Paper Reference(s)

# 6683/01

# Edexcel GCE

## Statistics S1

## Bronze Level B2

## 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 S1), the paper reference (6683), 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 7 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 |
|---|---|---|---|---|---|
| 73 | 67 | 61 | 53 | 47 | 42 |

**1.** A teacher asked a random sample of 10 students to record the number of hours of television, $t$, they watched in the week before their mock exam. She then calculated their grade, $g$, in their mock exam. The results are summarised as follows.

$$\sum t = 258 \qquad \sum t^2 = 8702 \qquad \sum g = 63.6 \qquad S_{gg} = 7.864 \qquad \sum gt = 1550.2$$

(*a*) Find $S_{tt}$ and $S_{gt}$.

**(3)**

(*b*) Calculate, to 3 significant figures, the product moment correlation coefficient between $t$ and $g$.

**(2)**

The teacher also recorded the number of hours of revision, $v$, these 10 students completed during the week before their mock exam. The correlation coefficient between $t$ and $v$ was –0.753.

(*c*) Describe, giving a reason, the nature of the correlation you would expect to find between $v$ and $g$.

**(2)**

**January 2013**

---

**2.** A bank reviews its customer records at the end of each month to find out how many customers have become unemployed, $u$, and how many have had their house repossessed, $h$, during that month. The bank codes the data using variables $x = \dfrac{u-100}{3}$ and $y = \dfrac{h-20}{7}$.

The results for the 12 months of 2009 are summarised below.

$$\sum x = 477 \qquad S_{xx} = 5606.25 \qquad \sum y = 480 \qquad S_{yy} = 4244 \qquad \sum xy = 23\,070$$

(*a*) Calculate the value of the product moment correlation coefficient for $x$ and $y$.

**(3)**

(*b*) Write down the product moment correlation coefficient for $u$ and $h$.

**(1)**

The bank claims that an increase in unemployment among its customers is associated with an increase in house repossessions.

(*c*) State, with a reason, whether or not the bank's claim is supported by these data.

**(2)**

**May 2012**

---

**3.** A biologist is comparing the intervals ($m$ seconds) between the mating calls of a certain species of tree frog and the surrounding temperature ($t$ °C). The following results were obtained.

| $t$ °C | 8 | 13 | 14 | 15 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|---|---|
| $m$ secs | 6.5 | 4.5 | 6 | 5 | 4 | 3 | 2 | 1 |

(You may use $\sum tm = 469.5$, $S_{tt} = 354$, $S_{mm} = 25.5$)

($a$) Show that $S_{tm} = -90.5$.

**(4)**

($b$) Find the equation of the regression line of $m$ on $t$ giving your answer in the form $m = a + bt$.

**(4)**

($c$) Use your regression line to estimate the time interval between mating calls when the surrounding temperature is 10 °C.

**(1)**

($d$) Comment on the reliability of this estimate, giving a reason for your answer.

**(1)**

**January 2013**

---

**4.** A second hand car dealer has 10 cars for sale. She decides to investigate the link between the age of the cars, $x$ years, and the mileage, $y$ thousand miles. The data collected from the cars are shown in the table below.

| Age, $x$ (years) | 2 | 2.5 | 3 | 4 | 4.5 | 4.5 | 5 | 3 | 6 | 6.5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Mileage, $y$ (thousands) | 22 | 34 | 33 | 37 | 40 | 45 | 49 | 30 | 58 | 58 |

[You may assume that $\sum x = 41$, $\sum y = 406$, $\sum x^2 = 188$, $\sum xy = 1818.5$ ]

($a$) Find $S_{xx}$ and $S_{xy}$.

**(3)**

($b$) Find the equation of the least squares regression line in the form $y = a + bx$. Give the values of $a$ and $b$ to 2 decimal places.

**(4)**

($c$) Give a practical interpretation of the slope $b$.

**(1)**

($d$) Using your answer to part ($b$), find the mileage predicted by the regression line for a 5 year old car.

**(2)**

**January 2008**

---

**5.** The probability function of a discrete random variable $X$ is given by

$$p(x) = kx^2, \qquad x = 1, 2, 3.$$

where $k$ is a positive constant.

($a$) Show that $k = \frac{1}{14}$.

**(2)**

Find

($b$) $P(X \geq 2)$,

**(2)**

($c$) $E(X)$,

**(2)**

($d$) $\text{Var}(1 - X)$.

**(4)**

**January 2010**

---

**6.** The blood pressures, $p$ mmHg, and the ages, $t$ years, of 7 hospital patients are shown in the table below.

| Patient | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| $t$ | 42 | 74 | 48 | 35 | 56 | 26 | 60 |
| $P$ | 98 | 130 | 120 | 88 | 182 | 80 | 135 |

$[\ \sum t = 341,\ \sum p = 833,\ \sum t^2 = 18\,181,\ \sum p^2 = 106\,397,\ \sum tp = 42\,948\ ]$

($a$) Find $S_{pp}$, $S_{tp}$ and $S_{tt}$ for these data.

**(4)**

($b$) Calculate the product moment correlation coefficient for these data.

**(3)**

($c$) Interpret the correlation coefficient.

**(1)**

($d$) Draw the scatter diagram of blood pressure against age for these 7 patients.

**(2)**

($e$) Find the equation of the regression line of $p$ on $t$.

**(4)**

($f$) Plot your regression line on your scatter diagram.

**(2)**

($g$) Use your regression line to estimate the blood pressure of a 40 year old patient.

**(2)**

**January 2010**

---

**7.** Tetrahedral dice have four faces. Two fair tetrahedral dice, one red and one blue, have faces numbered 0, 1, 2, and 3 respectively. The dice are rolled and the numbers face down on the two dice are recorded. The random variable $R$ is the score on the red die and the random variable $B$ is the score on the blue die.

($a$) Find P($R = 3$ and $B = 0$).

**(2)**

The random variable $T$ is $R$ multiplied by $B$.

($b$) Complete the diagram below to represent the sample space that shows all the possible values of $T$.

| | | | | |
|---|---|---|---|---|
| **3** | | | | |
| **2** | | 2 | | |
| **1** | 0 | | | |
| **0** | | | | |
| $B$ / $R$ | **0** | **1** | **2** | **3** |

**Sample space diagram of $T$**

**(3)**

The table below represents the probability distribution of the random variable $T$.

| $t$ | 0 | 1 | 2 | 3 | 4 | 6 | 9 |
|---|---|---|---|---|---|---|---|
| P($T = t$) | $a$ | $b$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $c$ | $\frac{1}{8}$ | $d$ |

($c$) Find the values of $a$, $b$, $c$ and $d$.

**(3)**

Find the values of

($d$) E($T$),

**(2)**

($e$) Var($T$).

**(4)**

**January 2008**

**TOTAL FOR PAPER: 75 MARKS**

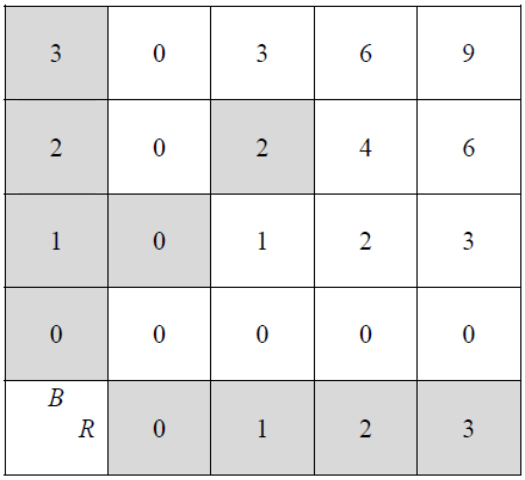
**END**

| Question Number | Scheme | Marks |
|---|---|---|
| **1.** (a) | $(S_{tt}) = 8702 - \frac{258^2}{10}$ or $(S_{gt}) = 1550.2 - \frac{258 \times 63.6}{10}$ | M1 |
| | $(S_{tt} =) 2045.6$ , $(S_{gt} =) -90.68$ awrt (2046), awrt – 90.7 | A1, A1 (3) |
| (b) | $r = \frac{-90.68}{\sqrt{2045.6 \times 7.864}} = -0.714956...$ awrt $-0.715$ | M1 A1 (2) |
| (c) | Positive | B1 |
| | e.g. high *v* corresponds to low *t* and low *t* corresponds to high *g* so expect high *v* to corresponds to high *g*<br>or expect more revision to result in a better grade | B1 (2) **[7]** |
| **2.** (a) | $\left[S_{xy} =\right] 23070 - \frac{477 \times 480}{12}$ $\left[= 3990\right]$ | B1 |
| | $r = \frac{"3990"}{\sqrt{5606.25 \times 4244}}$ | M1 |
| | $= 0.81799\ldots$ awrt 0.818 | A1 (3) |
| (b) | 0.818 | B1ft (1) |
| (c) | Positive correlation or value of *r* is close to 1 or value of $r > 0$ | B1 |
| | So there is support for the bank's claim | B1 |
| | or "increase in unemployment is accompanied by increase in house repossessions" | (2) **[6]** |

| Question Number | Scheme | Marks |
|---|---|---|
| **3.** (a) | $\sum t = 140$ (or $\bar{t} = 17.5$) and $\sum m = 32$ (or $\bar{m} = 4$) | B1 B1 |
| | $(S_{tm}) = 469.5 - \frac{"140" \times "32"}{8}$ | M1 |
| | $(S_{tm} =) -90.5$ | A1cso |
| | | (4) |
| (b) | $b = \frac{S_{tm}}{S_{tt}} = \frac{-90.5}{354}$ | M1 |
| | $b = -0.255649...$ (allow $\frac{181}{708}$) $-0.25$ or awrt $-0.26$ | A1 |
| | $a = \frac{"32"}{8} - b \times \frac{"140"}{8}$ | M1 |
| | So equation of the line is $\underline{m = 8.47 - 0.256t}$ | A1 |
| | | (4) |
| (c) | $(8.47 - 0.256 \times 10 =) 5.9...$ awrt $\underline{5.9}$ | B1 |
| | | (1) |
| (d) | Should be reliable since 10 is in the range (of the data) | B1 |
| | | (1) |
| | | **[10]** |
| **4.** (a) | $S_{xy} = 1818.5 - \frac{41 \times 406}{10}$, = 153.9 (could be seen in (b)) awrt 154 | M1, A1 |
| | $S_{xx} = 188 - \frac{41^2}{10}$ = 19.9 (could be seen in (b)) | A1 |
| | | (3) |
| (b) | $b = \frac{153.9}{19.9}$, = 7.733668… awrt 7.73 | M1, A1 |
| | $a$ = 40.6 − $b$×4.1(= 8.89796…) | M1 |
| | $y$ = 8.89 + 7.73$x$ | A1 |
| | | (4) |
| (c) | A typical car will travel 7700 miles every year | B1ft |
| | | (1) |
| (d) | $x$ = 5, $y$ = 8.89 + 7.73×5(= 47.5− 47.6) | M1 |
| | So mileage predicted is awrt 48000 | A1 |
| | | (2) |
| | | **[10]** |

| Question Number | Scheme | Marks |
|---|---|---|
| **5.** (a) | $k + 4k + 9k = 1$ | M1 |
| | $14k = 1$ | |
| | $k = \frac{1}{14}$ **given** cso | A1 |
| | | (2) |
| (b) | $P(X \geq 2) = 1 - P(X = 1)$ or $P(X = 2) + P(X = 3)$ | M1 |
| | $= 1 - k = \frac{13}{14}$ or 0.92857... awrt 0.929 | A1 |
| | | (2) |
| (c) | $E(X) = 1 \times k + 2 \times k \times 4 + 3 \times k \times 9$ or $36k$ | M1 |
| | $= \frac{36}{14} = \frac{18}{7}$ or $2\frac{4}{7}$ (or exact equivalent) | A1 |
| | | (2) |
| (d) | $\text{Var}(X) = 1 \times k + 4 \times k \times 4 + 9 \times k \times 9, -\left(\frac{18}{7}\right)^2$ | M1 M1 |
| | $\text{Var}(1 - X) = \text{Var}(X)$ | M1 |
| | $= \frac{19}{49}$ or 0.387755... awrt 0.388 | A1 |
| | | (4) |
| | | **[10]** |

| Question Number | Scheme | Marks |
|---|---|---|
| **6.** (a) | $S_{pp} = 106397 - \frac{833^2}{7} = 7270$ | M1 A1 |
| | $S_{tp} = 42948 - \frac{341 \times 833}{7} = 2369,$ | A1 A1 |
| | $S_{tt} = 18181 - \frac{341^2}{7} = 1569.42857...$ or $\frac{10986}{7}$ | (4) |
| (b) | $r = \frac{2369}{\sqrt{7270 \times 1569.42857..}}$ | M1 A1ft |
| | $= 0.7013375$ awrt (0.701) | A1 |
| | | (3) |
| (c) | (Pmcc shows positive correlation.) | |
| | Older patients have higher blood pressure | B1 |
| | | (1) |
| (d) | Points plotted correctly on graph: -1 each error or omission | B2 |
| | | (2) |
| (e) | $b = \frac{2369}{1569.42857..} = 1.509466...$ | M1 A1 |
| | $a = \frac{833}{7} - b \times \frac{341}{7} = 45.467413...$ | M1 |
| | $p = 45.5 + 1.51t$ | |
| | $b = \frac{2369}{1569.42857..} = 1.509466...$ | A1 |
| | | (4) |
| (f) | Line drawn with correct intercept, and gradient | B1ft B1 |
| | | (2) |
| (g) | $t = 40, p = 105.84...$ from equation or graph. awrt 106 | M1 A1 |
| | | (2) |
| | | **[18]** |

| Question Number | Scheme | Marks |
|---|---|---|
| **7.** (a) | $P(R=3 \cap B=0)=\frac{1}{4}\times\frac{1}{4},=\frac{1}{16}$ | M1, A1<br>(2) |
| (b) |  <br>3 0 3 6 9<br>2 0 2 4 6<br>1 0 1 2 3<br>0 0 0 0 0<br>B / R 0 1 2 3<br>All 0s<br>All 1,2,3s<br>All 4,6,9s | B1<br>B1<br>B1<br>(3) |
| (c) | $a=\frac{7}{16}$ | B1 |
| | $b=c=d=\frac{1}{16}$ | B1 B1<br>(3) |
| (d) | $\mathrm{E}(T)=\left(1\times\frac{1}{16}\right)+\left(2\times\frac{1}{8}\right)+\left(3\times\frac{1}{8}\right)+\left(4\times\frac{1}{16}\right)+\ldots$ | M1 |
| | $=2\frac{1}{4}$ or exact equivalent e.g. 2.25, $\frac{9}{4}$ | A1<br>(2) |
| (e) | $\mathrm{Var}(T)=\left(1^2\times\frac{1}{16}\right)+\left(2^2\times\frac{1}{8}\right)+\left(3^2\times\frac{1}{8}\right)+\left(4^2\times\frac{1}{16}\right)+,\ldots-\left(\frac{9}{4}\right)^2$ | M1 A1,<br>M1 |
| | $=\frac{49}{4}-\frac{81}{16}=7\frac{3}{16}\text{or}\frac{115}{16}$ (o.e.) awrt 7.19 | A1<br>(4)<br>**[14]** |

## Examiner reports

### Question 1

Part (a) proved an accessible opening to the paper and nearly all the candidates answered this correctly. Most also knew how to find $r$ in part (b) but a number still gave their final answer as –0.71 rather than the 3 significant figures requested. In part (c) many identified that the required correlation would be positive and gave a simple argument based on the context, although a few linked the variables $v$ and $g$ via the third variable $t$. A number of candidates seemed to misread this part of the question and gave a description of the negative correlation between $t$ and $v$.

### Question 2

This question was on familiar territory and was answered well. Most candidates could carry out the calculations in part (a) successfully although a small minority seemed unfamiliar with the formulae in the formula booklet and we had some using 23070 on the numerator of $r$. Some lost the accuracy mark for rounding to 0.82, or even 0.8, without first stating a more accurate value.

Part (b) was usually answered correctly but a significant minority launched into some complex decoding calculations or simply left it blank. The wording "write down" and the tariff of just 1 mark should indicate that complex calculations are not required.

In part (c) most recognised that a positive correlation suggested support for the bank's claim and scored both marks. Some just stated that the correlation was strong (failing to appreciate the importance of it being positive) and a few gave sociological rather than statistical reasons.

### Question 3

Part (a) was answered very well and only a handful of candidates did not secure the 4 marks here. Most knew how to find the equation of the regression line but sometimes candidates failed to use a sufficiently accurate value of $b$ to ensure that their value of $a$ was accurate to three significant figures and they therefore lost the final accuracy mark. A growing number of candidates are giving their coefficients as fractions, presumably because their calculators are set in this mode. Whilst such answers were accepted, they are not as useful as coefficients of a regression line and arguably not really appropriate in this branch of statistics.

Part (c) was answered very well but in part (d) some candidates' responses were vague: a comment that "it is reliable because it is in the range" was not accepted because "it" does not clearly refer to the temperature. Some candidates used the technical terms of "interpolation" or "not extrapolation" correctly and these were accepted.

**Question 4**

The first two parts of this question were answered very well. There were few problems encountered in parts (a) and (b) although $a = 8.91$ was a common error caused by using the rounded value of $b$ not a more accurate version. Most candidates adhered to the instruction to give their answers to 2 decimal places. Problems started though when the candidates were asked to interpret the equation. Many candidates simply said that mileage increases with age and few who mentioned the 7.7 value remembered the thousands. A simple response such as "the annual mileage is 7700 miles" or "each year a car travels 7700 miles" was rarely seen. In part (d) most could substitute $x = 5$ into their equation but once again the "thousand" was forgotten and the unlikely figure of 48 miles for a 5 year old car was all too common.

**Question 5**

Despite the compact nature of the probability function many candidates gave clear and fully correct solutions to this question. Part (a) was a "Show that" and candidates needed to make sure that they clearly used $\sum \mathrm{p}(x) = 1$ to form a suitable equation in $k$. Part (b) was often answered poorly as a number could not interpret $\mathrm{P}(X \geq 2)$ correctly and gave the answer of $\frac{5}{14}$ (from $\mathrm{P}(X \leq 2)$). Most could answer part (c) and many part (d) too but the usual errors arose here. Some forgot to subtract $\left(\mathrm{E}(X)\right)^2$ and there were a number of incorrect formulae for $\mathrm{Var}(1-X)$ seen such as: $-\mathrm{Var}(X)$, $1-\mathrm{Var}(X)$, $\left[\mathrm{Var}(X)\right]^2$ and $(-1)^2\,\mathrm{E}(X)$.

**Question 6**

This was a high scoring question for most candidates. The calculations in parts (a) and (b) were answered very well with very few failing to use the formulae correctly. Part (c) received a good number of correct responses but many still failed to interpret their value and simply described the correlation as strongly positive. The scatter diagram was usually plotted correctly and most knew how to calculate the equation of the regression line although some used $\mathrm{S}_{pp}$ instead of $\mathrm{S}_{tt}$ and some gave their final equation in terms of y and $x$ instead of $p$ and $t$. Plotting the line in part (f) proved quite challenging for many candidates and a number with the correct equation did not have the gradient correct. Part (g) was usually well done but some chose to use their graph rather than their equation of the line and lost the final accuracy mark.

**Question 7**

Although many candidates found part (a) straightforward there were some surprising incorrect responses. Some simply gave $\mathrm{P}(R = 3)$ and $\mathrm{P}(B = 0)$ without making any attempt to combine them, others added and a few multiplied but obtained the answer 1/8. Part (b) was almost always fully correct but rather surprisingly they did not always see the connection with part (c). Whilst many scored full marks here, some thought all 4 probabilities were equal or they guessed that they were 1/8 like the others. A few had half a page or more of equations with $a$, $b$, $c$ and $d$ in them and usually little progress towards the correct answers. The methods for parts (d) and (e) were generally well known and many scored both marks for the mean, but a mixture of arithmetic errors and the omission of the $-\mu^2$ often spoiled solutions to part (e).

## Statistics for S1 Practice Paper Bronze Level B2

| | | | | Mean score for students achieving grade: | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qu | Max Score | Modal score | Mean % | ALL | A* | A | B | C | D | E | U |
| **1** | 7 | 7 | 83 | 5.81 | 6.39 | 6.32 | 5.89 | 5.59 | 5.21 | 4.88 | 4.01 |
| **2** | 6 | | 80 | 4.78 | 5.87 | 5.74 | 5.43 | 5.04 | 4.57 | 3.97 | 2.68 |
| **3** | 10 | 9 | 83 | 8.25 | 9.44 | 9.13 | 8.48 | 8.05 | 7.48 | 7.11 | 5.12 |
| **4** | 10 | | 75 | 7.54 | | 8.54 | 7.60 | 7.25 | 6.67 | 6.18 | 4.32 |
| **5** | 10 | | 73 | 7.30 | | 9.07 | 8.09 | 7.02 | 5.87 | 4.46 | 2.12 |
| **6** | 18 | | 79 | 14.23 | | 15.83 | 14.71 | 13.64 | 12.59 | 11.19 | 8.86 |
| **7** | 14 | | 78 | 10.96 | | 13.24 | 11.31 | 9.24 | 8.15 | 7.14 | 4.30 |
| | **75** | | **78** | **58.87** | | **67.87** | **61.51** | **55.83** | **50.54** | **44.93** | **31.41** |