Paper Reference(s)

6691/01

Edexcel GCE

Statistics S3

Advanced Level

Wednesday 20 May 2015 – Morning

Time: 1 hour 30 minutes

Materials required for examination
Mathematical Formulae (Pink)

Items included with question papers
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 or symbolic differentiation/integration, or have retrievable mathematical formulae stored in them.

Instructions to Candidates

Write the name of the examining body (Edexcel), your centre number, candidate number, the unit title (Statistics S3), the paper reference (6691), your surname, initials 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.
The marks for the parts of questions are shown in round brackets, e.g. **(2)**.
There are 6 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 not gain full credit.

This publication may only be reproduced in accordance with Pearson Education Limited copyright policy.
©2015 Pearson Education Limited.

1. A mobile library has 160 books for children on its records. The librarian believes that books with fewer pages are borrowed more often. He takes a random sample of 10 books for children.

(*a*) Explain how the librarian should select this random sample.

(2)

The librarian ranked the 10 books according to how often they had been borrowed, with 1 for the book borrowed the most and 10 for the book borrowed the least. He also recorded the number of pages in each book. The results are in the table below.

Book	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Borrowing rank	1	2	3	4	5	6	7	8	9	10
Number of pages	50	212	115	80	301	90	356	283	152	317

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

(4)

(*c*) Test the librarian's belief using a 5% level of significance. State your hypotheses clearly.

(3)

---

2. A researcher believes that the mean weight loss of those people using a slimming plan as part of a group is more than 1.5 kg a year greater than the mean weight loss of those using the plan on their own. The mean weight loss of a random sample of 80 people using the plan as part of a group is 8.7 kg with a standard deviation of 2.1 kg. The mean weight loss of a random sample of 65 people using the plan on their own is 6.6 kg with a standard deviation of 1.4 kg.

(*a*) Stating your hypotheses clearly, test the researcher's claim. Use a 1% level of significance.

(8)

(*b*) For the test in part (*a*), state whether or not it is necessary to assume that the weight loss of a person using this plan has a normal distribution. Give a reason for your answer.

(2)

---

3. A nursery has 16 staff and 40 children on its records. In preparation for an outing the manager needs an estimate of the mean weight of the people on its records and decides to take a stratified sample of size 14.

(*a*) Describe how this stratified sample should be taken.

(3)

The weights, x kg, of each of the 14 people selected are summarised as

$$\sum x = 437 \text{ and } \sum x^2 = 26983.$$

(*b*) Find unbiased estimates of the mean and the variance of the weights of all the people on the nursery's records.

(4)

(*c*) Estimate the standard error of the mean.

(2)

The estimates of the standard error of the mean for the staff and for the children are 5.11 and 1.10 respectively.

(*d*) Comment on these values with reference to your answer to part (*c*) and give a reason for any differences.

(2)

---

4. The weights of bags of rice, X kg, have a normal distribution with unknown mean μ kg and known standard deviation σ kg. A random sample of 100 bags of rice gave a 90% confidence interval for μ of (0.4633, 0.5127).

(*a*) Without carrying out any further calculations, use this confidence interval to test whether or not $\mu = 0.5$.

State your hypotheses clearly and write down the significance level you have used.

(3)

A second random sample, of 150 of these bags of rice, had a mean weight of 0.479 kg.

(*b*) Calculate a 95% confidence interval for μ based on this second sample.

(6)

---

5. (i) The volume, B ml, in a bottle of *Burxton's* water has a normal distribution $B \sim N(325, 6^2)$ and the volume, H ml, in a bottle of *Hargate's* water has a normal distribution $H \sim N(330, 4^2)$.

Rebecca buys 5 bottles of *Burxton's* water and one bottle of *Hargate's* water.

Find the probability that the total volume in the 5 bottles of *Burxton's* water is more than 5 times the volume in the bottle of *Hargate's* water.

(5)

(ii) Two independent random samples X_1, X_2, X_3, X_4, X_5 and Y_1, Y_2, Y_3, Y_4, Y_5 are each taken from a normal population with mean μ and standard deviation σ.

(*a*) Find the distribution of the random variable $D = Y_1 - \overline{X}$.

(3)

(*b*) Hence show that $P(Y_1 > \overline{X} + \sigma) = 0.181$ correct to 3 decimal places.

(2)

Ankit believes that $P(U_1 > \overline{U} + \sigma) = 0.181$ correct to 3 decimal places, for **any** random sample U_1, U_2, U_3, U_4, U_5 taken from a normal population with mean μ and standard deviation σ.

(*c*) Explain briefly why the result from part (*b*) should not be used to confirm Ankit's belief.

(1)

(*d*) Find, correct to 3 decimal places, the actual value of $P(U_1 > \overline{U} + \sigma)$.

(6)

6.

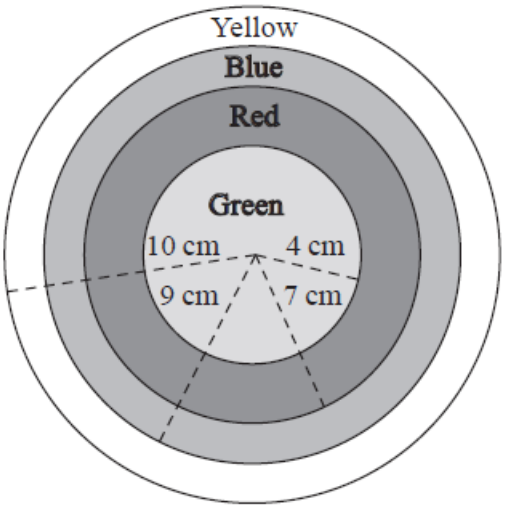

Figure 1

The sketch in Figure 1 represents a target which consists of 4 regions formed from 4 concentric circles of radii 4 cm, 7 cm, 9 cm and 10 cm. The regions are coloured as labelled in Figure 1.

A random sample of 100 children each choose a point on the target and their results are summarised in the table below.

Colour of region	Green	Red	Blue	Yellow
Frequency	22	39	25	14

Caitland is trying to model the distribution of the points chosen by the children. She defines the random variable D to be the distance, in cm, of a point from the centre of the target and assumes $D \sim \text{U}[0, 10]$.

(a) Stating your hypotheses clearly and using a 1% level of significance, test whether or not U[0, 10] is a suitable model for these data.

(9)

Henry claims that the points are randomly distributed over the target and the probability of a point being in any particular region is proportional to the area of that region.

He calculates expected frequencies and obtains the following table.

Colour of region	Green	Red	Blue	Yellow
Frequency	16	33	r	s

(b) Find the value of r and the value of s.

(3)

Henry obtained a test statistic of 6.188 and no groups were pooled.

(c) State what conclusion Henry should make about his claim.

(2)

Phoebe believes that the children chose the region of the target according to colour. She believes that boys and girls would favour different colours and splits the original data by gender to obtain the following table.

Observed frequencies

Colour of region	Green	Red	Blue	Yellow	**Total**
Boys	10	12	10	3	35
Girls	12	27	15	11	65

(*d*) State suitable hypotheses to test Phoebe's belief.

(1)

Phoebe calculated the following expected frequencies to carry out a suitable test.

Expected frequencies

Colour of region	Green	Red	Blue	Yellow
Boys	7.7	13.65	8.75	4.9
Girls	14.3	25.35	16.25	9.1

(*e*) Show how the value of 25.35 was obtained.

(1)

Phoebe carried out the test using 2 degrees of freedom and a 10% level of significance. She obtained a test statistic of 1.411.

(*f*) Explain clearly why Phoebe used 2 degrees of freedom.

(1)

(*g*) Stating your critical value clearly, determine whether or not these data support Phoebe's belief.

(2)

TOTAL FOR PAPER: 75 MARKS

END

June 2015
6691 S3
Mark Scheme

Question Number	Scheme	Marks
1. (a)	Label all the books from 1 – 160 (o.e.) Use random numbers to select the 10 books	B1 B1 (2)
(b)	<table><tr><td>Book</td><td>*A*</td><td>*B*</td><td>*C*</td><td>*D*</td><td>*E*</td><td>*F*</td><td>*G*</td><td>*H*</td><td>*I*</td><td>*J*</td></tr><tr><td>Borrow rank</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr><tr><td>Page rank</td><td>1</td><td>6</td><td>4</td><td>2</td><td>8</td><td>3</td><td>10</td><td>7</td><td>5</td><td>9</td></tr><tr><td>d^2</td><td>0</td><td>16</td><td>1</td><td>4</td><td>9</td><td>9</td><td>9</td><td>1</td><td>16</td><td>1</td></tr></table> M1 M1 $r_s = 1 - \frac{6\times 66}{10(100-1)}, [= 1-0.4]$ = 0.6 **0.6** M1,A1 (4)
(c)	H_0: $\rho = 0$ H_1: $\rho > 0$ Critical value is 0.5636 0.6 > cv so significant result and sufficient evidence to reject H_0 There is support for the librarian's belief or there is evidence of a correlation between the number of pages in a book and the number of times it is borrowed.	B1 B1 B1ft (3) **Total 9**

Notes

(a)

1st B1 for labelling\numbering\listing\using sampling frame of all 160 books

2nd B1 for use of random numbers\selection and mentioning the number 10

(b)

1st M1 for an attempt to rank the number of pages (at least 4 correct) Allow reverse ranks

2nd M1 for attempt at d^2 row (may be implied by sight of $\sum d^2 = 66$ or 264 for reverse ranks)

3rd M1 for use of the correct formula, follow through their $\sum d^2$ if clearly stated

If answer is not correct, a correct expression is required.

A1 for 0.6 (or -0.6 for reverse ranks)

(c)

1st B1 for both hypotheses in terms of ρ, one tail H_1 (compatible with ranks) Allow use of ρ_s

Hypotheses just in words e.g. "no correlation" score B0.

2nd B1 for cv of 0.5636 [If they have a two tail H_1 then allow 0.6485]

Allow $\pm$ for reverse ranking but must be same sign as r_s

If hypotheses are the wrong way around this must be B0 but 3rd B1 is possible.

3rd B1ft for a correct contextualised comment. Must mention "librarian" (or he)

or "number of pages" and "borrowing"

Follow through their r_s and their cv (provided it is |cv| <1)

Don't insist on the word "positive" or "negative" for a one-tailed test

Use of "association" is B0

Independent of 1st B1 so if $|r_s|$ > |cv| must say there is sufficient evidence of …….(o.e.)

and if $|r_s|$ < |cv| must say insufficient evidence of … (o.e.) regardless of their hypotheses

Question Number	Scheme	Marks
2. (a)	$H_0 : \mu_g - \mu_s = 1.5$ [g = in a group, s = on their own]	B1
	$H_1 : \mu_g - \mu_s > 1.5$	B1
	s.e. = $\sqrt{\frac{2.1^2}{80} + \frac{1.4^2}{65}} = \left[\sqrt{0.08527...}\right]$ = [0.292]	M1
	$z = \frac{8.7 - 6.6 - 1.5}{"\sqrt{\frac{2.1^2}{80} + \frac{1.4^2}{65}}"}$	dM1
	= 2.0546…. awrt 2.05(5)	A1
	cv 1% one tailed = 2.3263	B1
	Not significant, accept H_0	dM1
	Insufficient evidence that using plan as part of a group leads to weight loss of more than 1.5 kg than using plan on one's own or researcher's belief not supported	A1ft
		(8)
(b)	Since sample is large Central Limit Theorem (CLT) applies	B1
	No need to assume normal distribution	dB1
		(2)
		Total 10
	Notes	
(a)	1st & 2nd B1 for hypotheses. Accept μ_1, μ_2 or μ_A, μ_B etc if there is some indication of which is which e.g. $G \sim N(\mu_g, 8.7)$	
	1st M1 for an attempt at se with 3 out of 4 values correct. Condone switching 2.1 and 1.4 $\sqrt{\frac{2.1^2 \text{ or } 1.4^2}{80} + \frac{1.4^2 \text{ or } 2.1^2}{65}}$	
	2nd dM1 dependent on 1st M1 for a correct numerator(must have -1.5) and ft their se.	
	1st A1 for awrt 2.05	
	3rd B1 for ± 2.3263 or better seen or probability of awrt 0.02	
	3rd dM1 dep. on 1st M1 for a correct statement based on their normal cv and their test statistic	
	2nd A1ft for correct comment in context. Must mention "plan" and "group or individual" and "1.5" or "researcher" and "belief or claim"	
	NB Use of cv for difference in means D will have D = 1.5 + 2.3263 × s.e. = awrt 2.18 and requires sight of d = 2.1 with a comment for the 3rd M1	
(b)	1st B1 for mentioning "large samples" and "CLT"	
	2nd dB1 dependent on 1st B1 for stating **no need to** assume normality (since CLT assures it)	

Question Number	Scheme	Marks
3. (a)	Label staff (from 1 – 16) and children (from 1 – 40) Use random numbers to select 4 staff and 10 children	B1 B1 B1 (3)
(b)	$\bar{x} = \hat{\mu} =$ 31.2142… **awrt** **<u>31.2</u>** $s^2 = \frac{26983 - 14 \times "31.2..."^2}{13}$ = 1026.33… **awrt** **<u>1030</u>**	B1 M1 A1ft A1 (4)
(c)	$\frac{"\sqrt{1026.33...}"}{\sqrt{14}}$, = 8.562.. **awrt** **<u>8.56</u>**	M1, A1 (2)
(d)	The variation within each stratum is quite small (o.e.) **The difference in the means will be quite large**, (so variations from the overall mean will be large giving a larger overall s.e.)	B1 B1 (2) **Total 11**
	Notes	
(a)	1st B1 for labelling\numbering\listing staff <u>and</u> children 2nd B1 for use of random numbers or "randomly select" in <u>each group</u> (may be implied) 3rd B1 for selecting the correct number of staff <u>and</u> children e.g. randomly select 4 staff and 10 children scores 2nd and 3rd B marks since randomly selecting and the "each group" is implied,	
(b)	B1 for awrt 31.2 M1 for a correct expression ft their $\bar{x}$ and allow transcription error in $\sum x^2$ e.g. 29683 1st A1ft for a fully correct expression ft their $\bar{x}$ only 2nd A1 for awrt 1030	
(c)	M1 for attempting $\frac{"\text{their } s"}{\sqrt{14}}$ (must have 14) A1 for awrt 8.56	
(d)	1st B1 for a suitable comment about **variation (se)** suggesting that **variation (se) within strata is less than that overall** 2nd B1 for a suitable **reason** about **means**, pointing out that the individuals' weights will vary a lot from the overall mean and so overall s.e. will be higher.	

Question Number	Scheme	Marks
4. (a)	$H_0 : \mu = 0.5 \quad H_1 : \mu \neq 0.5$	B1
	(Significance level =)10%	dB1
	(0.5 is in the interval so not significant, accept H_0, can accept) $\mu = 0.5$	B1
		(3)
(b)	$1.6449 \times \frac{\sigma}{\sqrt{100}} = 0.0247$	M1
		B1
	$\sigma = 0.15016$ or $\frac{10 \times 0.0247}{1.6449}$ (awrt 0.15)	A1
	$0.479 \pm 1.96 \times \frac{"\sigma"}{\sqrt{150}}$	M1
		B1
	awrt **<u>(0.455, 0.503)</u>**	A1
		(6)
		Total 9
	Notes	
(a)	1st B1 for both hypotheses in terms of μ.	
	2nd dB1 for 10% but accept 5% if they have a one-tail test as H_1	
	3rd B1 for a correct comment leading to accepting H_0	
	Ignore any 'further calculations'.	
(b)	1st M1 for $z\frac{\sigma}{\sqrt{100}} = k$, using $n = 100$ and where $\lvert z \rvert > 1.5$ and $0.02 < k < 0.03$	
	1st B1 for 1.6449 or better in an attempt (could be $1.6449\sigma = k$ or even $1.6449\ \sigma^2 = k$)	
	1st A1 for a correct expression for σ e.g. awrt 0.15	
	2nd M1 for $\bar{x} \pm z \times \frac{\sigma}{\sqrt{150}}$ for any z (> 1) and ft their σ and allow $\bar{x} \in (0.4633, 0.5127)$	
	Allow use of letter σ without a value.	
	2nd B1 for 1.96 or better in an attempt (could be 1.96σ or even $1.96\ \sigma^2$)	
	2nd A1 for awrt 0.455 <u>and</u> awrt 0.503	

Question Number	Scheme	Marks
5 (i)	Let $R = B_1 + B_2 + B_3 + B_4 + B_5 - 5H$ so E(R) = -25 (o.e.)	B1
	Var(R) = $5\times 6^2 + 5^2 \times 4^2$ $\quad R$~N(-25, $\sqrt{580}^2$)	M1A1
	P(R > 0) = P($Z > \frac{0--25}{\sqrt{580}}$) = P($Z$ > 1.04), = 0.149619…(calc) <u>or</u> 0.1492 (tables)	dM1 A1 (5)
(ii)(a)	$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{5}\right)$	B1
	Var(D) = $\sigma^2 + "\frac{\sigma^2}{5}"\left[= \frac{6\sigma^2}{5}\right]$, so $\quad D \sim \mathrm{N}\left(0, \frac{6\sigma^2}{5}\right)$	M1, A1 (3)
(b)	P($Y_1 > \bar{X} + \sigma$) = P($D > \sigma$) = $\mathrm{P}\left(Z > \frac{\sigma}{\sqrt{\frac{6}{5}}\sigma}\right)$	M1
	= P(Z > 0.912…) = 0.181(3 dp) **(*)**	A1cso (2)
(c)	Since U_1 and $\bar{U}$ are not independent (so variance formula cannot be used)	B1
	Can be implied e.g. U_1 used to calculate $\bar{U}$, U_1 and $\bar{U}$ from same sample o.e.	(1)
(d)	Let $F = U_1 - \bar{U} = U_1 - \frac{(U_1+U_2+U_3+U_4+U_5)}{5}, = \frac{4U_1 - (U_2+U_3+U_4+U_5)}{5}$	M1, A1
	Var(F) = $\frac{4^2\sigma^2 + 4\sigma^2}{5^2}$ = $0.8\sigma^2$,so $\quad F$ ~ N(0, $0.8\sigma^2$)	dM1, A1
	P($F > \sigma$) = $\mathrm{P}\left(Z > \frac{\sigma}{\sigma\sqrt{0.8}}\right)$ = P(Z > 1.118..)	M1
	= 0.1314 (tables) or 0.131776…(calc) **awrt 0.131~0.132**	A1cso (6)
		Total 17

	Notes
(i)	1st B1 for E(R) = – 25 (or 25 if their R is defined the other way around) 1st M1 for an attempt at Var(R) = 5Var(B) + 25Var(H). Condone swapping of 6^2 and 4^2 1st A1 for normal and correct variance (ft their mean) 2nd dM1 for attempting the correct probability and standardising with their mean and sd. This mark is dependent on 1st M1 so if R is not being used or M0 for variance score M0 If their method is not crystal clear then they must be attempting P(Z > +ve value) o.e 2nd A1 for answer in the range [0.149, 0.150]
(ii)(a)	B1 for correct distribution of $\bar{X}$ (may be implied for a correct answer for D) M1 for correct attempt at Var(D) (ft their Var($\bar{X}$)) [A1 needs must be fully correct]
(ii)(b)	M1 for expressing probability in terms of D and standardising A1cso for seeing P(Z> 0.912..) or prob of 1 – 0.8186 (tables) or 0.180655…(calc)
(c)	B1 correct statement that should mention U_1 and $\bar{U}$
(d)	1st M1 for forming an expression in terms of U_1...U_5 only 1st A1 for collecting U_1 terms and getting in a form where Var($aX \pm bY$) can be used. 2nd dM1 for a correct expression for Var(their F). Dependent on 1st M1. 2nd A1 for a correct distribution for F 3rd M1 attempting a correct prob and standardising using their Var(F), σ must cancel 3rd A1cso for awrt 0.131 or 0.132

Question Number		Scheme	Marks
6.	**(a)**	H_0 : U[0, 10] is a suitable model H_1: U[0, 10] is not a suitable model	B1
		Values of *D* Expected Freq	B1 M1A1
		4^{th} or 5^{th} col	M1
		$\chi^2 = 13.65$	A1
		$\nu = 3, \quad \chi_3^{\,2}(1\%) = 11.345$	B1, B1
		[Reject H_0 ,] the uniform distribution over [0, 10] is not a suitable model	A1
			(9)
	(b)	Area $\propto \pi R^2$ so r = 81, – 49 = **32**	M1, A1
		s = 100 – "32" – 49 or 100 – 81 = **19**	B1ft
			(3)
	(c)	Not significant, Henry's model is suitable	M1, A1
			(2)
	(d)	H_0 : The colour\region chosen for the points is independent of gender(or no assoc')	B1
		H_1 : The colour\region chosen for the points is dependent on gender(or assoc')	
			(1)
	(e)	$\frac{39 \times 65}{100}$	B1
			(1)
	(f)	Expected frequency for Yellow and Boys is 4.9 < 5 so col. must be pooled/combined.	B1
		[This gives a 2×3 table so $\nu = (2-1) \times (3-1) = 2$]	
			(1)
	(g)	cv = 4.605	B1
		[Not significant] so the data do not support Phoebe's belief oe	B1
			(2)
			Total 19

	Notes
(a)	2^{nd} B1 for the correct values for *D* (can be implied by 40, 30, 20, and 10.) 1^{st} M1 for at least 2 expected frequencies or clear use of a correct formula e.g. 0.4*N* 1^{st} A1 for all the correct E_i 2^{nd} M1 for at least 2 correct calculations from 4^{th} or 5^{th} column 2^{nd} A1 for a test statistic of 13.65 (accept 13.7 to 3 sf) Awrt 13.7 only scores 2^{nd} B1M1A1M1A1 3^{rd} A1 for a correct conclusion rejecting the uniform model. Award provided their test statistic > 11.345
(b)	M1 for some attempt to use πR^2 to find *r*
(c)	M1 for a correct statement that it is not significant A1 for correctly stating that Henry's model is suitable o.e.
(d)	B1 Independence or association mentioned at least once if ditto marks used. Allow connection but not correlation.
(f)	B1 for recognising there is an *Ei* < 5 **and** need for pooling/combining oe
(g)	2^{nd}B1 for correctly stating that Phoebe's belief is not supported by the data oe (depends on their cv being > 1.411)