EXAMINATIONS

September 2000

Subject 101 — Statistical Modelling

EXAMINERS' REPORT

1 Mean = 92780/50 = £1855.60

25.5th value in order = (17.3 + 17.4)/2 = 17.35 so median amount = £1735

2 For simplicity change the units into thousands. The sum of the data is

$\Sigma x = 47 \times 28.47 + 24 \times 21.42 = 1{,}852.17.$

So, the mean salary is $1852.17/71 = 26.087$, or £26,087.

The sum of squares is given by

$\Sigma x^2 = \{46 \times 4.27^2 + 47 \times 28.47^2\} + \{23 \times 3.02^2 + 24 \times 21.42^2\} = 50{,}155.5.$

Thus, the standard deviation of the 71 values is

$$\sqrt{\frac{50155.5 - 1852.17^2/71}{70}} = 5.124, \text{ or £5,124.}$$

3 Can get answer directly from Green tables:

$$P(X < 20) = P(X \leq 19) = 0.18026 = 0.18$$

OR: use normal approximation with continuity correction.

$X \doteq N(24, \sqrt{24}^2)$

$P(X < 20) \rightarrow P(X < 19.5)$

$$= P\left(Z < \frac{19.5 - 24}{\sqrt{24}} = -0.92\right)$$

$= 1 - 0.82 = 0.18.$

Note: without continuity correction answer is 0.15 — only 1 mark.

4 $\mu = 0$, $n = 9$ so $\overline{X} / (S / 3) \sim t_8$ so $P(\overline{X} > S) = P(t_8 > 3)$,

which, from tables, is just less than 0.01

(*actually 0.0085, but not needed for the marks)*

5 $\bar{x} \pm 1.96\sqrt{\dfrac{\bar{x}}{500}}$ i.e. $0.168 \pm (1.96 \times 0.01833)$ i.e. 0.168 ± 0.036

6 (a) £

(b) £/year i.e. $£ \times \text{year}^{-1}$

(c) no units

7 *t* test is based on

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} : \text{here } n - 2 = 8$$

If *P*-value for this one-sided test is 0.025 then observed $t = 2.306$ ($t_{0.025,8}$ from Green tables)

$$\therefore \frac{r\sqrt{8}}{\sqrt{1-r^2}} = 2.306 \text{ and } r > 0$$

$$\therefore 5.318\,(1 - r^2) = 8r^2$$

$$\therefore r^2 = \frac{5.318}{13.318} = 0.3993 \quad \therefore r = 0.632$$

Q7 Comment: *An alternative approach is to use Fisher's transformation of r. The statistic is* $\tfrac{1}{2}\log\dfrac{1+r}{1-r}$ *, which, under* H_0*, has, approximately, a N(0,1/7) distribution. This approach gives r = 0.630 .*

8 Let Y = amount

Let X = 1, 2 for types I, II

$\therefore P(X = 1) = 0.25$, $P(X = 2) = 0.75$

$E(Y \mid X = 1) = 500$, $\text{Var}(Y \mid X = 1) = 100^2$

$E(Y \mid X = 2) = 300$, $\text{Var}(Y \mid X = 2) = 70^2$

} all given

$E(Y) = E(E(Y \mid X)) = 500(0.25) + 300(0.75)$

$= 125 + 225 = £350$

$V(Y) = E(V(Y \mid X)) + V(E(Y \mid X))$

$E(V(Y \mid X)) = 100^2(0.25) + 70^2(0.75)$

$= 2500 + 3675 = 6175$

$V(E(Y \mid X)) = 500^2(0.25) + 300^2(0.75) - 350^2$

$= 62500 + 67500 - 122500 = 7500$

$\therefore V(Y) = 6175 + 7500 = 13675$

$\therefore$ s.d.$(Y) = £116.9$

OR: $V(E(Y \mid X)) = 0.25(500 - 350)^2 + 0.75(300 - 350)^2 = 7500$

alternative method for $V(Y)$:

$E(Y^2 \mid X = 1) = 100^2 + 500^2 = 260000$

$E(Y^2 \mid X=2) = 70^2 + 300^2 = 94900$

$\therefore E(Y^2) = 0.25(260000) + 0.75(94900)$

$= 136175$

$\therefore V(Y) = 136175 - 350^2 = 13675$

$\therefore$ s.d.$(Y) = 116.9$

Q8 Comment: *Very few candidates seemed to be aware of the result V(Y) = E[V(Y | X)] + V[E(Y | X)] and how and when it should be used.*

9 $C(t) = \log M(t) = -\alpha\log(1-\theta t)$

$C'(t) = \alpha\theta(1-\theta t)^{-1}$, $C''(t) = \alpha\theta^2(1-\theta t)^{-2}$, $C'''(t) = 2\alpha\theta^3(1-\theta t)^{-3}$

$\therefore \kappa_2 = C''(0) = \alpha\theta^2$, $\kappa_3 = C'''(0) = 2\alpha\theta^3$ so coefficient is

$\kappa_3 / (\kappa_2)^{3/2} = 2\alpha\theta^3 / (\alpha\theta^2)^{3/2} = 2 / \sqrt{\alpha}$

Q9 Comment: *Many candidates used the MGF directly (and encountered some algebraic difficulties) rather than using the more convenient cumulant generating function.*

10 $\text{Cov}(X\ ,\ Y) = \text{Cov}(X\ ,\ X - Z) = \text{Cov}(X\ ,\ X) - \text{Cov}(X\ ,\ Z) = V(X) - 0 = 4$

$V(X) = 4$, $V(Y) = V(X) + V(Z) = 5$

so $\text{Corr}(X\ ,\ Y) = 4/(4 \times 5)^{1/2} = 0.894$

OR:

By independence, it is immediate that Y has mean and variance both equal to 5.

$$\text{Corr}(X\ ,\ Y) = \frac{E(XY) - [E(X)][E(Y)]}{\sqrt{Var(X)Var(Y)}} = \frac{E[X(X-Z)] - 25}{\sqrt{4 \times 5}}.$$

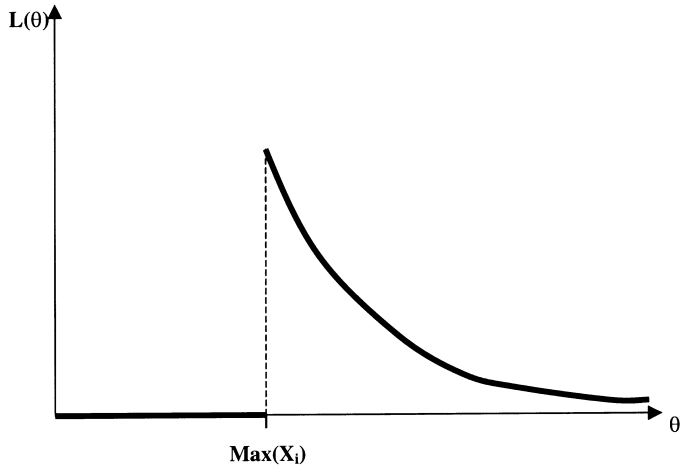
The independence of X and Z implies that $E(XZ) = 0$, and as

$E(X^2) = Var(X) + 5^2 = 29$, we obtain that the correlation is

$$\frac{29 - 0 - 25}{\sqrt{20}} = \frac{2}{\sqrt{5}} = 0.894.$$

11 (i) $$L(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{2x_i}{\theta^2} & : 0 < x_i < \theta \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{2^n \Pi x_i}{\theta^{2n}} & : \theta > \max(x_i) \\ 0 & \text{otherwise} \end{cases}$$



$\therefore$ maximum occurs at $\hat{\theta} = \max(X_i)$

(ii) Since $X_i < \theta$ for all i

then $\max(X_i) < \theta$ **always**

$\therefore\ E(\hat{\theta}) < \theta \quad \therefore \hat{\theta}$ is biased

Q11 Comment: *Few candidates appreciated the significance of the fact that the range of the variable included the unknown parameter θ and therefore the likelihood $L(\theta)$ has a discontinuity (above which it is decreasing).*

12 Assuming n will be large enough to make a normal approximation to the binomial acceptable, the confidence interval will be

$\hat{\theta} \pm 1.96\sqrt{\dfrac{\hat{\theta}(1-\hat{\theta})}{n}}$, where $\hat{\theta}$ is the proportion of the sample who support the party.

Although $\hat{\theta}$ is unknown, it can be estimated as 0.4, so the width of the interval will be approximately

$$2 \times 1.96\sqrt{\frac{0.4 \times 0.6}{n}} = \frac{1.9204}{\sqrt{n}}.$$

If this is to be less than 0.03 we must have that

$$n > \left(\frac{1.9204}{0.03}\right)^2 = 4097.7$$

We need to take n = 4,098.

(As n is large, this confirms that we were correct in assuming that a normal approximation would be appropriate.)

Notes: As this answer is approximate, approximations (such as replacing 1.96 with 2) are acceptable, as is rounding the final answer.

An acceptable alternative would be to estimate $\hat{\theta}$ by 0.415 rather than 0.4; this is the upper limit of a 95% confidence interval for θ. This leads to a value of n = 4,145.

13 $E(\bar{Y}_{i\cdot}) = \mu + \tau_i$ and $E(\bar{Y}_{\cdot\cdot}) = \mu$

$E(\bar{Y}_{i\cdot})\ E(\hat{\mu}) = E(\bar{Y}_{\cdot\cdot}) = \mu \quad \therefore$ unbiased

$E(\hat{\tau}_i) = E(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) = \mu + \tau_i - \mu = \tau_i \quad \therefore$ unbiased

$V(\hat{\mu}) = V(\bar{Y}_{..}) = \dfrac{\sigma^2}{kr}$

as $\bar{Y}_{..}$ is mean of kr r.v.'s each with var σ^2

$V(\hat{\tau}_i) = V(\bar{Y}_{i.} - \bar{Y}_{..})$

$= V(\bar{Y}_{i.}) + V(\bar{Y}_{..}) - 2Cov(\bar{Y}_{i.}, \bar{Y}_{..})$

$= \dfrac{\sigma^2}{r} + \dfrac{\sigma^2}{kr} - 2\dfrac{1}{r}.\dfrac{1}{kr}.r\sigma^2$

$= \dfrac{\sigma^2}{r} - \dfrac{\sigma^2}{kr} = \dfrac{(k-1)\sigma^2}{kr}$

Alternative method:

$V(\bar{Y}_{i.} - \bar{Y}_{..}) = V\left(\left(1 - \dfrac{1}{k}\right)\bar{Y}_{i.} - \dfrac{1}{k}\sum_{j \neq i}\bar{Y}_{j.}\right)$

$= \left(1 - \dfrac{1}{k}\right)^2 \dfrac{\sigma^2}{r} + \left(-\dfrac{1}{k}\right)^2 (k-1)\dfrac{\sigma^2}{r}$

$= \dfrac{\sigma^2}{k^2 r}\{(k-1)^2 + (k-1)\} = \dfrac{(k-1)\sigma^2}{kr}$

Q13 Comment: *This material was unfamiliar to most candidates.*

14 (i) $F_Y(y) = P(Y < y) = P(X^2 < y) = P(X < \sqrt{y}) = 1 - \exp(-y/\theta)$

which is the cdf of an exponential r.v. with mean θ.

(ii) (a) $f_X(x) = 2(x/\theta)\exp(-x^2/\theta)$

So $L(\theta) = k_1\,\theta^{-n}\exp[-(\Sigma x^2)/\theta]$ so $\log L = k_2 - n\log\theta - (\Sigma x^2)/\theta$

$d\log L / d\theta = -n/\theta + (\Sigma x^2)/\theta^2 = 0 \Rightarrow \hat{\theta} = \dfrac{1}{n}\Sigma X_i^2$

(b) From (i) $E(X^2) = E(Y) = \theta$ and $V(X^2) = V(Y) = \theta^2$

So $E(\hat{\theta}) = (1/n)\,\Sigma E(X_i^2) = (1/n)\,n\theta = \theta$, and

$V(\hat{\theta}) = (1/n^2)\,\Sigma V(X_i^2) = (1/n^2)\,n\theta^2 = \theta^2/n$

Now, $d^2\log L / d\theta^2 = n\theta^{-2} - 2(\Sigma x^2)\,\theta^{-3}$

so $-E(d^2\log L / d\theta^2) = -n\theta^{-2} + 2\theta^{-3}(n\theta) = n / \theta^2 = 1/V$

So $\hat{\theta}$ is unbiased and attains the *C–R* bound.

(c) $Y = X^2$ has mgf $(1 - \theta t)^{-1}$ so $n\hat{\theta} = \Sigma Y_i$ has mgf $(1 - \theta t)^{-n}$

So $2n\hat{\theta} / \theta$ has mgf $[1 - \theta(2t / \theta)]^{-n} = (1 - 2t)^{-n}$

which is mgf of χ^2 with $2n$ d.f.

(iii) (a) $\hat{\theta} = 485.9028/50 = 9.718$

(b) P(74.22 < $100\hat{\theta}$ / θ < 129.6) = 0.95 from tables of χ2 with 100 d.f.

$\Rightarrow$ 95% CI for θ is $(100\hat{\theta} / 129.6 , 100\hat{\theta} / 74.22)$

i.e. (971.8056 / 129.6 , 971.8056 / 74.22) i.e. (7.50,13.1)

15 (i) (a) Let critical value be *c*, such that we reject H_0 for $\overline{X} > c$ for a sample of size *n*.

P(type I error) = $P(\overline{X} > c)$ where $\overline{X} \sim N(100, 0.1^2 / n)$
$\therefore (c - 100) / (0.1 / \sqrt{n}) = 1.645$ …… (*)

P(type II error) = $P(\overline{X} < c)$ where $\overline{X} \sim N(100.5, 0.5^2 / n)$
$\therefore (c - 100.5) / (0.5 / \sqrt{n}) = -1.645$ …… (**)

Solving (*) and (**) for n gives $100 + 1.645(0.1/\sqrt{n}) = 100.5 - 1.645(0.5 / \sqrt{n})$
$\Rightarrow \sqrt{n} = 1.645 \times 0.6/0.5 = 1.974 \Rightarrow n = 3.9$
$\therefore$ sample of size $n = 4$ should be examined

(b) Let critical value be *c*

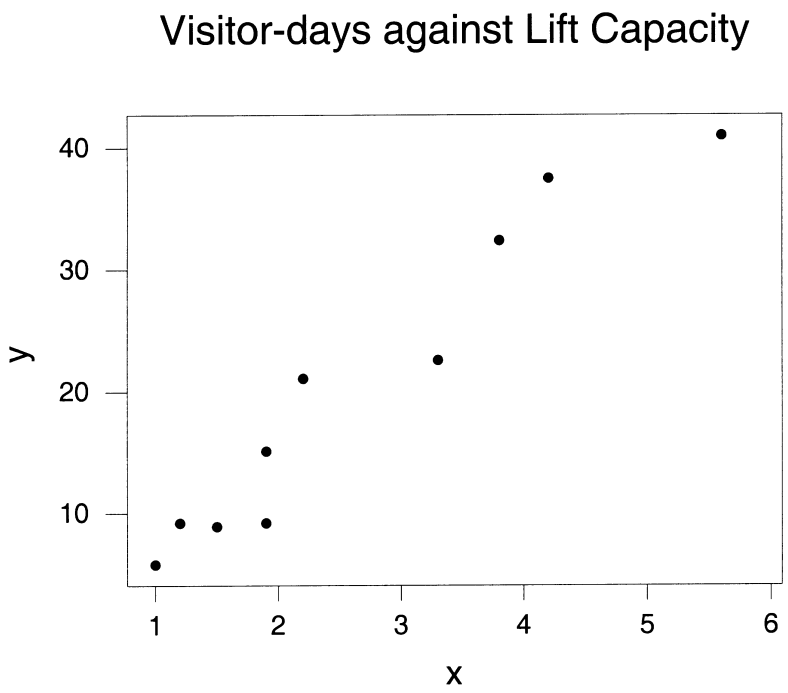
P(type I error) = 0.01 $\Rightarrow (c - 100) / (0.1 / \sqrt{4}) = 2.326$
$\therefore c = 100 + 2.326(0.1 / \sqrt{4}) = 100.1163$

Power = $1 - P$(type II error) = $1 - P[Z < (100.1163 - 100.5) / (0.5 / \sqrt{4})]$
= $1 - P(Z < -1.535) = P(Z < 1.535) = 0.9376$ i.e. Power = 93.8%

(ii) P-value = $P(\overline{X} > 100.1)$ where $\overline{X} \sim N(100, 0.1^2 / 10)$
= $P[Z > (100.1 - 100) / (0.1 / \sqrt{10})] = P(Z > 3.162) = 1 - 0.9992$
= 0.0008 i.e. 0.08%

Q15 Comment: *Most candidates understood the concepts of testing errors, but were unable to apply that knowledge.*

16 (i)



Seems to be a strong increasing linear relationship.

(ii) $S_{xx} = 91.08 - \frac{26.6^2}{10} = 20.324$

$S_{yy} = 5603.12 - \frac{202.8^2}{10} = 1490.336$

$S_{xy} = 707.58 - \frac{(26.6)(202.8)}{10} = 168.132$

$r = \frac{168.132}{\sqrt{(20.324)(1490.336)}} = 0.966$

Large and positive agreeing with comment above.

(iii) $y = \hat{\alpha} + \hat{\beta}x$ where

$$\hat{\beta} = \frac{168.132}{20.324} = 8.2726$$

$$\hat{\alpha} = \frac{202.8}{10} - 8.2726\left(\frac{26.6}{10}\right) = -1.73$$

$\Big\}$ $-1.73 + 8.273x$

(iv) (a) SSTOT $= S_{yy} = 1490.336$

SSRES $= S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$ (formula from green book)

$$= 1490.336 - \frac{(168.132)^2}{20.324} = 99.450$$

$\therefore$ SSREG $= 1390.886$

(b) $R^2 = \dfrac{\text{SSREG}}{\text{SSTOT}} = 0.933$ $= r^2$ from (ii)

(c) $\hat{\sigma}^2 = \dfrac{1}{8}(\text{SSRES}) = 12.43$

(v) If x increases by 0.5 (500 in '000's), then expected increase in y is $0.5\hat{\beta}$ $= 0.5(8.2726) = 4.136$ or 4136 visitor days

$$Var(0.5\hat{\beta}) = (0.5)^2\, Var(\hat{\beta}) = 0.25\frac{\hat{\sigma}^2}{S_{xx}}$$

$$= 0.25\frac{12.43}{20.324} = 0.153 \qquad \therefore \text{ s.d. } (0.5\hat{\beta}) = 0.391$$

$\therefore$ standard error of estimate is 391

Q16 Comment: *Nearly all candidates scored well on parts (i) – (iv), but most used an inappropriate expression for the standard error of* $0.5\hat{\beta}$.