Write your name here

| Surname | Other names |
| --- | --- |
| | |

**Pearson Edexcel**
**International**
**Advanced Level**

Centre Number

Candidate Number

# Statistics S3

## Advanced/Advanced Subsidiary

Wednesday 23 May 2018 – Morning
**Time: 1 hour 30 minutes**

Paper Reference
**WST03/01**

**You must have:**
Mathematical Formulae and Statistical Tables (Blue)

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.

*Turn over*

P51428A

1/1/1/1/

P 5 1 4 2 8 A 0 1 2 8

Pearson

June 2018 S3

Leave blank

1. A random sample of 9 footballers is chosen to participate in an obstacle course. The time taken, $y$ seconds, for each footballer to complete the obstacle course is recorded, together with the footballer's Body Mass Index, $x$. The results are shown in the table below.

| Footballer | Body Mass Index, $x$ | | Time taken to complete the obstacle course, $y$ seconds | | $d$ | $d^2$ |
|---|---|---|---|---|---|---|
| $A$ | 18.7 | 1 | 690 | 6 | 5 | 25 |
| $B$ | 19.5 | 2 | 801 | 9 | 7 | 49 |
| $C$ | 20.2 | 3 | 723 | 8 | 5 | 25 |
| $D$ | 20.4 | 4 | 633 | 2 | 2 | 4 |
| $E$ | 20.8 | 5 | 660 | 5 | 0 | 0 |
| $F$ | 21.9 | 6 | 655 | 4 | 2 | 4 |
| $G$ | 23.2 | 7 | 711 | 7 | 0 | 0 |
| $H$ | 24.3 | 8 | 642 | 3 | 5 | 25 |
| $I$ | 24.8 | 9 | 607 | 1 | 8 | 64 |
| | | | | | | 196 |

Russell claims, that for footballers, as Body Mass Index increases the time taken to complete the obstacle course tends to decrease.

(a) Find, to 3 decimal places, Spearman's rank correlation coefficient between $x$ and $y$.
**(5)**

(b) Use your value of Spearman's rank correlation coefficient to test Russell's claim. Use a 5% significance level and state your hypotheses clearly.
**(4)**

The product moment correlation coefficient for these data is –0.5594

(c) Use the value of the product moment correlation coefficient to test for evidence of a negative correlation between Body Mass Index and the time taken to complete the obstacle course. Use a 5% significance level.
**(2)**

(d) Using your conclusions to part (b) and part (c), describe the relationship between Body Mass Index and the time taken to complete the obstacle course.
**(1)**

(a) $1 - \dfrac{6\Sigma d^2}{n(n^2-1)}$

$= 1 - \dfrac{6(196)}{9(81-1)}$

$= -0.633$.

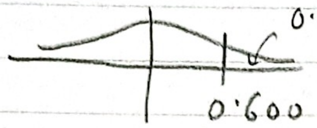
b) $H_0 : \rho = 0$

$H_1 : \rho < 0$.

From tables; critical value $= 0.600$.

Leave blank

**Question 1 continued**

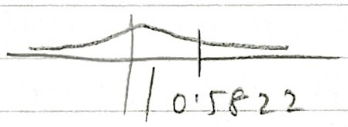



∴ 0.633 > 0.600 and lies in the critical region.

∴ Enough evidence to reject Ho and test is significant.

∴ Russel's claim is true.

(c) From tables;
r = + 0.5822




∴ 0.5594 < 0.5822
∴ doesn't lie in critical region

∴ there is no -ve correlation.

(d) The relationship between BMI and the time taken to complete the obstacle course is non linear.

Leave blank

**2.** A random sample of 75 packets of seeds is selected from a production line. Each packet contains 12 seeds. The seeds are planted and the number of seeds that germinate from each packet is recorded. The results are as follows.

| **Number of seeds that germinate from each packet** | 6 or fewer | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| **Number of packets** | 0 | 3 | 5 | 18 | 28 | 17 | 4 |

(a) Show that the probability of a randomly selected seed from this sample germinating is 0.82

**(2)**

A gardener suggests that a binomial distribution can be used to model the number of seeds that germinate from a packet of 12 seeds.

She uses a binomial distribution with the estimated probability 0.82 of a seed germinating. Some of the calculated expected frequencies are shown in the table below.

| **Number of seeds that germinate from each packet** | 6 or fewer | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| **Expected frequency** | $s$ | 2.80 | 7.97 | $r$ | 22.04 | 18.26 | 6.93 |

(b) Calculate the value of $r$ and the value of $s$, giving your answers correct to 2 decimal places.

**(3)**

(c) Test, at the 10% level of significance, whether or not these data suggest that the binomial distribution is a suitable model for the number of seeds that germinate from a packet of 12 seeds. State your hypotheses clearly and show your working.

**(7)**

(a) $f_i = P_i \times N$

$\frac{f_i}{\cancel{P_i}N} = \cancel{N} P_i$

$\frac{\Sigma fx}{\Sigma f} = \frac{(7\times3)+(8\times5)+(9\times18)+(10\times28)+(11\times17)+(12\times4)}{75}$

$= 9.84$

$\frac{9.84}{12} = 0.82$ as req.

(b) $\binom{n}{x} p^x (1-p)^{n-x}$

$\binom{12}{9}(0.82)^9(1-0.82)^3$

$r = \underline{16.13}$

$75 - (2.8 + 7.97 + 16.13 + 22.04 + 18.26 + 6.93)$

$s = 0.87$ //

Leave blank

**Question 2 continued**

(c) $H_0$: Binomial distribution is a suitable model.

$H_1$: Binomial distribution is not a suitable model.

| | 8 or less | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|
| $O_i$ | ~~8 or less~~ 8 | 18 | 28 | 17 | 4 |
| $E_i$ | 11.64 | ~~22.04~~ 16.13 | ~~18.26~~ 22.04 | 18.26 | 6.93 |

$$\sum \frac{O_i^2}{E_i} - N.$$

$= \underline{79.29} - 75 = \underline{\underline{4.29}}$

$\nu = 5 - 1 - 1 = 2$

From tables

$\chi^2 = 6.251$

$4.29 < 6.251$

$\therefore$ It doesn't fall in CR

$\therefore$ not enough evidence to reject $H_0$.

$\therefore$ Binomial distribution is a suitable model

Leave blank

**3.** Star Farm produces duck eggs. Xander takes a random sample of 20 duck eggs from Star Farm and their widths, $x$ cm, are recorded. Xander's results are summarised as follows.

$$\sum x = 92.0 \quad \sum x^2 = 433.4974$$

(a) Calculate unbiased estimates of the mean and the variance of the width of duck eggs produced by Star Farm.

**(3)**

Yinka takes an independent random sample of 30 duck eggs from Star Farm and their widths, $y$ cm, are recorded. Yinka's results are summarised as follows.

$$\sum y = 142.5 \quad \sum y^2 = 689.5078$$

(b) Treating the combined sample of 50 duck eggs as a single sample, estimate the standard error of the mean.

**(5)**

Research shows that the population of duck egg widths is normally distributed with standard deviation 0.71 cm.

The farmer claims that the mean width of duck eggs produced by Star Farm is greater than 4.5 cm.

(c) Using your combined mean, test, at the 5% level of significance, the farmer's claim. State your hypotheses clearly.

**(5)**

(a) $\bar{x} = \frac{\sum x}{n} = \frac{92}{20}$

$= \underline{4.6}$

$\sigma^2 = \frac{1}{19}(433.4974 - 20(4.6)^2)$

$= \underline{0.542}$

(b) Standard error $= \frac{\sigma}{\sqrt{n}}$.

$\sum x = 92$

$\sum y = 142.5$

let $\sum w = \sum x + \sum y$

$\sum w = 142.5 + 92 = \underline{234.5} \to \frac{234.5}{50}$

$= \underline{4.69}$

$\sum x^2 = 433.4974$

$\sum y^2 = 689.5078$

$\sum w^2 = 1123.0052$.

$\sigma^2 = \frac{1}{49}(1123.0052 - 50(4.69)^2)$

$\sigma^2 = 0.473$.

$\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.473}}{\sqrt{50}} = \underline{0.0973}$



DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA

Leave blank

**Question 3 continued**

(i) $W \sim N\left(4.5, \frac{0.71^2}{50}\right)$

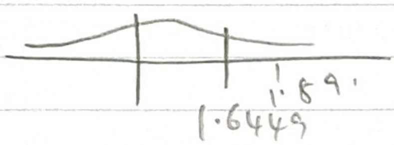
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{4.69 - 4.5}{\frac{0.71}{\sqrt{50}}}$$

$$= \underline{\underline{1.89}}$$

$H_0 : \mu = 4.5$

$H_1 : \mu > 4.5$.




$1.89 > 1.6449$.

$\therefore$ 1.89 lies in CR

$\therefore$ Test is significant

$\therefore$ enough evidence to reject $H_0$

$\therefore$ mean width of eggs is greater than 4.5cm.



Turn over

Leave blank

**4.** A company selects a random sample of five of its warehouses. The table below summarises the number of employees, in thousands, at each warehouse and the number of reported first aid incidents at each warehouse during 2017

| **Warehouse** | *A* | *B* | *C* | *D* | *E* | Total |
|---|---|---|---|---|---|---|
| **Number of employees, (in thousands)** | 2 | 1 | 3.8 | 3 | 2.2 | 12 |
| **Number of reported first aid incidents** | 15 | 10 | 40 | 26 | 23 | 114 |

The personnel manager claims that the mean number of reported first aid incidents per 1000 employees is the same at each of the company's warehouses.

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

**(7)**

Jean, the safety officer at warehouse *C*, kept a record of each reported first aid incident at warehouse *C* in 2017. Jean wishes to select a systematic sample of 10 records from warehouse *C*.

(b) Explain, in detail, how Jean should obtain such a sample.

**(2)**

$H_0$: Mean number of reported first aid incidents per 1000 employees is the same at each of the company's warehouses.

$H_1$: Mean number of reported first aid incidents per 1000 employees is not the same at each company warehouse.

Expected values:

A : $\frac{2 \times 114}{12} = 19$

B : $\frac{1 \times 114}{12} = 9.5$

C : $\frac{3.8 \times 114}{12} = 36.1$

D : $\frac{3 \times 114}{12} = 28.5$

E : $\frac{2.2 \times 114}{12} = 20.9$

DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA

Leave blank

**Question 4 continued**

| Warehouse | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|
| A | 15 | 19 | 0.84211 |
| B | 10 | 9.5 | 0.02632 |
| C | 40 | 36.1 | 0.42133 |
| D | 26 | 28.5 | 0.21930 |
| E | 23 | 20.9 | 0.21100 |
| | | Total | 1.720 |

$X^2 = \sum \frac{(O-E)^2}{E}$

$= 1.72$

$v = 5-1$

$= 4$

at 5% S.L. , CR: $X^2 \geqslant 9.488$

$1.72 < 9.488$

$\therefore$ Accept $H_0$. The manager's claim is supported

b) $\frac{40}{10} = 4$

Number each record from 1-40. Using the random number table, select a record between 1-4, then select every 4th record after the first number selected



P51428A01528

Turn over

Leave blank

**5.** A factory produces steel sheets whose weights, $X$ kg, have a normal distribution with an unknown mean $\mu$ kg and known standard deviation $\sigma$ kg.

A random sample of 25 sheets gave both a

- 95% confidence interval for $\mu$ of (30.612, 31.788)
- $c$% confidence interval for $\mu$ of (30.66, 31.74)

(a) Find the value of $\sigma$

**(3)**

(b) Find the value of $c$, giving your answer correct to 3 significant figures.

**(4)**

$X \sim N(\mu, \sigma^2)$

$\bar{x} \pm CI \cdot \frac{\sigma}{\sqrt{n}}$ $\quad n = 25.$

$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{25}}$

$\bar{X} + 1.96 \cdot \frac{\sigma}{5} = 31.788$

$\bar{X} - 1.96 \cdot \frac{\sigma}{5} = 30.612$

$3.92 \cdot \frac{\sigma}{5} = 1.176.$

$\sigma = 1.5$

(b) $\bar{x} = 1.96 \cdot \frac{1.5}{5} + 30.612$

$\bar{x} = 31.2$

$31.2 \pm CI \cdot \frac{1.5}{5}$

$31.2 + CI \cdot \frac{1.5}{5} = 31.74.$

$31.2 - CI \cdot \frac{1.5}{5} = 30.66.$

$2\left(CI \cdot \frac{\sigma}{\sqrt{n}}\right) = \text{width}.$

$2\left(CI \cdot \frac{1.5}{5}\right) = 1.08$

$CI = 1.8$


From tables

$2(1 - 0.9641) \times 100$

$100 - \text{ANS}$

$= 92.82$

Leave blank

**6.** The continuous random variable $Y$ is uniformly distributed over the interval

$$[a - 3, a + 6]$$

where $a$ is a constant.

A random sample of 60 observations of $Y$ is taken.

Given that $\overline{Y} = \dfrac{\sum_{i=1}^{60} Y_i}{60}$

(a) use the Central Limit Theorem to find an approximate distribution for $\overline{Y}$

**(3)**

Given that the 60 observations of $Y$ have a sample mean of 13.4

(b) find a 98% confidence interval for the maximum value that $Y$ can take.

**(4)**

$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\mu = \frac{\alpha+\beta}{2} = \frac{a-3+a+6}{2}$

$= a + 3/2$

$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(a+6-a+3)^2}{12}$

$= \frac{9^2}{12} = \frac{27}{4}$

$\frac{27}{4} \div 60 = \frac{9}{80}$

$\bar{Y} \sim N(a+3/2, 9/80)$

(b) $\bar{Y} = 13.4$

$13.4 \pm 2.3263 \cdot \sqrt{\frac{9}{80}}$

$= 13.4 \pm 2.3263 \cdot \frac{3\sqrt{5}}{20}$

$= (12.62, 14.18)$

$a + 3/2$

$12.62 < a+3/2 < 14.18$

$11.12 < a < 12.68$

Max value

$11.12+6 < a+6 < 12.68+6$

$[17.12, 18.68]$

Leave blank

**7.** (i) As part of a recruitment exercise candidates are required to complete three separate tasks. The times taken, $A$, $B$ and $C$, in minutes, for candidates to complete the three tasks are such that

$$A \sim \mathrm{N}(21, 2^2),\ B \sim \mathrm{N}(32, 7^2) \text{ and } C \sim \mathrm{N}(45, 9^2)$$

The time taken by an individual candidate to complete each task is assumed to be independent of the time taken to complete each of the other tasks.

A candidate is selected at random.

(a) Find the probability that the candidate takes a total time of more than 90 minutes to complete all three tasks.

**(4)**

(b) Find $\mathrm{P}(A > B)$

**(4)**

(ii) A simple random sample, $X_1$, $X_2$, $X_3$, $X_4$, is taken from a normal population with mean $\mu$ and standard deviation $\sigma$

Given that

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

and that

$$\mathrm{P}(X_1 > \bar{X} + k\sigma) = 0.1$$

where $k$ is a constant,

find the value of $k$, giving your answer correct to 3 significant figures.

**(7)**

(a) let $A+B+C \rightarrow W$

$P(W > 90)$

$E(W) = E(A) + E(B) + E(C)$

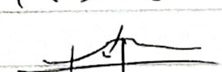
$= 21 + 32 + 45 = 98$

$Var(W) = Var(A) + Var(B) + Var(C) =$

$= 2^2 + 7^2 + 9^2 = 134$

$T \sim N(98, 134)$ $P\left(Z > \frac{90-98}{\sqrt{134}}\right)$

$P(Z > -0.691)$



$-0.691$

$= 0.7549$

b) $P(A - B > 0)$

$E(A) - E(B) = 21 - 32 = -11$

$Var(A) + Var(B) = 2^2 + 7^2$

$= 53$

$X \sim N(-11, 53)$

$P(X > 0)$

$P\left(Z > \frac{11}{\sqrt{53}}\right) = P(Z > 1.51)$

$1 - 0.9345$

$= 0.0655$

Leave blank

**Question 7 continued**

ii) $P(X_1 - \bar{X} > k\sigma) = 0.1$

$E(X_1 - \bar{X})$

$= E(X_1) - E(\bar{X})$

$= \mu - \mu = 0.$

$Var\cancel{(X_1) + Var}(\bar{X})$

$= \cancel{\sigma^2}$

$X_1 - \dfrac{X_1 + X_2 + X_3 + X_4}{4}.$

$\dfrac{3X_1 - (X_2 + X_3 + X_4)}{4}.$

$Var\left(\dfrac{3X_1 - (X_2 + X_3 + X_4)}{4}\right)$

$\dfrac{1}{16}\left[9Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4)\right]$

$= \dfrac{1}{16}\left[9\sigma^2 + 3\sigma^2\right)$

$= \dfrac{3}{4}\sigma^2$

$X_1 - \bar{X} \sim N(0, 0.75\sigma^2)$

$P\left(Z > \dfrac{k\sigma - 0}{\sqrt{\frac{3}{4}\sigma^2}}\right) = 0.1$

$\dfrac{k\sigma}{\sqrt{0.75}\sigma} = 1.2816$ (From tables)

$k = \underline{\underline{1.11}}$