THE ROYAL STATISTICAL SOCIETY

2004 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.

<u>Higher Certificate, Paper I, 2004. Question 1</u>

(i) (a) P(all four favour the complex) = $(0.6)^4$.
P(all four oppose the complex) = $(0.3)^4$.
P(all four are indifferent) = $(0.1)^4$.

So P(all four think alike) = $(0.6)^4 + (0.3)^4 + (0.1)^4 = 0.1378$.

(b) P(an individual is not opposed) = $0.6 + 0.1 = 0.7$.

So P(none of the four is opposed) = $(0.7)^4 = 0.2401$.

(c) Possible favourable results are *FFOI*, *FOOI*, *FOII*, in any order.

$P(FFOI) = (0.6)^2(0.3)(0.1) = 0.0108$
$P(FOOI) = (0.6)(0.3)^2(0.1) = 0.0054$
$P(FOII) = (0.6)(0.3)(0.1)^2 = 0.0018$

Each result can be arranged in $\dfrac{4!}{2!1!1!} = 12$ ways.

So overall probability is $12(0.0108 + 0.0054 + 0.0018) = 0.216$.

(d) From (a), P(all four in favour) = $(0.6)^4 = 0.1296$. From (b), P(none opposed) = 0.2401. So the required conditional probability is

$$0.1296/0.2401 = 0.5398.$$

(ii) The number in favour is binomially distributed with $n = 4$ and $p = 0.6$. So the expectation (mean) is $4 \times 0.6 = 2.4$ and the variance is $4 \times 0.6 \times 0.4 = 0.96$.

(iii) $P(\text{opposed}) = P(\text{opposed} \mid \text{young})P(\text{young}) + P(\text{opposed} \mid \text{older})P(\text{older})$

$$= (0.12 \times 0.25) + (p \times 0.75)$$

where $p = P(\text{opposed} \mid \text{older})$. But we are given that P(opposed) = 0.3. Hence $p = 0.36$.

(iv) In samples of one "young" and three "olders",

P(exactly one opposes) = P("young" opposes, "olders" do not)

+ P("young" does not oppose, <u>one</u> "older" opposes)

$= \{(0.12)(0.64)^3\} + \{3(0.88)(0.36)(0.64)^2\} = 0.03146 + 0.38928 = 0.4207.$

Higher Certificate, Paper I, 2004. Question 2

(i) (a) The moment generating function is

$$M_X(t) = E\left(e^{tX}\right) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty} e^{-\lambda} \frac{\left(\lambda e^t\right)^x}{x!} = \exp\left(\lambda e^t - \lambda\right) = \exp\left(\lambda\left(e^t - 1\right)\right)$$

(b) $E(X) = M_X'(0) = \left[\lambda e^t e^{\lambda\left(e^t-1\right)}\right]_{t=0} = \lambda$.

$$E\left(X^2\right) = M_X''(0) = \left[\frac{d}{dt}\left(\lambda e^t e^{\lambda\left(e^t-1\right)}\right)\right]_{t=0}$$

$$= \left[\lambda e^t . e^{\lambda\left(e^t-1\right)} \lambda e^t + e^{\lambda\left(e^t-1\right)} . \lambda e^t\right]_{t=0} = \lambda^2 + \lambda .$$

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

(Alternatively, the moments could be obtained from the power series expansion of $M_X(t)$.)

(Alternatively, though with comparatively lengthy algebra, the moments could be obtained directly by $E(X) = \Sigma x P(X = x)$ and $E(X^2) = \Sigma x^2 P(X = x)$; or, somewhat easier, use $E[X(X - 1)] = \Sigma x(x - 1)P(X = x)$ (this is λ^2) and then $\operatorname{Var}(X) = E[X(X - 1)] + E(X) - \{E(X)\}^2$.)

(c) The binomial distribution with parameters n and p may be approximated by the Poisson distribution with parameter np if n is large and p is small. As a "rule of thumb", $\frac{1}{2} \le np \le 10$ gives an indication of how large n should be and how small p should be. (If $np > 10$, a Normal approximation to the binomial may be better.)

(ii) Let X = number of wrong calculations. We have $X \sim B(200, 0.0075)$.

$$P(X = 1) = \binom{200}{1}(0.0075)(0.9925)^{199} = 200 \times 0.0075 \times 0.9925^{199} = 0.3353(2).$$

$$P(X = 4) = \binom{200}{4}(0.0075)^4 (0.9925)^{196}$$

$$= \frac{200 \times 199 \times 198 \times 197}{4 \times 3 \times 2 \times 1} \times 0.0075^4 \times 0.9925^{196} = 0.0468(0).$$

Continued on next page

(iii) We approximate using $X \sim$ Poisson($200 \times 0.0075 = 1.5$). With this,

$$P(X=1)=1.5e^{-1.5}=0.3347(0)$$

giving a percentage error of $\dfrac{100(0.33532-0.33470)}{0.33532}=0.18\%$, and

$$P(X=4)=\frac{e^{-1.5}(1.5)^4}{4!}=0.0470(7)$$

giving a percentage error of $\dfrac{100(0.04707-0.04680)}{0.04680}=0.58\%$.

[Note. These percentage errors might come out *slightly* differently if more accuracy is kept in the binomial and Poisson probabilities.]

Both approximations are remarkably accurate, with percentage errors well below 1%. The approximation for $X = 1$ (one wrong calculation) is the more accurate of the two. That approximation is an underestimate; the other is an overestimate.

<u>Higher Certificate, Paper I, 2004. Question 3</u>

Actual volume $X \sim$ N(1010, 8^2). Let $Z \sim$ N(0,1).

(i) $P(X < 1000) = P\left(Z < \frac{1000-1010}{8}\right) = P(Z < -1.25) = 0.1056$.

(ii) Let Y be the total volume in a 6-pack.

We have $Y \sim$ N(6×1010, 64 + 64 + 64 + 64 + 64 + 64), i.e. $Y \sim$ N(6060, 384).

$P(Y < 6000) = P\left(Z < \frac{6000-6060}{\sqrt{384}}\right) = P(Z < -3.06) = 0.0011$.

(Alternatively, could use $\bar{X} \sim$ N(1010, 64/6) and calculate $P(\bar{X} < 1000)$.)

This probability is considerably smaller than that in part (i). In practical terms, this is because there will be a tendency for heavier and lighter cartons in a 6-pack to balance each other out. Alternatively, in terms of probability distributions, consider X and $\bar{X}$: $\bar{X}$ has the same mean as X but only one-sixth of the variance, so less of the lower tail of the distribution of $\bar{X}$ is below the nominal volume of 1000.

(iii) The new volume $W \sim$ N(μ, 4^2), where μ is the new mean. So we have that $P(W < 1000) = P\left(Z < \frac{1000-\mu}{4}\right)$. We require that this probability must be no greater than 0.1056. Thus the cut-off point for Z is to be z = –1.25 (as before). Hence $\frac{1000-\mu}{4} = -1.25$, giving μ = 1005.

This means that 5 ml per carton could be saved, i.e. a cost saving per carton of $\frac{5}{1000} \times £1$. To recover the £200, the number of cartons required is $\frac{200}{5/1000}$ = 40000.

Higher Certificate, Paper I, 2004. Question 4

(i) The binomial distribution with parameters n, p can be approximated by $N(np, np(1-p))$ when n is large and p is not too near to 0 or 1. As a "rule of thumb", the approximation is likely to be good if both np and $np(1-p)$ are > 10.

Let $X \sim B(n, p)$ and let Φ denote the c.d.f. of $N(0, 1)$. Using a continuity correction, $P(X \le x) \approx \Phi\left(\dfrac{x+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$ and $P(X < x) \approx \Phi\left(\dfrac{x-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$.

The 95% confidence interval for p uses the estimated variance $\hat{p}(1-\hat{p})/n$, giving the approximate interval

$$\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The estimate of p is $\hat{p}=\dfrac{30}{50}=0.6$, so $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\dfrac{(0.6)(0.4)}{50}} = 0.0693$. Thus the approximate interval is

$$0.6-(1.96\times 0.0693)\ ,\ 0.6+(1.96\times 0.0693)$$

i.e. (0.464, 0.736).

(ii) P(neither hits) $= (1-p)^2$. Therefore P(at least 1 hit) $= 1-(1-p)^2 = p(2-p)$. We estimate this by $(0.6)(2-0.6) = 0.84$.

(iii) When $p = 0.464$ (lower limit of interval in part (i)), we have $p(2-p) = 0.713$. Similarly, when $p = 0.736$, we have $p(2-p) = 0.930$. Thus (0.713, 0.930) is the required interval.

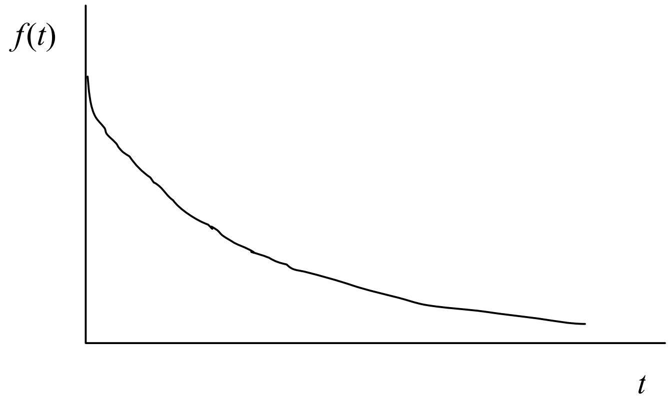
(iv) When n pairs are fired, P(all miss) $= [(1-p)^2]^n$, estimated by $(0.16)^n$. Hence $(0.16)^n < 0.0005$ is required. Solving this by taking logarithms to base 10, we have $n\log_{10}(0.16) < \log_{10}(0.0005)$, i.e. $-0.79588n < -3.30103$ which gives $n > 4.148$. so n must be at least 5.

Higher Certificate, Paper I, 2004. Question 5

(i) $$f(t) = \lambda e^{-\lambda t}, \qquad t > 0; \quad \lambda > 0$$

(a) Sketch of $f(t)$.

[***NOTE.*** *The curve should of course appear as a smooth decaying exponential; it might not do so, due to the limits of electronic reproduction.*]



(b) C.d.f. is $F(t) = P(T \le t) = \int_0^t \lambda e^{-\lambda v} dv = \lambda \left[-\frac{1}{\lambda} e^{-\lambda v} \right]_0^t = 1 - e^{-\lambda t}$.

(c) $P(a < T \le b) = F(b) - F(a) = e^{-\lambda a} - e^{\lambda b}$.

(ii) Assume all settlements of invoices are independent.

P(50 in first week) = $\{F(1)\}^{50} = \left(1 - e^{-\lambda}\right)^{50}$, because $T \le 1$ for all these 50.

Likewise, $1 < T \le 2$ for the 35 in the second week, so we have P(35 in second week) = $\{F(2) - F(1)\}^{35}$ = $\left(e^{-\lambda} - e^{-2\lambda}\right)^{35}$.

The remaining 15 have $T > 2$, which has probability $1 - P(T \le 2) = e^{-2\lambda}$, and thus P(15 after week 2) = $\left(e^{-2\lambda}\right)^{15}$.

The likelihood is therefore the product

$$L(\lambda) = k\left(1 - e^{-\lambda}\right)^{50} \left(e^{-\lambda} - e^{-2\lambda}\right)^{35} \left(e^{-2\lambda}\right)^{15}$$

where k is a constant of proportionality.

Continued on next page

Taking logarithms (base *e*),

$$\log L(\lambda) = \log k + 50\log\left(1-e^{-\lambda}\right) + 35\log\left\{e^{-\lambda}\left(1-e^{-\lambda}\right)\right\} + 15\log\left(e^{-2\lambda}\right)$$

$$= \log k + 85\log\left(1-e^{-\lambda}\right) - (35+30)\lambda \;=\; \log k + 85\log\left(1-e^{-\lambda}\right) - 65\lambda\,.$$

$$\therefore \frac{d}{d\lambda}\log L \;=\; \frac{85e^{-\lambda}}{1-e^{-\lambda}} - 65 \;=\; \frac{85}{e^{\lambda}-1} - 65\,.$$

Equating to zero, $85 = 65\left(e^{\lambda}-1\right)$ or $e^{\lambda} = 150/65$, so that $\hat{\lambda} = \log(150/65) = 0.836$.

[It is easy to check that this is indeed a maximum; e.g. $\dfrac{d^2}{d\lambda^2}\log L = -\dfrac{85}{\left(e^{\lambda}-1\right)^2} \;<\; 0$.]

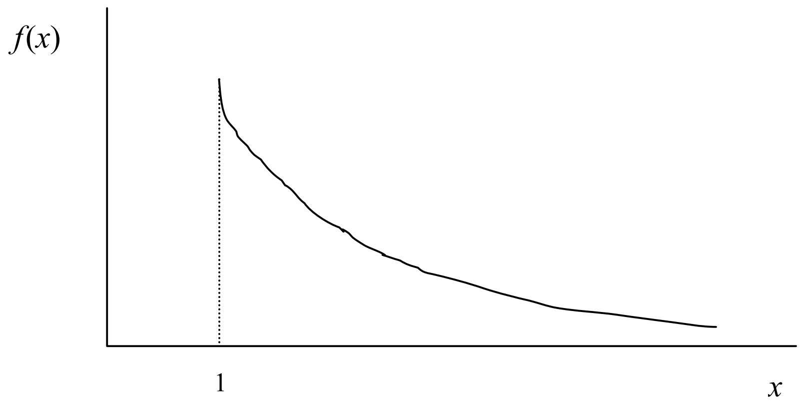
(iii) $1-e^{-0.836} = 0.5666$; $e^{-0.836} - e^{-1.672} = 0.43344 - 0.18787 = 0.2456$. Hence, out of 100 invoices, 56.66, 24.56 and 18.78 would be expected to be paid, on this model, in weeks 1, 2 and later. The actual numbers were 50, 35 and 15. The prediction for the second week is a long way from what happened, balanced by smaller discrepancies in the other two periods. This does not seem very satisfactory.

<u>Higher Certificate, Paper I, 2004. Question 6</u>

$$f(x) = \frac{k}{x^{k+1}}, \qquad x \geq 1; \quad k > 0$$

(i) Sketch of $f(x)$.

[***NOTE.*** *The curve should of course appear as a smooth curve; it might not do so, due to the limits of electronic reproduction.*]



C.d.f. is $F(x) = P(X \leq x) = \int_1^x \frac{k}{u^{k+1}}\,du = \left[-\frac{1}{u^k}\right]_1^x = 1 - \frac{1}{x^k}$ (for $x \geq 1$).

(ii) Median M has $\frac{1}{2} = F(m) = 1 - M^{-k}$, so $\frac{1}{2} = M^{-k}$ and hence $M = 2^{1/k}$.

Lower quartile Q_1 has $\frac{1}{4} = F(Q_1) = 1 - Q_1^{-k}$, so $\frac{3}{4} = Q_1^{-k}$, i.e. $Q_1 = (4/3)^{1/k}$.

Upper quartile Q_3 has $\frac{3}{4} = F(Q_3) = 1 - Q_3^{-k}$, so $Q_3 = (4)^{1/k}$.

Hence the semi-interquartile range is $\frac{1}{2}\left\{4^{1/k} - \left(\frac{4}{3}\right)^{1/k}\right\}$.

Continued on next page

(iii) $E(X)=\int_1^\infty xf(x)\,dx=\int_1^\infty \frac{k}{x^k}\,dx=\left[\frac{-k}{(k-1)x^{k-1}}\right]_1^\infty=\frac{k}{k-1}.$

$$E(X^2)=\int_1^\infty x^2 f(x)\,dx=\int_1^\infty \frac{k}{x^{k-1}}\,dx=\left[\frac{-k}{(k-2)x^{k-2}}\right]_1^\infty=\frac{k}{k-2}.$$

$$\therefore \mathrm{Var}(X)=E(X^2)-\{E(X)\}^2=\frac{k}{k-2}-\frac{k^2}{(k-1)^2}$$

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

$P(X>E(X))=\int_{k/(k-1)}^\infty \frac{k}{x^{k+1}}\,dx=\left[-\frac{1}{x^k}\right]_{k/(k-1)}^\infty=\left(\frac{k-1}{k}\right)^k$, or this can be written down directly from the c.d.f. found in part (i).

(iv) For the case $k=3$,

(a) $M=2^{1/3}$ in the units given, or £12599,

(b) mean = 3/2 in the units given, or £15000,

(c) inserting $X=10$, $P(X\le 10)=1-\frac{1}{10^3}$, so $P(X>10)=\frac{1}{10^3}$, i.e. 0.1%.

Higher Certificate, Paper I, 2004. Question 7

(i) $$E(X)=\int_0^\theta \frac{x}{\theta}dx=\frac{1}{\theta}\left[\frac{1}{2}x^2\right]_0^\theta=\frac{1}{2}\theta.$$

$$E(X^2)=\int_0^\theta \frac{x^2}{\theta}dx=\frac{1}{\theta}\left[\frac{1}{3}x^3\right]_0^\theta=\frac{1}{3}\theta^2.$$

$$\therefore \text{Var}(X)=E(X^2)-\{E(X)\}^2=\frac{1}{3}\theta^2-\left(\frac{1}{2}\theta\right)^2=\frac{1}{12}\theta^2.$$

(ii) P(longest offcut is $\le x$) = P(all n offcuts are $\le x$).

The c.d.f. for each X_i is $F(x)=P(X\le x)=\int_0^x \frac{du}{\theta}=\left[\frac{u}{\theta}\right]_0^x=\frac{x}{\theta}$, and the X_i are all independent. Therefore P(all n offcuts are $\le x$) = $\{F(x)\}^n=\left(\frac{x}{\theta}\right)^n$, and this is also P(longest offcut is $\le x$), i.e. the c.d.f. of the sample maximum $X_{(n)}$. Thus the p.d.f. of $X_{(n)}$ is the derivative of this, i.e. nx^{n-1}/θ^n. This is for the interval $(0, \theta)$.

$$\therefore E\left(X_{(n)}\right)=\int_0^\theta \frac{nx^n}{\theta^n}dx=\frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta=\frac{n\theta}{n+1}.$$

$$E\left(X_{(n)}{}^2\right)=\int_0^\theta \frac{nx^{n+1}}{\theta^n}dx=\frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\theta=\frac{n\theta^2}{n+2}.$$

$$\therefore \text{Var}\left(X_{(n)}\right)=E\left(X_{(n)}{}^2\right)-\left\{E\left(X_{(n)}\right)\right\}^2=\frac{n\theta^2}{n+2}-\frac{n^2\theta^2}{(n+1)^2}$$

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

Immediately we have $E\left(\frac{n+1}{n}X_{(n)}\right)=\theta$, so $\frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ.

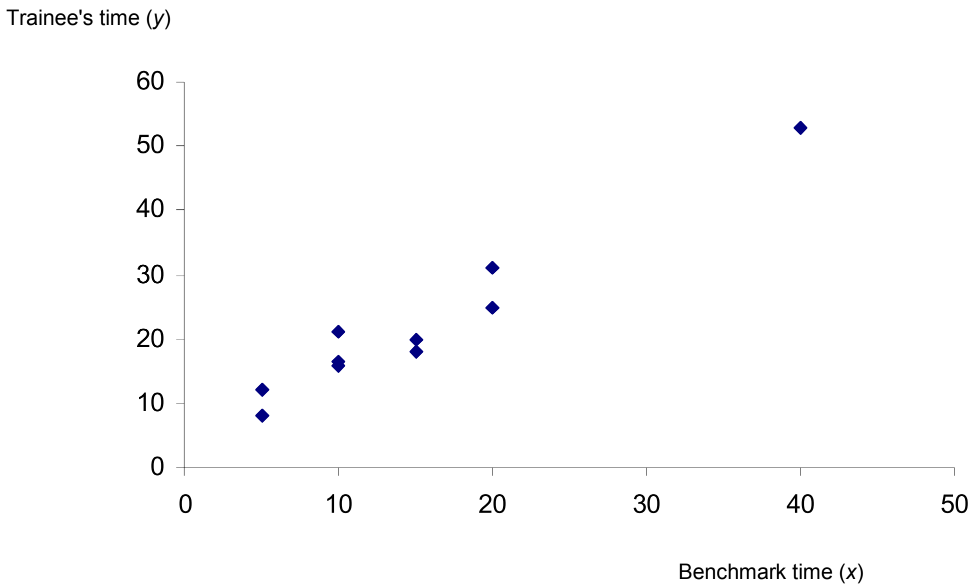
$$\text{Var}\left(\frac{n+1}{n}X(n)\right)=\frac{(n+1)^2}{n^2}\text{Var}\left(X_{(n)}\right)=\frac{(n+1)^2}{n^2}\frac{n\theta^2}{(n+1)^2(n+2)}=\frac{\theta^2}{n(n+2)}.$$

Continued on next page

(iii) We have (see part (i)) that $E(X) = \theta/2$. Thus the method of moments estimator of $\theta/2$ is $\bar{X}$, and so the method of moments estimator of θ is $2\bar{X}$ or $\frac{2}{n}\sum X_i$ as required.

$$\mathrm{Var}\left(\frac{2}{n}\sum X_i\right) = \mathrm{Var}\left(2\bar{X}\right) = 4\mathrm{Var}\left(\bar{X}\right) = \frac{4}{n}\mathrm{Var}\left(X\right) = \frac{4}{n}.\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

(i)



Simple linear regression analysis seems quite suitable.

(ii) The model is $y_i = \alpha + \beta x_i + e_i$, where $\{e_i\}$ are uncorrelated with zero mean and (constant) variance σ^2 (independent identically distributed $N(0, \sigma^2)$ for the purpose of undertaking statistical tests, as in part (iii)). Estimating by the method of least squares gives

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

where (standard notation)

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n},$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}.$$

We have

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{4440 - (150 \times 220/10)}{3200 - (150^2/10)} = \frac{1140}{950} = 1.20 \quad \text{and} \quad \hat{\alpha} = 22 - (1.20 \times 15) = 4,$$

so the line is

$y = 4 + 1.2x$.

Continued on next page

The total sum of squares is $S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^{\,2} - \dfrac{\left(\sum y_i\right)^2}{10} = 1440$.

The sum of squares for regression is $\hat{\beta} S_{xy}$ (or $S_{xy}^{\,2} / S_{xx}$) = 1368.

Therefore the residual sum of squares is 1440 – 1368 = 72.

This has 8 degrees of freedom, so the residual mean square ($\hat{\sigma}^2$) is 72/8 = 9.

The coefficient of determination R^2 = 1368/1440 = 0.95 (usually given as 95%).

(iii) The estimated variance of $\hat{\beta}$ is 9/950 = 0.009474. So the test statistic for testing the null hypothesis $\beta = 1$ is $\dfrac{1.2 - 1}{\sqrt{0.009474}}$ = 2.05, which we refer to t_8.

This is not significant at the 5% level, so the null hypothesis $\beta = 1$ cannot be rejected.

(iv) The model here is $y_i = bx_i + e_i$.

Estimating b by least squares, we minimise $\Omega = \sum_{i=1}^{n} (y_i - bx_i)^2$.

Differentiating with respect to b, we have $\dfrac{d\Omega}{db} = -2\sum (y_i - bx_i) x_i$.

Setting this equal to zero gives $\Sigma x_i y_i = \hat{b} \Sigma x_i^{\,2}$, i.e. $\hat{b} = \Sigma x_i y_i / \Sigma x_i^{\,2}$.

(Note that $\dfrac{d^2\Omega}{db^2} = 2\sum x_i^{\,2} > 0$, so this is a minimum.)

Thus we have $\hat{b}$ = 4440/3200 = 1.3875.