THE ROYAL STATISTICAL SOCIETY

2005 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

PAPER I – STATISTICAL THEORY

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Higher Certificate, Paper I, 2005. Question 1

(i) $\binom{49}{6} = \dfrac{49!}{6!\,43!} = \dfrac{49.48.47.46.45.44}{6.5.4.3.2.1} = 13983816.$

(ii) The exact distribution of X is binomial with n = 10,000,000 and p = 1/13983816.

Its Poisson approximation has parameter (mean) $\lambda = np = 0.7151124$.

(a) $P(X=0) = e^{-\lambda} = e^{-0.7151124} = 0.48914.$

(b) $P(X=1) = \lambda e^{-\lambda} = 0.34979.$

(iii) $\binom{31}{6} = \dfrac{31!}{6!\,25!} = \dfrac{31.30.29.28.27.26}{6.5.4.3.2.1} = 736281.$

Hence (using also the result of part (i)) we have

$$P(\text{winning set contains no number} > 31) = \frac{736281}{13983816} = 0.05265.$$

(iv) Let U be the number out of the 3,000,000 players choosing from (1, 2, …, 31) who match all 6 winning numbers, and let V be the number out of the 7,000,000 players choosing from all the numbers (1, 2, …, 49) who match all 6 winning numbers.

Then $Y = U + V$, where (given the winning numbers) U and V are independent.

(a) If all six winning numbers are in the list (1, 2, …, 31), then U is binomial with n = 3,000,000 and p = 1/736281, which is approximated by Poisson(np) i.e. Poisson (4.07453).

Similarly, V is binomial with n = 7,000,000 and p = 1/13983816, which is approximated by Poisson(0.50058).

Thus, using the result that Poisson(λ_1) + Poisson(λ_2) = Poisson($\lambda_1 + \lambda_2$) [where the two Poisson distributions are independent], we have that the approximate distribution of Y is Poisson(4.57511).

(b) In this case, U must be zero so we simply have $Y = V$, i.e. Poisson(0.50058) approximately.

Higher Certificate, Paper I, 2005. Question 2

Cycle ~ N(27, 6.25)

Bus ~ N(13, 20)   Walk$_1$ ~ N(7, 4)   Walk$_2$ ~ N(5, 1)

Car ~ N(23, 36)

The sum of independent N(μ_i, σ_i^2) distributions is N($\Sigma\mu_i$, $\Sigma\sigma_i^2$).

(i) The distribution of total journey time by bus is N(7+13+5, 4+20+1), i.e. N(25,25).

(ii) Cycle: $P\left(\mathrm{N}(27,6.25)<30\right)=\Phi\left(\dfrac{30-27}{\sqrt{6.25}}\right)=\Phi(1.2)=0.8849$.

Bus: $P\left(\mathrm{N}(25,25)<30\right)=\Phi\left(\dfrac{30-25}{\sqrt{25}}\right)=\Phi(1)=0.8413$.

Car: $P\left(\mathrm{N}(23,36)<30\right)=\Phi\left(\dfrac{30-23}{\sqrt{36}}\right)=\Phi(1.1667)=0.8783$.

Cycling is best, with a probability of 0.8849.

(iii) Cycle: $P\left(\mathrm{N}(27,6.25)>35\right)=1-\Phi\left(\dfrac{35-27}{\sqrt{6.25}}\right)=1-\Phi(3.2)=0.0007$.

Bus: $P\left(\mathrm{N}(25,25)>35\right)=1-\Phi\left(\dfrac{35-25}{\sqrt{25}}\right)=1-\Phi(2)=0.0228$.

Car: $P\left(\mathrm{N}(23,36)>35\right)=1-\Phi\left(\dfrac{35-23}{\sqrt{36}}\right)=1-\Phi(2)=0.0228$.

Again cycling is best, with a probability of 0.0007.

(iv) P(cycle) = 0.3 P(bus) = 0.3 P(car) = 0.4

$P\left(\text{cycle}\,|<30\right)=\dfrac{P\left(<30\,|\,\text{cycle}\right)P\left(\text{cycle}\right)}{P\left(<30\right)}$, and similarly for the other modes of travel.

$$P(<30)=P\left(<30\,|\,\text{cycle}\right)P\left(\text{cycle}\right)+P\left(<30\,|\,\text{bus}\right)P\left(\text{bus}\right)+P\left(<30\,|\,\text{car}\right)P\left(\text{car}\right)$$
$$=\left(0.8849\times0.3\right)+\left(0.8413\times0.3\right)+\left(0.8783\times0.4\right)$$
$$=0.26547+0.25239+0.35132 \;=\; 0.86918.$$

Hence $P\left(\text{cycle}\,|<30\right)=0.26547/0.86918=0.3054$

$P\left(\text{bus}\,|<30\right)=0.25239/0.86918=0.2904$

$P\left(\text{car}\,|<30\right)=0.35132/0.86918=0.4042$.

<u>Higher Certificate, Paper I, 2005. Question 3</u>

(i) $P(T > t) = P(\text{no failure in time } t) = P(X = 0) = e^{-\lambda t}$.

$\therefore$ $1 - F(t) = e^{-\lambda t}$ and so the pdf of T is $\frac{d}{dt}F(t) = -\frac{d}{dt}\left(e^{-\lambda t}\right) = \lambda e^{-\lambda t}$.

(ii) P(all n systems still functioning at time t) = $\prod_{i=1}^{n} P(T_i > t) = \prod_{i=1}^{n} e^{-\lambda t} = e^{-n\lambda t}$.

$\therefore$ the pdf of T_{min} is $-\frac{d}{dt}\left(e^{-n\lambda t}\right) = n\lambda e^{-n\lambda t}$ (for $t > 0$).

Thus T_{min} is exponential with parameter $n\lambda$.

(iii) P(all n systems have failed by time t) = $\prod_{i=1}^{n} P(T_i \le t) = \prod_{i=1}^{n}\left(1 - e^{-\lambda t}\right)$

$= \left(1 - e^{-\lambda t}\right)^n$, and this is $P(T_{\text{max}}) \le t$.

$\therefore$ the pdf of T_{max} is $\frac{d}{dt}\left\{\left(1 - e^{-\lambda t}\right)^n\right\} = n\lambda e^{-\lambda t}\left(1 - e^{-\lambda t}\right)^{n-1}$ (for $t > 0$).

(Note that this is not exponential.)

We now have $n = 10$ and $\lambda = 0.002$. We require t_1 and t_2 such that $1 - e^{-n\lambda t_1} = 0.05$ and $\left(1 - e^{-\lambda t_2}\right)^n = 0.95$.

$\therefore$ $1 - e^{-0.02 t_1} = 0.05$, giving $0.02t_1 = -\log(0.95) = 0.051293$ so that $t_1 = 2.565$.

Also, $1 - e^{-0.002 t_2} = (0.95)^{1/10} = 0.994884$, giving $0.002t_2 = -\log(0.005116) = -5.2754$ so that $t_2 = 2638$.

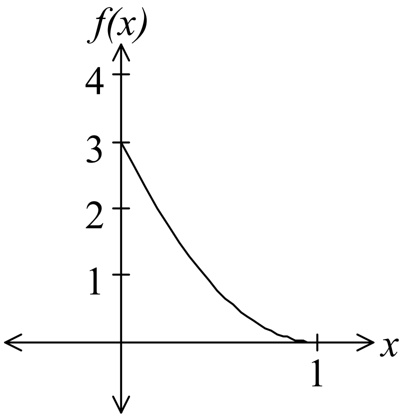
Higher Certificate, Paper I, 2005. Question 4

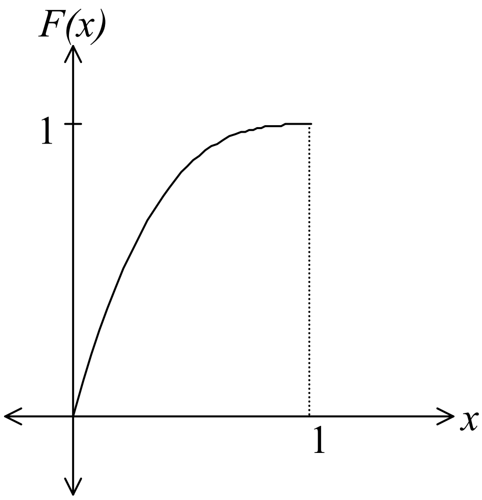
$$f(x) = \alpha(1-x)^{\alpha-1}, \qquad 0 < x < 1, \qquad \alpha > 0.$$

(i) $F(x) = \int_0^x \alpha(1-u)^{\alpha-1}\,du = \left[-(1-u)^{\alpha}\right]_0^x = 1-(1-x)^{\alpha}$ (for $0 < x < 1$ and $\alpha > 0$).

The median m is given by $F(m) = ½$, so we have $1 - (1 - m)^{\alpha} = ½$ or $(1 - m)^{\alpha} = ½$, so that $1 - m = 2^{-1/\alpha}$, i.e. $m = 1 - 2^{-1/\alpha}$.

When $\alpha = 3$, $f(x) = 3(1-x)^2$ and $F(x) = 1-(1-x)^3$ (in [0, 1]).





The solution to part (ii) is on the next page

(ii) $\quad L = \prod_{i=1}^{n}\left[\alpha(1-x_i)^{\alpha-1}\right] = \alpha^n \prod_{i=1}^{n}(1-x_i)^{\alpha-1}$.

Hence $\log L = n\log\alpha + (\alpha-1)\sum_{i=1}^{n}\log(1-x_i)$.

$\therefore \dfrac{d\log L}{d\alpha} = \dfrac{n}{\alpha} + \sum_{i=1}^{n}\log(1-x_i)$ which on setting equal to zero gives that the maximum likelihood estimate is $\hat{\alpha} = \dfrac{-n}{\sum_{i=1}^{n}\log(1-x_i)}$, as required. [Consideration of $\dfrac{d^2\log L}{d\alpha^2}$ (see below) confirms that this is a maximum.]

$\dfrac{d^2\log L}{d\alpha^2} = -\dfrac{n}{\alpha^2}$. Hence, using the result quoted in the question, $\hat{\alpha}$ is approximately Normally distributed with mean α and variance $\dfrac{\alpha^2}{n}$. We estimate the variance by $\dfrac{\hat{\alpha}^2}{n}$, so that we have $\hat{\alpha} \sim \mathrm{N}\left(\alpha,\ \dfrac{\hat{\alpha}^2}{n}\right)$, approximately.

Hence an approximate 90% confidence interval is given by

$$0.90 \approx P\left(-1.645 < \frac{\hat{\alpha}-\alpha}{\hat{\alpha}/\sqrt{n}} < 1.645\right),$$

leading to the interval $\left(\hat{\alpha} - \dfrac{1.645\hat{\alpha}}{\sqrt{n}}\ ,\quad \hat{\alpha} + \dfrac{1.645\hat{\alpha}}{\sqrt{n}}\right)$.

For the given sample, we have $n = 5$ and the values of $1 - x_i$ are 0.88, 0.57, 0.93, 0.13 and 0.71. Therefore

$$\Sigma\log(1 - x_i) = -0.1278 - 0.5621 - 0.0726 - 2.0402 - 0.3425 = -3.1452$$

giving $\hat{\alpha} = \dfrac{5}{3.1452} = 1.5897$.

Also, $\dfrac{1.645\hat{\alpha}}{\sqrt{n}} = \dfrac{1.645\times 1.5897}{\sqrt{5}} = 1.1695$, so the confidence interval is (0.420, 2.759).

<u>Higher Certificate, Paper I, 2005. Question 5</u>

(i) We have $Y \sim \mathrm{B}(n, p)$, so $P(Y=y) = \dfrac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$ (for $y = 0, 1, 2, \ldots, n$ and $0 < p < 1$). The likelihood L is simply $P(Y = y)$.

Hence $\log L = \text{constant} + y \log p + (n-y)\log(1-p)$.

$\therefore \dfrac{d \log L}{dp} = \dfrac{y}{p} - \dfrac{n-y}{1-p}$ which on setting equal to zero gives that the maximum likelihood estimate is $\hat{p} = \dfrac{y}{n}$. [Consideration of $\dfrac{d^2 \log L}{dp^2}$ confirms that this is a maximum.]

$\mathrm{Var}(\hat{p}) = \dfrac{1}{n^2}\mathrm{Var}(Y) = \dfrac{1}{n^2} np(1-p) = \dfrac{p(1-p)}{n}$. We may estimate p by $\hat{p}$ in this and thus obtain an estimate of the standard error of $\hat{p}$ as $\mathrm{SE}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

For $n = 100$ and $y = 20$, we have $\hat{p} = 0.2$ and $\mathrm{SE}(\hat{p}) = \sqrt{\dfrac{0.2 \times 0.8}{100}} = 0.04$.

(ii) P("yes") is given by P(takes drugs <u>and</u> coin shows "takes drugs") + P(does not take drugs <u>and</u> coin shows "does not take drugs").

Hence $\theta = P(\text{"yes"}) = 0.75p + 0.25(1-p) = 0.25 + 0.5p$.

$Z \sim \mathrm{B}(n, \theta)$, so from part (i) we have $\hat{\theta} = z/n$ and $\mathrm{SE}(\hat{\theta}) = \sqrt{\hat{\theta}(1-\hat{\theta})/n}$.

We have $p = 2\theta - ½$, so the MLE of p is $\tilde{p} = 2\hat{\theta} - \frac{1}{2}$.

Thus $\mathrm{SE}(\tilde{p}) = 2\,\mathrm{SE}(\hat{\theta}) = 2\sqrt{\hat{\theta}(1-\hat{\theta})/n}$.

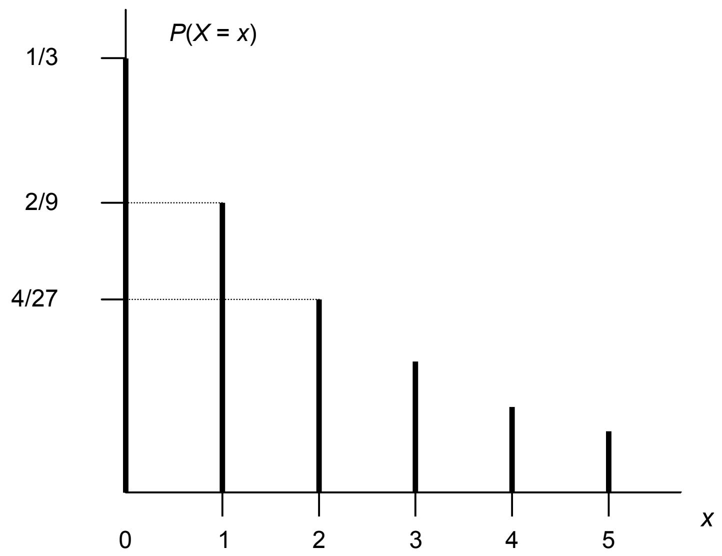
For $n = 100$ and $z = 45$, we have $\hat{\theta} = 0.45$ and so $\tilde{p} = 0.4$ and $\mathrm{SE}(\tilde{p}) = 0.0995$.

(iii) Both p and the standard error are estimated to be larger by the second survey. The larger p is plausible, as some people are likely not to admit to taking drugs when asked directly as in the first survey. There is a much better expectation of truthful answers in the second survey. We should not claim that the first is better just because the SE is smaller; the first survey is likely to be biased. (Is it clear what the journalist means by "reliable"?)

Higher Certificate, Paper I, 2005. Question 6

Probability mass function: $f(x) = (1-p)^x p, \quad x = 0, 1, 2, \ldots, \quad 0 < p < 1$.

(i)



(ii) Probability generating function $G(s)$ is

$$G(s) = E\left[s^X\right] = \sum_{x=0}^{\infty} s^x (1-p)^x p = p\sum_{x=0}^{\infty}\{(1-p)s\}^x = \frac{p}{1-(1-p)s}$$

(requires $|s| < 1/(1-p)$ for convergence).

The mean is given by $E[X] = G'(1)$. We have $G'(s) = p\dfrac{1-p}{\{1-(1-p)s\}^2}$ and inserting $s = 1$ gives $G'(1) = \dfrac{p(1-p)}{p^2}$, i.e. the mean is $\dfrac{1-p}{p}$.

The variance is given by $\text{Var}(X) = G''(1) + \text{mean} - \text{mean}^2$. $G''(s) = \dfrac{2p(1-p)^2}{\{1-(1-p)s\}^3}$, so that $G''(1) = \dfrac{2(1-p)^2}{p^2}$. Thus the variance is given by $\dfrac{2(1-p)^2}{p^2} + \dfrac{1-p}{p} - \left(\dfrac{1-p}{p}\right)^2$

$$= \frac{1}{p^2}\left\{(1-p)^2 + p(1-p)\right\} = \frac{1-p}{p^2}.$$

The solution to parts (iii) and (iv) is on the next page

(iii) $P(X \geq x) = \sum_{r=x}^{\infty} p(1-p)^r$ [geometric series] $= \dfrac{p(1-p)^x}{1-(1-p)} = (1-p)^x$ (for x = 0, 1, 2, …). We now use $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ and take the event A as "$X \geq l + m$" and the event B as "$X \geq l$", so that $A \cap B = A$. Thus

$$P(X \geq l+m | X \geq l) = \frac{(1-p)^{l+m}}{(1-p)^l} = (1-p)^m = P(X \geq m).$$

This is the "lack of memory" property of a geometric distribution.

(iv) By independence, $P(Z \geq z) = P(X \geq z)P(Y \geq z) = (1-p)^z (1-\theta)^z$.

$$P(Z = z) = P(Z \geq z) - P(Z \geq z+1) = \{(1-p)(1-\theta)\}^z - \{(1-p)(1-\theta)\}^{z+1}$$

$$= \{(1-p)(1-\theta)\}^z (1-1+p+\theta-p\theta) = \{(1-p)(1-\theta)\}^z (p+\theta-p\theta)\text{, for } z = 0, 1, \ldots .$$

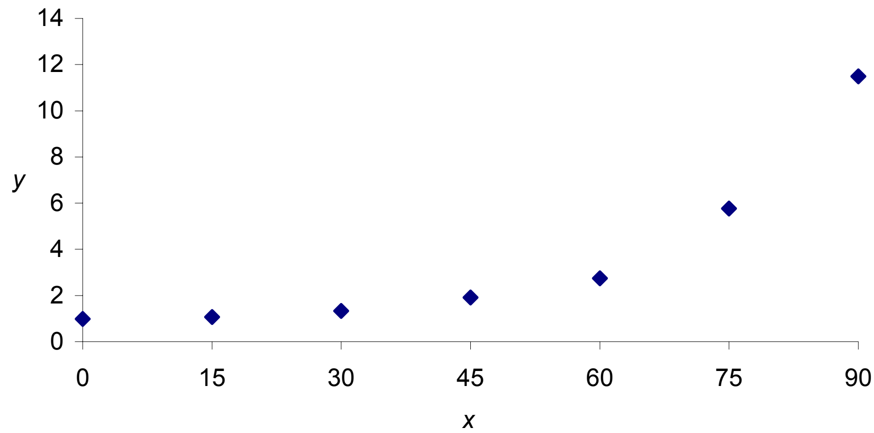
This is a geometric distribution as given at the start of the question with p replaced by $p + \theta - p\theta$. Hence, from part (ii),

$$E[Z] = \frac{1-p-\theta+p\theta}{p+\theta-p\theta}, \qquad \mathrm{Var}(Z) = \frac{1-p-\theta+p\theta}{(p+\theta-p\theta)^2}.$$

Higher Certificate, Paper I, 2005. Question 7

Note There are many equivalent forms of the formulae that are required to be stated. The basic expressions are $\Sigma(x-\bar{x})(y-\bar{y})$, $\Sigma(x-\bar{x})^2$ and $\Sigma(y-\bar{y})^2$. Convenient computing expressions are $\Sigma xy-(\Sigma x \Sigma y)/n$ and similarly for the others. [Where appropriate, numerators and denominators of fractions could both be multiplied by n (7) to avoid possible slight inaccuracies caused by rounding when dividing by 7.]

Part (i)



$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}}=\frac{\Sigma xy-\Sigma x\Sigma y/n}{\sqrt{\left(\Sigma x^2-(\Sigma x)^2/n\right)\left(\Sigma y^2-(\Sigma y)^2/n\right)}}$$

$$=\frac{1773.795-(315\times 25.305/7)}{\sqrt{(20475-315^2/7)(180.474-25.305^2/7)}}=\frac{635.07}{748.784}=0.848\,.$$

This indicates a strong linear association between x and y, but nevertheless the scatter diagram clearly suggests that the relationship is curved.

Part (ii)

We have $\dfrac{1}{y}=\dfrac{1-\sin^2 x}{a}+\dfrac{\sin^2 x}{b}=\dfrac{1}{a}+\sin^2 x\left(\dfrac{1}{b}-\dfrac{1}{a}\right)$, i.e. $Y=\dfrac{1}{a}+\left(\dfrac{1}{b}-\dfrac{1}{a}\right)X$ where X and Y are as given.

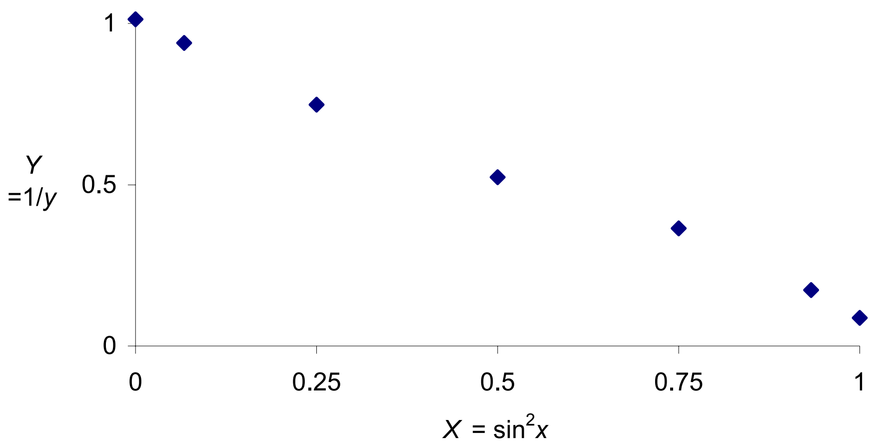
So $A=\dfrac{1}{a}$ and $B=\dfrac{1}{b}-\dfrac{1}{a}$.

The solution to part (iii) is on the next page

Part (iii)

$X = \sin^2 x$	0	0.067	0.250	0.500	0.750	0.933	1
$Y = 1/y$	1.013	0.940	0.748	0.523	0.365	0.173	0.087

[Note. Summary statistics for these are given in the question.]



The scatter diagram indicates that the relationship between X and Y is very close to linear, with little scatter about a straight line. Linear regression should be suitable.

We have $\bar{X} = 3.5/7 = 0.500$ and $\bar{Y} = 3.849/7 = 0.550$.

So the fitted linear regression is $Y = A + BX$ where $B = \dfrac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$ and $A = \bar{Y} - B\bar{X}$.

Carrying out the calculations as in part (i), we get

$$B = \frac{1.03385 - (3.5 \times 3.849/7)}{2.75 - (3.5^2/7)} = -\frac{0.89065}{1} = -0.89065$$

and hence $A = 0.550 + (0.89065)(0.500) = 0.9953$.

Thus the corresponding estimates of a and b are given by $\hat{a} = \dfrac{1}{0.9953} = 1.0047$ and $\hat{b} = \left(B + \dfrac{1}{\hat{a}}\right)^{-1} = \dfrac{1}{0.10465} = 9.556$.

The correlation coefficient for X and Y is

$$\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}} = \frac{-0.89065}{1 \times \sqrt{\Sigma(Y-\bar{Y})^2}} = \frac{-0.89065}{\sqrt{2.9136 - (3.849^2/7)}} = -0.9975.$$

This is an even stronger indication of a linear relationship than in part (i), and we can see from the scatter diagram that the relationship appears almost purely linear (very little random scatter, certainly no curved component).

Higher Certificate, Paper I, 2005. Question 8

(i) The marginal distributions of X and Y are as shown, appended to the table.

		Values of Y			Marginal distribution of X
		0	1	2	
Values of X	–1	1/6	1/12	1/12	1/3
	0	1/12	1/6	1/12	1/3
	1	1/12	1/12	1/6	1/3
Marginal distribution of Y		1/3	1/3	1/3	

Hence $E[X] = 0$ and $E[Y] = 1$ (by symmetry; by noting that X and Y are both discrete uniform; or by explicit calculation).

$\text{Var}(X) = \Sigma(x-0)^2 P(X=x) = \{(-1)^2 + 0^2 + 1^2\}/3 = 2/3$.

$\text{Var}(Y)$ can be calculated similarly; or, since $Y = X + 1$, we have $\text{Var}(Y) = \text{Var}(X)$.

(ii) The conditional distributions of Y for each value of X are as follows.

$Y =$	0	1	2
$X = -1$	1/2	1/4	1/4
0	1/4	1/2	1/4
1	1/4	1/4	1/2

Hence $E[Y \mid X=-1] = (0)(½) + (1)(¼) + (2)(¼) = 3/4$.

Similarly, $E[Y \mid X = 0] = 1$ and $E[Y \mid X = 1] = 5/4$.

Thus we have $E[Y \mid X = x] = 1 + \frac{x}{4}$.

(iii) $E[XY] = (-1)(0)(1/6) + (-1)(1)(1/12) + \ldots.. + (1)(2)(1/6) = 1/6$.

$\therefore \text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{1}{6} - (0)(1) = 1/6$.

$$\therefore \text{Corr}(X, Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} = \frac{1/6}{2/3} = \frac{1}{4}.$$

X and Y are not independent: their correlation (or covariance) is non-zero. (In fact we saw in part (i) that they are linearly related: $Y = X + 1$.)

(iv) $Z = X^3 + (Y-1)^3$. The values of Z and their probabilities are as shown:

$Y =$	0	1	2
$X = -1$	$Z = -2$; $p = 1/6$	$Z = -1$; $p = 1/12$	$Z = 0$; $p = 1/12$
0	$Z = -1$; $p = 1/12$	$Z = 0$; $p = 1/6$	$Z = 1$; $p = 1/12$
1	$Z = 0$; $p = 1/12$	$Z = 1$; $p = 1/12$	$Z = 2$; $p = 1/6$

Thus we have

Values of Z	–2	–1	0	1	2
Probabilities	1/6	1/6	1/3	1/6	1/6