1. As part of a statistics project, Gill collected data relating to the length of time, to the nearest minute, spent by shoppers in a supermarket and the amount of money they spent. Her data for a random sample of 10 shoppers are summarised in the table below, where *t* represents time and £*m* the amount spent over £20.

t(minutes)	£m
15	-3
23	17
5	-19
16	4
30	12
6	-9
32	27
23	6
35	20
27	6
Total =212	**Total=61**

(a) Write down the actual amount spent by the shopper who was in the supermarket for 15 minutes. **(1)**

(b) Calculate *Stt*, *Smm* and *Stm*.
(You may use $\Sigma t^2 = 5478$ $\Sigma m^2 = 2101$ $\Sigma tm = 2485$)

(6)

(c) Calculate the value of the product moment correlation coefficient between *t* and *m*.

(3)

(d) Write down the value of the product moment correlation coefficient between *t* and the actual amount spent. Give a reason to justify your value.

(2)

On another day Gill collected similar data. For these data the product moment correlation coefficient was 0.178

(e) Give an interpretation to both of these coefficients. **(2)**
(f) Suggest a practical reason why these two values are so different. **(1)**

1a) m= number over £20

Therefore -3 represents a spend of £17.

b) Using

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{tm} = 2485 - \frac{\sum t \sum m}{n} = 2485 - \frac{212 \times 61}{10} = 1191.8$$

Using

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

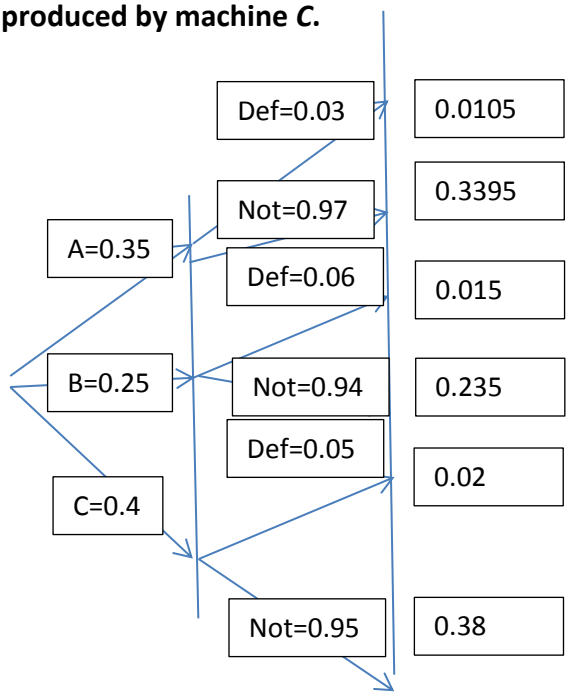
$$S_{tt} = 5478 - \frac{(212)^2}{10} = 983.6$$

$$S_{mm} = 2101 - \frac{(61)^2}{10} = 1728.9$$

c) Using $r = \frac{S_{xy}}{\sqrt{(S_{xx} \times S_{yy})}}$	$r = \frac{1191.8}{\sqrt{983.6 \times 1728.9}} = 0.9139$
d) As the product moment correlation coefficient is a ratio it remains the same.	$r = 0.9139$
e) First study	Strong positive correlation so the longer spent shopping the more spent.
Second study	Weak correlation so same amount spent over different periods of time.
f)	The first shoppers are likely to be doing their weekly shop whereas the second shoppers are likely to be browsing.

2. In a factory, machines *A*, *B* and *C* are all producing metal rods of the same length. Machine *A* produces 35% of the rods, machine *B* produces 25% and the rest are produced by machine *C*. Of their production of rods, machines *A*, *B* and *C* produce 3%, 6% and 5% defective rods respectively.

(a) Draw a tree diagram to represent this information. (3)
(b) Find the probability that a randomly selected rod is
(i) produced by machine *A* and is defective,
(ii) is defective. (5)
(c) Given that a randomly selected rod is defective, find the probability that it was produced by machine *C*. (3)



b) (i) From tree =0.0105

(ii) Add all defective=0.0105+0.015+0.02 =0.0455

c)

$$P(C|D) = \frac{P(C \cap D)}{P(D)}$$

So it is 0.02/0.0455 =0.439=43.9%

3. The random variable *X* has probability function

$$\boldsymbol{P(X = x) = \frac{(2x-1)}{36} \qquad x = 1, 2, 3, 4, 5, 6}$$

(a) Construct a table giving the probability distribution of *X*. (3)

Find

(b) P(2 < *X* ≤ 5), (2)

(c) the exact value of E(*X*). (2)

(d) Show that Var(*X*) = 1.97 to 3 significant figures. (4)

(e) Find Var(2 – 3*X*). (2)

a) Simply put the six values of x into the formula.

x	1	2	3	4	5	6
P(X=x)	$\frac{1}{36}$	$\frac{3}{36} = \frac{1}{12}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36} = \frac{1}{4}$	$\frac{11}{36}$

b) Greater than 2 and up to and equal to 5 is P(X=3)+P(X=4)+P(X=5)

$$= \frac{5}{36} + \frac{7}{36} + \frac{9}{36} = \frac{21}{36} = \frac{7}{12}$$

c)

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

$$E(X) = 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + 3 \times \frac{5}{36} + 4 \times \frac{7}{36} + 5 \times \frac{9}{36} + 6 \times \frac{11}{36}$$

$$E(X) = \frac{1}{36} + \frac{6}{36} + \frac{15}{36} + \frac{28}{36} + \frac{45}{36} + \frac{66}{36} = \frac{161}{36}$$

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

$$E(X^2) = \frac{1}{36} + \frac{12}{36} + \frac{45}{36} + \frac{112}{36} + \frac{225}{36} + \frac{396}{36} = \frac{791}{36}$$

$$Var(X) = \frac{791}{36} - \left(\frac{161}{36}\right)^2 = 1.9714\ (5.s.f)$$

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

$$Var(2 - 3X) = 3^2Var(X) = 9 \times 1.97 = 17.73$$

4. Summarised below are the distances, to the nearest mile, travelled to work by a random sample of 120 commuters.

Distance (to the nearest mile)	Number of Commuters
0-9	10
10-19	19
20-29	43
30-39	25
40-49	8
50-59	6
60-69	5
70-79	3
80-89	1

For this distribution,

(a) describe its shape, **(1)**

(b) use linear interpolation to estimate its median. **(2)**

The mid-point of each class was represented by *x* and its corresponding frequency by *f* giving

$$\sum fx = 3550 \text{ and } \sum fx^2 = 138020$$

(c) Estimate the mean and the standard deviation of this distribution. **(3)**

One coefficient of skewness is given by

$$\frac{3(mean - median)}{standard\ deviation}$$

(d) Evaluate this coefficient for this distribution. **(3)**

(e) State whether or not the value of your coefficient is consistent with your description in part (a). Justify your answer. **(2)**

(f) State, with a reason, whether you should use the mean or the median to represent the data in this distribution. **(2)**

(g) State the circumstance under which it would not matter whether you used the mean or the median to represent a set of data. **(1)**

a) The data is weighted towards the earlier results so is positively skewed.

b) The median will be the 60th commuter

60th falls in the 20-29 class. The first 29 commuters take us to an average of 19.5 then we have to go 31/43 through the next class therefore

$Median = 19.5 + \frac{31}{43}(10) = 26.7093$ = 26.7 (1.d.p)

c) Using

$$\mu = \frac{\sum fx}{n} \qquad \mu = \frac{3550}{120} = 29.6$$

$$\sigma^2 = \frac{\sum f x^2}{n} - \mu^2 \qquad \sigma^2 = \frac{138020}{120} - (29.6)^2 \qquad \sigma = 16.6$$

d)

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(29.6 - 26.7)}{16.6} = 0.520\ (3.d.p)$$

e) This is a positive number and therefore represents a positive skew as in a).

f) For very skewed data use Median as it is less affected by outliers.

g) If there is no skew or if the data is normally distributed the median will equal the mean and it won't make any difference which is used.

5. A teacher recorded, to the nearest hour, the time spent watching television during a particular week by each child in a random sample. The times were summarised in a grouped frequency table and represented by a histogram. One of the classes in the grouped frequency distribution was 20–29 and its associated frequency was 9. On the histogram the height of the rectangle representing that class was 3.6 cm and the width was 2 cm.

(a) Give a reason to support the use of a histogram to represent these data. (1)

(b) Write down the underlying feature associated with each of the bars in a histogram. (1)

(c) Show that on this histogram each child was represented by 0.8 cm^2. (3)

The total area under the histogram was 24 cm^2.
(d) Find the total number of children in the group. (2)

a) Histograms can be used for continuous data such as time and for frequency grouped data as in the question.

b) The area of the histogram bar is proportional to the frequency of that class.

c) $Area = 3.6 \times 2 = 7.2cm^2\ which\ represents\ 9\ kids$

Each child is therefore represented by= $\frac{7.2}{9} = 0.8cm^2$

d) Total number $=\frac{24}{0.8} = 30$ children.

6. (a) Give two reasons to justify the use of statistical models. (2)

It has been suggested that there are 7 stages involved in creating a statistical model. They are summarised below, with stages 3, 4 and 7 missing.
Stage 1. The recognition of a real-world problem.
Stage 2. A statistical model is devised.
Stage 3.
Stage 4.
Stage 5. Comparisons are made against the devised model.
Stage 6. Statistical concepts are used to test how well the model describes the real-world problem.
Stage 7.
(b) Write down the missing stages. (3)

a) Using statistical models of real world problems tend to be simpler, cheaper and quicker to use. They can be more easily modified to new situations and allow us to predict further outcomes.

b) Stage 3: Model used to make predictions
Stage 4: Data collected
Stage 7: Model is refined

7. The measure of intelligence, IQ, of a group of students is assumed to be Normally distributed with mean 100 and standard deviation 15.

(a) Find the probability that a student selected at random has an IQ less than 91. (4)

The probability that a randomly selected student has an IQ of at least 100 + *k* is 0.2090.

(b) Find, to the nearest integer, the value of *k*.

a) $N(\mu, \sigma^2)$ $N(100, 15^2)$

91 is less than 100 and for normal distribution we need greater than 100. But the probability of less than 91 is the same as the probability of greater than 109. Which is 1 minus probability of less than 109. Therefore

$P(X < 91)$ $= 1 - P\left(Z < \frac{109 - 100}{15}\right)$
$= 1 - P(Z < 0.6)$

Read from tables $= 1 - 0.7257 = 0.2743$

b) At least 100+k means greater than or equal to 100+k

$P(X > 100 + k) = 1 - P(X < 100 + k) = 0.2090$

$P(X < 100 + k) = 0.791 \quad z = 0.81$

$\frac{100 + k - 100}{15} = 0.81 \qquad k = 12.$