefficient for each of the sets of data.

The values were

0.68 –0.79 0.08

Write down, with a reason, which value corresponds to which scatter diagram.

(6)

1.

Diagram A : r = - 0.79; As x increases, y decreases. Data is negatively correlated.

Diagram B : r = 0.08; No real pattern. Data has nearly no correlation at all

Diagram C : r = 0.68; As x increases, y increases. Data is positively correlated.

2. The following table summarises the distances, to the nearest km, that 134 examiners travelled to attend a meeting in London.

Distance	Number of examiners	Distance of Class boundaries	Class Width	Frequency Density
41-45	4	40.5-45.49	5	0.8
46-50	19	45.5-50.49	5	3.8
51-60	53	50.5-60.49	10	5.3
61-70	37	60.5-70.49	10	3.7
71-90	15	70.5-90.49	20	0.75
91-150	6	90.50-150.49	60	0.1

(a) Give a reason to justify the use of a histogram to represent these data.

(1)

(b) Calculate the frequency densities needed to draw a histogram for these data.

(DO NOT DRAW THE HISTOGRAM)

(2)

(c) Use interpolation to estimate the median *Q*2, the lower quartile *Q*1, and the upper quartile *Q*3 of these data.

(4)

The mid-point of each class is represented by *x* and the corresponding frequency by *f*.

Calculations then give the following values

$$\boldsymbol{\Sigma fx = 8379.5 \text{ and } \Sigma fx^2 = 557489.75}$$

(d) Calculate an estimate of the mean and an estimate of the standard deviation for these data.

(4)

One coefficient of skewness is given by

$$\boldsymbol{\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}}$$

(e) Evaluate this coefficient and comment on the skewness of these data.

(4)

(f) Give another justification of your comment in part (e).

(1)

2 a) A histogram can be used when the data is continuous in this example distance is continuous.

b) Using $F.D = \frac{Freq}{Class\ width}$

As the distance is to the nearest km note that the classes boundaries are shown in the table, along with the class widths. And the calculated F.D.

c) For Q_2 is 134/2 = 67 to work out which class it will fall in. IN this class it is the third class.

$$Q_2 = lower\ class\ boundary + \frac{\frac{n}{2} - \sum f(before\ median\ class)}{f\ of\ median\ class} \times class\ width$$

$$Q_2 = 50.5 + \frac{67-23}{53} \times 10 = 58.8 \text{ (1.d.p)}$$

For Q_1 is 134/4=33.5 which is also class 3.

$$Q_1 = 50.5 + \frac{33.5 - 23}{53} \times 10 = 52.48\ (2.d.p)$$

For Q_3 is 3x134/4=100.5 which is class 4.

$$Q_3 = 60.5 + \frac{100.5 - 76}{37} \times 10 = 67.12\ (2.d.p)$$

d) Using $\bar{x} = \frac{\sum fx}{n}$

$$\bar{x} = \frac{8379.5}{134} = 62.53\ (2.d.p)$$

Using $s = \sqrt{(\frac{\sum fx^2}{n} - \bar{x}^2)}$ | $s = \sqrt{\frac{557489.75}{134} - 62.53^2} = 15.82\ (2.d.p)$

e) Using given formula | $\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{67.1 - 2(58.8) + 52.5}{67.1 - 52.5} = 0.136$

Which is a small +ve skew.

f) For a +ve skew mean>median | $62.53 > 58.8$ which is also true so the data is positively skewed.

3. A long distance lorry driver recorded the distance travelled, *m* miles, and the amount of fuel used, *f* litres, each day. Summarised below are data from the driver's records for a random sample of 8 days.

The data are coded such that *x* = *m* – 250 and *y* = *f* – 100.

$$\Sigma x = 130,\ \Sigma y = 48,\ \Sigma xy = 8880,\ S_{xx} = 20\,487.5$$

(a) Find the equation of the regression line of *y* on *x* in the form *y* = *a* + *bx*. **(6)**

(b) Hence find the equation of the regression line of *f* on *m*. **(3)**

(c) Predict the amount of fuel used on a journey of 235 miles. **(1)**

3 a) For a regression line of this form find b first then a.

First calculate S_{xy}. $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ | $S_{xy} = 8880 - \frac{130 \times 48}{8} = 8100$

Using $b = \frac{S_{xy}}{S_{xx}}$ | $b = \frac{8100}{20487.5} = 0.395363$

The value of a is then given by $a = \bar{y} - b\bar{x}$ | $a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{48}{8} - 0.395363 \times \frac{130}{8} = -0.4246488$

Therefore the regression line is | $y = -0.425 + 0.395x\ \ (3.s.f)$

b) Substitute in *x* = *m* – 250 and *y* = *f* – 100. | $f - 100 = -0.425 + 0.395(m - 250)$

$$f = 0.395363m - (250 \times 0.395363) + 100 - 0.4246488$$

Use all significant figures for calculation.	$f = 0.395m + 0.735\ \ (3.s.f)$
c) Simply find f when m=235. Substitute in	$f = 0.395363 \times (235) + 0.735 = 93.65\ (2.d.p)$

4. Aeroplanes fly from City *A* to City *B*. Over a long period of time the number of minutes delay in take-off from City *A* was recorded. The minimum delay was 5 minutes and the maximum delay was 63 minutes. A quarter of all delays were at most 12 minutes, half were at most 17 minutes and 75% were at most 28 minutes. Only one of the delays was longer than 45 minutes.

An outlier is an observation that falls either 1.5× (interquartile range) above the upper quartile or 1.5× (interquartile range) below the lower quartile.

(a) On the graph paper opposite draw a box plot to represent these data.

(7)

(b) Comment on the distribution of delays. Justify your answer.

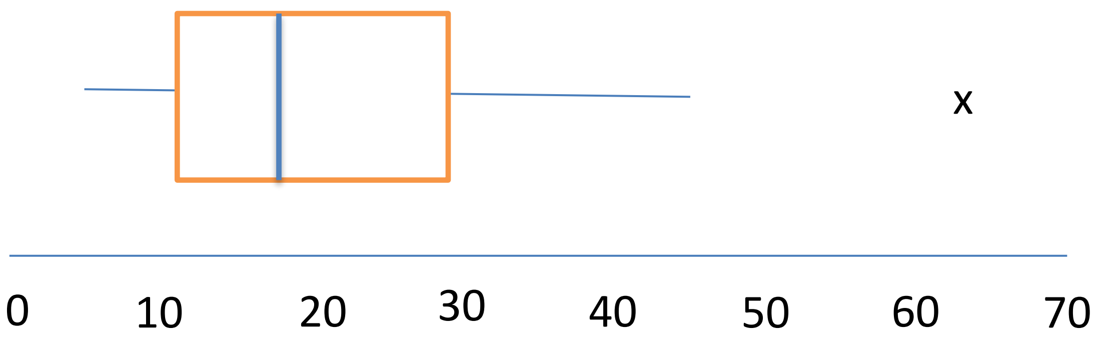
(2)

(c) Suggest how the distribution might be interpreted by a passenger who frequently flies from City *A* to City *B*.

(1)

4 a) The information given is such that the range is 5 mins to 63 mins. $Q_1 = 12, Q_2 = 17, Q_3 = 28.$

Work out outliers	$1.5(Q_3 - Q_1) = 1.5\,(28 - 12) = 24$
Upper region is above	$Q_3 + 24 = 52\quad 63\ is\ an\ outlier$
Lower region is below	$Q_1 - 24 = -12\quad no\ outliers$



b) This is asking for a comment on skewness. Carry out $Q_3 - Q_2 = 11$ and $Q_2 - Q_1 = 5$ and as $Q_3 - Q_2 > Q_2 - Q_1$ the data is positively skewed. This can clearly be seen by the the box plot i.e. the line is left of centre.

c) The data is positively skewed so most delays are small and there are infrequent longer delays. Most passengers would be relatively happy with that.

5. The random variable *X* has probability function where *k* is a constant.

$$P(X = x) = \begin{cases} kx & x = 1, 2, 3 \\ k(x+1) & x = 4, 5 \end{cases}$$

(a) Find the value of *k*.

(2)

(b) Find the exact value of E(*X*).

(2)

(c) Show that, to 3 significant figures, Var(*X*) = 1.47.

(4)

(d) Find, to 1 decimal place, Var(4 – 3*X*).

(2)

5 a) Fill out a probability table

x	1	2	3	4	5
P(X=x)	k	2k	3k	5k	6k

All the probabilities must equal 1. Therefore

$k + 2k + 3k + 5k + 6k = 17k = 1 \qquad k = \frac{1}{17}$

b) Using $E(X) = \sum xP(X = x)$

$E(X) = 1k + 2 \times 2k + 3 \times 3k + 4 \times 5k + 5 \times 6k$

$E(X) = k + 4k + 9k + 20k + 30k = 64k = \frac{64}{17}$

c) Using $Var(X) = E(X^2) - (E(X))^2$

$E(X^2) = 1k + 4 \times 2k + 9 \times 3k + 16 \times 5k + 25 \times 6k$

$E(X^2) = k + 8k + 27k + 80k + 150k = 266k = \frac{266}{17}$

$Var(X) = \frac{266}{17} - (\frac{64}{17})^2 = 1.474\ (3.d.p)$

d) Using $Var(aX + b) = a^2 Var(X)$

$Var(4 - 3X) = 9Var(X) = 9 \times 1.47 = 13.2\ (1.d.p)$

6. A scientist found that the time taken, *M* minutes, to carry out an experiment can be modelled by a normal random variable with mean 155 minutes and standard deviation 3.5 minutes.
Find
(a) $P(M > 160)$.

(3)

(b) $P(150 \le M \le 157)$.

(4)

(c) the value of *m*, to 1 decimal place, such that $P(M \le m) = 0.30$.

(4)

6 a) Normal distribution can be written as $M \sim N(155, 3.5^2)$

For $P(M > 160) = 1 - P(M < 60)$

$$= 1 - P\left(z < \frac{160 - 155}{3.5}\right) = 1 - P(z < 1.43)$$

Read $z < 1.43$ from tables

$$= 1 - 0.9236 = 0.0764$$

b)

$P(M < 150) = P(M > 160)$ in a)

$$P(150 \le M \le 157) = P(M < 157) - P(M < 150)$$

$$= P\left(z < \frac{157 - 155}{3.5}\right) - 0.0764$$

$$= P(z < 0.57) - 0.0764$$

$$= 0.7157 - 0.0764 = 0.6393$$

c) Working backwards with 0.3 – read z from tables

$$P(M \le m) = 0.30 \qquad P\left(z < \frac{X - 155}{3.5}\right) = 0.5244$$

$$z = -0.524 \times 3.5 + 155 = 153.2$$

7. In a school there are 148 students in Years 12 and 13 studying Science, Humanities or Arts subjects. Of these students, 89 wear glasses and the others do not. There are 30 Science students of whom 18 wear glasses. The corresponding figures for the Humanities students are 68 and 44 respectively.

A student is chosen at random.

Find the probability that this student

(a) is studying Arts subjects,

(4)

(b) does not wear glasses, given that the student is studying Arts subjects.

(2)

Amongst the Science students, 80% are right-handed. Corresponding percentages for Humanities and Arts students are 75% and 70% respectively.

A student is again chosen at random.

(c) Find the probability that this student is right-handed.

(3)

(d) Given that this student is right-handed, find the probability that the student is studying Science subjects.

(3)

7 a) Fill out figures into a table as follows:

	Glasses	No Glasses	Total
Science	18	(30-18)= 12	**30**
Arts	(89-18-44)=27	(59-12-24)=23	**(148-68-30)=50**
Humanities	44	(68-44)=24	**68**
Total	**89**	**(148-89)=59**	**148**

7a) Is studying arts is therefore $\frac{50}{148} = 0.3378$

b) As it is from arts it is out of 50 and does not wear glasses is therefore 23 $\frac{23}{50} = 0.46$

c) Probability right handed is

$$= 0.8 \times \frac{30}{148} + 0.75 \times \frac{68}{148} + 0.7 \times \frac{50}{148}$$
$$= 0.743\ (3.d.p)$$

d) Given that the student is right handed then chances of science is

$$= \frac{0.8 \times \frac{30}{148}}{0.743} = 0.218\ (3.d.p)$$