

Paper Reference(s)

**6684/01**

# Edexcel GCE

## Statistics S2

### Gold Level G2

**Time: 1 hour 30 minutes****Materials required for examination papers**

Mathematical Formulae (Green)

**Items included with question**

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 S2), the paper reference (6684), 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 8 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
65	55	47	36	26	17

1. A bag contains a large number of counters. A third of the counters have a number 5 on them and the remainder have a number 1.

A random sample of 3 counters is selected.

(a) List all possible samples. (2)

(b) Find the sampling distribution for the range. (3)

---

2. A test statistic has a distribution  $B(25, p)$ .

Given that

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5,$$

(a) find the critical region for the test statistic such that the probability in each tail is as close as possible to 2.5%. (3)

(b) State the probability of incorrectly rejecting  $H_0$  using this critical region. (2)

---

3. Bacteria are randomly distributed in a river at a rate of 5 per litre of water. A new factory opens and a scientist claims it is polluting the river with bacteria. He takes a sample of 0.5 litres of water from the river near the factory and finds that it contains 7 bacteria. Stating your hypotheses clearly test, at the 5% level of significance, the claim of the scientist. (7)
-

4. Past records suggest that 30% of customers who buy baked beans from a large supermarket buy them in single tins. A new manager questions whether or not there has been a change in the proportion of customers who buy baked beans in single tins. A random sample of 20 customers who had bought baked beans was taken.

(a) Using a 10% level of significance, find the critical region for a two-tailed test to answer the manager's question. You should state the probability of rejection in each tail which should be less than 0.05.

(5)

(b) Write down the actual significance level of a test based on your critical region from part (a).

(1)

The manager found that 11 customers from the sample of 20 had bought baked beans in single tins.

(c) Comment on this finding in the light of your critical region found in part (a).

(2)

---

5. In a game, players select sticks at random from a box containing a large number of sticks of different lengths. The length, in cm, of a randomly chosen stick has a continuous uniform distribution over the interval  $[7, 10]$ .

A stick is selected at random from the box.

(a) Find the probability that the stick is shorter than 9.5 cm.

(2)

To win a bag of sweets, a player must select 3 sticks and wins if the length of the longest stick is more than 9.5 cm.

(b) Find the probability of winning a bag of sweets.

(2)

To win a soft toy, a player must select 6 sticks and wins the toy if more than four of the sticks are shorter than 7.6 cm.

(c) Find the probability of winning a soft toy.

(4)

---

6. A company has a large number of regular users logging onto its website. On average 4 users every hour fail to connect to the company's website at their first attempt.

(a) Explain why the Poisson distribution may be a suitable model in this case. (1)

Find the probability that, in a randomly chosen **2 hour** period,

- (b) (i) all users connect at their first attempt,  
 (ii) at least 4 users fail to connect at their first attempt. (5)

The company suffered from a virus infecting its computer system. During this infection it was found that the number of users failing to connect at their first attempt, over a 12 hour period, was 60.

- (c) Using a suitable approximation, test whether or not the mean number of users per hour who failed to connect at their first attempt had increased. Use a 5% level of significance and state your hypotheses clearly. (9)
- 

7. (a) Define the critical region of a test statistic. (2)

A discrete random variable  $X$  has a Binomial distribution  $B(30, p)$ . A single observation is used to test  $H_0 : p = 0.3$  against  $H_1 : p \neq 0.3$

- (b) Using a 1% level of significance find the critical region of this test. You should state the probability of rejection in each tail which should be as close as possible to 0.005. (5)

- (c) Write down the actual significance level of the test. (1)

The value of the observation was found to be 15.

- (d) Comment on this finding in light of your critical region. (2)
-

8. The continuous random variable  $X$  has probability density function given by

$$f(x) = \begin{cases} \frac{3}{32}(x-1)(5-x) & 1 \leq x \leq 5, \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Sketch  $f(x)$  showing clearly the points where it meets the  $x$ -axis. (2)
- (b) Write down the value of the mean,  $\mu$ , of  $X$ . (1)
- (c) Show that  $E(X^2) = 9.8$ . (4)
- (d) Find the standard deviation,  $\sigma$ , of  $X$ . (2)

The cumulative distribution function of  $X$  is given by

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{32}(a - 15x + 9x^2 - x^3) & 1 \leq x \leq 5 \\ 1 & x > 5 \end{cases}$$

where  $a$  is a constant.

- (e) Find the value of  $a$ . (2)
- (f) Show that the lower quartile of  $X$ ,  $q_1$ , lies between 2.29 and 2.31. (3)
- (g) Hence find the upper quartile of  $X$ , giving your answer to 1 decimal place. (1)
- (h) Find, to 2 decimal places, the value of  $k$  so that

$$P(\mu - k\sigma < X < \mu + k\sigma) = 0.5. \quad (2)$$

**TOTAL FOR PAPER: 75 MARKS**

**END**



Question Number	Scheme	Marks
<p>3</p>	<p><u>One tail test</u>  <u>Method 1</u>  <math>H_0 : \lambda = 5 (\lambda = 2.5)</math>  <math>\mu</math>  <math>H_1 : \lambda &gt; 5 (\lambda &gt; 2.5)</math></p> <p><math>X \sim \text{Po} (2.5)</math></p> <p><math>P(X \geq 7) = 1 - P(X \leq 6)</math>  <math>= 1 - 0.9858</math>  <math>= 0.0142</math></p> <p><math>0.0142 &lt; 0.05</math></p> <p>(Reject <math>H_0</math>.) There is significant evidence at the 5% significance level that the factory is <u>polluting the river</u> with bacteria.</p> <p><u>or</u>  The scientists claim is justified</p>	<p>may use <math>\lambda</math> or  B1  B1  M1  may be implied  M1  [ <math>P(X \geq 5) = 1 - 0.8912 = 0.1088</math> ] att <math>P(X \geq 7)</math> <math>P(X \geq 6)</math>  <math>P(X \geq 6) = 1 - 0.9580 = 0.0420</math>  A1  CR <math>X \geq 6</math> awrt 0.0142  M1  7 <math>\geq 6</math> or 7 is in critical region or 7 is significant  B1  (7)  Total 7</p>
	<p><u>Method 2</u>  <math>H_0 : \lambda = 5 (\lambda = 2.5)</math>  <math>H_1 : \lambda &gt; 5 (\lambda &gt; 2.5)</math></p> <p><math>X \sim \text{Po} (2.5)</math></p> <p><math>P(X &lt; 7)</math>  <math>= 0.9858</math></p> <p><math>0.9858 &gt; 0.95</math></p> <p>(Reject <math>H_0</math>.) There is significant evidence at the 5% significance level that the factory is <u>polluting the river</u> with bacteria.</p> <p><u>or</u>  The scientists claim is justified</p>	<p>may use <math>\lambda</math> or <math>\mu</math>  B1  B1  M1  may be implied  att <math>P(X &lt; 7)</math> <math>P(X &lt; 6)</math>  [ <math>P(X &lt; 5) = 0.8912</math> ]  <math>P(X &lt; 6) = 0.9580</math>  M1 A1  CR <math>X \geq 6</math> wrt 0.986  M1  7 <math>\geq 6</math> or 7 is in critical region or 7 is significant  B1  (7)</p>



Question Number	Scheme	Marks
4.	(a) $X \sim B(20, 0.3)$	M1
	$P(X \leq 2) = 0.0355$	A1
	$P(X \leq 9) = 0.9520$ so $P(X \geq 10) = 0.0480$	A1
	Therefore the critical region is $\{X \leq 2\} \cup \{X \geq 10\}$	A1 A1 (5)
(b)	$0.0355 + 0.0480 = 0.0835$ awrt (0.083 or 0.084)	B1 (1)
(c)	11 is in the critical region there is evidence of a <u>change/ increase</u> in the <u>proportion/number</u> of <u>customers</u> <u>buying single tins</u>	B1ft B1ft (2)
		<b>(8 marks)</b>

5.	(a)	$\frac{9.5 - 7}{10 - 7}$ $= \frac{5}{6}$	M1 A1 awrt 0.833	(2)
	(b)	$P(\text{Longest} > 9.5) = 1 - P(\text{all} < 9.5) = 1 - \left(\frac{5}{6}\right)^3$ $= \frac{91}{216}$ or 0.421	M1 A1	(2)
	(c)	$P(\text{a stick} < 7.6) = \frac{0.6}{3} = 0.2$ Let $Y = \text{number of sticks (out of 6)} < 7.6$ then $Y \sim B(6, 0.2)$ $P(Y > 4) = 1 - P(Y \leq 4)$ $= 1 - 0.9984$ $= 0.0016$ or $\frac{1}{625}$	B1 M1 M1 A1	(4)
				<b>8</b>

Question Number	Scheme	Marks
6.	<p>(a) Connecting occurs at random/independently, singly or at a constant rate</p> <p>(b) Po (8)</p> <p>(i) <math>P(X = 0) = 0.0003</math></p> <p>(ii) <math>P(X \geq 4) = 1 - P(X \leq 3)</math>  <math>= 1 - 0.0424</math>  <math>= 0.9576</math></p> <p>(c) <math>H_0 : \lambda = 4</math> (48) <math>H_1 : \lambda &gt; 4</math> (48)  <math>N(48, 48)</math></p> <p>Method 1</p> $P(X \geq 59.5) = P\left(Z \geq \frac{59.5 - 48}{\sqrt{48}}\right)$ $= P(Z \geq 1.66)$ $= 1 - 0.9515$ $= 0.0485$ <p>Method 2</p> $\frac{x - 0.5 - 48}{\sqrt{48}} = 1.6449$ $x = 59.9$ <p><math>0.0485 &lt; 0.05</math>  Reject <math>H_0</math>. Significant. 60 lies in the Critical region  The number of failed connections at the first attempt has increased.</p>	<p>B1 (1)</p> <p>B1</p> <p>M1A1</p> <p>M1</p> <p>A1 (5)</p> <p>B1</p> <p>M1 A1</p> <p>M1 M1 A1</p> <p>A1</p> <p>M1</p> <p>A1 ft (9)</p> <p>[15]</p>

Question Number	Scheme	Marks
7.	<p>(a) The set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test.</p> <p>(b) <math>X \sim B(30, 0.3)</math>  <math>P(X \leq 3) = 0.0093</math>  <math>P(X \leq 2) = 0.0021</math>  <math>P(X \geq 16) = 1 - 0.9936 = 0.0064</math>  <math>P(X \geq 17) = 1 - 0.9979 = 0.0021</math>  Critical region is <math>(0 \leq) x \leq 2</math> or <math>16 \leq x (\leq 30)</math></p> <p>(c) Actual significance level <math>0.0021 + 0.0064 = 0.0085</math> or 0.85%</p> <p>(d) 15 (it) is not in the critical region  not significant  No significant evidence of a change in <math>p = 0.3</math>  accept <math>H_0</math>, (reject <math>H_1</math>)  <math>P(x \geq 15) = 0.0169</math></p>	<p>B1</p> <p>B1</p> <p>(2)</p> <p>M1</p> <p>A1</p> <p>A1</p> <p>A1A1</p> <p>(5)</p> <p>B1</p> <p>(1)</p> <p>Bft 2, 1, 0</p> <p>(2)</p> <p>Total [10]</p>

Question Number	Scheme	Marks
<b>8.</b> <b>(a)</b> <b>(b)</b> <b>(c)</b> <b>(d)</b> <b>(e)</b> <b>(f)</b> <b>(g)</b> <b>(h)</b>	$\cap$ shape which does not go below the $x$ -axis [condone missing parities]	B1
	Graph must end at the points (1,0) and (5,0) and the points labelled at 1 and 5	B1
	$E(X) = 3$ (by symmetry)	B1
	$[E(X^2)] = \int x^2 f(x) dx = \frac{3}{32} \int (6x^3 - x^4 - 5x^2) dx$	M1
	$= \frac{3}{32} \left[ \frac{6x^4}{4} - \frac{x^5}{5} - \frac{5x^3}{3} \right]_1^5$	A1
	$= \frac{3}{32} \left( \left[ \frac{6 \times 625}{4} - 625 - \frac{625}{3} \right] - \left[ \frac{6}{4} - \frac{1}{5} - \frac{5}{3} \right] \right) = 9.8 \text{ (*)}$	M1 A1 cso (4)
	s.d. = $\sqrt{9.8 - E(X)^2}$ , = 0.8944...	M1 A1 (2)
	$F(1) = 0 \Rightarrow \frac{1}{32}(a - 15 + 9 - 1) = 0$ , leading to <u><math>a = 7</math></u>	M1 A1 (2)
	$F(2.29) = 0.2449\dots$ , $F(2.31) = 0.2515\dots$ Since $F(q_1) = 0.25$ and these values are either side of 0.25 then $2.29 < q_1 < 2.31$	M1 A1 A1 (3)
	Since the distribution is symmetric $q_3 = 5 - 1.3 = \underline{3.7}$	cao B1 (1)
We know $P(q_1 = 2.3 < X < 3.7 = q_3) = 0.5$ so $k\sigma = 0.7$ so $k = \frac{0.7}{0.894\dots} = 0.7826\dots = \text{awrt } \underline{0.78}$	M1 A1 (2)	

**17**

## Examiner reports

### Question 1

This question was answered well by many candidates. In part (a) the vast majority of candidates were able to score one mark and many scored both marks as they could list all 8 combinations.

Part (b) caused some candidates problems as they failed to recognise that the range was the difference between the largest and smallest values. Some candidates found the sampling distribution of the mean or the sum rather than the range. Others thought that all samples were equally likely and hence obtained the wrong probabilities.

Those candidates who could identify the range usually went on to calculate the correct corresponding probabilities. A common error was to give the range as 0 and 1.

### Question 2

Part (a) was a routine question with many fully correct answers being seen. However, a minority of candidates were unable to write down the critical regions correctly. The two most common errors were to write down the regions in terms of probabilities, e.g.  $P(X \leq 7)$  or to simply write  $X = 7$  and  $X = 18$ .

In part (b) many candidates were distracted by the word “incorrectly”. We reject the null hypothesis ‘incorrectly’ when it is in fact true. In this case, this means that  $p$  really is 0.5, so that we can use the distribution  $B(25, 0.5)$  to work out probabilities. This is the distribution that had already been used in part (a), so that all that needed to be done was to simply add the two probabilities that were used to identify the two parts of the critical region in part (a):  $0.0216 + 0.0216 = 0.0432$ . It was all too common to see correct working followed by incorrect and irrelevant work that invalidated the whole response. The most common incorrect postscripts were  $0.05 - 0.0432 = 0.0068$  and  $1 - 0.0432 = 0.9568$ .

### Question 3

The majority of candidates found this question straightforward. They were most successful if they used the probability method and compared it with 0.05. Those who attempted to use 95% were less successful and this is not a recommended route for these tests.

Most candidates knew how to specify the hypotheses with most candidates using 2.5 rather than 5. Some candidates used  $p$ , or did not use a letter at all, in stating their hypotheses, but most of the time they used  $\lambda$ . A minority found  $P(X=7)$  and some worked with  $Po(5)$ .

If using the critical region method, not all candidates showed clearly, either their working, or a comparison with the value of 7 and the CR  $X > 7$ . A sizeable minority of candidates failed to put their conclusion back into the given context. Reject  $H_0$  is not sufficient.

### Question 4

This was a very well answered question. Candidates were able to use binomial tables and gave the answer to the required number of decimal places. As in previous years there were some candidates who confused the critical region with the probability of the test statistic being in that region but this error has decreased. Candidates were able to describe the acceptance of the hypothesis in context although sometimes it would be better if they just repeated the wording from the question which would help them avoid some of the mistakes seen. There were still a few candidates who did not give a reason in context at all.

In part (a) many candidates failed to read this question carefully assuming it was identical to similar ones set previously. Most candidates correctly identified  $B(20,0.3)$  to earn the method mark and many had 0.0355 written down to earn the first A mark, although in light of their subsequent work, this may often have been accidental. A majority of candidates did not gain the second A mark as they failed to respond to the instruction “state the probability of rejection in each case”. In the more serious cases, candidates had shown no probabilities from the tables, doing all their work mentally, only writing their general strategy: “ $P(X \leq c) < 0.05$ ”. Whilst many candidates were able to write down the critical region using the correct notation there are still some candidates who are losing marks they should have earned, by writing  $P(X \leq 2)$  for the critical region  $X \leq 2$

Part (b) was usually correct.

Part (c) provided yet more evidence of candidates who had failed to read the question: “in the light of your critical region”. Some candidates chose not to mention the critical region and a number of those candidates who identified that 11 was in the critical region did not refer to the manager’s question.

### Question 5

Part (a) was routine and the vast majority of candidates demonstrated familiarity with the probability density function of a rectangular distribution.

Part (b) required some careful initial thought that eluded a large majority of the candidates. The successful solutions fell into two camps. On the one hand were those candidates who identified the Binomial distribution  $X \sim B\left(3, \frac{5}{6}\right)$  and then evaluated the probability  $P(X \geq 1)$  by conventional means. On the other hand there were candidates who worked ‘from first principles’ and produced either an elaborate tree diagram or a list of all possible outcomes. This latter approach (‘from first principles’) is of course valid, but takes longer and is more susceptible to error.

There were many complete and correct solutions to part (c). Finding the length of a stick shorter than 7.6 cm was straightforward for most but many attempts reflected candidates' lack of understanding of what the question required of them in this part. Successful candidates realised it was a binomial situation using  $B(6, p)$  and used tables to find  $1 - P(X \leq 4)$  or calculated  $P(X = 5) + P(X = 6)$ . Common errors included misinterpretation of  $P(X > 4)$  as  $1 - P(X \leq 3)$  or even  $P(X = 4)$  and a small minority of candidates gave the answer as

$\left(\frac{1}{5}\right)^4 = \frac{1}{625}$  which gained no marks as it is an incorrect method.

### Question 6

The majority of candidates were familiar with the technical terms in part (a), but failed to establish any context.

Part (b) was a useful source of marks for a large proportion of the candidates. The only problems were occasional errors in detail. In part (i) a few did not spot the change in time scale and used  $Po(4)$  rather than  $Po(8)$ . Some were confused by the wording and calculated  $P(X = 8)$  rather than  $P(X = 0)$ . The main source of error for (ii) was to find  $1 - P(X \leq 4)$  instead of  $1 - P(X \leq 3)$ .

In part (c) the Normal distribution was a well-rehearsed routine for many candidates with many candidates concluding the question with a clear statement in context.

The main errors were

- Some other letter (or none) in place of  $\lambda$  or  $\mu$
- Incorrect Normal distribution: e.g.  $N(60, 60)$
- Omission of (or an incorrect) continuity correction
- Using 48 instead of 60
- Calculation errors

A minority of candidates who used the wrong distribution (usually Poisson) were still able to earn the final two marks in the many cases when clear working was shown. This question was generally well done with many candidates scoring full marks.

### Question 7

Part (a) tested candidates' understanding of the critical region of a test statistic and responses were very varied, with many giving answers in terms of a 'region' or 'area' and making no reference to the null hypothesis or the test being significant. Many candidates lost at least one mark in part (b), either through not showing the working to get the probability for the upper critical value, i.e.  $1 - P(X \leq 15) = P(X \geq 16) = 0.0064$ , or by not showing any results that indicated that they had used  $B(30, 0.3)$  and just writing down the critical regions, often incorrectly. A minority of candidates still write their critical regions in terms of probabilities and lose the final two marks. Responses in part (c) were generally good with the majority of candidates making a comment about the observed value and their critical region. A small percentage of responses contained contradictory statements.

## Question 8

The overall response to part (a) was disappointing. The shape of the graph is assumed knowledge from GCSE mathematics. If  $f(x)$  is a quadratic expression, then the graph of  $y = f(x)$  is a parabola, whose shape is ‘known’. Many candidates made a ‘table of values’ which they then used to plot the graph which then consisted of straight lines. Occasionally the correct parabola shape was seen with 1 and 5 marked on the  $x$ -axis but the graph continued below the  $x$ -axis indicating that the probability could be negative for values outside of the given range.

Although part (b) was intended to be straightforward, many candidates failed to see the symmetry of the probability density function (even with a correct diagram), and pursued integration to find the mean. Fortunately, most attempts were successful. It should be noted that in general if there is 1 mark for a question the answer should be able to be written down without any long calculations.

Part (c) was well answered by the majority of candidates, with responses ranging from the very concise to lengthy. Nearly all candidates were able to multiply the brackets and integrate successfully, although there were occasional sign errors. Sufficient working for the substitution of the limits was nearly always shown. Some candidates preferred to find  $\text{Var}(X)$  first, only to add  $[E(X)]^2$  at the end.

Some candidates obtained an answer of almost 9.8, this is despite the fact that most modern calculators will work using exact fractions. It was in fact possible to work with exact decimals, the penultimate stage of working being  $9.765625 - (-0.034375)$ . However, some candidates chose to approximate the earlier decimals, resulting in a final answer that differed from 9.8.

Part (d) was usually answered correctly, with a significant minority giving their answer in surd form.

In part (e) the obvious methods, in this case solving either  $F(1) = 0$  or  $F(5) = 1$ , were not universally adopted. For those who adopted the latter approach, a surprising proportion struggled to solve the linear equation:  $\frac{1}{32}(a - 75 + 225 - 125) = 1$ . Many candidates integrated  $f(x)$ , or occasionally  $F(x)$ , with mixed success.

The solution to part (f) had been well rehearsed by candidates, although some failed to give an adequate explanation for the final mark. An ideal correct response such as “0.25 is between 0.245 and 0.252, so the median must lie between 2.29 and 2.31” was rarely seen. A sizeable minority pursued the solution of  $F(x) = \frac{1}{4}$ . Those who used a sign change method were sometimes successful, and a few candidates used their graphical calculators to state all three roots, before selecting the required one.

In part (g), many candidates did not appreciate the symmetric nature of the probability density function. However, those who did usually produced a correct answer of 3.7. Other candidates resorted to a repetition of their method for finding the lower quartile.

In part (h) those candidates who saw that the given probability referred to the IQR of the distribution often produced a neat solution. More often, however, candidates abandoned their solution, usually after much algebraic effort.

## Statistics for S2 Practice Paper Gold 2

Qu	Max Score	Modal score	Mean %	Mean average scored by candidates achieving grade:							
				ALL	A*	A	B	C	D	E	U
1	5		62.4	3.12	4.01	3.62	2.55	2.03	1.66	1.31	0.85
2	5		67.2	3.36	4.38	4.07	3.54	2.98	2.35	1.65	0.82
3	7		64.9	4.54		5.90	5.08	3.90	2.77	1.80	0.76
4	8		62.4	4.99		6.45	5.18	4.25	3.27	2.18	1.14
5	8		59.5	4.76	6.80	6.07	4.88	3.93	2.92	1.84	0.99
6	15		70.2	10.53	13.77	12.86	11.06	9.10	6.63	4.69	2.36
7	10		64.5	6.45		6.64	5.50	4.26	2.76	1.58	0.79
8	17		56.2	9.56	14.37	12.56	9.54	7.32	5.53	3.61	2.38
	<b>75</b>		<b>63.1</b>	<b>47.31</b>		<b>58.17</b>	<b>47.33</b>	<b>37.77</b>	<b>27.89</b>	<b>18.66</b>	<b>10.09</b>