Paper Reference(s)

6684/01

Edexcel GCE

Statistics S2

Gold Level G4

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 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
60	49	36	28	20	13

1. Explain what you understand by

(*a*) a population,

(1)

(*b*) a statistic.

(1)

A researcher took a sample of 100 voters from a certain town and asked them who they would vote for in an election. The proportion who said they would vote for Dr Smith was 35%.

(*c*) State the population and the statistic in this case.

(2)

(*d*) Explain what you understand by the sampling distribution of this statistic.

(1)

---

2. A bag contains a large number of 1p, 2p and 5p coins.

50% are 1p coins

20% are 2p coins

30% are 5p coins

A random sample of 3 coins is chosen from the bag.

(*a*) List all the possible samples of size 3 with median 5p.

(2)

(*b*) Find the probability that the median value of the sample is 5p.

(4)

(*c*) Find the sampling distribution of the median of samples of size 3.

(5)

---

3. A rectangle has a perimeter of 20 cm. The length, X cm, of one side of this rectangle is uniformly distributed between 1 cm and 7 cm.

Find the probability that the length of the longer side of the rectangle is more than 6 cm long.

(5)

---

4. A bag contains a large number of coins:

75% are 10p coins,

25% are 5p coins.

A random sample of 3 coins is drawn from the bag.

Find the sampling distribution for the median of the values of the 3 selected coins.

(7)

---

5. The continuous random variable T is used to model the number of days, t, a mosquito survives after hatching.

The probability that the mosquito survives for more than t days is

$$\frac{225}{(t+15)^2}, \quad t \geq 0.$$

(*a*) Show that the cumulative distribution function of T is given by

$$\text{F}(t) = \begin{cases} 1 - \dfrac{225}{(t+15)^2}, & t \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(1)

(*b*) Find the probability that a randomly selected mosquito will die within 3 days of hatching. **(2)**

(*c*) Given that a mosquito survives for 3 days, find the probability that it will survive for at least 5 more days.

(3)

A large number of mosquitoes hatch on the same day.

(*d*) Find the number of days after which only 10% of these mosquitoes are expected to survive.

(4)

---

6. Frugal bakery claims that their packs of 10 muffins contain on average 80 raisins per pack. A Poisson distribution is used to describe the number of raisins per muffin.

A muffin is selected at random to test whether or not the mean number of raisins per muffin has changed.

(*a*) Find the critical region for a two-tailed test using a 10% level of significance. The probability of rejection in each tail should be less than 0.05.

(4)

(*b*) Find the actual significance level of this test.

(2)

The bakery has a special promotion claiming that their muffins now contain even more raisins.

A random sample of 10 muffins is selected and is found to contain a total of 95 raisins.

(*c*) Use a suitable approximation to test the bakery's claim. You should state your hypotheses clearly and use a 5% level of significance.

(8)

7.

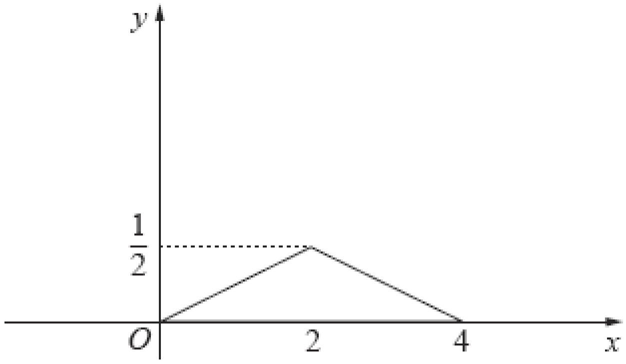


Figure 1

Figure 1 shows a sketch of the probability density function f(x) of the random variable X. The part of the sketch from x = 0 to x = 4 consists of an isosceles triangle with maximum at (2, 0.5).

(a) Write down E(X).

(1)

The probability density function f(x) can be written in the following form.

$$f(x) = \begin{cases} ax & 0 \leq x < 2 \\ b - ax & 2 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

(b) Find the values of the constants a and b.

(2)

(c) Show that σ, the standard deviation of X, is 0.816 to 3 decimal places.

(7)

(d) Find the lower quartile of X.

(3)

(e) State, giving a reason, whether P($2 - \sigma < X < 2 + \sigma$) is more or less than 0.5

(2)

8. In a large restaurant an average of 3 out of every 5 customers ask for water with their meal.

A random sample of 10 customers is selected.

(*a*) Find the probability that

(i) exactly 6 ask for water with their meal,

(ii) less than 9 ask for water with their meal.

(5)

A second random sample of 50 customers is selected.

(*b*) Find the smallest value of n such that

$$P(X < n) \geq 0.9,$$

where the random variable X represents the number of these customers who ask for water.

(3)

TOTAL FOR PAPER: 75 MARKS

END

Question Number	Scheme	Marks
1. (a)	A population is collection of all items	B1 (1)
(b)	(A random variable) that is a function of the sample which contains no unknown quantities/parameters.	B1 (1)
(c)	The voters in the town	B1
	Percentage/proportion voting for Dr Smith	B1 (2)
(d)	Probability Distribution of those voting for Dr Smith from all possible samples (of size 100)	B1 (1)
		[5]

2. (a)	(5,5,5) or (1,5,5) or (2,5,5)	B1
	(5,5,5) (5,5,1) (5,1,5) (1,5,5) (5,5,2) (5,2,5) (2,5,5) or (5,5,5) **and** (5,5,1) $(\times 3)$ **and** (5,5,2) $(\times 3)$	B1 (2)
(b)	(5,5,5) $\left(\frac{3}{10}\right)^3 = \frac{27}{1000} = 0.027$	B1
	(5,5,1) $3\times\frac{1}{2}\times\left(\frac{3}{10}\right)^2 = \frac{135}{1000}$ *or* $\frac{27}{200} = 0.135$	M1
	(5,5,2) $3\times\frac{1}{5}\times\left(\frac{3}{10}\right)^2 = \frac{54}{1000} = \frac{27}{500} = 0.054$	
	$P(M=5) = \left(\frac{3}{10}\right)^3 + 3\times\frac{1}{2}\times\left(\frac{3}{10}\right)^2 + 3\times\frac{1}{5}\times\left(\frac{3}{10}\right)^2 = \frac{27}{125} = 0.216$ oe	A1A1 (4)
(c)	$P(M=1) = (0.5)^3 + 3(0.5)^2(0.2) + 3(0.5)^2(0.3)$	M1
	$= 0.5$	A1
	$P(M=2) = \left(\frac{1}{5}\right)^3 + 3\times\left(\frac{1}{5}\right)^2\times\frac{1}{2} + 3\times\left(\frac{1}{5}\right)^2\times\frac{3}{10} + 6\times\frac{1}{2}\times\frac{1}{5}\times\frac{3}{10}$	M1
	$= 0.284$ or $\frac{71}{250}$ oe	A1
	(table below)	A1 (5)
		[11]

m	1	2	5
$P(M=m)$	0.5	0.284	0.216

Question Number	Scheme			Marks
3.	$P(X > 6) = \frac{1}{6}$	$P(4 < X < 6) = \frac{1}{3}$	$P(X > 6) = \frac{1}{6}$	B1 M1
	$P(X < 4) = \frac{1}{2}$		Y~U[3,9] $P(Y > 6) = \frac{1}{2}$	A1
	total = $\frac{1}{6}+\frac{1}{2}=\frac{2}{3}$	$1-\frac{1}{3}=\frac{2}{3}$	total = $\frac{1}{6}+\frac{1}{2}=\frac{2}{3}$	M1dep B A1 (5)
				[5]

4.	Attempt to write down combinations — at least one seen	M1
	(5,5,5), (5,5,10) any order (10,10,5) any order, (10,10,10)	A1
	(5,10,5), (10,5,5), (10,5,10), (5,10,10), — all 8 cases considered. May be implied by 3 * (10,5,10) and 3 * (5,5,10)	A1
	median 5 and 10	B1
	Median = 5 P(M = m) = $\left(\frac{1}{4}\right)^3+3\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)=\frac{10}{64}$ = 0.15625 — add at least two prob using ¼ and ¾. identified by having same median of 5 or 10 Allow no 3 for M	M1 A1
	Median = 10 P(M = m) = $\left(\frac{3}{4}\right)^3+3\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)=\frac{54}{64}$ = 0.84375	A1 (7) Total 7

Question Number	Scheme	Marks
5 **(a)**	$P(T > t) = \frac{225}{(t+15)^2}$ $P(T \le t) = 1 - P(T > t)$ $= 1 - \frac{225}{(t+15)^2}$ $F(t) = \begin{cases} 1 - \frac{225}{(t+15)^2} & t \ge 0 \\ 0 & \text{otherwise.} \end{cases}$	B1 **(1)**
(b)	$P(T < 3) = 1 - \frac{225}{(3+15)^2}$ $= \frac{11}{36}$ or 0.30555… awrt 0.306	M1 A1 **(2)**
(c)	$P(T > 8 \mid T > 3) = \frac{P(T > 8)}{P(T > 3)}$ $= \frac{\frac{225}{23^2}}{\frac{225}{18^2}}$ $= \frac{324}{529}$ or 0.612.. awrt 0.612 / 0.6125	M1 M1 A1 **(3)**
(d)	$1 - F(t) = 0.1$ $\frac{225}{(t+15)^2} = 0.1$ or $1 - \frac{225}{(t+15)^2} = 0.9$ $\frac{225}{0.1} = (t+15)^2$ $t = \sqrt{\frac{225}{0.1}} - 15$ t =32.4, also accept 32/33	M1 A1 M1 A1 **(4)** **Total 10**

Question Number	Scheme	Marks
6.	[X = the number of raisins in a mini-muffin]	
(a)	$X \sim \text{Po}(8)$	B1
	e.g. $P(X \le 3) = 0.0424$, $P(X \le 13) = 0.9658$ so $P(X \ge 14) = 0.0342$	M1
	So Critical Region is $X \le 3$ or $X \ge 14$	A1 A1
		(4)
(b)	$0.0424 + 0.0342$	M1
	= **<u>0.0766</u>** (or better)	A1
		(2)
(c)	$H_o: \lambda = 8$ (or μ=80) $H_1: \lambda > 8$ (or $\mu > 80$)	B1
	[R = no. of raisins in 10 muffins. $R \sim \text{Po}(80)$.] Use $Y \sim N(80, 80)$	M1A1
	$P(R \ge 95) \simeq P(Y \ge 94.5)$	M1
	$= P\left(Z > \frac{94.5-80}{\sqrt{80}}\right)$	M1
	$= P(Z > 1.62...) = 1 - 0.9474 =$ awrt **<u>0.053</u>**	A1
	Probability is greater than 0.05 so not significant (accept H_0)	M1
	Insufficient evidence to support the <u>bakery's claim</u>	A1cso
	<u>Or</u> insufficient evidence of an increase in the (mean) number of <u>raisins</u> per <u>muffin</u>	**(8)**
		[14]

7. **(a)**	$E(X) = 2$ (by symmetry)	B1 (1)
(b)	$0 \le x < 2$, gradient $= \frac{\frac{1}{2}}{2} = \frac{1}{4}$ and equation is $y = \frac{1}{4}x$ so $a = \frac{1}{4}$	B1
	$b - \frac{1}{4}x$ passes through (4, 0) so $b = 1$	B1 (2)
(c)	$E(X^2) = \int_0^2 \left(\frac{1}{4}x^3\right) dx + \int_2^4 \left(x^2 - \frac{1}{4}x^3\right) dx$	M1 M1
	$= \left[\frac{x^4}{16}\right]_0^2 + \left[\frac{x^3}{3} - \frac{x^4}{16}\right]_2^4$	A1
	$= 1 + \frac{64-8}{3} - \frac{256-16}{16} = 4\frac{2}{3}$ or $\frac{14}{3}$	M1 A1
	$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{14}{3} - 2^2$, $= \frac{2}{3}$ (so $\sigma = \sqrt{\frac{2}{3}} = 0.816$) **(*)**	M1 A1cso (7)
(d)	$P(X \le q) = \int_0^q \frac{1}{4}x\, dx = \frac{1}{4}$, $\frac{q^2}{2} = 1$ so $q = \sqrt{2} = 1.414$ awrt 1.41	M1 A1, A1 (3)
(e)	$2 - \sigma = 1.184$ so $2 - \sigma$, $2 + \sigma$ is wider than IQR, therefore greater than 0.5	M1, A1 (2)
		(15 marks)

Question Number	Scheme			Marks
8	Let X be the random variable the number of customers asking for water.			
(a) (i)	$X \sim B(10,0.6)$	$Y \sim B(10,0.4)$		B1
	$P(X = 6) = (0.6)^6(0.4)^4 \frac{10!}{6!4!}$	$P(Y = 4) = (0.4)^4(0.6)^6 \frac{10!}{6!4!}$		M1
	= 0.2508…	= 0.2508	awrt 0.251	A1
(ii)	$X \sim B(10,0.6)$	$Y \sim B(10,0.4)$		
	$P(X < 9) = 1 – (P(X = 10) + P(X = 9))$ $= 1 - (0.6)^{10} - (0.6)^9(0.4)^1 \frac{10!}{9!1!}$	$P(X < 9) = 1 – P(Y \le 1)$ $= 1 – 0.0464$		M1
	= 0.9536…	= 0.9536…	awrt 0.954	A1
				(5)
(b)	$X \sim B(50,0.6)$			M1
	$Y \sim B(50,0.4)$			
	$P(X < n) \ge 0.9$			
	$P(Y > 50 – n) \ge 0.9$	or $P(X < 34) = 0.8439$ awrt 0.844		
	$P(Y \le 50 – n) \le 0.1$	$P(X < 35) = 0.9045$ awrt 0.904/0.905		M1
	$50 – n \le 15$			
	$n \ge 35$			
	$n = 35$			A1
				(3) **Total 8**

Examiner reports

Question 1

This was poorly done with very few candidates scoring full marks. Those candidates who had learnt standard definitions fared better than those who used their own understanding of the terms because they were less likely to leave out vital elements of the definitions. Even those who answered parts (a) and (b) correctly were then unable to apply these definitions in context.

In part (a) a large majority of candidates omitted to mention “all”, or its equivalent.

Part (b) was well answered because many candidates used a standard definition. The most common errors were using “population” instead of “sample and omitting “no unknown parameters”.

In part (c) a substantial number of candidates were confused about “the population in this case”. Many thought it to be the sample of 100 voters. Others were closer to the truth with “all the residents of the town”, but did not earn the mark because they had failed to distinguish between registered voters and residents. The statistic was more easily identified.

Part (d) was poorly answered with many candidates having no idea what a sampling distribution was and those that did being unable to put it into context. The sampling distribution of a proportion is arguably one of the hardest to get a grip on and articulate convincingly.

Question 2

The opening question, testing the sampling distribution of the median, proved to be both challenging and a good discriminator for candidates. Some responses contained lengthy lists of many, if not all, 27 possible permutations of the possible samples. However, some candidates were able to express their probabilities concisely.

In part (a), many candidates were able to obtain at least one of the two marks available, but common errors included failing to give all three permutations of the triples (1, 5, 5) and (2, 5, 5) when asked to list all possible samples to include extra triples such as (1, 5, 2) in the mistaken belief that this has a median of 5 because it has been written as the “middle number” in the triple.

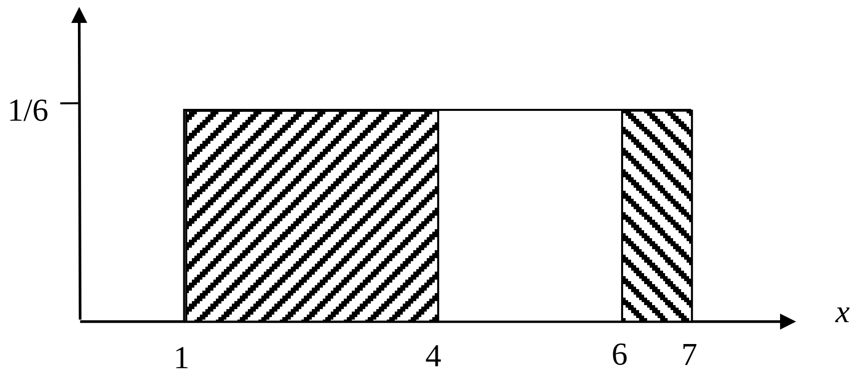
In part (b), a large proportion of candidates were able to obtain each of the probabilities (0.3)3 with either (0.5)(0.3)2 or (0.2)(0.3)2, though the factors of 3 for the latter two cases were often absent.

Part (c) was more testing. The most efficient approach was to calculate $P(M =1)$ and then use $P(M = 2) = 1 – P(M = 1) – P(M = 5)$. However, much more often, candidates tried to calculate $P(M = 2)$ directly, some with success, but often their calculation omitted either some or all of the 6 permutations of (1, 2, 5). Some weaker candidates made attempts at the sampling distribution of the mean.

It was also interesting to notice that a significant proportion of candidates who obtained the correct probability of 0.216 in part (b), either did not include it in their sampling distribution in part (c) or gave a different value in part (c). It was also apparent that the vast majority of candidates do not check their work as their the three probabilities they gave did not add up to 1.

Question 3

A minority of students got this completely correct and those that did often showed minimal working with only a diagram as evidence of their method. A major source of error was treating the distribution as a discrete uniform rather than a continuous uniform. Most students, however, simply worked out $P(X > 6)$ and gave this as their answer, unaware that there was more to the question than this. The easiest and perhaps most successful solutions came from candidates who drew a diagram and realised they needed $P(X < 4) + P(X > 6)$.



Question 4

This question differentiated between candidates. It was disappointing to see the majority of candidates had no idea how to find the probability distribution of the median although quite a few specified the definition of a probability distribution. Several were able to write down the combinations but did not understand how to find the median and construct the probability distribution. There were few completely accurate solutions.

Question 5

Many candidates found parts (a) and (c) difficult.

Part (a) was testing candidates' understanding of the cumulative distribution function, $F(t)$. It perhaps helps to think visually, in terms of area, but ultimately the algebraic definition $F(t) = P(T \leq t)$ is required.

In part (b) nearly all candidates were able to calculate $P(T < 3) = 11/36$. The response to part (c) was disappointing with very many candidates scoring no marks. Of those who realised that a conditional probability was required the most common error was to find $\frac{P(X \geq 5)}{P(X \geq 3)}$ or $\frac{P(X \geq 5)}{P(X \leq 3)}$. Part (d) required candidates to solve a quadratic equation. Any written method would have been acceptable, but by far the easiest was to rearrange the equation to obtain $(t + 15)^2 = 2250$ and then to square root both sides. Many candidates multiplied out the brackets and then either used the formula or completed the square.

Question 6

Many candidates were able to find the correct critical region for part (a) although the upper region, as usual, caused more problems. The most common error was to write the upper CR as $X \geq 13$. A few candidates gave their CR as probability statement and some lost all the marks as they used Po(10) instead of Po(8).

The concept of the actual significance level was well understood even if they didn't always manage to obtain the correct answer.

Part (c) had many good solutions although for some clearly there was great confusion between the test statistic, 95, and the null value of the population parameter, 8 or 80. This was more evident amongst those who made the fundamental mistake of altering the test statistic and dividing 95 by 10 creating a different hypothesis test altogether. The most common error was to use N(8, 8)

Candidates seemed to know that they needed to use a continuity correction but many candidates applied it incorrectly. The final M mark saved many candidates as they could give a correct statement based on their probability.

Contextual statements were generally very good and candidates seemed to have improved in this area. However those candidates that took the route involving the 'bakery's claim' were more successful than those who tried to write a contextualised statement involving 'raisins' and 'muffins'. If the second approach was taken and the mark was not awarded it was usually because the word 'muffin' was not used.

Question 7

A minority of candidates achieved a high rate of success on this question.

In part (a) most candidates were able to write E(X) = 2 without difficulty.

A variety of methods were seen in part (b). The method of the mark scheme was seen, perhaps only from a minority of candidates. Many candidates preferred to use calculus: $\int \mathrm{f}(x)\mathrm{d}x = 1$. However, the use of calculus requires more subtlety and sensitivity than was available to many of the candidates. Answers of a = ½ and b = 2 seemed to be not uncommon, resulting from the incorrect methods: $\int_0^2 ax \, \mathrm{d}x = 1$ and $\int_2^4 (b - ax) \, \mathrm{d}x = 1$.

There were candidates who obtained the correct answers using calculus, but it often took considerable working, in contrast to the expected method.

There were some candidates who obtained full marks to part (c) with solutions that were confident, fluent and accurate. Furthermore, many of these responses were also efficient: four or five lines of working provided a solution that was not just correct but contained all the required details. However, a wide variety of alternative responses were also seen. Some were indeed correct, but inefficient. Other candidates used an incorrect strategy. Some candidates only worked with the domain $2 \le x \le 4$. Others worked with both domains, but wanted to keep the domains separate, resulting in two separate versions of $\mathrm{Var}(X)$:

$$\mathrm{Var}(X) = \int_0^2 x^2 \frac{1}{4} x \, \mathrm{d}x - \left(\int_0^2 x \frac{1}{4} x \, \mathrm{d}x \right)^2 \text{ and } \mathrm{Var}(X) = \int_2^4 x^2 \left(x - \frac{1}{4} x \right) \mathrm{d}x - \left(\int_2^4 x \left(x - \frac{1}{4} x \right) \mathrm{d}x \right)^2$$

Some candidates then calculated the average of these two versions of the variance.

Many candidates also found E(X) from scratch in this part rather than using the answer they had in part (a). Not only did this waste time, but whilst they often had it correct in part (a) they gained an incorrect value by integration in this part which they then went on to use.

It must be noted that where the answer to a question is given, marks cannot be gained by restating this without sufficient working. Some attempts were made to describe the answer as proven even though no real working had been done.

A reasonable number of correct solutions to part (d) were seen. Some candidates went so far as to specify fully the cumulative distribution function before using the correct part to find the lower quartile. Even though this extra work was not required, strictly speaking, it did provide these candidates with a ready method for part (e).

It would appear that whilst most candidates attempted part (e), their responses consisted of a simple statement, usually “greater than 0.5” together with an irrelevant reason. A tiny minority of candidates responded in the manner intended. A few provided a clear diagram to illustrate this same argument. However, the majority of successful candidates preferred to evaluate the probability. This was not straightforward, except for those who had already obtained a full and correct version of the cumulative distribution function. Part (e) seemed to challenge all but the most able candidates.

Question 8

There were many correct answers to part (a). However, this was not a unanimous response. The theme of this question was probability distributions that do not appear in the tables. There were many candidates whose strategy was to work around this problem by using a Poisson approximation to the Binomial. For this particular Binomial distribution, $X \sim B(10, 0.6)$, it is not appropriate to use a Poisson approximation. The candidates who were confident using the formula for Binomial probability were most successful, obtaining the required probabilities in (i) and (ii) without the use of the cumulative probability tables. The candidates who felt they must use the tables and considered the related distribution $Y \sim B(10,0.4)$ were not quite as successful. Not all candidates made the necessary amendments: for instance, in (a)(ii) we require $P(X \le 8)$, but cannot just look up $P(Y \le 8)$ in the tables. It is essential to realise that X and Y are related by $X + Y = 10$. We can therefore replace X by $10 - Y$, and rewrite $P(X \le 8)$ as $P(Y \ge 2)$, and then evaluate this using the standard strategy of $1 - P(Y \le 1)$.

In part (b) many candidates appreciated that they needed to use either B(50, 0.6) or B(50, 0.4). However, many were unable to proceed further. Success in part (b) relied on using the relationship $X + Y = 50$ so that X can be replaced by $50 - Y$. After some algebraic manipulation, the inequality in the question, $P(X < n) \ge 0.9$, is transformed into $P(Y \le 50 - n) \le 0.1$. This proved elusive to the great majority of candidates and the final two marks were often not awarded. Very common incorrect answers were 15, 16, 24, 25, 26 and 34.

Statistics for S2 Practice Paper Gold 4

				Mean average scored by candidates achieving grade:							
Qu	**Max Score**	**Modal score**	**Mean %**	**ALL**	**A***	**A**	**B**	**C**	**D**	**E**	**U**
1	5		33.0	1.65	2.45	2.05	1.58	1.26	1.00	0.86	0.72
1	11	11	49.4	5.43	9.02	7.67	5.34	3.57	2.46	1.61	0.81
3	5		41.0	2.05	3.46	2.73	1.87	1.56	1.17	0.95	0.57
4	7		43.1	3.02		4.36	2.86	2.16	1.55	1.07	0.65
5	10	6	49.0	4.86	6.74	5.69	3.71	2.98	2.14	1.56	0.89
6	14		70.3	9.84	12.53	11.20	8.57	7.34	5.43	2.95	2.02
7	15		50.8	7.62		11.27	7.34	5.38	3.69	2.47	1.43
8	8		49.6	3.97	5.62	4.86	3.86	3.36	2.80	2.38	1.40
	75		**51.2**	**38.44**		**49.83**	**35.13**	**27.61**	**20.24**	**13.85**	**8.49**