Paper Reference(s)

6683/01

Edexcel GCE

Statistics S1

Gold Level G2

Time: 1 hour 30 minutes

Materials required for examination papers	**Items included with question**
Mathematical Formulae (Green)	Nil

Candidates may use any calculator allowed by the regulations of the Joint Council for Qualifications. Calculators must not have the facility for symbolic algebra manipulation, differentiation and integration, or have retrievable mathematical formulas stored in them.

Instructions to Candidates

Write the name of the examining body (Edexcel), your centre number, candidate number, the unit title (Statistics S1), the paper reference (6683), your surname, initials and signature.

Information for Candidates

A booklet 'Mathematical Formulae and Statistical Tables' is provided.
Full marks may be obtained for answers to ALL questions.
There are 7 questions in this question paper. The total mark for this paper is 75.

Advice to Candidates

You must ensure that your answers to parts of questions are clearly labelled.
You must show sufficient working to make your methods clear to the Examiner. Answers without working may gain no credit.

Suggested grade boundaries for this paper:

A*	A	B	C	D	E
59	52	45	38	32	26

1. On a particular day the height above sea level, x metres, and the mid-day temperature, y °C, were recorded in 8 north European towns. These data are summarised below

$$S_{xx} = 3\,535\,237.5 \qquad \sum y = 181 \qquad \sum y^2 = 4305 \qquad S_{xy} = -23\,726.25$$

(*a*) Find S_{yy} .

(2)

(*b*) Calculate, to 3 significant figures, the product moment correlation coefficient for these data.

(2)

(*c*) Give an interpretation of your coefficient.

(1)

A student thought that the calculations would be simpler if the height above sea level, h, was measured in kilometres and used the variable $h = \frac{x}{1000}$ instead of x.

(*d*) Write down the value of S_{hh}.

(1)

(*e*) Write down the value of the correlation coefficient between h and y.

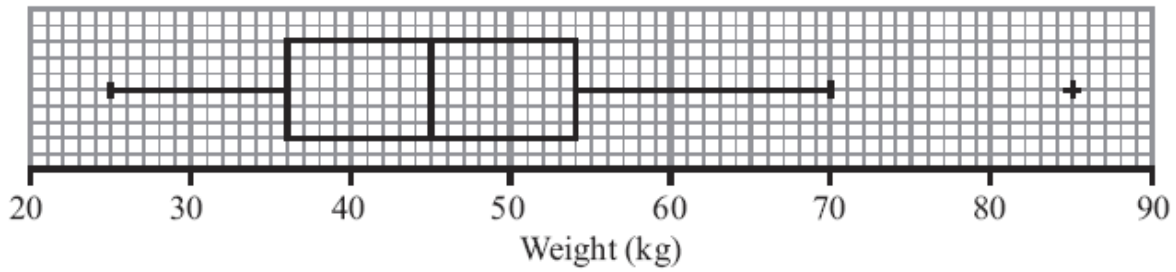
(1)

May 2011

---

2. The box plot in Figure 1 shows a summary of the weights of the luggage, in kg, for each musician in an orchestra on an overseas tour.

Figure 1



The airline's recommended weight limit for each musician's luggage was 45 kg.

Given that none of the musician's luggage weighed exactly 45 kg,

(*a*) state the proportion of the musicians whose luggage was below the recommended weight limit.

(1)

A quarter of the musicians had to pay a charge for taking heavy luggage.

(*b*) State the smallest weight for which the charge was made.

(1)

(*c*) Explain what you understand by the + on the box plot in Figure 1, and suggest an instrument that the owner of this luggage might play.

(2)

(*d*) Describe the skewness of this distribution. Give a reason for your answer.

(2)

One musician of the orchestra suggests that the weights of the luggage, in kg, can be modelled by a normal distribution with quartiles as given in Figure 1.

(*e*) Find the standard deviation of this normal distribution.

(4)

June 2007

---

3. The histogram in Figure 1 shows the time taken, to the nearest minute, for 140 runners to complete a fun run.

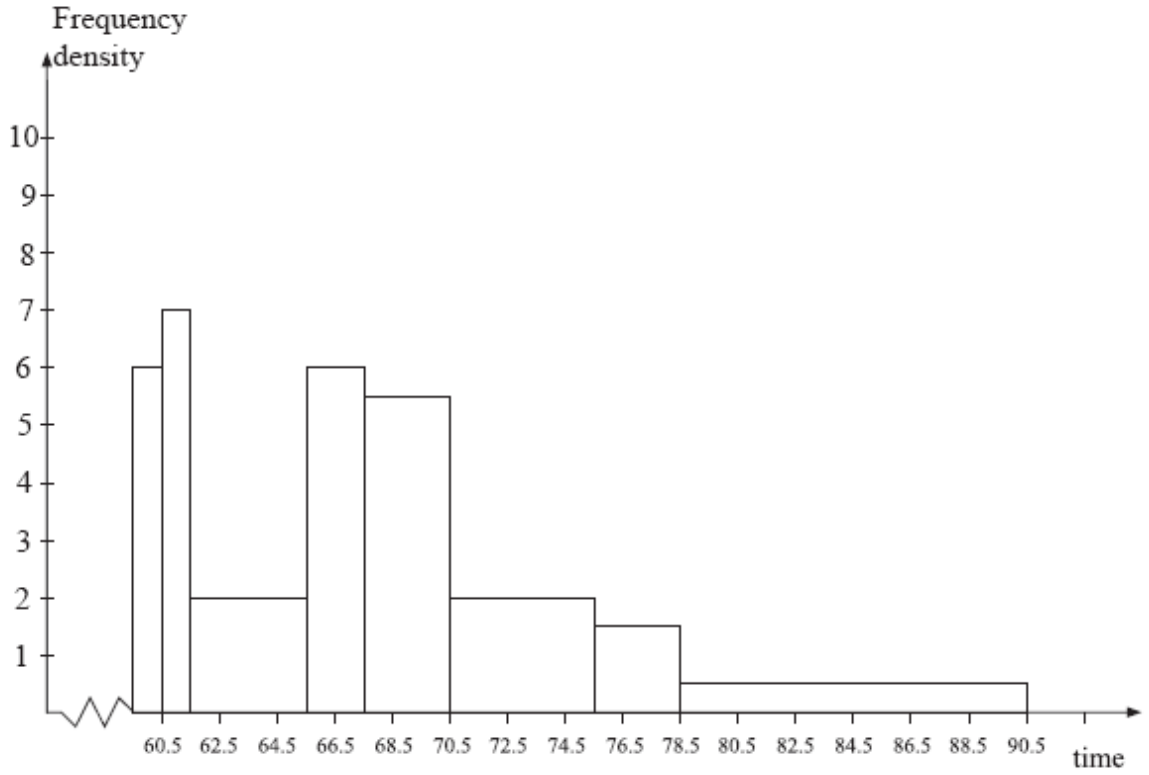


Figure 1

Use the histogram to calculate the number of runners who took between 78.5 and 90.5 minutes to complete the fun run.

(5)

January 2008

4. The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\sum ft^2 = 134281.25$]

(a) Estimate the mean and standard deviation of these data.

(5)

(b) Use linear interpolation to estimate the value of the median.

(2)

(c) Show that the estimated value of the lower quartile is 18.6 to 3 significant figures.

(1)

(d) Estimate the interquartile range of this distribution.

(2)

(e) Give a reason why the mean and standard deviation are not the most appropriate summary statistics to use with these data.

(1)

The person timing the exam made an error and each student actually took 5 minutes less than the times recorded above. The table below summarises the actual times.

Time (minutes) t	6 – 15	16 – 20	21 – 25	26 – 30	31 – 40	41 – 55
Number of students f	62	88	16	13	11	10

(f) Without further calculations, explain the effect this would have on each of the estimates found in parts (a), (b), (c) and (d).

(3)

May 2013

5. A researcher measured the foot lengths of a random sample of 120 ten-year-old children. The lengths are summarised in the table below.

Foot length, *l*, (cm)	**Number of children**
$10 \leq l < 12$	5
$12 \leq l < 17$	53
$17 \leq l < 19$	29
$19 \leq l < 21$	15
$21 \leq l < 23$	11
$23 \leq l < 25$	7

(*a*) Use interpolation to estimate the median of this distribution.

(2)

(*b*) Calculate estimates for the mean and the standard deviation of these data.

(6)

One measure of skewness is given by

$$\text{Coefficient of skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

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

(3)

Greg suggests that a normal distribution is a suitable model for the foot lengths of ten-year-old children.

(*d*) Using the value found in part (*c*), comment on Greg's suggestion, giving a reason for your answer.

(2)

May 2009

6.

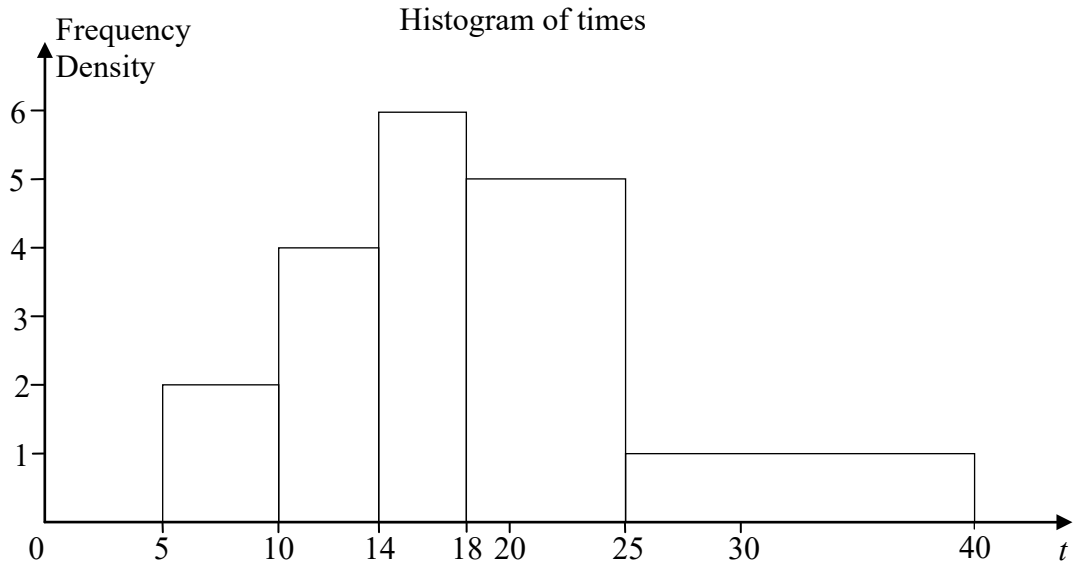


Figure 2

Figure 2 shows a histogram for the variable t which represents the time taken, in minutes, by a group of people to swim 500 m.

(a) Copy and complete the frequency table for t.

t	5 – 10	10 – 14	14 – 18	18 – 25	25 – 40
Frequency	10	16	24		

(2)

(b) Estimate the number of people who took longer than 20 minutes to swim 500 m.

(2)

(c) Find an estimate of the mean time taken.

(4)

(d) Find an estimate for the standard deviation of t.

(3)

(e) Find the median and quartiles for t.

(4)

One measure of skewness is found using $\dfrac{3(\text{mean} - \text{median})}{\text{standard deviation}}$.

(f) Evaluate this measure and describe the skewness of these data.

(2)

May 2007

7. The weights of bags of popcorn are normally distributed with mean of 200 g and 60% of all bags weighing between 190 g and 210 g.

(*a*) Write down the median weight of the bags of popcorn.

(1)

(*b*) Find the standard deviation of the weights of the bags of popcorn.

(5)

A shopkeeper finds that customers will complain if their bag of popcorn weighs less than 180 g.

(*c*) Find the probability that a customer will complain.

(3)

January 2008

TOTAL FOR PAPER: 75 MARKS

END

Question Number	Scheme	Marks
1. (a)	$S_{yy} = 4305 - \frac{181^2}{8}$	M1
	$= \underline{209.875}$ **(awrt 210)**	A1
		(2)
(b)	$r = \frac{(-)23726.25}{\sqrt{3535237.5 \times "209.875"}}$	M1
	$= \underline{-0.87104\ldots}$ **(awrt −0.871)**	A1
		(2)
(c)	Higher towns have lower temperature or temp. decreases as height increases	B1
		(1)
(d)	$S_{hh} = 3.5352375$ (**awrt 3.54**) (condone 3.53)	B1
		(1)
(e)	$r = \underline{-0.87104\ldots}$ **(awrt −0.871)**	B1ft
		(1)
		[7]
2. (a)	$\frac{1}{2}$	B1
		(1)
(b)	54	B1
		(1)
(c)	+ is an 'outlier' or 'extreme value'	B1
	Any heavy musical instrument or a statement that the instrument is heavy	B1
		(2)
(d)	$Q_3 - Q_2 = Q_2 - Q_1$	B1
	so symmetrical or no skew	B1
		(2)
(e)	$P(W < 54) = 0.75$ (or $P(W > 54) = 0.25$)	M1
	$\frac{54-45}{\sigma} = 0.67$	M1B1
	$\sigma = 13.43.....$	A1
		(4)
		[10]

Question Number	Scheme	Marks
3.	<table><tr><td>Width</td><td>1</td><td>1</td><td>4</td><td>2</td><td>3</td><td>5</td><td>3</td><td>12</td></tr><tr><td>Freq. Density</td><td>6</td><td>7</td><td>2</td><td>6</td><td>5.5</td><td>2</td><td>1.5</td><td>0.5</td></tr></table>	M1
	0.5 × 5 12 or 6	A1
	Total area is (1 × 6) + (1 × 7) + (4 × 2) + …, = 70	
	$(90.5-78.5)\times\frac{1}{2}\times\frac{140}{\text{their}70}$	M1
	"70 seen anywhere"	B1
	Number of runners is 12	A1 **[5]**
4. (a)	$\sum ft = 4837.5$ (allow 4838 or 4840)	B1
	Mean = $\frac{"4837.5"}{200} = 24.1875$ awrt **<u>24.2</u>** or $\frac{387}{16}$	M1 A1
	$\sigma = \sqrt{\frac{134281.25}{200} - \left(\frac{4837.5}{200}\right)^2}$	M1
	= 9.293 …..... (accept s = 9.32) awrt **<u>9.29</u>**	A1 (5)
(b)	$Q_2 = [20.5] + \frac{(100/100.5-62)}{88}\times 5 = 22.659...$ awrt **<u>22.7</u>**	M1 A1 (2)
(c)	$Q_1 = 10.5 + \frac{(50/50.25)}{62}\times 10 [= 18.56]$ (*****) (n + 1 gives 18.604…)	B1 cso (1)
(d)	Q_3= 25.5 (Use of n + 1 gives 25.734…)	B1
	IQR = 6.9 (Use of n + 1 gives 7.1)	B1 ft (2)
(e)	The data is skewed (condone "negative skew")	B1 (1)
(f)	Mean decreases and st. dev. remains the same. (from(a))	B1
	The median and quartiles would decrease. ((b)(c))	B1
	The IQR would remain unchanged (from (d))	B1 (3) **[14]**

Question Number	Scheme	Marks
5. (a)	$Q_2 = 17 + \left(\frac{60-58}{29}\right) \times 2$	M1
	= 17.1 (17.2 if use 60.5) awrt 17.1 (or17.2)	A1 (2)
(b)	$\sum fx = 2055.5$ $\sum fx^2 = 36500.25$	B1 B1
	Evidence of attempt to use midpoints with at least one correct	M1
	Mean = 17.129… awrt 17.1	B1
	$\sigma = \sqrt{\frac{36500.25}{120} - \left(\frac{2055.5}{120}\right)^2}$	M1
	= 3.28 (s = 3.294) awrt 3.3	A1 (6)
(c)	$\frac{3(17.129 - 17.1379...)}{3.28} = -0.00802$ Accept 0 or awrt 0.0	M1 A1
	No skew/ slight skew	B1 (3)
(d)	The skewness is very small. Possible.	B1 B1 (2) **[13]**

Question Number	Scheme	Marks
6. (a)	18-25 group, area=7x5=35	B1
	25-40 group, area=15x1=15	B1 (2)
(b)	(25-20)x5+(40-25)x1=40	M1A1 (2)
(c)	Mid points are 7.5, 12, 16, 21.5, 32.5	M1
	$\sum f = 100$	B1
	$\frac{\sum ft}{\sum f} = \frac{1891}{100} = 18.91$	M1A1 (4)
(d)	$\sigma_t = \sqrt{\frac{41033}{100} - \bar{t}^2}$ $\sqrt{\frac{n}{n-1}\left(\frac{41033}{100} - \bar{t}^2\right)}$ alternative OK	M1 M1
	$\sigma_t = \sqrt{52.74...} = 7.26$	A1 (3)
(e)	$Q_2 = 18$ or 18.1 if (n+1) used	B1
	$Q_1 = 10 + \frac{15}{16} \times 4 = 13.75$ or 15.25 numerator gives 13.8125	M1A1
	$Q_3 = 18 + \frac{25}{35} \times 7 = 23$ or 25.75 numerator gives 23.15	A1 (4)
(f)	0.376…	B1
	Positive skew	B1∫ (2) **[17]**
7. (a)	200 or 200g	B1 (1)
(b)	P(190 < X < 210) = 0.6 or P(X < 210) = 0.8 or P(X > 210) = 0.2	M1
	Correct use of 0.8 or 0.2	A1
	$Z = (\pm)\frac{210-200}{\sigma}$	M1
	$\frac{10}{\sigma} = 0.8416$ 0.8416	B1
	$\sigma = 11.882129…$ awrt 11.9	A1 (5)
(c)	$P(X < 180) = P\left(Z < \frac{180-200}{\sigma}\right)$	
	$= P(Z < -1.6832)$	
	$= 1 - 0.9535$	
	$= 0.0465$ or awrt 0.046	(3) **[9]**

Examiner reports

Question 1

Part (a) was answered well with only a small minority using $4305 - \left(\frac{181}{8}\right)^2$. Substitution into the formula for *r* was carried out successfully but a number of candidates gave their final answer to only 2 significant figures instead of the standard 3 significant figures we look for on S1. Most candidates now realised that the instruction "interpret" requires a contextualised comment but there were a number of nonsensical comments such as "temperature increases as sea level decreases" which gained no credit. Most candidates knew that coding had no effect on the correlation coefficient and picked up the mark for part (e) but very few scored the mark for part (d) with the commonest error being to divide by 1000. It appears that the effect of coding is being remembered as a fact rather than being deduced from an understanding of the structure of the formula.

Question 2

Parts (a), (b) and (c) were generally well done, although in part (c) there were many with strange ideas of heavy instruments. In part (d) the majority of candidates were able to make a credible attempt at this with most giving one of the two possible solutions with a reason. The majority used the median and quartiles to find that the distribution was symmetrical. The use of the words ‘symmetrical skew’, similar to ‘fair bias’, is all too often seen but was accepted. Equal, even or normal skew were also often seen and were given no credit.

Part (e) was attempted successfully by a minority of candidates. A large number of candidates did not understand the distinction between z-values and probabilities. A lot gave 0.68 as z-value leading to the loss of the accuracy mark. Others tried to put various values into standard deviation formulae.

Question 3

The common error here was to assume that frequency equals the area under a bar, rather than using the relationship that the frequency is proportional to the area under the bar. Many candidates therefore ignored the statement in line 1 of the question about the histogram representing 140 runners and simply gave an answer of 12 × 0.5 = 6. A few candidates calculated the areas of the first 7 bars and subtracted this from 140, sadly they didn’t think to look at the histogram and see if their answer seemed reasonable. Those who did find that the total area was 70 usually went on to score full marks. A small number of candidates had difficulty reading the scales on the graph and the examiners will endeavour to ensure that in any future questions of this type such difficulties are avoided.

Question 4

A small number of candidates still failed to calculate the mean correctly in part (a). For some this was due to errors with the midpoints but the more extreme errors involved dividing by 6 rather than 200 or using the class widths rather than the mid-points. The standard deviation formula still causes problems for some: forgetting the square root and failing to divide $\sum ft^2$ by 200 were common errors and some candidates used their rounded value of the mean and lost the final accuracy mark as their answer was not accurate to 3 significant figures. The calculation of the median in part (b) was answered well but applying the same principles to

part (c) caused difficulties for some with many of those attempting the $(n + 1)$ approach using 50.5 instead of 50.25 and others using incorrect end points. In part (d) a few spotted that Q_3 was on the class boundary and gave the value of 25.5 but others encountered similar problems to those with Q_1 but most were able to find their interquartile range. Part (e) was not answered well with many mentioning "continuous data" or "extreme values" and only a few stating that their data was skewed. Most candidates scored some marks in part (f) but many failed to secure all the marks because they did not deal with all of the estimates; in particular the standard deviation was often omitted.

Question 5

Very few candidates got full marks for this question, being unable to perform the calculations for grouped data, although the mean caused the least problems. Those candidates with good presentation particularly those who tabulated their workings tended to fare better. In spite of the well defined groups many candidates subtracted or added 0.5 to the endpoints or adjusted the midpoints to be 0.5 less than the true value with the majority getting part (a) incorrect as a result. As usual all possible errors were seen for the calculation of $\sum fx^2$ i.e. $(\sum fx)^2$, $\sum(fx)^2$, $\sum f^2x$ and $\sum x^2$. Use of 17.1 for the mean in the calculation of the standard deviation led to the loss the accuracy mark. Candidates are once again reminded not to use rounded answers in subsequent calculations even though they usually gain full marks for the early answer. The comment in part (c) was often forgotten perhaps indicating that candidates are able to work out the figures but do not know what they mean, although many did appreciate in part (d) that there is no skew in a normal distribution. As opposed to question 1, correlation was often mentioned instead of skewness although again this is becoming less common.

Question 6

Many candidates started well with this question, but a large number of inaccurate answers were seen for the latter parts. Part (a) was usually correct and part (b) was generally done well. In part (c) there were a lot of mistakes in finding midpoints and also Σf. Most knew the correct method for finding the mean, but rather fewer knew how to find the standard deviation in part (d) although most remembered to take the square root. Part (e) was very badly answered, with the majority unable to interpolate correctly which was often due to wrong class boundaries and / or class widths. In part (f), although the majority got an incorrect numerical value, most picked up the mark for interpreting their value correctly.

Question 7

Most candidates knew that mean = median for a normal distribution and wrote down the correct value, others obtained this by calculating (190 + 210)/2. In part (b) many were able to illustrate a correct probability statement on a diagram and most knew how to standardize but the key was to identify the statement $P(X < 210) = 0.8$ (or equivalent) and then use the tables to find the z value of 0.8416 and this step defeated the majority. Some used the "large" table and obtained the less accurate $z = 0.84$ but this still enabled them to score all the marks except the B1 for quoting 0.8416 from tables. In part (c) most were able to score some method marks for standardizing using their value of σ (provided this was positive!) and then attempting 1 – the probability from the tables. As usual the candidates' use of the notation connected with a normal distribution was poor: probabilities and z values were frequently muddled.

Statistics for S1 Practice Paper Gold Level G2

				Mean score for students achieving grade:							
Qu	Max Score	Modal score	Mean %	ALL	A*	A	B	C	D	E	U
1	7		66	4.63	5.97	5.55	4.97	4.59	4.25	3.94	3.18
2	10		48	4.81		7.16	5.15	4.23	3.62	3.15	2.27
3	5		49	2.45		3.02	2.19	1.86	1.53	1.55	1.05
4	14	0	53	7.40	12.10	11.32	9.17	7.43	5.88	4.41	2.10
5	13		50	6.50		9.67	7.41	6.04	4.73	3.64	1.89
6	17		53	9.07		13.14	10.12	8.41	7.05	6.00	4.00
7	9		48	4.33		6.57	4.55	2.87	2.31	1.31	0.75
	75		**52**	**39.19**		**56.43**	**43.56**	**35.43**	**29.37**	**24.00**	**15.24**