THE ROYAL STATISTICAL SOCIETY

2001 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, 2001. Question 1

(i) A and B are independent. So $P(\overline{A} \cap \overline{B}) = P(\overline{A})P(\overline{B})$.

$P(\overline{A}) = 1 - P(A) = \frac{1}{3}$; $P(\overline{B}) = \frac{1}{2}$; so $P(\overline{A} \cap \overline{B}) = \frac{1}{6}$.

(ii) $\overline{A} \cap \overline{B} = [(\overline{A} \cap \overline{B}) \cap C] \cup [(\overline{A} \cap \overline{B}) \cap \overline{C}]$,

with the two events $[(\overline{A} \cap \overline{B}) \cap C]$ and $[(\overline{A} \cap \overline{B}) \cap \overline{C}]$ being disjoint.

Hence $P(\overline{A} \cap \overline{B} \cap \overline{C}) = P(\overline{A} \cap \overline{B}) - P(\overline{A} \cap \overline{B} \cap C) = \frac{1}{6} - \frac{1}{10} = \frac{1}{15}$.

(iii) $B \cap C = (A \cap B \cap C) \cup (\overline{A} \cap B \cap C)$, these being disjoint.

Further, $(\overline{A} \cap B \cap C) \cup (\overline{A} \cap B \cap \overline{C}) = (\overline{A} \cap B)$.

Hence
$$\begin{aligned} P(B \cap C) &= P(A \cap B \cap C) + P(\overline{A} \cap B) - P(\overline{A} \cap B \cap \overline{C}) \\ &= \frac{1}{4} + \left(\frac{1}{3} \times \frac{1}{2}\right) - \left[P(\overline{A} \cap \overline{C}) - P(\overline{A} \cap \overline{B} \cap \overline{C})\right] \end{aligned}$$
since $\overline{A}, B$ are independent.

But A, C are also independent and so are $\overline{A}, \overline{C}$.

Therefore $P(B \cap C) = \frac{1}{4} + \frac{1}{6} - \left(\frac{1}{3} \times \frac{2}{5}\right) + \frac{1}{15} = \frac{5}{12} - \frac{1}{15} = \frac{63}{12 \times 15} = \frac{7}{20}$.

(iv) $P(B \mid C) = \dfrac{P(B \cap C)}{P(C)} = \dfrac{7/20}{3/5} = \dfrac{7}{12}$.

(v) $P(A \mid B \cap C) = \dfrac{P(A \cap B \cap C)}{P(B \cap C)} = \dfrac{1/4}{7/20} = \dfrac{5}{7}$.

(vi) $P(A \cap B \mid A \cap C) = \dfrac{P(A \cap B \cap C)}{P(A \cap C)} = \dfrac{\frac{1}{4}}{\frac{2}{3} \times \frac{3}{5}} = \dfrac{5}{8}$

(A, C are independent).

<u>Higher Certificate, Paper I, 2001. Question 2</u>

(a) Fix the position of M_1 (suppose him to be the host).

	M_1 (1)		Label the positions clockwise.
(6)		(2)	M_2, M_3, W_1, W_2, W_3 may be arranged in $5! = 120$ ways.
(5)		(3)	
	(4)		

(i) M_2, M_3 must occupy (3) and (5); W_1, W_2, W_3 may occupy the other places in 3! ways, making $2\times 3!$ arrangements. The probability is then $\frac{12}{120}=\frac{1}{10}$.

(ii) M_2, M_3 must occupy (2) and (6) <u>or</u> (2) and (3) <u>or</u> (5) and (6). In each case, M_2, M_3 can be placed in two orders, making 6 positions altogether for the three men. The women may again fill the remaining places in 3! ways.

The probability is $\frac{6\times 6}{120}=\frac{3}{10}$.

(iii) EITHER $1-\frac{1}{10}-\frac{3}{10}=\frac{3}{5}$, because this is the only other arrangement possible besides (i) and (ii);

OR by having M_2 in (2), M_3 in (4) or (5); M_2 in (6), M_3 in (3) or (4); M_2 in (3), M_3 in (4); M_2 in (4), M_3 in (5); or any of these with M_2, M_3 interchanged, giving 12 positionings of the men. There are again 3! orders for the women, so the probability is $\frac{6\times 12}{120}=\frac{3}{5}$.

(b) (i) Event D is "has disease", T is "tests positive".

$$P(D\,|\,T)=\frac{P(D\cap T)}{P(T)}=\frac{P(T\,|\,D)P(D)}{P(T)}=\frac{P(T\,|\,D)P(D)}{P(T\,|\,D)P(D)+P(T\,|\,\bar{D})P(\bar{D})}$$
$$=\frac{p_1p_0}{p_1p_0+(1-p_2)(1-p_0)}.$$

(ii) $$\frac{0.95\times 0.005}{(0.95\times 0.005)+(0.05\times 0.995)}=\frac{0.00475}{0.0545}=0.0872.$$

The error rates in the clinical tests are large compared to the chance of having the disease, so the calculated probability is very small.

Higher Certificate, Paper I, 2001. Question 3

(a) (i) $P(S \geq 2300) = 1 - P(S < 2300) = 1 - \Phi\left(\dfrac{2300-2000}{300}\right)$

$= 1 - \Phi(1) = 1 - 0.8413 = 0.1587$.

$P(H \geq 2300) = 1 - \Phi\left(\dfrac{2300-2500}{125}\right) = 1 - \Phi(-1.6)$

$= 1 - 0.0548 = 0.9452$.

(ii) $P(S > H) = P(S - H > 0)$, where $(S - H)$ is N(–500, 90000+15625) i.e. N(–500, 325^2). Hence $P(S > H) = \Phi\left(\dfrac{-500}{325}\right) = \Phi(-1.5385) = 0.0620$.

(b) (i) The lifetime X is S with probability 0.6 and H with probability 0.4.

Hence $E[X] = 0.6 \times 2000 + 0.4 \times 2500 = 2200$ hrs.

$$P(X > 2600) = P(X > 2600 \mid S)P(S) + P(X > 2600 \mid H)P(H)$$

$$= \Phi\left(-\frac{600}{300}\right) \times 0.6 + \Phi\left(-\frac{100}{125}\right) \times 0.4$$

(using the appropriate tail areas from Normal tables)

$$= 0.6\,\Phi(-2) + 0.4\,\Phi(-0.8) = 0.6 \times 0.02275 + 0.4 \times 0.2119$$

$$= 0.01365 + 0.08476 = 0.09841.$$

(ii) $P(H \mid \; > 2600) = \dfrac{P(X > 2600 \mid H)P(H)}{P(X > 2600)} = \dfrac{0.08476}{0.09841} = 0.8613$.

(iii) $\bar{X}$ will have mean 2200. It is not Normally distributed but we way apply the Central Limit Theorem if we know its variance. In large samples we may take $\bar{X}$ as approximately N(2200, 346.8^2), so that

$$P(\bar{X} > 2300) = 1 - \Phi\left(\frac{2300-2200}{346.8/\sqrt{100}}\right) = 1 - \Phi\left(\frac{100}{34.68}\right) = \Phi(-2.8835) = 0.002.$$

<u>Higher Certificate, Paper I, 2001. Question 4</u>

An easy method is to consider X as $\sum X_i$, where X_i are a set of n Bernoulli variables with $P(X_i = 1) = p, P(X_i = 0) = (1 - p)$. Then $E[X_i] = p,$ so $E[X] = np$.

Also $E\left[X_i^2\right] = p,$ so $\operatorname{Var}(X_i) = p - p^2$ and $\operatorname{Var}(X) = n\left(p - p^2\right) = npq.$

ALTERNATIVELY: $$E[X] = \sum_{x=0}^{n} x\binom{n}{x}p^x q^{n-x} = \sum_{x=1}^{n} \frac{n!\,p^x q^{n-x}}{(x-1)!(n-x)!}$$

$$= np\sum_{x=1}^{n}\binom{n-1}{x-1}p^{x-1}q^{(n-1)-(x-1)} = np.$$

Similarly, $\operatorname{Var}(\mathrm{X}) = E\left[X(X-1)\right] + E[X] - \left(E[X]\right)^2$, and we have

$$E\left[X(X-1)\right] = \sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x q^{n-x} = \sum_{x=2}^{n} x(x-1)\binom{n}{x}p^x q^{n-x}$$

$$= n(n-1)p^2\sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}q^{(n-2)-(x-2)} = n(n-1)p^2,$$

and hence $\operatorname{Var}(\mathrm{X}) = n(n-1)p^2 + np - n^2p^2 = np - np^2 = npq$.

PGFs or MGFs could also be used.

(i) (a) $0.75^{16} \approx 0.0100226 = 0.0100$ approx.

(b) $1 - P$(no one gets all 16 right), probability is $1 - \left\{1 - 0.75^{16}\right\}^{12}$
$= 1 - \{0.9899774\}^{12} = 0.1139$.

(ii) $P(B - A > 0)$ can be studied using a Normal approximation to the difference $B - A$, i.e. $\mathrm{N}\left(16\{0.5 - 0.75\}, 16\{(0.5\times0.5) + (0.75\times0.25)\}\right)$, i.e. $\mathrm{N}(-4, 7)$.

The probability is found as $P\left(B - A > \frac{1}{2}\right)$ using a continuity correction since $B - A$ takes discrete values.

Hence it is $1-\Phi\left(\dfrac{0.5-(-4)}{\sqrt{7}}\right)=\Phi\left(-\dfrac{4.5}{\sqrt{7}}\right)=\Phi(-1.7008)\approx 0.0445$.

[Note: this would be 0.0653 without the continuity correction.]

(iii) $E\left[\bar{X}\right]=E\left[X\right]=np=16\times 0.75=12$ in set A.

Similarly, $E\left[\bar{Y}\right]=16\times 0.5=8$ in set B.

There are 12 students in A and 25 in B, so that

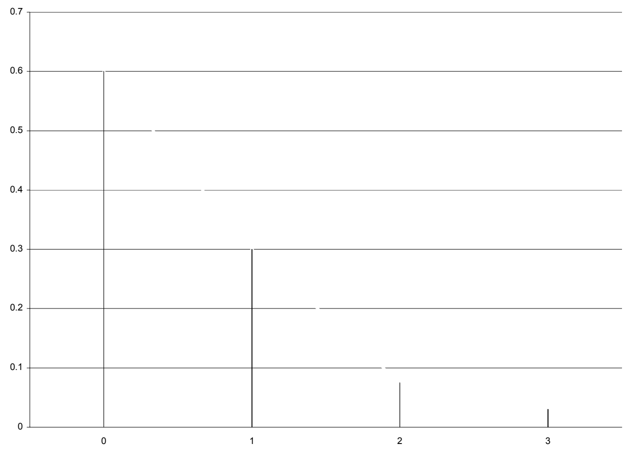
$$\operatorname{Var}\left(\bar{X}\right)=\frac{16\times 0.75\times 0.25}{12}=\frac{1}{4} \quad \text{in set } A$$

$$\operatorname{Var}\left(\bar{Y}\right)=\frac{16\times 0.5\times 0.5}{25}=\frac{4}{25} \quad \text{in set } B.$$

Higher Certificate, Paper I, 2001. Question 5

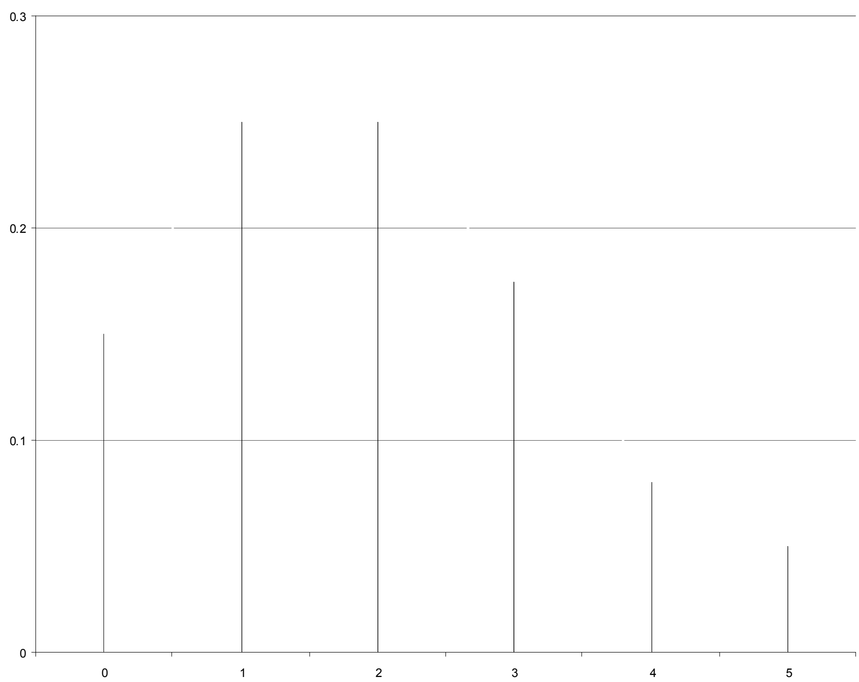
$$\frac{p(x+1)}{p(x)} = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}.\frac{x!}{e^{-\lambda}\lambda^{x}} = \frac{\lambda}{x+1} \quad \text{for } x = 0, 1, 2, \dots .$$

For $\lambda = \frac{1}{2}$, $p(0) = 0.60653$; so $p(1) = 0.30327$, $p(2) = 0.07582$, $p(3) = 0.01264$.



Graph of $p(x)$ for λ = ½.

For $\lambda = 2$, $p(0) = 0.13534$;
so $p(1) = 0.27067 = p(2)$, $p(3) = 0.18045$, $p(4) = 0.09022$, $p(5) = 0.03609$.



Graph of $p(x)$ for λ = 2.

$$M_X(t) = E\left[e^{Xt}\right] = \sum_{x=0}^{\infty} \frac{e^{xt}e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda}.e^{\lambda e^t} = \exp\left(\lambda\left\{e^t - 1\right\}\right).$$

$\dfrac{\partial M}{\partial t} = \lambda e^t e^{\lambda\left(e^t-1\right)}$; put $t = 0$ and this is λ, which is therefore $E[X]$.

$\dfrac{\partial^2 M}{\partial t^2} = \left(\lambda^2 e^{2t} + \lambda e^t\right)\left(e^{\lambda\left(e^t-1\right)}\right)$; put $t = 0$ and this is $\lambda^2+\lambda$, but it is also $E\left[X^2\right]$.

Hence $\quad \mathrm{Var}(\mathrm{X}) = E\left[X^2\right] - \left(E[X]\right)^2 = \lambda^2 + \lambda - (\lambda)^2 = \lambda$.

$\dfrac{\partial^3 M}{\partial t^3} = \left(\lambda^3 e^{3t} + 3\lambda^2 e^{2t} + \lambda e^t\right)\left(e^{\lambda\left(e^t-1\right)}\right) = \lambda^3 + 3\lambda^2 + \lambda$ at $t = 0$. This is $E\left[X^3\right]$.

Now, $E\left[(X-\lambda)^3\right] = E\left[X^3\right] - 3\lambda E\left[X^2\right] + 3\lambda^2 E[X] - \lambda^3$

$$= \lambda^3 + 3\lambda^2 + \lambda - 3\lambda\left(\lambda^2 + \lambda\right) + 3\lambda^2.\lambda - \lambda^3 = \lambda.$$

For $Y = X_1 + X_2 + \ldots + X_n$ we have

$$E\left[e^{Yt}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{X_i t}\right] = \left(M_X(t)\right)^n = \exp\left[\lambda n\left(e^t - 1\right)\right].$$

This is the mgf of a Poisson distribution with parameter λn and so $E[Y]$ = Var(Y) = λn.

$$P(Y \geq 40) = 1 - P(Y \leq 39) \approx 1 - \Phi\left(\frac{39.5-25}{5}\right),$$

using continuity correction, and $\mu = \lambda n = 25$. This is $1 - \Phi(2.9) = 0.00187$.

With a positively skew distribution, the Normal approximation is likely to underestimate the probability in the right hand tail and so we expect this answer to be <u>less</u> than the true value.

Higher Certificate, Paper I, 2001. Question 6

$$L(p)=\prod_{i=1}^{n}\left(q^{x_i-1}p\right)=p^n q^{\sum x_i-n}=\left(\frac{p}{1-p}\right)^n(1-p)^{n\bar{x}}$$

$$\ln L=n\ln p-n\ln(1-p)+n\bar{x}\ln(1-p)$$

$\dfrac{\partial}{\partial p}(\ln L)=\dfrac{n}{p}+\dfrac{n}{1-p}-\dfrac{n\bar{x}}{1-p}=0$ when $\dfrac{1}{\hat{p}}=\dfrac{-1+\bar{x}}{1-\hat{p}}$ which gives $\hat{p}=\dfrac{1}{\bar{x}}$ as m.l.estimate.

$\dfrac{\partial^2(\ln L)}{\partial p^2}=-\dfrac{n}{p^2}-\dfrac{n(\bar{x}-1)}{(1-p)^2}$ which is < 0, confirming the maximum.

$$E\left[\frac{\partial^2\ln L}{\partial p^2}\right]=\frac{n}{p^2}+\frac{n}{(1-p)^2}\left[E(\bar{X})-1\right]=\frac{n}{p^2}+\frac{n}{(1-p)^2}\left(\frac{1}{p}-1\right)=\frac{n}{p^2}+\frac{n}{p(1-p)}$$

$$=\frac{n}{p^2(1-p)}$$

Hence $\operatorname{Var}(\hat{p})\approx\dfrac{p^2(1-p)}{n}$.

$\sum fx=448,\ \sum f=56,\ \hat{p}=\dfrac{56}{448}=0.125$.

$\operatorname{Var}(\hat{p})=\dfrac{0.125^2\times0.875}{56}=0.0002441,\quad \mathrm{SE}(\hat{p})=0.015625$.

Approximate 95% confidence interval for p is $\hat{p}\pm1.96\,\mathrm{SE}(\hat{p})$, which is

$0.125\pm1.96\times0.015625$, i.e. 0.125 ± 0.030625 or (0.0944, 0.1556).

When $p=\dfrac{1}{6}$, we have $\dfrac{1}{\bar{X}}\sim \mathrm{N}\left(\dfrac{1}{6},\dfrac{5}{216\times56}\right)$, i.e. N(0.1667, 0.00041336); therefore

the probability of obtaining $\hat{p}\le0.125$ is approximately

$$\Phi\left(\frac{0.125-0.1667}{0.020331}\right)=\Phi(-2.0496)=-0.0202.$$

The confidence interval for p did not include $\dfrac{1}{6}$; also now the probability being very small is consistent with rejecting a null hypothesis that $p=\dfrac{1}{6}$, i.e. that the die is fair.

<u>Higher Certificate, Paper I, 2001. Question 7</u>

(a) $P(X_i \le x) = F(x)$ for $i = 1, 2, ..., n$.

(i) $$F_{X_{\max}}(x) = P(X_1 \le x,\ X_2 \le x,\ ...,\ X_n \le x) = \prod_{i=1}^{n} P(X_i \le x)$$
$$= \left[F(x)\right]^n$$

(ii) For $X_{\min} \ge x$, we require <u>every</u> X_i to be $\ge x$.

Now, $P(X_i \le x) = F(x)$, so $P(X_i \ge x) = 1 - F(x)$, for all i.

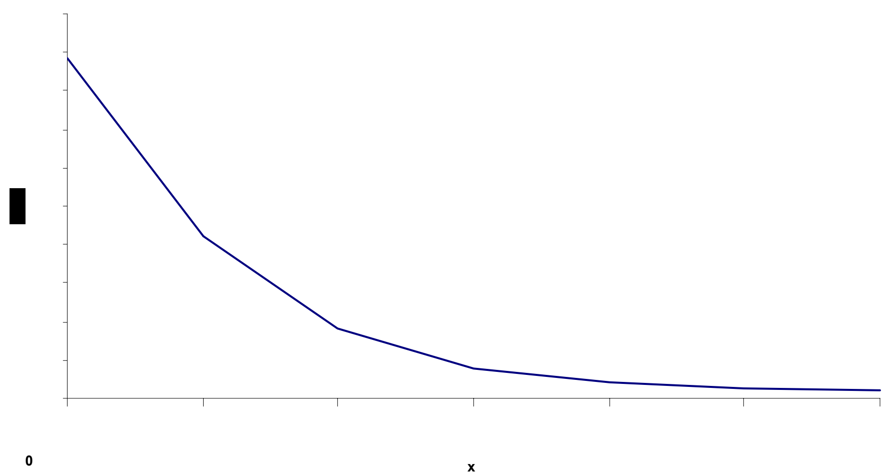
$$\therefore F_{X_{\min}}(x) = P(X_{\min} \le x) = 1 - P(X_{\min} \ge x) = 1 - P(X_1 \ge x, ..., X_n \ge x)$$
$$= 1 - \left[1 - F(x)\right]^n$$

(iii) Pdfs are derivatives of cdfs:

$$f_{X_{\max}} = n\left[F(x)\right]^{n-1} f(x)$$

$$f_{X_{\min}} = n\left[1 - F(x)\right]^{n-1} f(x), \qquad \text{since } \frac{\partial F(x)}{\partial x} = f(x).$$

(b)



$$F(x) = \int_0^x \frac{\alpha}{(1+u)^{\alpha+1}}\,du = \left[-\frac{1}{(1+u)^{\alpha}} \right]_0^x = 1 - \frac{1}{(1+x)^{\alpha}}.$$

Median M is such that $F(M) = \frac{1}{2}$. So we have

$$1 - \frac{1}{(1+M)^{\alpha}} = \frac{1}{2} \quad \text{or} \quad \frac{1}{(1+M)^{\alpha}} = \frac{1}{2} \quad \text{or } 2 = (1+M)^{\alpha} \quad \text{or} \quad M = 2^{(1/\alpha)} - 1.$$

Using (a)(ii), $F_{X_{\min}} = 1 - \left(\frac{1}{(1+x)^{\alpha}} \right)^n = 1 - \frac{1}{(1+x)^{n\alpha}}$, also Pareto but with α replaced by $n\alpha$.

The median of $X_{\min}$ is then $2^{\frac{1}{n\alpha}} - 1$, which is $2^{1/n} - 1$ if $\alpha = 1$. We require $2^{1/n} - 1 < 0.1$ or $2^{1/n} < 1.1$, i.e. $\frac{1}{n} \ln 2 < \ln 1.1$, giving $n > \frac{\ln 2}{\ln 1.1} = \frac{0.6931}{0.0953} = 7.27$. Hence $n \geq 8$.

<u>Higher Certificate, Paper I, 2001. Question 8</u>

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where y is the response (observation) and x the value of the explanatory variable on that unit; $\{\varepsilon_i\}$ is a set of independent, identically distributed random variables with mean 0 and the same variance σ^2. Usually they are assumed Normal as a basis for inference. x_i is assumed "fixed", not "random".

(i) (a) There is an increasing trend, and the relationship between y and x appears curvilinear.

(b) In simple regression R^2 is the square of the correlation, r, between x and y. In general it is the proportion of the variance of y which can be explained by the dependence of y on all explanatory variables $\{x_i\}$ in the model; hence it is the square of the correlation between $\hat{y}$ and y.

In the ANOVA , $\dfrac{\text{regression SS}}{\text{total SS}} = \dfrac{269.33}{335.37} = 0.803, \text{ or } 80.3\%$.

(c) As x (% operating capacity) increases by 1 unit so y (profit) increases by 0.31562 units.

A 95% confidence interval is $0.31562 \pm t_{10} \times 0.04942$, which is 0.31562 ± 0.11011 or $(0.2055, 0.4257)$.

(d) Values of profit have been predicted for capacity 25%, 50% and 75%. The confidence intervals for these predictions are those given; but note that 25% is far outside the range of available data (hence the remark about extreme x values). A 95% confidence interval is an interval which should cover the <u>true</u> y at a given x with probability 0.95, based on the fitted linear regression.

(ii) (a) The logarithmic plot shows that a linear regression in these units is a much better fit. There is still an increasing trend.

(b) $\log_{10}(\text{profits}) = -0.519 + 0.0177(\text{capacity})$

i.e. $\text{profits} = 10^{-0.519+0.0177(\text{capacity})}$.

(c) For capacity = 25, the 95% limits are –0.2731 and +0.1210, and the actual prediction is –0.0760, in $\log_{10}$ units. Anti-logging these (i.e. raising 10 to these powers) we find the 95% limits are 0.5332 and 1.3213.

The 'prediction' is 0.8395.

These limits do not overlap the limits on the previous model.

The prediction now is for a small profit, compared with a loss on the previous model.

(iii) The scatter plots indicate that the logarithmic model is preferred, and so do the plots of residuals which show a random pattern (as compared with a systematic, curved one for the previous model). R^2 also higher (92.4%) on the log model.

But a log model cannot predict negative profits – i.e. losses – which are quite possible in general though not for these data if used within the range of x values given.

Extrapolation down to 25% is well outside the data and so is not reliable on any model.