2008-01-15 OCR STATS 1 Q01

1 **(i)** The letters A, B, C, D and E are arranged in a straight line.

(a) How many different arrangements are possible? [2]

(b) In how many of these arrangements are the letters A and B next to each other? [3]

(ii) From the letters A, B, C, D and E, two different letters are selected at random. Find the probability that these two letters are A and B. [2]

1(i) (a) $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

(b) AB or BA

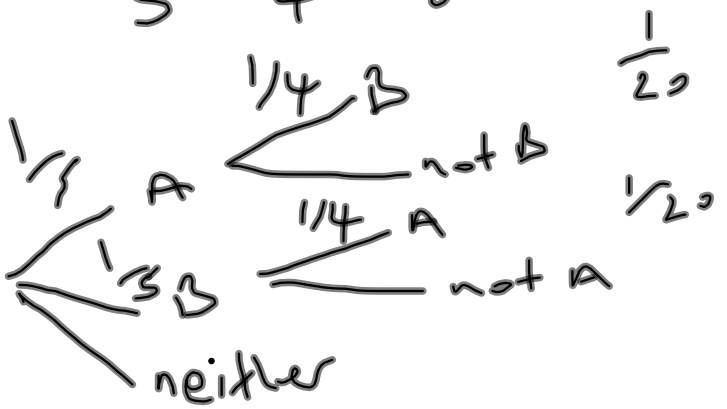
ABxxx etc.

$4! = 4 \times 3 \times 2 \times 1 = 24$

BAxxx

so $24 \times 2 = 48$

(ii) $\frac{1}{5} \times \frac{1}{4} = \frac{1}{20}$ $\frac{1}{20} \times 2 = \frac{1}{10}$



2008-01-15 OCR STATS 1 Q02

2 A random variable T has the distribution Geo$(\frac{1}{5})$. Find

(i) P$(T = 4)$, [2]

(ii) P$(T > 4)$, [2]

(iii) E(T). [1]

2(i) $P(T=4)$ $T \sim Geo(\frac{1}{5})$

$P(T=4) = \left(\frac{4}{5}\right)^3 \times \frac{1}{5} = \frac{64}{625}$ ✓ or 0·102 (3sf)

(ii) $P(T>4) = \left(\frac{4}{5}\right)^4 = \frac{256}{625}$ ✓ or 0.410 (3sf)

(iii) $E(T) = \frac{1}{1/5} = 5$ ✓ $= E(x) = \frac{1}{p}$

2008-01-15 OCR STATS 1 Q03

3 A sample of bivariate data was taken and the results were summarised as follows.

$$n = 5 \quad \Sigma x = 24 \quad \Sigma x^2 = 130 \quad \Sigma y = 39 \quad \Sigma y^2 = 361 \quad \Sigma xy = 212$$

(i) Show that the value of the product moment correlation coefficient r is 0.855, correct to 3 significant figures. [2]

(ii) The ranks of the data were found. One student calculated Spearman's rank correlation coefficient r_s, and found that $r_s = 0.7$. Another student calculated the product moment coefficient, R, of these ranks. State which one of the following statements is true, and explain your answer briefly.

(A) $R = 0.855$
(B) $R = 0.7$
(C) It is impossible to give the value of R without carrying out a calculation using the original data.

[2]

(iii) All the values of x are now multiplied by a scaling factor of 2. State the new values of r and r_s. [2]

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$= \frac{24.8}{\sqrt{14.8 \times 56.8}}$$

$$= 0.855 \text{ (3sf)}$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$= 212 - \frac{(24 \times 39)}{5}$$

$$= 24.8$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$= 130 - \frac{24^2}{5}$$

$$= 14.8$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{5}$$

$$= 361 - \frac{39^2}{5}$$

$$= 56.8$$

(ii) The PRODUCT MOMENT CORRELATION COEFFICIENT of RANKS is the same thing as the SPEARMAN'S RANK CORRELATION COEFFICIENT so

$R = 0.7$ ✓

(iii) $r = 0.855$ (3sf) ✓

$r_s = 0.7$ ✓

scaling the data does not change the PMCC or SRCC ✓

2008-01-15 OCR STATS 1 Q04

4 A supermarket has a large stock of eggs. 40% of the stock are from a firm called Eggzact. 12% of the stock are brown eggs from Eggzact.

An egg is chosen at random from the stock. Calculate the probability that

(i) this egg is brown, given that it is from Eggzact, [2]

(ii) this egg is from Eggzact and is not brown. [2]

(i) $P(\text{Eggzact}) = 0.4$

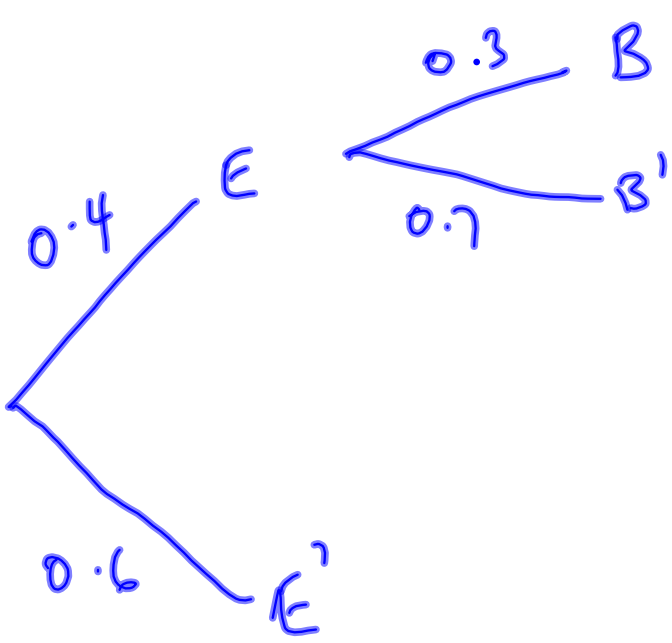
$P(\text{Eggzact and Brown}) = 0.12$

$P(A \cap B) = P(A)P(B|A)$

$P(E \cap B) = P(E)P(B|E)$

$0.12 = 0.4 \times P(B|E)$

$P(B|E) = \frac{0.12}{0.4} = \frac{3}{10}$ ✓


0.3
B
0.12
E
0.4
B'
0.7
0.6
E'

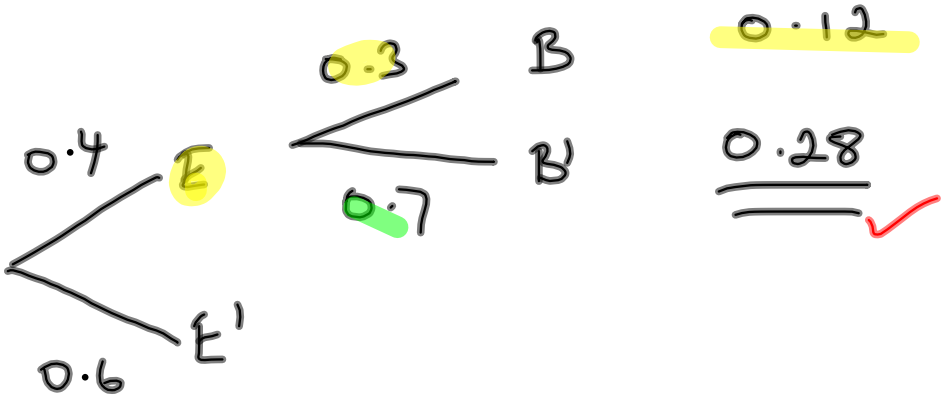
100 eggs

40 eggs exact

12 eggs exact + brown

$$\frac{12}{40} = 0.3$$

(ii)



0.12

0.28

2008-01-15 OCR STATS 1 Q05

5 **(i)** 20% of people in the large town of Carnley support the Residents' Party. 12 people from Carnley are selected at random. Out of these 12 people, the number who support the Residents' Party is denoted by U.

Find

(a) $P(U \leqslant 5)$, [2]

(b) $P(U \geqslant 3)$. [3]

(ii) 30% of people in Carnley support the Commerce Party. 15 people from Carnley are selected at random. Out of these 15 people, the number who support the Commerce Party is denoted by V.

Find $P(V = 4)$. [3]

B(n, p)

5(i) $U \sim B(12, 0.2)$

(a) $P(U \leq 5) = 0.9806$ ✓

(b) $P(U \geq 3) = 1 - P(U \leq 2)$

$= 1 - 0.5583$

$= 0.4417$ ✓ $= 0.442$ (3sf)

(ii) $V \sim B(15, 0.3)$

$P(V = 4) = \binom{15}{4} \times 0.3^4 \times 0.7^{11}$

$= 0.219$ (3sf) ✓

2008-01-15 OCR STATS 1 Q06

6 The probability distribution for a random variable Y is shown in the table.

y	1	2	3
$P(Y = y)$	0.2	0.3	0.5

(i) Calculate $E(Y)$ and $Var(Y)$. [5]

Another random variable, Z, is independent of Y. The probability distribution for Z is shown in the table.

z	1	2	3
$P(Z = z)$	0.1	0.25	0.65

One value of Y and one value of Z are chosen at random. Find the probability that

(ii) $Y + Z = 3$, [3]

(iii) $Y \times Z$ is even. [3]

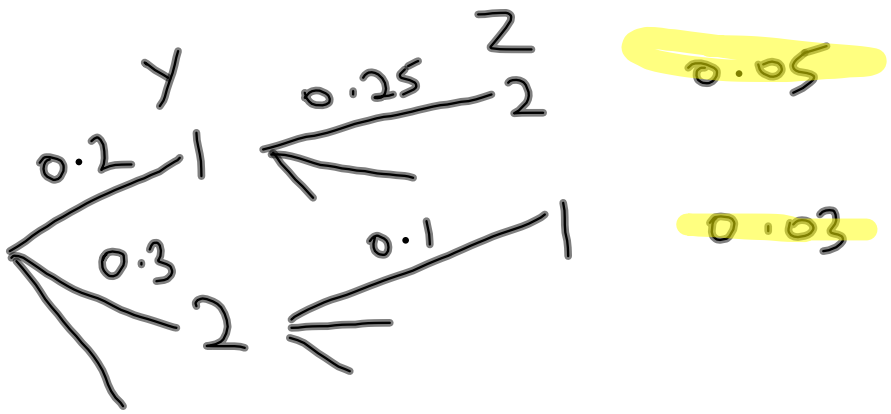
(i) $E(Y) = \sum y \cdot p_i = \underline{2.3}$ ✓

$Var(Y) = \sum y^2 p_i - \mu^2 = 5.9 - 2.3^2$

$= \underline{0.61}$ ✓

y	1	2	3	Σ
$P(Y=y)$	0.2	0.3	0.5	
yp	0.2	0.6	1.5	2.3
y^2	1	4	9	
y^2p	0.2	1.2	4.5	5.9

6(ii)

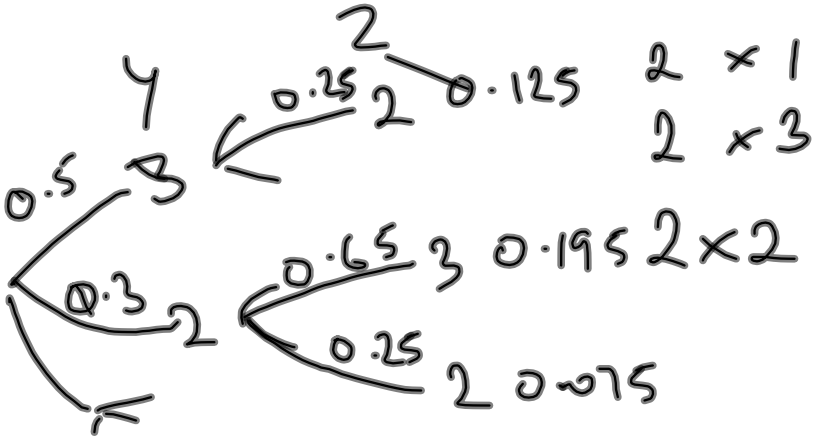


$P(Y+Z=3) = 0.05 + 0.03$

$= 0.08$ ✓

(iii) $Y \times 2 =$ even if

Y	Z	
1	2	0.05
3	2	0.125
2	1	0.03
2	3	0.195
2	2	0.075
		0.475



$P(Y \times 2 = \text{even}) = 0.475$ ✓

7 **(i)** Andrew plays 10 tennis matches. In each match he either wins or loses.

(a) State, in this context, two conditions needed for a binomial distribution to arise. [2]

(b) Assuming these conditions are satisfied, define a variable in this context which has a binomial distribution. [1]

(ii) The random variable X has the distribution $B(21, p)$, where $0 < p < 1$.

Given that $P(X = 10) = P(X = 9)$, find the value of p. [5]

7(i)(a) 1. the result of each tennis match is independent from any other tennis match ✓

2. the probability of winning any tennis match is constant ✓

(b) $X \sim B$(matches played, chance of winning)

X ~ the number of matches won.

7(ii) $X \sim B(21, p)$

$P(X=10) = P(X=9)$

$\binom{21}{10} \times p^{10} \times q^{11} = \binom{21}{9} \times p^{9} \times q^{12}$

$352716 \times p^{10} \times q^{11} = 293930 \times p^{9} \times q^{12}$

$(\div 293930)$ $\frac{6}{5} \times p^{10} \times q^{11} = p^{9} \times q^{12}$

$(\div p^{9})$ $\frac{6}{5} \times p \times q^{11} = q^{12}$

$(\div q^{11})$ $\frac{6}{5} \times p = q$

$\frac{6}{5} \times p = 1 - p$

$(\times 5)$ $6p = 5(1-p)$

$6p = 5 - 5p$

$11p = 5$

$p = \frac{5}{11}$

8 The stem-and-leaf diagram shows the age in completed years of the members of a sports club.

Male		Female
8 8 7 6	1	6 6 6 7 7 8 8 9
7 6 5 5 3 3 2 1	2	1 3 3 4 5 7 8 8 9 9
9 8 4 4 3	3	2 3 3 4 7
5 2 1	4	0 1 8
9 0	5	0

Key: 1 | 4 | 0 represents a male aged 41 and a female aged 40.

(i) Find the median and interquartile range for the males. [3]

(ii) The median and interquartile range for the females are 27 and 15 respectively. Make two comparisons between the ages of the males and the ages of the females. [2]

(iii) The mean age of the males is 30.7 and the mean age of the females is 27.5, each correct to 1 decimal place. Give one advantage of using the median rather than the mean to compare the ages of the males with the ages of the females. [1]

A record was kept of the number of hours, X, spent by each member at the club in a year. The results were summarised by

$$n = 49, \quad \Sigma(x - 200) = 245, \quad \Sigma(x - 200)^2 = 9849.$$

(iv) Calculate the mean and standard deviation of X. [6]

8(i) $n = 4+8+5+3+2 = 22$

$\frac{n+1}{2} = \frac{23}{2} = 11.5$

11^{th} value = 26

12^{th} value = 27

median = $\frac{1}{2}(26+27) = 26.5$ ✓

LQ $\frac{n+1}{4} = \frac{23}{4} = 5.75$

5^{th} value = 21

6^{th} value = 22

$LQ = \frac{(22\times3)+21}{4} = 21.75$ ✓

UQ $\frac{3(n+1)}{4} = \frac{3\times23}{4} = 17.25$

17^{th} value = 39

18^{th} value = 41

$UQ = \frac{(3\times39)+41}{4}$

$= 39.5$ ✓

IQR = UQ − LQ

$= 39.5 - 21.75$

$= 17.75$ ✓

8(ii) The median ages for men (26.5) and women (27) are very close. ✓

The interquartile range for men (17.75) is 2.75 years wider than for women.

This suggests men join younger and leave slightly older than women. ✓

8(iii) Medians are not as affected by outliers as means. There is one male member aged 59 who is much older than the rest of the males at the club. His age has caused the mean value to be increased as compared to the mean age of women.

8(iv) $n = 49$ $\sum(x-200) = 245$

$\sum(x-200)^2 = 9849$

* go over coding

$$\bar{x} = \frac{\sum(x-200)}{n} + 200 = \frac{245}{49} + 200$$

$$= 205 \text{ hrs} \checkmark$$

$$\sigma^2 = \frac{\sum(x-200)^2}{n} - (\bar{x}-200)^2$$

$$= \frac{9849}{49} - 5^2 = 176$$

$$\sigma = \sqrt{176} = 13.266\,499\,16$$

$$\approx 13.3 \text{ hrs (3s.f.)} \checkmark$$

9 It is thought that the pH value of sand (a measure of the sand's acidity) may affect the extent to which a particular species of plant will grow in that sand. A botanist wished to determine whether there was any correlation between the pH value of the sand on certain sand dunes, and the amount of each of two plant species growing there. She chose random sections of equal area on each of eight sand dunes and measured the pH values. She then measured the area within each section that was covered by each of the two species. The results were as follows.

	Dune	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
	pH value, x	8.5	8.5	9.5	8.5	6.5	7.5	8.5	9.0
Area, y cm^2, covered	Species *P*	150	150	575	330	45	15	340	330
	Species *Q*	170	15	80	230	75	25	0	0

The results for species *P* can be summarised by

$$n = 8, \quad \Sigma x = 66.5, \quad \Sigma x^2 = 558.75, \quad \Sigma y = 1935, \quad \Sigma y^2 = 711\,275, \quad \Sigma xy = 17\,082.5.$$

(i) Give a reason why it might be appropriate to calculate the equation of the regression line of y on x rather than x on y in this situation. [1]

(ii) Calculate the equation of the regression line of y on x for species *P*, in the form $y = a + bx$, giving the values of a and b correct to 3 significant figures. [4]

(iii) Estimate the value of y for species *P* on sand where the pH value is 7.0. [2]

The values of the product moment correlation coefficient between x and y for species *P* and *Q* are $r_P = 0.828$ and $r_Q = 0.0302$.

(iv) Describe the relationship between the area covered by species *Q* and the pH value. [1]

(v) State, with a reason, whether the regression line of y on x for species *P* will provide a reliable estimate of the value of y when the pH value is

(a) 8, [1]

(b) 4. [1]

(vi) Assume that the equation of the regression line of y on x for species *Q* is also known. State, with a reason, whether this line will provide a reliable estimate of the value of y when the pH value is 8. [1]

9(i) Because it is thought that the area of vegetation covered is affected by (dependent on) the pH value of the soil.

(ii) $y = a + bx$ $\quad b = \frac{S_{xy}}{S_{xx}}$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$= 17082.5 - \frac{66.5 \times 1935}{8}$$

$$= 997.8125$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= 558.75 - \frac{66.5^2}{8}$$

$$= 5.96875$$

$$b = \frac{997.8125}{5.96875} = 167.172774 9$$

$$= \underline{167} \ (3\text{sf})$$

$$a = \bar{y} - b\bar{x}$$

$$\bar{y} = \frac{1935}{8} = 241.875$$

$$\bar{x} = \frac{66.5}{8} = 8.3125$$

$$a = 241.875 - (167.172774 9 \times 8.3125)$$

$$= -1147.748691 = \underline{-1150} \ (3\text{sf})$$

$$y = -1150 + 167x$$

(iii) $x = 7$

$y = -1150 + (167 \times 7)$

$= \underline{19}$

(iv) There is very little correlation between pH value and coverage of plant Q.

(v)(a) pH of 8 is within the range of values measured so will give a reasonable estimate. And r value is high for plant P.

(b) pH of 4 is outside that range so will not give a good estimate as this would be EXTRAPOLATING from the data.

vi) a regression line for Q on pH won't give good estimate because the r value is very low so there isn't a good correlation to work with.