Paper Reference(s)

6683/01

Edexcel GCE

Statistics S1

Bronze Level B1

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
74	68	62	55	48	43

1. A random sample of 50 salmon was caught by a scientist. He recorded the length l cm and weight w kg of each salmon.

The following summary statistics were calculated from these data.

$\sum l = 4027 \qquad \sum l^2 = 327\,754.5 \qquad \sum w = 357.1 \qquad \sum lw = 29\,330.5 \qquad S_{ww} = 289.6$

(*a*) Find S_{ll} and S_{lw} .

(3)

(*b*) Calculate, to 3 significant figures, the product moment correlation coefficient between l and w.

(2)

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

(1)

January 2011

---

2. The 19 employees of a company take an aptitude test. The scores out of 40 are illustrated in the stem and leaf diagram below.

	2|6 means a score of 26	
0	7	(1)
1	88	(2)
2	4468	(4)
3	2333459	(7)
4	00000	(5)

Find

(*a*) the median score,

(1)

(*b*) the interquartile range.

(3)

The company director decides that any employees whose scores are so low that they are outliers will undergo retraining.

An outlier is an observation whose value is less than the lower quartile minus 1.0 times the interquartile range.

(*c*) Explain why there is only one employee who will undergo retraining.

(2)

(*d*) Draw a box plot to illustrate the employees' scores.

(3)

January 2010

---

3. The discrete random variable X can take only the values 2, 3, 4 or 6. For these values the probability distribution function is given by

x	2	3	4	6
$\mathrm{P}(X = x)$	$\frac{5}{21}$	$\frac{2k}{21}$	$\frac{7}{21}$	$\frac{k}{21}$

where k is a positive integer.

(*a*) Show that $k = 3$.

(2)

Find

(*b*) F(3),

(1)

(*c*) E(X),

(2)

(*d*) E(X^2),

(2)

(*e*) Var ($7X - 5$).

(4)

January 2012

4. In a study of how students use their mobile telephones, the phone usage of a random sample of 11 students was examined for a particular week.

The total length of calls, y minutes, for the 11 students were

$$17, 23, 35, 36, 51, 53, 54, 55, 60, 77, 110$$

(*a*) Find the median and quartiles for these data.

(3)

A value that is greater than $Q_3 + 1.5 \times (Q_3 - Q_1)$ or smaller than $Q_1 - 1.5 \times (Q_3 - Q_1)$ is defined as an outlier.

(*b*) Show that 110 is the only outlier.

(2)

(*c*) Draw a box plot for these data indicating clearly the position of the outlier.

(3)

The value of 110 is omitted.

(*d*) Show that S_{yy} for the remaining 10 students is 2966.9

(3)

These 10 students were each asked how many text messages, x, they sent in the same week. The values of S_{xx} and S_{xy} for these 10 students are $S_{xx} = 3463.6$ and $S_{xy} = -18.3$.

(*e*) Calculate the product moment correlation coefficient between the number of text messages sent and the total length of calls for these 10 students.

(2)

A parent believes that a student who sends a large number of text messages will spend fewer minutes on calls.

(*f*) Comment on this belief in the light of your calculation in part (*e*).

(1)

January 2009

5. A person's blood group is determined by whether or not it contains any of 3 substances A, B and C.

A doctor surveyed 300 patients' blood and produced the table below.

Blood contains	**No. of Patients**
only C	100
A and C but not B	100
only A	30
B and C but not A	25
only B	12
A, B and C	10
A and B but not C	3

(*a*) Draw a Venn diagram to represent this information.
(4)

(*b*) Find the probability that a randomly chosen patient's blood contains substance C.
(2)

Harry is one of the patients. Given that his blood contains substance A,

(*c*) find the probability that his blood contains all 3 substances.
(2)

Patients whose blood contains none of these substances are called universal blood donors.

(*d*) Find the probability that a randomly chosen patient is a universal blood donor.
(2)

May 2008

6. The following shows the results of a survey on the types of exercise taken by a group of 100 people.

65 run
48 swim
60 cycle
40 run and swim
30 swim and cycle
35 run and cycle
25 do all three

(*a*) Draw a Venn Diagram to represent these data.

(4)

Find the probability that a randomly selected person from the survey

(*b*) takes none of these types of exercise,

(2)

(*c*) swims but does not run,

(2)

(*d*) takes at least two of these types of exercise.

(2)

Jason is one of the above group.

Given that Jason runs,

(*e*) find the probability that he swims but does not cycle.

(3)

January 2012

7. A teacher took a random sample of 8 children from a class. For each child the teacher recorded the length of their left foot, f cm, and their height, h cm. The results are given in the table below.

f	23	26	23	22	27	24	20	21
h	135	144	134	136	140	134	130	132

(You may use $\sum f = 186$ $\quad \sum h = 1085$ $\quad S_{ff} = 39.5$ $\quad S_{hh} = 139.875$ $\quad \sum fh = 25\,291$)

(*a*) Calculate S_{fh}.

(2)

(*b*) Find the equation of the regression line of h on f in the form $h = a + bf$.
Give the value of a and the value of b correct to 3 significant figures.

(5)

(*c*) Use your equation to estimate the height of a child with a left foot length of 25 cm.

(2)

(*d*) Comment on the reliability of your estimate in part (*c*), giving a reason for your answer.

(2)

The left foot length of the teacher is 25 cm.

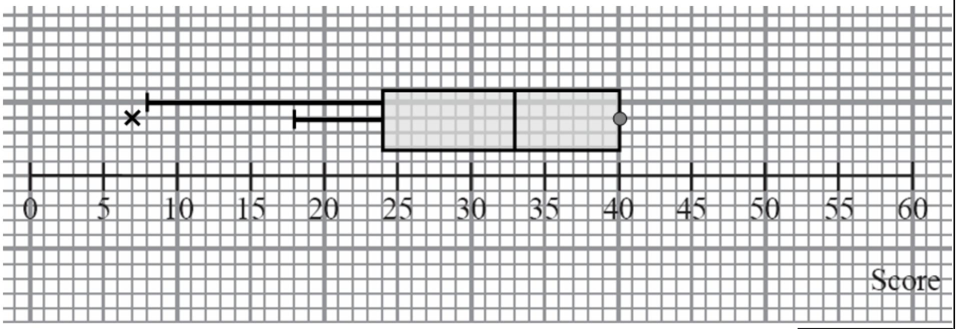
(*e*) Give a reason why the equation in part (*b*) should not be used to estimate the teacher's height.

(1)

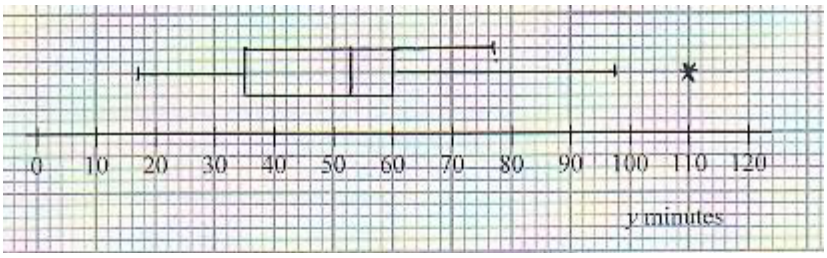
May 2011

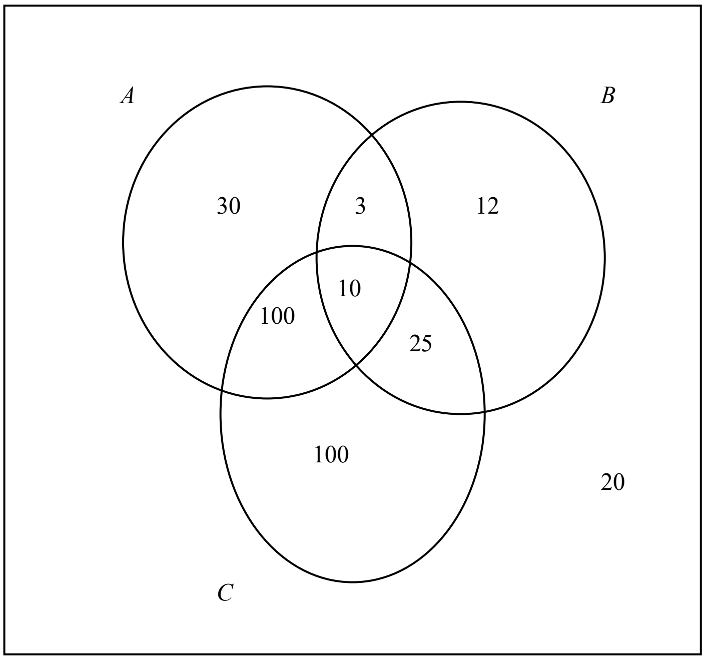
TOTAL FOR PAPER: 75 MARKS

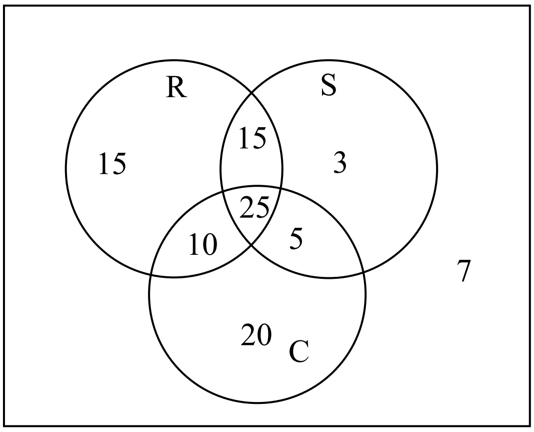
END

Question Number	Scheme	Marks
1. (a)	$S_{ll} = 327754.5 - \frac{4027^2}{50} = 3419.92$	M1 A1
	$S_{lw} = 29330.5 - \frac{357.1\times 4027}{50} = 569.666$	A1
		(3)
(b)	$r = \frac{569.666}{\sqrt{3419.92\times 289.6}} = 0.572$ awrt 0.572 or 0.573	M1 A1
		(2)
(c)	As the length of the salmon increases the weight increases	B1ft
		(1) **[6]**
2. (a)	Median is 33	B1
		(1)
(b)	$Q_1 = 24, Q_3 = 40, \text{IQR} = 16$	B1, B1, B1
		(3)
(c)	$Q_1 - \text{IQR}=24-16=8$	M1
	So 7 is only outlier	A1ft
		(2)
(d)		
	Box Outlier Whisker	B1ft B1 B1ft
		(3) **[9]**

Question Number	Scheme	Marks
3. (a)	$\frac{5}{21}+\frac{2k}{21}+\frac{7}{21}+\frac{k}{21}=1$	M1
	$\frac{12+3k}{21}=1$	
	$k=3$ * AG required for both methods	A1 (2)
(b)	$\frac{11}{21}$	B1 (1)
(c)	$\text{E}(X)=2\times\frac{5}{21}+3\times\frac{6}{21}+4\times\frac{7}{21}+6\times\frac{1}{7}$	M1
	$=3\frac{11}{21}$ or $\frac{74}{21}$ or awrt 3.52	A1 (2)
(d)	$\text{E}(X^2)=2^2\times\frac{5}{21}+3^2\times\frac{6}{21}+4^2\times\frac{7}{21}+6^2\times\frac{1}{7}$	M1
	$=14$	A1 (2)
(e)	$\text{Var}(X)=14-\left(3\frac{11}{21}\right)^2$	M1
	$=1\frac{257}{441}$ or $\frac{698}{441}$ or awrt 1.6	A1
	$\text{Var}(7X-5)=49\ \text{Var}(X)$	M1
	$=77\frac{5}{9}$ or $\frac{698}{9}$ or awrt 77.6	A1 (4) **[11]**

Question Number	Scheme	Marks
4. (a)	$Q_2 = 53, \quad Q_1 = 35, \quad Q_3 = 60$	B1, B1, B1 (3)
(b)	$Q_3 - Q_1 = 25 \Rightarrow Q_1 - 1.5\times 25 = -2.5$ (no outlier)	M1
	$Q_3 + 1.5\times 25 = 97.5$ (so 110 is an outlier)	A1 (2)
(c)		M1 A1ft A1ft (3)
(d)	$\sum y = 461, \sum y^2 = 24\ 219 \therefore\ \mathrm{S}_{yy} = 24219 - \frac{461^2}{10}$, = 2966.9 (*)	B1 B1 B1cso (3)
(e)	$r = \frac{-18.3}{\sqrt{3463.6\times 2966.9}}$ or $\frac{-18.3}{3205.64...} = -0.0057$	M1
	awrt –0.006 or -6×10^{-3}	A1 (2)
(f)	*r* suggests correlation is close to zero so parent’s claim is not justified	B1 (1) **[14]**

Question Number	Scheme	Marks
5. (a)	 A B C 30 3 12 10 100 25 100 20	
	3 closed intersecting curves with labels	M1
	100, 100, 30, 12,10, 3, 25	A1A1
	Box	B1
		(2)
(b)	P(Substance C) $=\dfrac{100+100+10+25}{300}=\dfrac{235}{300}=\dfrac{47}{60}$ or exact equivalent	M1 A1ft
		(2)
(c)	$\mathrm{P}(\text{ All 3 }\lvert A)=\dfrac{10}{30+3+10+100}=\dfrac{10}{143}$ or exact equivalent	M1 A1ft
		(2)
(d)	$\mathrm{P}(\text{Universal donor})=\dfrac{20}{300}=\dfrac{1}{15}$ or exact equivalent	M1 A1 cao
		(2) **[10]**

Question Number	Scheme	Marks
6. (a)		
	3 closed curves and 25 in correct place	M1
	15,10,5	A1
	15,3,20	A1
	Labels *R, S ,C* and box	B1
		(4)
(b)	7/100 or 0.07	M1 A1
		(2)
(c)	(3+5)/100 = 2/25 or 0.08	M1 A1
		(2)
(d)	(25+15+10+5)/100 = 11/20 or 0.55	M1 A1
		(2)
(e)	$\mathrm{P}\left(S \cap C' \middle\| R\right) = \dfrac{P\left(S \cap C' \cap R\right)}{P(R)}$	M1
	$= \dfrac{15}{65}$	A1
	$= \dfrac{3}{13}$ or exact equivalents	A1
		(3) **[13]**
7. (a)	$\left(\mathrm{S}_{fh} =\right) 25291 - \dfrac{186 \times 1085}{8}$	M1
	$=$ <u>64.75</u> (accept 64.8)	A1
		(2)
(b)	$b = \dfrac{\text{"64.75"}}{39.5},$ $=$ <u>1.6392….</u> (awrt 1.6)	M1, A1
	$a = \dfrac{1085}{8} - b \times \dfrac{186}{8},$ $=$ <u>97.512…</u> (awrt 97.5)	M1, A1
	<u>$h = 97.5 + 1.64f$</u>	A1ft
		(5)
(c)	$h = 97.5 + 1.64 \times 25$, $=$ <u>138~139</u>	M1, A1
		(2)
(d)	Should be reliable, since 25 cm(or *f* or footlength) is within the range of the data	B1, B1
		(2)
(e)	Line is for children – a different equation would apply to adults <u>or</u> Children are still growing, height will increase more than foot length	B1
		(1) **[12]**

Examiner reports

Question 1

This proved to be a friendly starter for most candidates with many scoring all 5 marks in part (a) and (b). Most errors here were arithmetic such as writing S_{ll} = 596.666... rather than 569.666... or accuracy problems in part (b) where an answer of 0.57 was often seen, rather than the 3sf accuracy that we look for. A minority still have difficulty in using the printed formulae and $S_{ll} = 327754.5 - \left(\frac{4027}{50}\right)^2$ or $\sum l \times \sum w$ instead of $\sum lw$ in S_{lw} were sometimes seen.

Part (c) caused problems for many candidates who simply wrote "positive correlation" but did not interpret this statement in the context by mentioning that longer salmon usually weigh more. Some candidates tended to "overstate" their conclusion by implying that as a salmon grows it gets longer (not strictly true in this instance as the study was of 50 different salmon not one salmon at 50 different time intervals) and others referred to a proportionate relationship such as "for every cm increase in length the salmon weighs 0.572 kg more". Whilst such indiscretions were overlooked for the single mark on this occasion, these examples should provide useful points of discussion for teachers with future cohorts of candidates.

Question 2

Parts (a) and (b) were answered very well although a few candidates gave the upper quartile as 39 or 39.5 (usually as a result of incorrectly rounding $\frac{3n}{4}$) however the follow through marks meant that no further penalty need occur. A few found the upper and lower quartiles but failed to give the interquartile range. Most found the limit for an outlier using the given definition, although a few used $1.5 \times \text{IQR}$, and went on to make a suitable comment about the one employee who needed retraining. There were some excellent box plots seen with all the correct features clearly present but a number failed to plot the outlier appropriately and simply drew their lower whisker to 7. A not insignificant minority were confused by the absence of an upper whisker and felt the need to add one usually at $Q_3 + \text{IQR}$.

Question 3

Part (a) was answered well with a large majority setting out the solution as expected. A small number tried to verify the value, but most only did the substitution and did not say that it showed $k = 3$, thus losing the final accuracy mark.

Part (b) was poorly answered with a large number finding P(3) instead. A small number gave the answer as an inexact decimal instead of a fraction.

Part (c) and part (d) were both well answered with complete methods shown. Only a few candidates confused $[\text{E}(X)]^2$ with $\text{E}(X^2)$. In part (e) some of those candidates who got $\text{E}(X^2)$ wrong still got $\text{Var}(X)$ right here, as they did not realise the link and started again. Most realised that they needed to find $\text{Var}(X)$ but many did not know the link with $\text{Var}(7X-5)$. Some candidates worked out $7\,\text{Var}(X) - 5$, some $7\,\text{Var}(X)$ and others $52\,\text{Var}(X)$. The result for $\text{Var}(7X-5)$ was often not awarded the final accuracy mark as some candidates had used rounded answers in their working.

Question 4

This question was usually answered well. In part (b) some did not realise that they needed to check the lower limit as well in order to be sure that 110 was the only outlier. Part (c) was answered very well although some lost the last mark because there was no gap between the end of their whisker and the outlier. Part (d) was answered very well and most gave the correct values for $\sum y$ and $\sum y^2$ in the appropriate formula. A few tried to use the $\sum\left(y-\bar{y}\right)^2$ approach but this requires all 10 terms to be seen for a complete "show that" and this was rare.

Part (e) was answered well although some gave the answer as -5.7 having forgotten the 10^{-3}, or failed to interpret their calculator correctly. Many candidates gave comments about the correlation being small or negative in part (f) but they did not give a clear reason for rejecting the parent's belief. Once again the interpretation of a calculated statistic caused difficulties.

Question 5

A lot of fully correct Venn Diagrams were seen in part (a) although it was surprising the number who resorted to decimals rather than just using straightforward fractions; this often led to loss of many accuracy marks. A significant minority had negative numbers in their Venn diagram and saw nothing wrong in this when converting them to probabilities later in the question. Fewer candidates forgot the box this time. Part (c) proved to be the only difficult part, as many candidates struggled with the concept of conditional probability, and many denominators of 300 were seen.

Question 6

Construction of the Venn diagram was nearly always correct. Occasional errors were mainly the omission of the box and failure to subtract frequencies accurately. Unfortunately, several candidates left the region for $R \cap S \cap C$ so small that it was extremely difficult to decipher the number written there.

In part (b) there were relatively few incorrect solutions. Occasionally an incorrect subtraction from 100 to find $n(R' \cap S' \cap C')$ was seen.

Part (c) and part (d) were very well answered by the majority of candidates. However, in part (d), $\frac{30}{100}$ was not an uncommon response, with the central frequency of 25 being omitted. This stems from a failure to understand the phrase "at least" in the question. Conditional probability in part (e) continues to be a problem for many candidates. Perhaps greater emphasis on the restricted sample space would produce better and quicker rewards.

Question 7

Parts (a) and (b) were answered well by the majority of candidates. Only a small minority used $\frac{S_{fh}}{S_{hh}}$ for *b* and there were few cases of the incorrect sign being used for finding *a*. Premature rounding and a failure to write their final answer in terms of *f* and *h* to 3significant figures meant a number lost the final accuracy mark. Part (c) was answered very well but in part (d) the candidates comments were often a little confused. The question was looking for a comment that the value of the independent variable *f* was within the range of the data and therefore the estimate should be reliable. A number of candidates seemed to focus their comments on the value of the dependent variable *h* and others were just a little vague referring to "it" was within the range of the data. There were many clear and thorough answers to part (e) that showed the candidates had a good understanding of the limitations of the equation they had calculated (namely that it was based on data for children not adults). Questions of this type are simply looking for the candidates to engage with the context and give a sensible comment.

Statistics for S1 Practice Paper Bronze Level B1

				Mean score for students achieving grade:							
Qu	Max Score	Modal score	Mean %	ALL	A*	A	B	C	D	E	U
1	6		86	5.13	5.50	5.46	5.18	4.95	4.75	4.32	3.54
2	9		84	7.55		8.42	7.81	7.15	6.37	5.51	3.86
3	11		80	8.79	10.75	10.20	9.30	8.45	7.62	7.00	4.17
4	14		83	11.60		12.59	11.80	10.93	10.22	9.35	6.12
5	10		78	7.83		9.21	8.17	7.49	6.97	6.69	5.85
6	13		75	9.79	12.11	11.45	9.98	8.96	8.10	7.21	5.03
7	12		76	9.14	11.32	10.98	10.19	9.49	8.79	7.81	5.03
	75		**80**	**59.83**		**68.31**	**62.43**	**57.42**	**52.82**	**47.89**	**33.60**