Write your name here

Surname	Other names

Pearson Edexcel GCE

Centre Number | Candidate Number

Statistics S3

Advanced/Advanced Subsidiary

Wednesday 25 May 2016 – Morning
Time: 1 hour 30 minutes

Paper Reference
6691/01

You must have:
Mathematical Formulae and Statistical Tables (Pink)

Total Marks

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 formulae stored in them.

Instructions

- Use **black** ink or ball-point pen.
- If pencil is used for diagrams/sketches/graphs it must be dark (HB or B). Coloured pencils and highlighter pens must not be used.
- **Fill in the boxes** at the top of this page with your name, centre number and candidate number.
- Answer **all** questions and ensure that your answers to parts of questions are clearly labelled.
- Answer the questions in the spaces provided – *there may be more space than you need.*
- You should show sufficient working to make your methods clear. Answers without working may not gain full credit.
- 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

- The total mark for this paper is 75.
- The marks for each question are shown in brackets – *use this as a guide as to how much time to spend on each question.*

Advice

- Read each question carefully before you start to answer it.
- Try to answer every question.
- Check your answers if you have time at the end.

1. (*a*) State two reasons why stratified sampling might be a more suitable sampling method than simple random sampling.

(2)

(*b*) State two reasons why stratified sampling might be a more suitable sampling method than quota sampling.

(2)

(Total 4 marks)

---

2. A new drug to vaccinate against influenza was given to 110 randomly chosen volunteers. The volunteers were given the drug in one of 3 different concentrations, A, B and C, and then were monitored to see if they caught influenza. The results are shown in the table below.

	A	B	C
Influenza	12	29	9
No influenza	15	23	22

Test, at the 10% level of significance, whether or not there is an association between catching influenza and the concentration of the new drug. State your hypotheses and show your working clearly. You should state your expected frequencies to 2 decimal places.

(Total 10 marks)

---

3. (*a*) Describe when you would use Spearman's rank correlation coefficient rather than the product moment correlation coefficient to measure the strength of the relationship between two variables.

(1)

A shop sells sunglasses and ice cream. For one week in the summer the shopkeeper ranked the daily sales of ice cream and sunglasses. The ranks are shown in the table below.

	Sun	Mon	Tues	Weds	Thurs	Fri	Sat
Ice cream	6	4	7	5	3	2	1
Sunglasses	6	5	7	2	3	4	1

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

(3)

(*c*) Test, at the 5% level of significance, whether or not there is a positive correlation between sales of ice cream and sales of sunglasses. State your hypotheses clearly.

(4)

The shopkeeper calculates the product moment correlation coefficient from his raw data and finds $r = 0.65$.

(*d*) Using this new coefficient, test, at the 5% level of significance, whether or not there is a positive correlation between sales of ice cream and sales of sunglasses.

(2)

(*e*) Using your answers to part (*c*) and part (*d*), comment on the nature of the relationship between sales of sunglasses and sales of ice cream.

(1)

(Total 11 marks)

---

4. The weights of eggs are normally distributed with mean 60 g and standard deviation 5 g. Sairah chooses 2 eggs at random.

(*a*) Find the probability that the difference in weight of these 2 eggs is more than 2 g.

(5)

Sairah is packing eggs into cartons. The weight of an empty egg carton is normally distributed with mean 40g and standard deviation 1.5 g.

(*b*) Find the distribution of the total weight of a carton filled with 12 randomly chosen eggs.

(3)

(*c*) Find the probability that a randomly chosen carton, filled with 12 randomly chosen eggs, weighs more than 800 g.

(2)

(Total 10 marks)

---

5. A doctor claims there is a higher mean lung capacity in people who exercise regularly compared to people who do not exercise regularly. He measures the lung capacity, x, of 35 people who exercise regularly and 42 people who do not exercise regularly. His results are summarised in the table below.

	n	$\bar{x}$	s^2
Exercise regularly	35	26.3	12.2
Do not exercise regularly	42	24.8	10.1

(*a*) Test, at the 5% level of significance, the doctor's claim. State your hypotheses clearly.
(6)

(*b*) State any assumptions you have made in testing the doctor's claim.
(2)

The doctor decides to add another person who exercises regularly to his data. He measures the person's lung capacity and finds $x = 31.7$.

(*c*) Find the unbiased estimate of the variance for the sample of 36 people who exercise regularly. Give your answer to 3 significant figures.
(4)

(Total 12 marks)

6. An airport manager carries out a survey of families and their luggage. Each family is allowed to check in a maximum of 4 suitcases. She observes 50 families at the check-in desk and counts the total number of suitcases each family checks in. The data are summarised in the table below.

Number of suitcases	0	1	2	3	4
Frequency	6	25	12	6	1

The manager claims that the data can be modelled by a binomial distribution with $p = 0.3$.

(*a*) Test the manager's claim at the 5% level of significance. State your hypotheses clearly. Show your working clearly and give your expected frequencies to 2 decimal places.

(8)

The manager also carries out a survey of the time taken by passengers to check in. She records the number of passengers that check in during each of 100 five-minute intervals.

The manager makes a new claim that these data can be modelled by a Poisson distribution. She calculates the expected frequencies given in the table below.

Number of passengers	0	1	2	3	4	5 or more
Observed frequency	5	40	31	18	6	0
Expected frequency	16.53	29.75	r	s	7.23	3.64

(*b*) Find the value of r and the value of s giving your answers to 2 decimal places.

(3)

(*c*) Stating your hypotheses clearly, use a 1% level of significance to test the manager's new claim.

(6)

(Total 17 marks)

7. A restaurant states that its hamburgers contain 20% fat. Paul claims that the mean fat content of their hamburgers is less than 20%. Paul takes a random sample of 50 hamburgers from the restaurant and finds that they contain a mean fat content of 19.5% with a standard deviation of 1.5%.

You may assume that the fat content of hamburgers is normally distributed.

(*a*) Find the 90% confidence interval for the mean fat content of hamburgers from the restaurant.

(4)

(*b*) State, with a reason, what action Paul should recommend the restaurant takes over the stated fat content of their hamburgers.

(2)

The restaurant changes the mean fat content of their hamburgers to μ % and adjusts the standard deviation to 2%. Paul takes a sample of size n from this new batch of hamburgers. He uses the sample mean $\overline{X}$ as an estimator of μ.

(*c*) Find the minimum value of n such that $\mathrm{P}(|\overline{X} - \mu| < 0.5) \geq 0.9$.

(5)

(Total 11 marks)

TOTAL FOR PAPER: 75 MARKS

June 2016
6691 Statistics S3
Mark Scheme

Question Number	Scheme	Marks
1(a) e.g.	Analyse / find estimates for a particular **subgroup** of the population. Stratified guarantees **representation** of all **groups**, srs does not. Observe **relationships between subgroups** – srs does not guarantee equal or proportionate representation. Rare or extreme cases as part of a **small subgroups** can be represented **proportionately** in stratified i.e. stratified represents the structure of the population– srs does not allow this. Stratified typically require **large sample size** compared to srs due to lower variability within subgroups compared to entire population.	
	Any 2 distinct reasons	B1B1
		(2)
(b) e.g.	It (a stratified sample) is **not biased** as the members are chosen randomly. You can **estimate the sampling errors** (for a stratified sample) It (a stratrified sample) gives **more accurate estimates** as it is a random process. A quota sample may be (**interviewer / process**) **biased**. It's **not possible to estimate/find the sampling errors** for a quota sample (whereas you can for a stratified sample)	
	Any 2 distinct reasons	B1B1
		(2)
		Total 4
Notes	Award B1B1 two correct, B1B0 one correct. Allow 'it' for 'stratified'. Do not award marks for vague responses such as 'cheap', 'easy' 'quick' 'random' etc. Mentioning 'sampling frame' alone is not sufficient for a mark. Mentioning 'non-response are not recorded' alone is not sufficient for a mark.	

Question Number	Scheme	Marks
2	H_0 : **Drug concentration** and catching **influenza** are independent / not associated H_1 : **Drug concentration** and catching **influenza** are not independent / associated	B1

	A	B	C	
Influenza	$\frac{50\times27}{110}=12.272...$	$\frac{50\times52}{110}=23.636...$	$\frac{50\times31}{110}=14.090...$	50
No influenza	$\frac{60\times27}{110}=14.727...$	$\frac{60\times52}{110}=28.363...$	$\frac{60\times31}{110}=16.909...$	60
	27	52	31	110

M1A1

O	E	$\frac{(O-E)^2}{E}$	$\frac{O^2}{E}$
12	12.272…	0.0060…	11.7333…
29	23.636…	1.2171…	35.5807…
9	14.090…	1.8392…	5.7483…
15	14.727…	0.0050…	15.2777…
23	28.363…	1.0142…	18.6506…
22	16.909…	1.5327…	28.6236…

M1A1

$\sum\frac{(O-E)^2}{E}=5.6145...$ or $\sum\frac{O^2}{E}-N=115.62...-110=5.6145...$ awrt 5.61-5.62 — A1

$\nu=(3-1)(2-1)=2$, $\chi_2^2(10\%)=4.605$ — B1B1ft

Reject H_0 — M1

Drug concentration and catching **influenza** are not independent / are associated. — A1cao

(10)

Total 10

Notes

B1 hyps correct way around

M1 for correct expression at least once

A1 **all seen** and correct 2dp or better. Can be implied by test statistic of awrt 5.61-5.62.

M1 either method at least one correct

A1 at least 3 correct values. Can be implied by test statistic of awrt 5.61-5.62

A1 awrt 5.61-5.62

B1 cao

B1 follow through their ν

M1 must correctly reject / not reject the null hypothesis based on their test stat and cv oe

Question Number	Scheme	Marks
3 (a)	Variables do not have a (joint) normal distribution Relationship is not linear The given data is ordinal — Any 1	B1 (1)
(b)	(table below)	
	$\sum d^2 = 14$	M1
	$r_s = 1 - \frac{6 \times 14}{7(49-1)} = 0.75$	M1A1 (3)
(c)	$H_0 : \rho = 0$, $H_1 : \rho > 0$	B1
	5% cv 0.7143	B1
	Reject H_0	M1
	Evidence of **positive** correlation between sales of **ice cream** and sales of **sunglasses.**	A1cao (4)
(d)	(5% cv 0.6694) Accept H_0	M1
	Insufficient evidence of **positive** correlation between sales of ice cream and sales of sunglasses.	A1cao (2)
(e)	Suggests relationship might be non-linear.	B1 (1)
		Total 11

Day	Sun	Mon	Tues	Weds	Thurs	Fri	Sat
Ice-cream	6	4	7	5	3	2	1
Sunglasses rank	6	5	7	2	3	4	1
d	0	-1	0	3	0	-2	0
d^2	0	1	0	9	0	4	0

Notes

(a)	Accept 'already ranked' oe for ordinal Accept one variable is not normally distributed
(b)	M1 attempt to find d, d^2 and sum. may be implied by sight of $\sum d^2 = 14$ 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 0.75 cao
(c)	1st B1for both hypotheses in terms of ρ, one tail. Allow use of ρ_s. **Only award** if no errors seen in hypotheses in part(c) **and part(d)** Hypotheses just in words e.g. "no correlation" score B0. B1 0.7143 cao M1 must correctly reject / not reject the null hypothesis based on their test stat and cv oe A1 Conclusion must mention ice cream and sunglasses
(d)	M1 for not rejecting / accepting null hyp A1 must mention ice cream and sunglasses

Question Number	Scheme	Marks
4 (a)	X_i be rv 'weight of i^{th} randomly chosen egg'	
	$E(X_1 - X_2) = 0$	B1
	$Var(X_1 - X_2) = 2\times5^2 = 50$	B1
	$P(\lvert X_1 - X_2\rvert > 2) = 2P(X_1 - X_2 > 2)$	M1
	$= 2P(Z > \frac{2}{\sqrt{50}})$	dM1
	$= 2P(Z > 0.2828...)$	
	$= 2(1-0.6103)=0.7794$ awrt 0.777-0.779	A1
		(5)
(b)	$W = C + X_1 + X_2 + ... + X_{12}$	
	$E(W) = 40 + 12\times60 = 760$	B1
	$Var(W) = 1.5^2 + 12\times5^2$	M1
	$= 302.25$	A1
	Distribution is $N(760, 302.25)$	
		(3)
(c)	$P(W > 800) = P\left(Z > \frac{800-760}{\sqrt{302.25}}\right)$	M1
	$= 1 - P(Z < 2.3007...)$	
	$= 0.0107$ awrt 0.0107	A1
		(2)
		Total 10
Notes (a)	B1 for 0	
	B1 for 50	
	M1 for $\lvert X_1 - X_2\rvert > 2$ seen. Accept $X_1 - X_2 > 2$ provided a subsequent doubling of the probability is seen. i.e. 0.3897 x 2. dM1 standardise with their 0 and their $\sqrt{50}$ dependent on previous M. A1 awrt 0.777-0.779	
(b)	B1 for 760	
	M1 requires squares	
	A1 cao	
(c)	Must be finding correct probability (ie $P(W > 800)$ or $P(Z > 2.3007...)$ etc) and standardise with 800 and their 760 and their $\sqrt{302.25}$ A1 awrt 0.0107 from correct working.	

Question Number	Scheme	Marks
5(a)	$H_0: \mu_e = \mu_n$, $H_1: \mu_e > \mu_n$	B1
	$z = \dfrac{26.3-24.8}{\sqrt{\dfrac{12.2}{35}+\dfrac{10.1}{42}}} = \dfrac{1.5}{\sqrt{0.58904...}} = \dfrac{1.5}{0.76749...}$	M1M1
	$z = 1.9544...$ awrt 1.95	A1
	Critical value is 1.6449	B1
	Reject H_0 . Doctor's claim is supported.	A1
		(6)
(b)	Either assume $\bar{X}$ has a normal distribution (for both samples) or assume sample sizes are large enough to use CLT Assume individual **results** are independent Assume $\sigma^2 = s^2$ for **both** populations or a single general population	B1 B1
		(2)
(c)	$\bar{x} = \left(\dfrac{35\times 26.3+31.7}{36} = \dfrac{952.2}{36}\right)26.45$	B1
	For $n = 35$, $\sum x^2 = 34\times 12.2 + 35\times 26.3^2 (= 24623.95)$	M1
	For $n = 36$, $s^2 = \dfrac{25628.84 - 36\times 26.45^2}{35} = 12.661...$ awrt 12.7	dM1A1
		(4)
		Total 12
Notes (a)	Both hyps, one tailed only oe. Accept μ_1, μ_2 or μ_A, μ_B etc if there is some indication of which is which. M1 for correct method for standard error M1 for whole expression A1 awrt 1.95 B1 1.6449 or $p = 0.974...$ (>0.95) A1 must mention doctor and claim or description of claim that includes '**mean** lung capacity' and 'exercise'.	
ALT (a)	M1 for $\sqrt{\dfrac{12.2}{35}+\dfrac{10.1}{42}}$ M1 for $1.6449 = \dfrac{c}{\sqrt{\dfrac{12.2}{35}+\dfrac{10.1}{42}}}$ A1 for awrt $c = 1.26$ seen B1 1.5	
(c)	M1 Attempt $\sum x^2 = 34\times 12.2 + 35\times 26.3^2$ or $\sum (x-\bar{x})^2 = 34\times 12.2 + 35(26.45-26.3)^2 (= 415.5875)$ dM1 $s^2 = \dfrac{\sum x^2 + 31.7^2 - 36\times 26.45^2}{35}$ or $s^2 = \dfrac{415.5875 + (31.7-26.45)^2}{35}$ A1 awrt 12.7	

Question Number	Scheme	Marks
6(a)	H_0 : Binomial with $p = 0.3$ is a good fit. H_1 : Binomial with $p = 0.3$ is not a good fit.	B1

	0	1	2 or more
Observed	6	25	19
Expected	50x0.2401 =12.005 or **12.01** or 12.00	50x0.4116 =20.58	50x0.2646+50x0.0756+50x0.0081 =13.23+3.78+0.405 =17.415 or **17.41** or 17.42
$\frac{(O-E)^2}{E}$	3.003751	0.949291	0.144256
$\frac{O^2}{E}$	2.998751	30.36929	20.72926

M1A1

$\sum\frac{(O-E)^2}{E} = 4.097...$ or $\sum\frac{O^2}{E} - N = 54.097... - 50 = 4.097...$ awrt 4.09-4.1(0) — dM1A1

$\nu = 3 - 1 = 2$ — B1ft

$\chi_2^2(5\%) = 5.991 \ (> 4.1(0))$ — B1ft

Insufficient evidence to reject H_0 (Accept H_0)

Binomial with $p = 0.3$ is a good fit. — A1

(8)

(b)

$\bar{x} = \frac{40 + 62 + 54 + 24}{100} = 1.8$ — B1 cao

$r = 26.78$ — B1 cao

$s = 16.07$ — B1 cao

(3)

(c)

H_0 : Poisson is a good fit. H_1 : Poisson is not a good fit. — B1

	0	1	2	3	4 or more
Observed	5	40	31	18	6
Expected	16.53	29.75	26.78	16.07	10.87
$\frac{(O-E)^2}{E}$	$\frac{11.53^2}{16.53} = 8.042...$	$\frac{10.25^2}{29.75} = 3.532...$	$\frac{4.22^2}{26.78} = 0.665...$	$\frac{1.93^2}{16.07} = 0.232...$	$\frac{4.87^2}{10.87} = 2.182...$
$\frac{O^2}{E}$	$\frac{5^2}{16.53} = 1.512...$	$\frac{40^2}{29.75} = 53.782...$	$\frac{31^2}{26.78} = 35.885...$	$\frac{18^2}{16.07} = 20.162...$	$\frac{6^2}{10.87} = 3.312...$

$\sum\frac{(O-E)^2}{E} = 14.65 - 14.66$ or $\sum\frac{O^2}{E} - N = 114.65 - 100 = 14.65 - 14.66$ — M1A1

$\nu = 5 - 1 - 1 = 3$ — B1 cao

$\chi_3^2(1\%) = 11.345 \ (< 14.65)$ — B1ft

Sufficient evidence to reject H_0

Poisson is not a good fit. — A1 cao

(6)

Total 17

Notes (a)

B1 both including $p = 0.3$
M1 with some combined columns and at least one E correct to 2sf
A1 all correct to 2dp and total of expected values is 50.
dM1 either method
A1 awrt 4.09-4.1(0)
B1 ft their columns -1
B1 ft their
A1 cao

(c)

B1 no parameters included
M1 either method

	B1 ft their ν	
Question Number	Scheme	Marks
7(a)	$19.5 \pm 1.6449 \times \frac{1.5}{\sqrt{50}}$	M1B1
	$= (19.151..., 19.848...)$ awrt 19.2, awrt 19.8	A1A1
		(4)
(b)	CI does not contain 20 oe	M1
	Fast Food restaurant statement is too high; they should reduce the stated value.	A1
		(2)
(c)	$\text{P}(\lvert \bar{X} - \mu \rvert < 0.5) = 0.9$	
	$\frac{0.5}{\frac{2}{\sqrt{n}}} = 1.6449$	M1A1
	$n = \left(2 \times \frac{1.6449}{0.5}\right)^2 = 43.29...$	dM1A1
	Sample size required is 44	A1
		(5)
		Total 11
Notes		
(a)	M1 correct with their z i.e. $19.5 \pm (z \text{ value}) \times \frac{1.5}{\sqrt{50}}$	
	B1 for 1.6449	
	A1 awrt 19.2, A1 awrt 19.8(5)	
(b)	M1 Require 20 compared to their interval	
	A1 Accept statement that relates to 20 being above the interval.	
(c)	M1 $\frac{0.5}{\frac{2}{\sqrt{n}}} = z \text{ value}$ or equivalent expression	
	A1 All correct	
	dM1 Attempt to solve $\frac{0.5}{\frac{2}{\sqrt{n}}} = \text{their } z \text{ value}$	
	A1 awrt 43.3	
	A1 44 cao	