THE ROYAL STATISTICAL SOCIETY

2003 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.

Higher Certificate, Paper I, 2003. Question 1

(i) (a) Any of 0, 1, …, 9 can occur in each of the six positions, so the number is $10^6 = 1000000$.

(b) $10 \times 9 \times 8 \times 7 \times 6 \times 5 = 151200$, since no repetition is allowed.

[Alternatively, $\dfrac{10!}{(10-6)!} = \dfrac{10!}{4!}$ as above.]

(c) There are $\binom{10}{6}$ choices of six different digits from ten, each of which can be used in only one of its possible orders. So the number is

$$\binom{10}{6} = \frac{10!}{6!4!} = 210 .$$

(d) Here there are $\binom{10}{3}$ choices of digits, for each of which there are $\dfrac{6!}{2!2!2!} = 90$ orders, so the number is $\binom{10}{3} \times 90 = 10800$.

(ii) (a) There are $3! = 6$ possible orders for the first three digits, and 1 order (the reverse order of the first three digits) for the last three. So there are 6 codes.

(b) If one digit is used 4 times in a palindromic code, another must be used twice. These digits may be chosen in 3 and 2 ways respectively, i.e. in 6 ways for the pair. Once the digits have been chosen, only 3 patterns are possible; for example, say the digits are 1 and 2, then the possible patterns are 1 1 2 2 1 1, 1 2 1 1 2 1 and 2 1 1 1 1 2. So the total number of codes is $6 \times 3 = 18$.

(c) There are 3 choices of digit and only one possible pattern for each; so there are 3 codes.

(iii) Using the patterns from part (ii), there are $\binom{10}{3}$ choices of digits for (a), each giving 6 codes, i.e. $\binom{10}{3} \times 6 = 720$ altogether. For (b), there are $\binom{10}{2}$ choices of digits, i.e. 45, with 3 patterns as above, in which the two digits can be used in 2 ways (4 of 1 and 2 of 2 *or* vice versa), giving $45 \times 3 \times 2 = 270$ ways. And (c) can occur in 10 ways (because any digit can be used 6 times). Thus the total is $720 + 270 + 10 = 1000$ ways.

ALTERNATIVELY, the first three positions may each be filled in 10 ways, and then the whole sequence is determined, so there are $10^3 = 1000$ ways.

Higher Certificate, Paper I, 2003. Question 2

$$P(A)=\frac{2}{3} \qquad P(B)=\frac{1}{2} \qquad P(C)=\frac{1}{4}$$

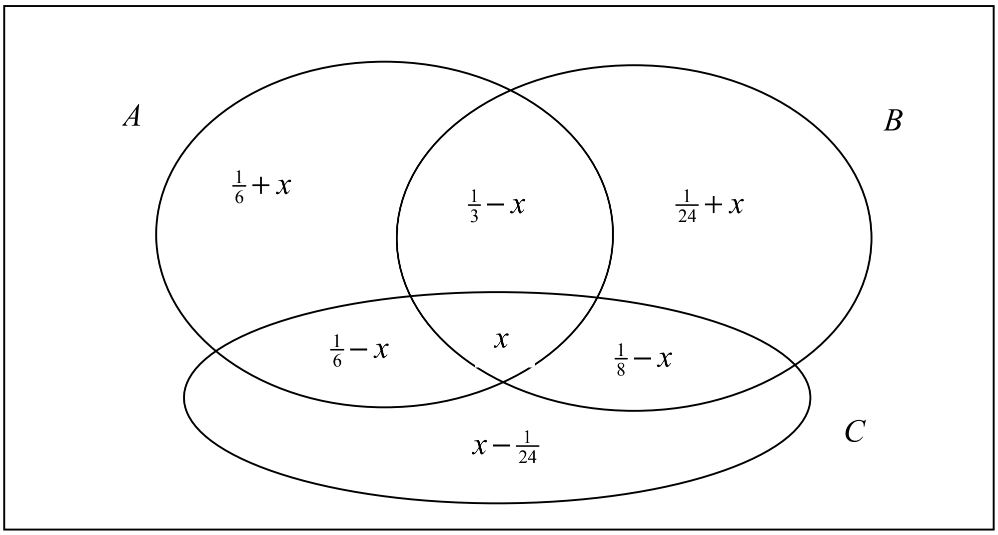
(i) (a) By the given independence,

$$P\left(A\cap\overline{B}\cap\overline{C}\right)=P(A)P\left(\overline{B}\right)P\left(\overline{C}\right)=\frac{2}{3}\left(1-\frac{1}{2}\right)\left(1-\frac{1}{4}\right)=\frac{1}{4}\ .$$

(b) $$P\left(A\cap\overline{C}\,\middle|\,A\cap\overline{B}\right)=\frac{P\left(\left(A\cap\overline{C}\right)\cap\left(A\cap\overline{B}\right)\right)}{P\left(A\cap\overline{B}\right)}=\frac{P\left(A\cap\overline{C}\cap\overline{B}\right)}{P\left(A\cap\overline{B}\right)}$$

$$=\frac{\frac{1}{4}}{\frac{2}{3}\left(1-\frac{1}{2}\right)}=\frac{3}{4}\ .$$

(ii)



Using pairwise independence, the value of $P(A\cap B)$ is $P(A)P(B)$, etc, and hence the values $\frac{1}{3}-x$, $\frac{1}{8}-x$ and $\frac{1}{6}-x$ are found. The others follow using $P(A)$, $P(B)$ and $P(C)$.

(a) $P\left(A\cap\overline{B}\cap\overline{C}\right)=\frac{1}{6}+x$ from the diagram.

(b) $$P\left(A\cap\overline{C}\,\middle|\,A\cap\overline{B}\right)=\frac{P\left(\left(A\cap\overline{C}\right)\cap\left(A\cap\overline{B}\right)\right)}{P\left(A\cap\overline{B}\right)}=\frac{P\left(A\cap\overline{C}\cap\overline{B}\right)}{P\left(A\cap\overline{B}\right)}$$

$$=\frac{\frac{1}{6}+x}{\left(\frac{1}{6}+x\right)+\left(\frac{1}{6}-x\right)}=3x+\frac{1}{2}\ .$$

(c) $$P(A\cup B\cup C)=\frac{19}{24}+x\ .$$

Since all probabilities must lie in [0,1], we have $x\geq\frac{1}{24}$ and $x\leq\frac{1}{8}$, i.e.

$$\frac{1}{24}\leq x\leq\frac{1}{8}\ .$$

Higher Certificate, Paper I, 2003. Question 3

(i) (a) $X + Y_A$ is N(10 + 15 , 12 + 16) i.e. N(25, 28).

(b) $X + Y_B$ is N(10 + 12 , 12 + 9) i.e. N(22, 21).

(ii) If X is the same for both, we require $P(Y_A < Y_B)$, i.e. $P(Y_A - Y_B < 0)$.

$Y_A - Y_B$ is N(15 – 12 , 16 + 9) i.e. N(3, 25).

$P(Y_A - Y_B < 0) = \Phi\left(\frac{0-3}{5}\right)$ where (as usual) Φ denotes the cdf of the N(0, 1) distribution. From tables $\Phi(-0.6) = 1 - \Phi(0.6) = 0.2743$.

(iii) Writing $W_A = X + Y_A$ and $W_B = X + Y_B$, we require $P(W_A < W_B)$, i.e. $P(W_A - W_B < 0)$.

$W_A - W_B$ is N(25 – 22 , 28 + 21) i.e. N(3, 49).

$$P(W_A - W_B < 0) = \Phi\left(\frac{0-3}{7}\right) = \Phi\left(-\frac{3}{7}\right) = 0.3341.$$

(iv) $\bar{W}_A$ is $N\left(25, \frac{28}{16}\right)$ and $\bar{W}_B$ is $N\left(22, \frac{21}{16}\right)$.

Let $U = \bar{W}_A - \bar{W}_B$; then U is $N\left(3, \frac{49}{16}\right)$, and we require

$$P(U < 0) = \Phi\left(\frac{0-3}{\frac{7}{4}}\right) = \Phi\left(-\frac{12}{7}\right) = \Phi(-1.7143) = 0.0432.$$

Higher Certificate, Paper I, 2003. Question 4

$$f(x) = kx^2(1-x), \qquad 0 \le x \le 1$$

(i) $$\int_0^1 k\left(x^2 - x^3\right)dx = k\left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 = k\left(\frac{1}{3} - \frac{1}{4}\right) = \frac{k}{12},$$

which must be equal to 1. So $k = 12$.

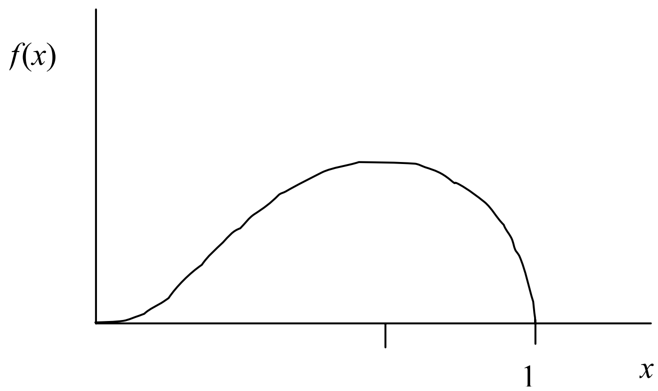
(ii) $$\frac{df(x)}{dx} = \frac{d}{dx}\left(12x^2 - 12x^3\right) = 24x - 36x^2 = 12x(2-3x)$$

which is zero for $2 - 3x = 0$ [and for $x = 0$, but this is clearly not the mode (i.e. not the maximum of $f(x)$)], i.e. $x = 2/3$. To check that this *is* the maximum (i.e. the mode), we can consider the second derivative:-

$$\frac{d^2 f(x)}{dx^2} = 24 - 72x, \text{ which is clearly} < 0 \text{ at } x = 2/3.$$

Hence the mode is at $x = 2/3$, and the graph of $f(x)$ is as shown. [***NOTE.*** *The curve should of course appear smooth; it might not do so, due to the limits of electronic reproduction.*]

[At the mode, $f(x) = 12(2/3)^2(1/3) = 16/9$.]



Continued on next page

(iii) $$E(X) = \int_0^1 xf(x)\,dx = \int_0^1 12\left(x^3 - x^4\right)dx = 12\left[\frac{x^4}{4} - \frac{x^5}{5}\right]_0^1$$

$$= 12\left(\frac{1}{4} - \frac{1}{5}\right) = \frac{12}{20} = \frac{3}{5}\ .$$

$$E\left(X^2\right) = \int_0^1 x^2 f(x)\,dx = \int_0^1 12\left(x^4 - x^5\right)dx = 12\left[\frac{x^5}{5} - \frac{x^6}{6}\right]_0^1$$

$$= 12\left(\frac{1}{5} - \frac{1}{6}\right) = \frac{12}{30} = \frac{2}{5}\ .$$

So $\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \dfrac{2}{5} - \dfrac{9}{25} = \dfrac{1}{25}$.

(iv) The cumulative distribution function is $F(x) = \int_0^x 12\left(u^2 - u^3\right)du$

$$= \left[12\left(\frac{u^3}{3} - \frac{u^4}{4}\right)\right]_0^x = 4x^3 - 3x^4 = x^3(4 - 3x), \quad \text{for } 0 \le x \le 1\ .$$

The mean is $\frac{3}{5}$ and the standard deviation is $\frac{1}{5}$. We require $P\left(\frac{2}{5} < X < \frac{4}{5}\right)$. This can be found by integrating the pdf between $\frac{2}{5}$ and $\frac{4}{5}$ or, directly, as

$$F\left(\frac{4}{5}\right) - F\left(\frac{2}{5}\right) = \left(\frac{4}{5}\right)^3\left(\frac{8}{5}\right) - \left(\frac{2}{5}\right)^3\left(\frac{14}{5}\right) = \frac{64\times 8\ -\ 8\times 14}{625} = \frac{16}{25}\ .$$

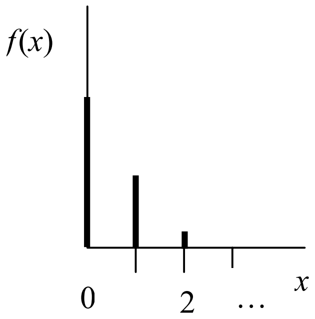
Higher Certificate, Paper I, 2003. Question 5

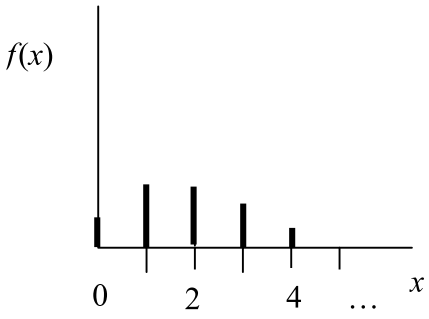
Poisson distribution: $f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$. Expectation = variance = λ.

(i) $\lambda = 0.5$: $f(0) = e^{-0.5} = 0.6065$, $f(1) = 0.3033$, $f(2) = 0.0758$, … . Expectation = variance = 0.5.

$\lambda = 2$: $f(0) = 0.1353$, $f(1) = 0.2707$, $f(2) = 0.2707$, $f(3) = 0.1804$, $f(4) = 0.0902$, … . Expectation = variance = 2.

Sketches are as shown.





(ii) Likelihood $L = \prod_{i=1}^{n} \dfrac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \dfrac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\Pi x_i!}$.

Taking logarithms to base e,

$$\log L = -n\lambda + (\Sigma x_i)\log\lambda - \log(\Pi x_i!) .$$

Differentiating, $\dfrac{d\log L}{d\lambda} = -n + \dfrac{\Sigma x_i}{\lambda}$; setting this equal to 0 gives the solution $\hat{\lambda}_{ML} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$. We have $\dfrac{d^2\log L}{d\lambda^2} = -\dfrac{\Sigma x_i}{\lambda^2} < 0$, confirming that this is a maximum.

The central limit theorem gives $\bar{X} \sim$ (approx) $\mathrm{N}(\lambda, \lambda/n)$, so we have (approximately)

$$P\left(\lambda - 1.96\sqrt{\frac{\lambda}{n}} \le \bar{X} \le \lambda + 1.96\sqrt{\frac{\lambda}{n}}\right) = 0.95$$

Continued on next page

or

$$P\left(\bar{X} - 1.96\sqrt{\frac{\lambda}{n}} \le \lambda \le \bar{X} + 1.96\sqrt{\frac{\lambda}{n}} \right) = 0.95 .$$

Hence, inserting the observed value $\bar{x}$ and, further, using $\hat{\lambda}_{ML} = \bar{x}$ as an estimate for the underlying variance, an approximate 95% confidence interval for λ is

$$\bar{x} - 1.96\sqrt{\frac{\bar{x}}{n}} \ , \ \bar{x} + 1.96\sqrt{\frac{\bar{x}}{n}} \ .$$

(iii) $n = 400, \quad \Sigma x_i = 2500; \quad \bar{x} = 6.25$.

So the approximate 95% confidence interval is

$$6.25 - 1.96\sqrt{\frac{6.25}{400}} \ , \ 6.25 + 1.96\sqrt{\frac{6.25}{400}}$$

i.e. 6.005 , 6.495 .

Now using $\Sigma x_i^2 = 25600$, we have that the sample variance s^2 is

$$s^2 = \frac{1}{399}\left(25600 - \frac{(2500)^2}{400} \right) = \frac{9975}{399} = 25.00 \ .$$

Using s^2 in the confidence interval gives the interval as

$$6.25 - 1.96\sqrt{\frac{25.00}{400}} \ , \ 6.25 + 1.96\sqrt{\frac{25.00}{400}}$$

i.e. 5.76 , 6.74 .

This interval is twice as wide – because s^2 is four times the size of $\bar{x}$ – which suggests that a Poisson assumption is <u>not</u> valid.

Higher Certificate, Paper I, 2003. Question 6

$E(Y) = np \qquad \text{Var}(Y) = np(1-p)$

(i) Binomial with $n = 48$, $p = ¼$.

(ii) Score is distributed $36 + \text{B}(12, ¼)$.

(a) Hence mean correct is $36 + (12/4) = 39$ and variance is $12 \times ¼ \times ¾ = 9/4$.

(b) Number wrong is distributed B(12, ¾).

(c) The required probability is $1 - P(0) - P(1) - P(2)$ based on the B(12, ¼) distribution. This is

$$1-\left(\frac{3}{4}\right)^{12}-12\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^{11}-\frac{12\times 11}{2}\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right)^{10}$$

$$= 1-0.031676-0.126705-0.232293 = 0.6093 .$$

(iii) Number of correct answers for A is distributed as $27 + \text{B}(21, ¼)$.
Number of correct answers for B is distributed as $28 + \text{B}(20, ¼)$.
Number of correct answers for C is distributed as $30 + \text{B}(18, ¼)$.

Means are $27 + (21/4) = 32¼$, $28 + (20/4) = 33$, $30 + (18/4) = 34½$ respectively.

Variances are $(21)(¼)(¾) = 63/16$, $(20)(¼)(¾) = 60/16 = 15/4$, $(18)(¼)(¾) = 54/16 = 27/8$ respectively.

So overall mean is $\frac{1}{3}(32.25+33+34.5) = 33.25$,

and variance of overall mean is $\frac{1}{9}\left(\frac{63}{16}+\frac{15}{4}+\frac{27}{8}\right) = 1.2292$.

Continued on next page

$$P(A|29)=\frac{P(29|A)P(A)}{\sum_{i=A,B,C} P(29|i)P(i)}; \qquad P(i)=\frac{1}{3} \text{ for } i=A,B,C \ .$$

$$P(29|A)=P\left[\mathrm{B}\left(21,\tfrac{1}{4}\right)=2\right]=\frac{21\times 20}{2}\left(\frac{3}{4}\right)^{19}\left(\frac{1}{4}\right)^{2}$$

$$P(29|B)=P\left[\mathrm{B}\left(20,\tfrac{1}{4}\right)=1\right]=20\left(\frac{3}{4}\right)^{19}\left(\frac{1}{4}\right)$$

$$P(29|C)=P\left[\mathrm{B}\left(18,\tfrac{1}{4}\right)=-1\right]=0$$

[Note. C must get at least the 30 he knows, so it must follow that $P(C|29)=0$, which is true if $P(29|C)=0$.]

So $$P(A|29) = \frac{\frac{1}{2}.21.20\left(\frac{3}{4}\right)^{19}\left(\frac{1}{4}\right)^{2}}{\frac{1}{2}.21.20\left(\frac{3}{4}\right)^{19}\left(\frac{1}{4}\right)^{2} + 20\left(\frac{3}{4}\right)^{19}\left(\frac{1}{4}\right)}$$

$$= \frac{\frac{21}{32}}{\frac{21}{32}+\frac{1}{4}} = \frac{21}{29} = 0.7241$$

and similarly

$$P(B|29) = \frac{\frac{1}{4}}{\frac{21}{32}+\frac{1}{4}} = \frac{8}{29} = 0.2759$$

(and $P(C|29) = 0$, see above).

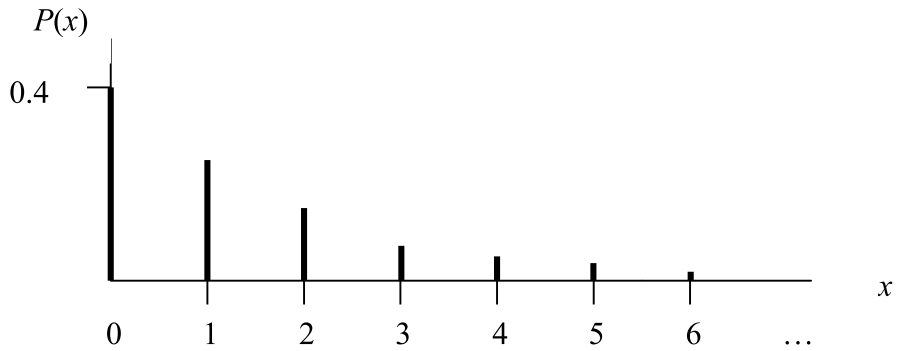
Higher Certificate, Paper I, 2003. Question 7

(i) $P(X = x) = (1 - p)(1 - p)\ldots(1 - p)p$, for $x = 0, 1, 2, \ldots$.

$F \quad F \quad\quad F \quad S$

----- x times -----

When $p = 0.4$, $P(0) = 0.4$, $P(1) = 0.24$, $P(2) = 0.144$, $P(3) = 0.0864$, $P(4) = 0.0518$, $P(5) = 0.0311$, $P(6) = 0.0187$, … .



(ii) The probability generating function of X is

$$G(s) = E\left(s^X\right) = \sum_{i=1}^{\infty} s^{x_i} p_i = \sum_{i=1}^{\infty} s^{x_i} p\left(1-p\right)^{x_i} = p\sum_{i=1}^{\infty} t^{x_i} \quad \text{where } t = (1-p)s$$

$$= p\left(1+t+t^2+t^3+\ldots\right) = p\left(1-t\right)^{-1}$$

$$= \frac{p}{1-\left(1-p\right)s} .$$

The mean is given by $G'(1)$ and the variance by $G''(1) + G'(1) - [G'(1)]^2$, where the differentiation is with respect to s.

$$G'(s) = \frac{p\left(1-p\right)}{\left\{1-\left(1-p\right)s\right\}^2} , \quad \text{so} \quad \text{mean } = G'(1) = \frac{p\left(1-p\right)}{p^2} = \frac{1-p}{p} .$$

$$G''(s) = \frac{2p\left(1-p\right)^2}{\left\{1-\left(1-p\right)s\right\}^3} , \quad \text{so} \quad G''(1) = \frac{2p\left(1-p\right)^2}{p^3} = \frac{2\left(1-p\right)^2}{p^2} .$$

Hence the variance is $\dfrac{2\left(1-p\right)^2}{p^2} + \dfrac{1-p}{p} - \dfrac{\left(1-p\right)^2}{p^2} = \dfrac{\left(1-p\right)^2}{p^2} - \dfrac{1-p}{p} = \dfrac{1-p}{p^2}$.

Continued on next page

(iii) We have $Y = X + 1$.

So $P(Y = y) = p(1-p)^{y-1}$, for $y = 1, 2, 3, \ldots$.

The probability generating function of Y can be obtained by a similar method to that used for X above, or it can be written down using the "linear transformation" result for probability generating functions:

Pgf of Y is $s^b G(as)$ with $a = 1$ and $b = 1$, i.e. $\dfrac{ps}{1-(1-p)s}$.

Mean of Y = (mean of X) + 1 = $1 + \dfrac{1-p}{p} = \dfrac{1}{p}$.

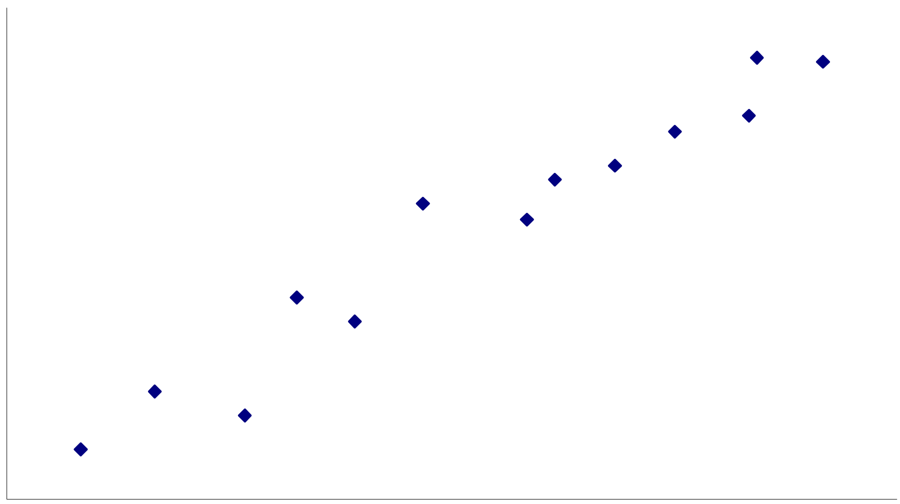
Variance of Y = variance of X.

Higher Certificate, Paper I, 2003. Question 8

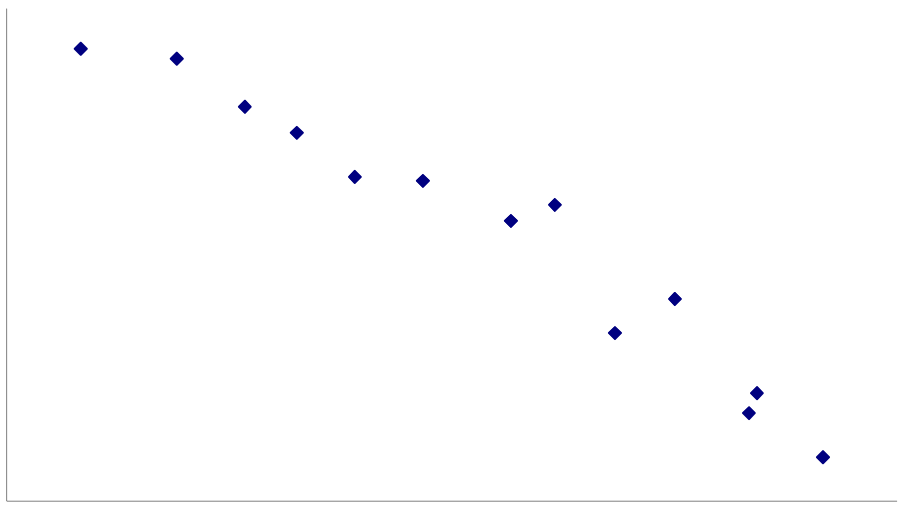
(i) $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

This explains the strength of linear relationship between the x_i and y_i, with $r = \pm 1$ showing linearity and $r = 0$ showing no linear relationship. The underlying X and Y are both random variables.

(a) r near to $+1$, small amount of scatter about an (increasing) linear relationship

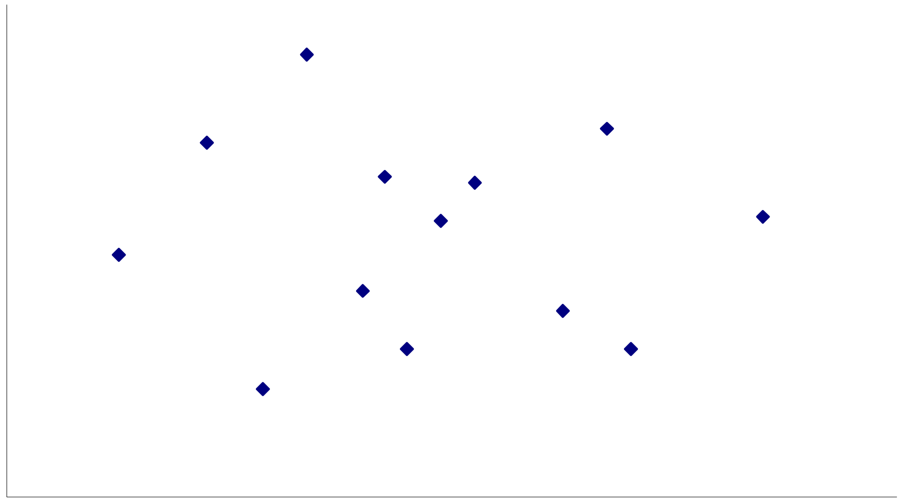


(b) r near to -1, y decreases as x increases, otherwise as in (a)

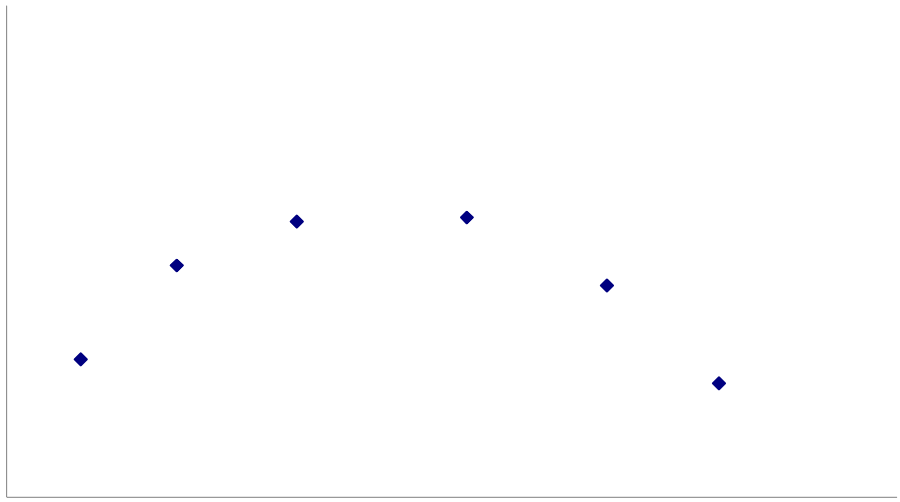


Continued on next page

(c) Independent data ($r \approx 0$)



(d) Non-linear relationship, e.g. $y = x^2$



(ii) (a) Simple linear regression of y = cholesterol on x = age. y is the dependent variable, x the independent. Assume a linear relationship underlying the data, $Y_i = a + bx_i + \varepsilon_i$, where the $\{\varepsilon_i\}$ are independent identically distributed $N(0, \sigma^2)$ random variables with σ^2 constant for all i.

(b) $r = \sqrt{(0.323)} = 0.568$ for 'chol' and 'age'.
$r = \sqrt{(0.940)} = 0.970$ for 'newchol' and 'newage'.

The latter consists of the 8 data points omitting the observation at $x = 27$ which seems very far from the roughly linear pattern of the rest. Omitting it has made a linear relationship seem much more plausible. Subject number 2 has very high cholesterol for his age.

Continued on next page

(c) Using the "constant" row in either set of output, the constant term is not significantly different from 0. A model omitting *a* could perhaps be used.

This would imply cholesterol 0 at age 0, which might not be very sensible – but we do not actually have data in that region, so we cannot claim that a linear relationship still holds.

(d) There is a tendency towards a curved relationship even when the very "unusual" observation at age 27 is omitted. The fit of a line without that observation is however much better than with it, and the diagnostic plots, of residuals and Normal probability, seem acceptable.