Paper Reference(s)

6684/01

Edexcel GCE

Statistics S2

Gold Level G3

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 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 9 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
63	53	42	34	26	20

1. A factory produces components. Each component has a unique identity number and it is assumed that 2% of the components are faulty. On a particular day, a quality control manager wishes to take a random sample of 50 components.

(*a*) Identify a sampling frame.

(1)

The statistic F represents the number of faulty components in the random sample of size 50.

(*b*) Specify the sampling distribution of F.

(2)

---

2. A student takes a multiple choice test. The test is made up of 10 questions each with 5 possible answers. The student gets 4 questions correct. Her teacher claims she was guessing the answers. Using a one tailed test, at the 5% level of significance, test whether or not there is evidence to reject the teacher's claim.

State your hypotheses clearly.

(6)

---

3. A test statistic has a Poisson distribution with parameter λ.

Given that

$$H_0: \lambda = 9,\ H_1: \lambda \neq 9,$$

(*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)

---

4. A random sample $X_1, X_2, \ldots X_n$ is taken from a population with unknown mean μ and unknown variance σ^2. A statistic Y is based on this sample.

(*a*) Explain what you understand by the statistic Y.

(2)

(*b*) Explain what you understand by the sampling distribution of Y.

(1)

(*c*) State, giving a reason which of the following is **not** a statistic based on this sample.

(i) $\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n}$ (ii) $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$ (iii) $\sum_{i=1}^{n} X_i^2$

(2)

5.

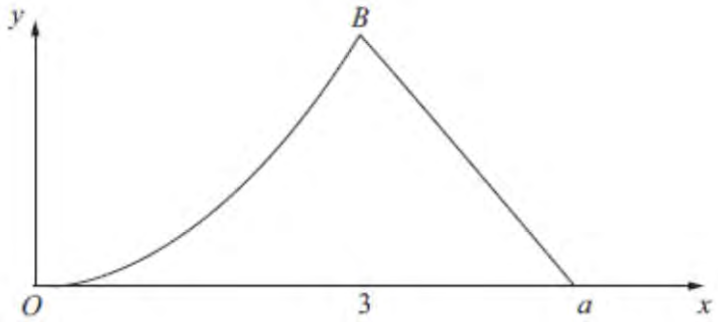

Figure 1

Figure 1 shows a sketch of the probability density function f(x) of the random variable X.

For $0 \leq x \leq 3$, f(x) is represented by a curve OB with equation f(x) = kx^2, where k is a constant.

For $3 \leq x \leq a$, where a is a constant, f(x) is represented by a straight line passing through B and the point (a, 0).

For all other values of x, f(x) = 0.

Given that the mode of X = the median of X, find

(a) the mode,

(1)

(b) the value of k,

(4)

(c) the value of a.

(3)

Without calculating E(X) and with reference to the skewness of the distribution

(d) state, giving your reason, whether E(X) < 3, E(X) = 3 or E(X) > 3.

(2)

---

6. In a village shop the customers must join a queue to pay. The number of customers joining the queue in a 10 minute interval is modelled by a Poisson distribution with mean 3.

Find the probability that

(*a*) exactly 4 customers join the queue in the next 10 minutes,

(2)

(*b*) more than 10 customers join the queue in the next 20 minutes.

(3)

When a customer reaches the front of the queue the customer pays the assistant. The time each customer takes paying the assistant, T minutes, has a continuous uniform distribution over the interval [0, 5]. The random variable T is independent of the number of people joining the queue.

(*c*) Find P($T > 3.5$).

(1)

In a random sample of 5 customers, the random variable C represents the number of customers who took more than 3.5 minutes paying the assistant.

(*d*) Find P($C \geq 3$).

(3)

Bethan has just reached the front of the queue and starts paying the assistant.

(*e*) Find the probability that in the next 4 minutes Bethan finishes paying the assistant and no other customers join the queue.

(4)

---

7. A bag contains a large number of balls.

65% are numbered 1

35% are numbered 2

A random sample of 3 balls is taken from the bag.

Find the sampling distribution for the range of the numbers on the 3 selected balls.

(6)

---

8. (*a*) Explain what you understand by

(i) a hypothesis test,

(ii) a critical region.

(3)

During term time, incoming calls to a school are thought to occur at a rate of 0.45 per minute. To test this, the number of calls during a random 20 minute interval, is recorded.

(*b*) Find the critical region for a two-tailed test of the hypothesis that the number of incoming calls occurs at a rate of 0.45 per 1 minute interval. The probability in each tail should be as close to 2.5% as possible.

(5)

(*c*) Write down the actual significance level of the above test.

(1)

In the school holidays, 1 call occurs in a 10 minute interval.

(*d*) Test, at the 5% level of significance, whether or not there is evidence that the rate of incoming calls is less during the school holidays than in term time.

(5)

---

9. A cloth manufacturer knows that faults occur randomly in the production process at a rate of 2 every 15 metres.

(*a*) Find the probability of exactly 4 faults in a 15 metre length of cloth.

(2)

(*b*) Find the probability of more than 10 faults in 60 metres of cloth.

(3)

A retailer buys a large amount of this cloth and sells it in pieces of length x metres. He chooses x so that the probability of no faults in a piece is 0.80.

(*c*) Write down an equation for x and show that $x = 1.7$ to 2 significant figures.

(4)

The retailer sells 1200 of these pieces of cloth. He makes a profit of 60p on each piece of cloth that does not contain a fault but a loss of £1.50 on any pieces that do contain faults.

(*d*) Find the retailer's expected profit.

(4)

TOTAL FOR PAPER: 75 MARKS

END

Question Number		Scheme	Marks
1.	**(a)**	The list of ID numbers	**B1** **(1)**
	(b)	$F \sim B(50,0.02)$	**B1 B1** **(2)** **3**

2. **(a)**	$H_0 : p = 0.2$ $H_1 : p > 0.2$ B1 Under H_0 , $X \sim$ Bin(10,0.2) B1 $P(X \geq 4)$ $=1- P(X \leq 3)$ OR $P(X < 4) = 0.9672$ $= 1 - 0.8791$ $P(X > 5) = 0.0328$ $= 0.1209$ CR $X > 5$ A1 0.1209 > 0.05. Insufficient evidence to reject H_0 so teacher's claim is supported.	**B1** **B1** **M1** **A1** **M1 A1ft** **[6]**

3.	**(a)**	$X \sim$ Po (9) may be implied by calculations in part a or b	**M1**
		$P(X \leq 3) = 0.0212$ $P(X \geq 16) = 0.0220$	
		CR $X \leq 3$; $\cup$ $X \geq 16$	**A1; A1** **(3)**
	(b)	P(rejecting Ho) = 0.0212 + 0.0220	
		= 0.0432 or 0.0433	**M1** **A1 cao (2)** **(5 marks)**

4.	**(a)**	*A statistic* is a function of $X_1, X_2, ...X_n$ that does not contain any unknown parameters	B1 B1 (2)
	(b)	The probability distribution of Y or the distribution of all possible values of Y (o.e.)	B1 1)
	(c)	Identify (ii) as not a statistic Since it contains unknown parameters μ and σ.	B1 dB1 (2) **(5 marks)**

Question Number	Scheme	Marks
5. **(a)**	Mode = 3 from graph	B1 (1)
(b)	$\int_0^3 kx^2 \, dx = 0.5 \quad \Rightarrow \left[\frac{kx^3}{3}\right]_0^3 = 0.5$	M1 A1
	So $\frac{27k}{3} - 0 = 0.5 \Rightarrow k = \frac{1}{18}$ (using median = 3)	M1d A1 (4)
(c)	Height of triangle = $\frac{1}{18} \times 3^2 = \frac{1}{2}$	B1ft
	Area of triangle = $\frac{1}{2} \times (a-3) \times \frac{1}{2} = \frac{1}{2}$	M1
	so $a = 5$ cao	A1 (3)
(d)	From graph distribution is negative skew (left tail is longer)	B1
	$\mu <$ median for negative skew so E(X) < 3	B1d (2)
	[N.B. E(X) = $2\frac{23}{24}$]	**10**

6.	[X = number of customers joining the queue in the next 10 mins ~Po(3)]	
(a)	P($X = 4$) = P($X \leq 4$) – P($X \leq 3$) or $\frac{e^{-3}3^4}{4!}$	M1
	0.8153 – 0.6472 = 0.1681 or 0.1680313... (awrt **0.168**)	A1 **(2)**
(b)	Y [= number of customers joining the queue in the next 20 mins] ~ Po(6)	B1
	P($Y > 10$) = 1 – P($Y \leq 10$)	M1
	= 1 – 0.9574 = 0.0426(209...) (awrt **0.0426**)	A1 **(3)**
(c)	P($T > 3.5$) = **0.3**	B1 **(1)**
(d)	C ~ B(5, 0.3)	M1
	P($C \geq 3$) = 1 – P($C \leq 2$)	M1
	= 1 – 0.8369 = 0.1631 (0r 0.16308..) (awrt **0.163**)	A1 **(3)**
(e)	P(Bethan is served in < 4 minutes) = 0.8 (o.e.)	B1
	J = number joining the queue in 4 mins has J ~ Po(1.2)	M1
	P($J = 0$) = $e^{-1.2} = 0.30119...$	A1
	P(Bethan is served and $J = 0$) = $0.8 \times e^{-1.2} = 0.240955...$ (awrt **0.241**)	A1 **(4)** **[13]**

Question Number	Scheme	Marks
7.	Attempt to write down combinations — at least one seen	**M1**
	(1,1,1), (1,1,2) any order (1,2,2) any order, (2,2,2) — no extra combinations	**A1**
	Range 0 and 1 — 0 and 1 only	**B1**
	[P(range = 0) =] $(0.65)^3 + (0.35)^3$ — either range	**M1**
	$= 0.3175$ or $\frac{127}{400}$	**A1 cao**
	[P(range = 1) =] $(0.35)^2(0.65) \times 3 + (0.65)^2(0.35) \times 3$	
	$= 0.6825$ or $\frac{273}{400}$	**A1 cao**
		(6)
		Total 6

8.		
(a) (i)	a <u>population parameter</u> proposed by <u>the null hypothesis compared with an alternative hypothesis.</u>	**B1**
(ii)	The critical region is the <u>range of values</u> **or** <u>a test statistic or region where the test is significant</u> that would lead <u>to the rejection of H_0.</u>	**B1g** **B1h** **(3)**
(b)	Let X represent the number of incoming calls : X ~ Po(9)	**B1**
	From table	
	$P(X \geq 16) = 0.0220$	**M1 A1**
	$P(x \leq 3) = 0.0212$	**A1**
	Critical region ($x \leq 3$ or $x \geq 16$)	**B1** **(5)**
(c)	Significance level = 0.0220 + 0.0212	**B1**
	$= 0.0432$ or 4.32%	**(1)**
(d)	$H_0 : \lambda = 0.45$; $H_1 : \lambda < 0.45$ (accept : $H_0 : \lambda = 4.5$; $H_1 : \lambda < 4.5$)	**B1**
	Using X ~ Po(4.5)	**M1**
	$P(X \leq 1) = 0.0611$ CR $X < 0$ awrt 0.0611	**A1**
	0.0611 > 0.05. $1 \geq 0$ or 1not in the critical region	
	There is evidence to Accept H0 or it is not significant	**M1**
	There is no evidence that there are less calls during school holidays.	**B1cao** **(5)**

Question Number	Scheme	Marks
9. **(a)**	$X \sim \text{Po}(2)$ $\quad \text{P}(X=4) = \dfrac{\text{e}^{-2}\times 2^4}{4!} = 0.0902$ awrt 0.09	M1 A1 (2)
(b)	$Y \sim \text{Po}(8)$	B1
	$\text{P}(Y>10) = 1 - \text{P}(Y \le 10) = 1 - 0.8159 = 0.18411\ldots$ awrt 0.184	M1 A1 (3)
(c)	F = no. of faults in a piece of cloth of length x $\quad F \sim \text{Po}(x \times \frac{2}{15})$	
	$\text{e}^{-\frac{2x}{15}} = 0.80$	M1 A1
	$\text{e}^{-\frac{2}{15}\times 1.65} = 0.8025...,\quad \text{e}^{-\frac{2}{15}\times 1.75} = 0.791...$	M1
	These values are either side of 0.80 therefore x = 1.7 to 2 sf	A1 (4)
(d)	Expected number with no faults $= 1200\times 0.8 = 960$	M1
	Expected number with some faults $= 1200\times 0.2 = 240$	A1
	So expected profit $= 960\times 0.60 - 240\times 1.50,$ $= £216$	M1, A1 (4)
		(13 marks)

Examiner reports

Question 1

In part (a) many candidates had learnt standard definitions but were generally unable to apply these definitions in context. Although the word 'list' was evident in the majority of answers it often referred to the components or the sample and not the identity numbers.

Many candidates failed to appreciate what was required in part (b), merely stating the standard requirements of a sampling distribution. i.e. all the values taken by the variate and their associated probabilities. Indeed in some cases it was appreciated that B(50,0.02) was required, but unfortunately this appeared in part (a) and went unrewarded. A small number of candidates took 2% as 0.2.

Question 2

The overall response to this question was disappointing. A common incorrect 'alternative hypothesis' of $p < 0.2$ was frequently seen implying that 'not guessing' is an inferior strategy to 'guessing'. Other common errors included the use of $p = 0.4$ instead of $p = 0.2$ and using $P(X = 4)$ or $P(X \leq 4)$ instead of $P(X \geq 4)$. There were also problems with candidates' conclusions. It was fairly common for candidates to provide complete and correct responses to the entire question except the final contextual conclusion. Their correct statement "do not reject the null hypothesis" was often followed by an incorrect comment in context such as: "So she was not guessing" or "So reject the teacher's claim".

Question 3

Whilst many candidates knew what they were doing in part (a) they lost marks because they left their answers as $P(X \leq 3)$ etc and did not define the critical regions. A few candidates were able to get the figures 0.0212 and 0.0220 but then did not really understand what this meant in terms of the critical value. A critical region of $X \geq 15$ was common.

Part (b) was poorly answered. The wording "incorrectly rejecting H_0" confused many candidates. They often managed to get to 0.432 but then they took this away from 0.5 or occasionally 1. It was not uncommon for this to be followed by a long paragraph trying to describe what they had done.

Question 4

This question was either answered very well with some text book solutions, although it seemed that only a minority of candidates earned all five marks, or badly with some strange descriptions. A reasonable number of candidates responded with comments that were very close to those in the mark scheme: evidence possibly of deliberate preparation and learning whilst others had internalised the concepts and provided responses in their own words. Whilst these responses might not have matched the 'official' answers, they nevertheless captured the essence of the concepts and were fully acceptable. There was confusion with the definition of statistics and parameters and part (b) was often attempted badly with candidates not knowing the definition of a probability distribution. On the whole this was one of the worst answered questions in the paper.

In part (a) candidates gave various definitions sometimes all muddled up. Not many candidates gave clear definitions but a common error was candidates writing "***any*** *function*" or "*no* ***other*** *quantities*".

In part (b) again the candidates had mixed success. A significant minority scored marks by knowing that a sampling distribution involved all possible values of the statistic and their associated probabilities.

In part (c) many could identify (ii) correctly and a variety of reasons were seen. This part seemed to be done well even by candidates who could not answer part (a).

It was interesting to see that a relatively large proportion of candidates who earned both marks for part (c), were unable to achieve either of the two marks in part (a). There was a connection between parts (a) and (c) that many candidates failed to recognise. If those candidates who wrote "(ii) is not a statistic because it has unknown parameters" had then reflected on their responses to parts (a) and (c), they could then have gone back to modify their answer to (a) in order to earn more marks.

Question 5

Many candidates found this question challenging and there were few fully correct answers. Part (a) was well answered.

In part (b), almost all candidates correctly integrated the given probability density function, between the limits of 0 and 3. Unfortunately, the majority equated their definite integral to 1, instead of $\frac{1}{2}$ showing a misunderstanding of the concept that the total area is equal to 1.

In part (c), equating the area of the triangle to ½ proved to be surprisingly demanding for many candidates, although there were a significant percentage of concise solutions. A minority opted for finding the equation of the line segment, before integrating it. This in itself was rather complex especially if they chose B as the point on the line rather than (a, 0). This then had to be integrated leading to a lot of complex algebra, which rarely resulted in a correct solution. Some candidates clearly guessed the answer to (a) by looking at the diagram; this was awarded no marks.

Part (d) was very poorly answered. Only a small percentage of candidates stated that the distribution was negatively skewed. The vast majority were under the illusion that as median = mode then the mean must be the same value, thus drawing the conclusion that it was symmetrical despite evidence to the contrary on the diagram drawn on the question paper.

Question 6

This question allowed candidates to score many marks and only part (e) seemed to cause problems.

In part (a) some used the incorrect value for λ (usually 6) or calculated P($X \leq 5$) – P($X \leq 4$) or just calculated P($X \leq 4$). In part (b) the errors included using the incorrect value for λ (usually 9) or calculating 1 – P($X \leq 9$).

Part (c) was generally correct with only a very few putting 0.7.

In part (d) some candidates decided to calculate values rather than use the tables in (d) even though they had done so successfully in (b). The main error was to calculate 1 – P($X \leq 1$).

In part (e) it was evident that many candidates did not understand the context of the question clearly enough. Others correctly identified 0.8 but nothing else whilst some identified Po(1.2) and used it correctly Those who calculated both generally went on to score full marks but a few candidates lost the final mark as they either added the values to give a probability > 1 or gave the answer as 0.24 and not 0.241.

Question 7

Virtually all candidates were able to write down the eight combinations of 1s and 2s and although most candidates were able to calculate the probability of each of these outcomes, the meaning of "range" seems to have eluded all but a minority of candidates. The key to the question was remembering that the 'range' of a list of numbers is obtained by subtracting the lowest number from the highest.

Many candidates obtained the sampling distribution of (in descending order of popularity) the mean, the total score, the median or the mode. Some candidates gave the sampling distribution of two (or more) of these statistics, without ever considering the range.

Of those candidates who considered the range, the most common error was to state that the range was either 1 or 2.

Question 8

This question appeared to be difficult for many candidates with a large proportion achieving less than half the available marks.

(a) The majority of candidates were unable to give an accurate description of a hypothesis test as a method of deciding between 2 hypotheses. There were more successful definitions of a critical region but many candidates achieved only 0 or 1 of the 3 available marks. Common errors included too much re-use of the word region without any expansion on it. Even those who could complete the rest of the question with a great deal of success could not describe accurately what they were actually doing.

(b) Although most of those attempting this part of the question realised that a Poisson distribution was appropriate there was a sizeable number who used a Binomial distribution. Again, the most common problem was in expressing and interpreting inequalities in order to identify the critical regions. Many found the correct significance level but struggled to express the critical region correctly. Answers with 15 were common and some candidates even decided that 4 to 15 was the CR.

(c) Those candidates that identified the correct critical regions were almost always able to state the significance level correctly, as were some who had made errors in stating these regions. Some still gave 5% even with part (b) correct.

Candidates who had used a Binomial distribution in part (b), and many of those who had not, used p instead of λ in stating the hypotheses and went on to obtain a Binomial probability in this part of the question. In obtaining $P(X \leq 1)$ some used Po(9) from part (b) instead of Po(4.5). Most of those achieving the correct statement (failing to reject H_0) were able to place this in a suitable contextualised statement. There were some candidates who still tried to find $P(X=1)$ rather than $P(X \leq 1)$.

Question 9

Parts (a) and (b) were completed successfully by most candidates. The most common errors seen were using the wrong Poisson parameter or identifying the incorrect probability in part (b).

Part (c) proved to be a good discriminator with only those with good mathematical skills able to attain all the marks for this part of the question. Few candidates used the method given on the mark scheme and chose to use natural logarithms instead. Whilst this is an accepted method this knowledge is not expected at S2 and full marks were gained by the most able candidates using the given method.

In part (d) a few candidates seemed confused by this with some using 1.7 or 2/15 as a probability rather than the 0.8 given in the question, and far too many seemed unable to use 60p and £1.50 correctly when calculating the profit.

Statistics for S2 Practice Paper Gold 3

Qu	Max Score	Modal score	Mean %	Mean average scored by candidates achieving grade: ALL	A*	A	B	C	D	E	U
1	3		52.3	1.57	1.91	1.79	1.56	1.42	1.24	1.12	0.77
2	6		65.2	3.91	4.60	4.43	2.95	2.53	1.53	1.09	0.36
3	5		57.6	2.88		3.61	3.00	2.54	1.94	1.24	0.63
4	5		44.8	2.24		2.99	2.08	1.64	1.30	1.13	0.71
5	10		48.8	4.88	7.42	6.22	4.60	3.87	3.34	2.70	1.79
6	13		78.2	10.16	12.02	11.01	9.76	8.40	7.39	6.22	4.06
7	6		51.2	3.07	4.60	3.95	2.99	2.33	1.85	1.39	0.79
8	14		54.7	7.66		10.55	6.43	4.06	3.56	1.83	0.59
9	13		62.8	8.17		10.61	8.29	6.86	5.67	4.69	2.41
	75		**59.4**	**44.54**		**55.16**	**41.66**	**33.65**	**27.82**	**21.41**	**12.11**