Paper Reference(s)

6691/01

Edexcel GCE

Statistics S3

Advanced Level

Wednesday 17 June 2009 – Afternoon

Time: 1 hour 30 minutes

Materials required for examination	Items included with question papers
Mathematical Formulae (Orange or 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

In the boxes on the answer book, write the name of the examining body (Edexcel), your centre number, candidate number, the unit title (Statistics S3), the paper reference (6691), your surname, other name and signature.
Values from the statistical tables should be quoted in full. When a calculator is used, the answer should be given to an appropriate degree of accuracy.

Information for Candidates

A booklet 'Mathematical Formulae and Statistical Tables' is provided.
Full marks may be obtained for answers to ALL questions.
This paper has 8 questions.
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 not gain full credit.

M34278A

1. A telephone directory contains 50 000 names. A researcher wishes to select a systematic sample of 100 names from the directory.

(*a*) Explain in detail how the researcher should obtain such a sample.

(2)

(*b*) Give one advantage and one disadvantage of

(i) quota sampling,

(ii) systematic sampling.

(4)

---

2. The heights of a random sample of 10 imported orchids are measured. The mean height of the sample is found to be 20.1 cm. The heights of the orchids are normally distributed.

Given that the population standard deviation is 0.5 cm,

(*a*) estimate limits between which 95% of the heights of the orchids lie,

(3)

(*b*) find a 98% confidence interval for the mean height of the orchids.

(4)

A grower claims that the mean height of this type of orchid is 19.5 cm.

(*c*) Comment on the grower's claim. Give a reason for your answer.

(2)

---

3. A doctor is interested in the relationship between a person's Body Mass Index (BMI) and their level of fitness. She believes that a lower BMI leads to a greater level of fitness. She randomly selects 10 female 18 year-olds and calculates each individual's BMI. The females then run a race and the doctor records their finishing positions. The results are shown in the table.

Individual	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
BMI	17.4	21.4	18.9	24.4	19.4	20.1	22.6	18.4	25.8	28.1
Finishing position	3	5	1	9	6	4	10	2	7	8

(*a*) Calculate Spearman's rank correlation coefficient for these data.

(5)

(*b*) Stating your hypotheses clearly and using a one tailed test with a 5% level of significance, interpret your rank correlation coefficient.

(5)

(*c*) Give a reason to support the use of the rank correlation coefficient rather than the product moment correlation coefficient with these data.

(1)

---

4. A sample of size 8 is to be taken from a population that is normally distributed with mean 55 and standard deviation 3. Find the probability that the sample mean will be greater than 57.

(5)

---

5. The number of goals scored by a football team is recorded for 100 games. The results are summarised in Table 1 below.

Number of goals	**Frequency**
0	40
1	33
2	14
3	8
4	5

Table 1

(*a*) Calculate the mean number of goals scored per game.

(2)

The manager claimed that the number of goals scored per match follows a Poisson distribution. He used the answer in part (*a*) to calculate the expected frequencies given in Table 2.

Number of goals	**Expected Frequency**
0	34.994
1	r
2	s
3	6.752
≥ 4	2.221

Table 2

(*b*) Find the value of r and the value of s giving your answers to 3 decimal places.

(3)

(*c*) Stating your hypotheses clearly, use a 5% level of significance to test the manager's claim.

(7)

6. The lengths of a random sample of 120 limpets taken from the upper shore of a beach had a mean of 4.97 cm and a standard deviation of 0.42 cm. The lengths of a second random sample of 150 limpets taken from the lower shore of the same beach had a mean of 5.05 cm and a standard deviation of 0.67 cm.

(*a*) Test, using a 5% level of significance, whether or not the mean length of limpets from the upper shore is less than the mean length of limpets from the lower shore. State your hypotheses clearly.

(8)

(*b*) State two assumptions you made in carrying out the test in part (*a*).

(2)

---

7. A company produces climbing ropes. The lengths of the climbing ropes are normally distributed. A random sample of 5 ropes is taken and the length, in metres, of each rope is measured. The results are given below.

120.3 120.1 120.4 120.2 119.9

(*a*) Calculate unbiased estimates for the mean and the variance of the lengths of the climbing ropes produced by the company.

(5)

The lengths of climbing rope are known to have a standard deviation of 0.2 m. The company wants to make sure that there is a probability of at least 0.90 that the estimate of the population mean, based on a random sample size of n, lies within 0.05 m of its true value.

(*b*) Find the minimum sample size required.

(6)

---

8. The random variable A is defined as

$$A = 4X - 3Y$$

where $X \sim N(30, 3^2)$, $Y \sim N(20, 2^2)$ and X and Y are independent.

Find

(*a*) E(A),

(2)

(*b*) Var(A).

(3)

The random variables Y_1, Y_2, Y_3 and Y_4 are independent and each has the same distribution as Y. The random variable B is defined as

$$B = \sum_{i=1}^{4} Y_i \ .$$

(*c*) Find P($B > A$).

(6)

TOTAL FOR PAPER: 75 MARKS

END

Question Number	Scheme	Marks
Q1 (a)	Randomly select a number between 00 and 499 (001 and 500) select every 500th person	B1 B1 (2)
(bi)	<u>Quota</u> **Advantage:** <u>Representative</u> sample can be achieved (with small sample size) <u>Cheap</u> (costs kept to a minimum) <u>**not**</u> "quick" Administration relatively <u>easy</u> **Disadvantage** Not possible to estimate sampling errors (due to lack of randomness) Not a random process Judgment of interviewer can affect choice of sample – <u>bias</u> Non-response not recorded Difficulties of defining controls e.g. social class	B1 B1 (2)
(bii)	<u>Systematic</u> **Advantage:** <u>Simple</u> or <u>easy</u> to use <u>**not**</u> "quick" or "cheap" or "efficient" It is suitable for large <u>samples</u> (not populations) **Disadvantage** Only random if the ordered list is (truly) random Requires a list of the population <u>or</u> must assign a number to each member of the pop.	B1 B1 (2) **[6]**
Q2 (a)	Limits are $20.1 \pm 1.96 \times 0.5$ <u>**(19.1, 21.1)**</u>	M1 B1 A1cso (3)
(b)	98 % confidence limits are $20.1 \pm 2.3263 \times \frac{0.5}{\sqrt{10}}$ <u>**(19.7, 20.5)**</u>	M1 B1 A1A1 (4)
(c)	The growers claim is not correct Since 19.5 does not lie in the interval (19.7, 20.5)	B1 dB1 (2) **[9]**

Question Number	Scheme	Marks

Q3 (a)

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
BMI	1	6	3	8	4	5	7	2	9	10
or	10	5	8	3	7	6	4	9	2	1
Finishing position	3	5	1	9	6	4	10	2	7	8
d^2	4	1	4	1	4	1	9	0	4	4

M1

$\sum d^2 = 32 \quad (298)$ — M1

$r_s = 1 - \dfrac{6\times 32}{10\times 99}$ — M1 A1ft

$= 0.80606\ldots \quad (-0.80606) \qquad \text{accept } \pm\dfrac{133}{165}$ **<u>awrt $\pm$ 0.806</u>** — A1 (5)

(b)

$H_0 : \rho = 0,\ H_1 : \rho > 0,$ — B1 B1

Critical value is $(\pm)0.5636$ — B1

(0.806 > 0.5636 therefore) in critical region/ reject H_0 — M1

The lower the BMI the higher the position in the race./ support for doctors belief — A1ft (5)

(c)

The position is already ranked OR Position is not Normally distributed — B1 (1)

[11]

Q4

$X \sim N(55, 3^2)$ therefore $\overline{X} \sim N\left(55, \dfrac{9}{8}\right)$ — B1 B1

$P(\overline{X} > 57) = P\left(Z > \dfrac{57-55}{\sqrt{\frac{9}{8}}}\right) \quad = P(Z > 1.8856\ldots)$ — M1

$= 1 - 0.9706$ — M1

$= 0.0294$ **<u>0.0294~0.0297</u>** — A1

[5]

Question Number	Scheme	Marks
Q5 (a)	$\lambda = \dfrac{0\times40+1\times33+2\times14+3\times8+4\times5}{100} = 1.05$	M1 A1 (2)
(b)	Using Expected frequency = 100 × P($X = x$) = 100 × $\dfrac{e^{-1.05}1.05^x}{x!}$ gives	M1
	$r = 36.743$ awrt 36.743 or 36.744	A1
	$s = 19.290$ 19.29 or awrt 19.290	A1 (3)
(c)	H_0 : Poisson distribution is a suitable model H_1 : Poisson distribution is not a suitable model	B1
	Number of goals / Frequency / Expected frequency: 0 / 40 / 34.994; 1 / 33 / 36.743; 2 / 14 / 19.290; 3 / 8 / 6.752; ≥ 4 / 5 / 2.221; (3 and ≥ 4 combined: 8.972443)	M1
	$\upsilon = 4 - 1 - 1 = 2$	B1ft
	CR : $\chi^2_2(0.05) > 5.991$	B1
	$\sum\dfrac{(O-E)^2}{E} = \dfrac{(40-34.9937)^2}{34.9937} + \ldots + \dfrac{(13-8.972443)^2}{8.972443}$	M1
	[=0.7161…+0.3813…+1.4508…+1.80789..] = 4.356. (ans in range 4.2 – 4.4)	A1
	Not in critical region Number of goals scored can follow a Poisson distribution / managers claim is justified	A1 ft (7)
		[12]
Q6 (a)	μ_U ~ mean length of upper shore limpets, μ_L ~ mean length of lower shore limpets	
	$H_0 : \mu_u = \mu_L$ $H_1 : \mu_u < \mu_L$ both	B1
	s.e. = $\sqrt{\dfrac{0.42^2}{120} + \dfrac{0.67^2}{150}}$	M1
	= 0.0668	A1
	$z = \dfrac{5.05 - 4.97}{0.0668} = (\pm)1.1975$ awrt ± **<u>1.20</u>**	dM1 A1
	Critical region is $z \geq 1.6449$, or probability = awrt (0.115 or 0.116) $z = \pm 1.6449$	B1
	(1.1975 < 1.6449) therefore not in critical region / accept H_0/not significant (or $P(Z \geq 1.1975) = 0.1151$, $0.1151 > 0.05$ or z not in critical region)	M1
	There is no evidence that the limpets on the upper shore are shorter than the limpets	A1

The table used in Q5 (c):

Number of goals	Frequency	Expected frequency	
0	40	34.994	
1	33	36.743	
2	14	19.290	
3	8	6.752	8.972443
≥ 4	5	2.221	

Question Number	Scheme	Marks
	on the lower shore.	(8)
(b)	Assume the populations or variables are independent Standard deviation of sample = standard deviation of population [Mention of Central Limit Theorem does NOT score the mark]	B1 B1 (2) **[10]**
Q7 (a)	Estimate of Mean = $\frac{600.9}{5}$ = 120.18	M1A1
	Estimate of Variance = $\frac{1}{4}\{72216.31 - \frac{600.9^2}{5}\}$ or $\frac{0.148}{4}$ = 0.037	M1 A1ft A1 (5)
(b)	P(-0.05 < $\mu - \hat{\mu}$ < 0.05) = 0.90 or $P(-0.05 < \overline{X} - \mu < 0.05) = 0.90$ [$\le$ is OK]	B1
	$\frac{0.05}{\frac{0.2}{\sqrt{n}}} = 1.6449$	M1 A1
	$n = \frac{1.6449^2 \times 0.2^2}{0.05^2}$	dM1
	n = 43.29…	A1
	n = 44	A1 (6) **[11]**
Q8 (a)	E(4X-3Y)=4E(X) – 3E(Y)	M1
	= 4×30 – 3 ×20 = 60	A1 (2)
(b)	Var(4X-3Y) = 16 Var (X) + 9 Var (Y) 16 or 9; adding	M1; M1
	= 16 × 9 + 9 × 4 = 180	A1 (3)
(c)	E(B) = 80	B1
	Var (B) = 16	B1
	E(B - A) = 20 E(B)-E(A)	M1
	Var (B - A) = 196 ft on 180 and 16	A1ft
	P (B - A >0) = $P\left(Z > \frac{-20}{\sqrt{196}}\right) = [P(Z > -1.428...)]$ stand. using their mean and var	dM1
	= 0.923 … awrt 0.923 – 0.924	A1 (6) **[11]**