EXAMINATIONS

September 2001

Subject 101 — Statistical Modelling

EXAMINERS' REPORT

1 $n = 100$. So median = $50\frac{1}{2}th$ observation = 2

$Q_1 = 25\frac{1}{2}th$ observation = 1

$Q_3 = 75\frac{1}{2}th$ observation = 3 $\quad\therefore$ IQR = 3 – 1 = 2

(same answers using alternative definitions)

2 As X_1 and X_2 are independent $M_{X_1+X_2}(t) = M_{X_1}(t).M_{X_2}(t)$

$= e^{\mu_1(e^t-1)}.e^{\mu_2(e^t-1)}$ using formula in Green book

$= e^{(\mu_1+\mu_2)(e^t-1)}$

$\therefore$ $X_1 + X_2$ is Poisson with mean $(\mu_1 + \mu_2)$

3 $\bar{X} \sim N$ with mean 0 so $P(\bar{X} > 0) = 0.5$

$4S^2/\sigma^2 \sim \chi^2$ with 4 d.f. i.e. $4S^2 \sim \chi^2$ with 4 d.f.

$\therefore\ P\left[\sum_{i=1}^{5}\left(X_i - \bar{X}\right)^2 < 9.488\right] = P(\chi_4{}^2 < 9.488) = 0.95$

$\bar{X}$ and S^2 are independent, so probability required = $0.5 \times 0.95 = 0.475$

4 Fitted regression line:

$\hat{y} = \hat{\alpha} + \hat{\beta}x$

But least squares estimate of α is

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

Therefore $\hat{y} = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x})$

$\Rightarrow \hat{y} = \bar{y}$ when $x = \bar{x}$. Hence line passes through point $(\bar{x}, \bar{y})$.

5 $V(Y) = E[V(Y|X)] + V[E(Y|X)]$

$= E(X + 12) + V(15X + 20)$

$= E(X) + 12 + [15^2 \times V(X)]$

$= 10 + 12 + 225(10) = 2272$

6 (i) X = number of claims arising

$\therefore X \sim$ binomial with $n = 200$, $p = 0.015$

Use Poisson approximation

$\therefore X \approx$ Poisson with $\lambda = 200(0.015) = 3$

Using Green book tables (or otherwise)

$P(X > 10) = 1 - P(X \leq 10) = 1 - 0.99971 = 0.00029$

The normal approximation is not as appropriate as the Poisson approximation for a bit (200, 0.015) distribution, which is quite skewed.

(ii) X = number of claims arising $\therefore X \sim$ binomial with $n = 2000$, $p = 0.015$

Could use Poisson approximation $\therefore X \approx$ Poisson with $\lambda = 2000(0.015) = 30$

This is beyond the scope of the Green Book tables, and direct calculation would be awkward. So use Normal approximation

$\therefore X \approx \mathrm{N}(2000(0.015)\ ,\ 2000(0.015)(0.985)) = N(30\ ,\ 29.55)$

$\therefore P(\mathrm{X} > 40) \equiv P(\mathrm{X} > 40.5)$ using continuity correction

$$\approx P(Z > \frac{40.5 - 30}{\sqrt{29.55}} = 1.93) = 1 - 0.973 = 0.027$$

7 (i)

$$f(x) = \begin{cases} kx(1 - ax^2), & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

To be a pdf $f(x) \geq 0$ for $0 \leq x \leq 1 \Rightarrow (1 - ax^2) \geq 0$ since $k > 0$

$\Rightarrow 1 \geq ax^2$ for $x \leq 1 \Rightarrow a \leq 1$.

Also $\int_0^1 f(x)dx = 1 \Rightarrow k\int_0^1 \left(x - \alpha x^3\right)dx = 1$

$$\Rightarrow k\left[\frac{1}{2}x^2 - \frac{1}{4}\alpha x^4\right]_0^1 = 1 \Rightarrow k\left(\frac{1}{2} - \frac{\alpha}{4}\right) = 1 \Rightarrow k\left(\frac{2-\alpha}{4}\right) = 1$$

$\Rightarrow k = 4/(2 - \alpha)$

(ii) $\alpha = 1 \Rightarrow k = 4$; $f(x) = 4x(1 - x^2)$,

$$E(X) = \int_0^1 4x^2(1-x^2)dx$$

$$= \left[\frac{4}{3}x^3 - \frac{4}{5}x^5\right]_0^1 = \frac{4}{3} - \frac{4}{5} = \frac{8}{15}.$$

8 $X \sim N(28,2^2)$ $\quad Y \sim N(25,1^2)$

Require $P(X - Y > 5)$

where $X - Y \sim N(3,5)$

i.e. $P\left(Z > \frac{5-3}{\sqrt{5}}\right) = P(Z > 0.894) = 0.186.$

9 (i) $E(S_W^{\,2}) = \alpha E(S_1^{\,2}) + (1-\alpha)E(S_2^{\,2})$

$= \alpha\sigma^2 + (1 - \alpha)\sigma^2 = \sigma^2$ (using unbiasedness of sample variance)

Therefore $S_W^{\,2}$ is unbiased for σ^2. MSE = $Var(S_W^{\,2})$ since unbiased.

MSE = $Var(S_W^{\,2}) = \alpha^2 Var(S_1^{\,2}) + (1-\alpha)^2 Var(S_2^{\,2})$

$$= 2\sigma^4\left(\frac{\alpha^2}{n_1 - 1} + \frac{(1-\alpha)^2}{n_2 - 1}\right) \qquad \left(\text{using } Var(S_i^{\,2}) = \frac{2\sigma^4}{n_i - 1};\ i = 1,2\right)$$

(ii) $$\frac{dMSE}{d\alpha} = 2\sigma^4\left(\frac{2\alpha}{n_1 - 1} + \frac{-2(1-\alpha)}{n_2 - 1}\right)$$

Setting equal to zero gives

$$\frac{\alpha}{n_1 - 1} - \frac{1-\alpha}{n_2 - 1} = 0 \Rightarrow (n_2 - 1)\,\alpha - (n_1 - 1)(1 - \alpha) = 0$$

Thus giving $\alpha = \dfrac{n_1 - 1}{n_1 + n_2 - 2}$ which clearly minimises MSE.

10 (i) $S = \sum_{i=1}^{n} \left(Y_i - E(Y_i)\right)^2 = \sum_{i=1}^{n} \left(Y_i - \lambda x_i\right)^2$

$\dfrac{dS}{d\lambda} = -2\sum x_i(Y_i - \lambda x_i)$ Setting to 0 $\Rightarrow \tilde{\lambda} = \dfrac{\sum x_i Y_i}{\sum x_i^{\,2}}$

(ii) $L(\lambda) = e^{-\lambda \sum x_i} \prod \left(\lambda x_i\right)^{Y_i} \times \text{const.}$ so $\log L = -\lambda \sum x_i + \left(\sum Y_i\right) \log \lambda + \text{const.}$

$\therefore \dfrac{d \log L}{d\lambda} = -\sum x_i + \dfrac{1}{\lambda} \sum Y_i$ Setting to 0 $\Rightarrow \hat{\lambda} = \dfrac{\sum Y_i}{\sum x_i}$

(iii) $E(\tilde{\lambda}) = \dfrac{1}{\sum x_i^{\,2}} E\left(\sum x_i Y_i\right) = \dfrac{1}{\sum x_i^{\,2}} \sum x_i \lambda x_i = \lambda$ hence unbiased

$E(\hat{\lambda}) = \dfrac{1}{\sum x_i} E\left(\sum Y_i\right) = \dfrac{1}{\sum x_i} \sum \lambda x_i = \lambda$ hence unbiased

Some candidates did not appear to understand that the two methods of deriving estimators could produce different estimators.

11 (i) 129.1 + 109.8 + 123.5 = 362.4 , 1,534.37 + 1,109.88 + 1,401.73 = 4,045.98

$SS_T = 4{,}045.98 - 362.4^2/33 = 66.17$

So $SS_B = 66.17 - 48.24 = 17.93$ **

Table is:

Source of variation	d.f.	SS	MSS
Between companies	2	17.93	8.97
Residual	30	48.24	1.61
	32	66.17	

(ii) $F = 8.97/1.61 = 5.57$ on 2,30 d.f.

P-value is less than 0.01, so reject null hypothesis.

There is strong evidence that there are differences among the (population) means of the sums insured for the three companies.

** OR: Calculate SS_B directly as

$SS_B = (129.1^2 + 109.8^2 + 123.5^2) / 11 - 362.4^2/33 = 17.93$

12 H_0 : this year's pattern is the same as last year's v. H_1 : not the same

Under H_0, the expected frequencies are:

$120 \times 0.184 ; 0.703 ; 0.113 = 22.08 ; 84.36 ; 13.56$

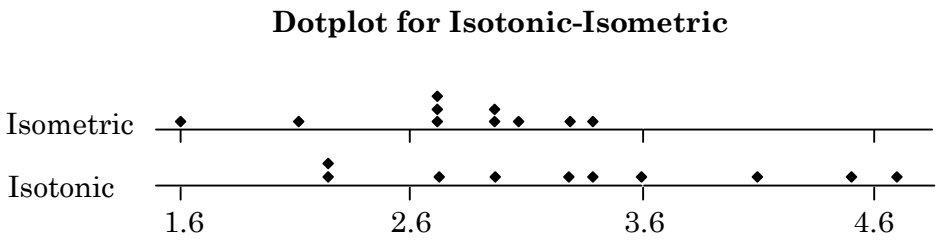
o_i	e_i	$(o-e)^2/e$
15	22.08	2.270
87	84.36	0.083
18	13.56	1.454
		3.807 on 2 df

5% point from χ^2_2 is 5.991. So cannot reject H_0 at 5% level.

These data provide no evidence to suggest that this year's pattern differs from that of last year.

A few candidates worked with percentages of claims instead of numbers of claims (when using the chi-squared goodness-of-fit statistic, one must work with observed and expected frequencies). Such work received few if any marks.

13 (i) (a) Plot for Isotonic-Isometric exercise methods:



Normality seems OK for each data set.

Let X_A, X_B be reductions in measurements from the isometric and isotonic methods, respectively.

A: $\sum x_A = 27.6, \sum x_A{}^2 = 78.90 ; \bar{x}_A = 2.76, s_A{}^2 = 0.3027, n_A = 10$

B: $\sum x_B = 33.7, \sum x_B{}^2 = 120.53 ; \bar{x}_B = 3.37, s_B{}^2 = 0.7734, n_B = 10$

(b) $s_A{}^2 = 0.3027 ; s_B{}^2 = 0.7734$

$H_0 : \sigma_A{}^2 = \sigma_B{}^2 ; H_1 : \sigma_A{}^2 \neq \sigma_B{}^2$

$$F = \frac{0.7734}{0.3027} = 2.56 \text{ on 9,9 d.f.}$$

Upper 5% point is 3.179, so p-value > 0.10.

Therefore do not reject H_0.

(c) The pooled sample variance

$$s_p^{\,2} = \left\{(n_A - 1)s_A^{\,2} + (n_B - 1)s_B^{\,2}\right\} / (n_A + n_B - 2)$$

$= \{(9 \times 0.3027) + (9 \times 0.7734)\}/18 = 0.538.$

The test statistic $t = (\bar{x}_B - \bar{x}_A) / \sqrt{\left\{s^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\right\}}$

$$= (3.37 - 2.76) / \sqrt{\left\{0.538\left(\frac{1}{10} + \frac{1}{10}\right)\right\}}$$

$= 1.86.$

$H_0 : \delta = 0$; $H_1 : \delta > 0$ δ : mean difference in reduction in abdomen measurements ($\mu_B - \mu_A$).

A one-sided test is appropriate.

There are 18 d.f. The upper 5% point of t_{18} is 1.734. Thus the probability value is less than 0.05. There is sufficient evidence, at the 5% level, to suggest that the isotonic method (B) is more effective in reducing abdomen measurement.

(ii) (a) Two-sided 95% confidence interval for $\mu_B - \mu_A$:

$$(\bar{x}_B - \bar{x}_A) \pm t_{18}(2.5\%) \sqrt{\left\{s^2\left(\frac{1}{10} + \frac{1}{10}\right)\right\}}$$

$$(3.37 - 2.76) \pm 2.101 \sqrt{\left\{0.538\left(\frac{2}{10}\right)\right\}}$$

$0.61 \pm 0.69 = (-0.08, 1.3)$

[Note that this just includes zero.]

(b) $(n_A + n_B - 2)\frac{S_p^{\,2}}{\sigma^2} \sim \chi^2_{n_A + n_B - 2}$ Here $n_A + n_B - 2 = 18$

95% confidence interval for σ^2 (common variance)

$$\left(\frac{18s^2}{\chi^2_{18}(0.025)}, \frac{18s^2}{\chi^2_{18}(0.975)}\right)$$

$$\left(\frac{18(0.538)}{31.53}, \frac{18(0.538)}{8.231}\right)$$

Taking square-roots gives the 95% confidence interval for the common standard deviation σ as $\left(\sqrt{0.31}, \sqrt{1.18}\right) = (0.55\,, 1.08)$

Some candidates used a "paired samples" approach in part (ii)(a). This was quite inappropriate.

14 (i) (a) In Model M1 we have a basic model for the *initial bp* of the whole population of young male athletes, with mean μ, and the mean *bp* increases by α after using the stimulant.

(*Note:* E(*follow-up bp*) $= E(Y) = E[E(Y \mid X)] = E[X + \alpha] = \mu + \alpha$)

Model M2 extends M1 by allowing for a different initial mean for each athlete (μ_i).

Model M3 extends M2 by allowing for a different mean increase in *bp* for each athlete (α_i).

(*Note:* In all three models we have a single population variance for *initial bp* and a single, but different, variance for *follow-up bp*.)

(*Note:* V(*follow-up bp*) $= V(Y) = V[E(Y \mid X)] + E[V(Y \mid X)] = \sigma_1^2 + \sigma_2^2$)

(b) For 10 athletes, M3 has 22 unknown parameters — but we only have 20 data points. So estimation of parameters is impossible.

(ii) (a) *Initial bp*: $\Sigma x = 1191$, $\Sigma x^2 = 142471$ so $\bar{x} = 119.1$, $s^2 = 69.211$

$t_9(0.025) = 2.262$

$\therefore$ 95% CI for μ is $119.1 \pm \{2.262 \times (69.211/10)^{1/2}\}$ i.e. 119.1 ± 5.95

i.e. (113.15 , 125.05)

(b) *Follow-up bp* : $\Sigma x = 1264$ $\therefore \bar{x} = 126.4$ so $\hat{\alpha} = 126.4 - 119.1 = 7.3$

(iii) Use the differences (follow-up less initial) for each athlete:

d_i : 7, 4, 11, 10, 14, 5, 8, 7, –2, 9

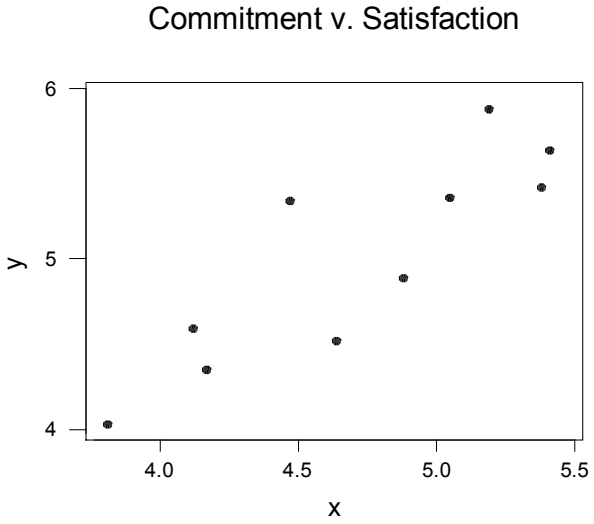
$\Sigma d = 73$, $\Sigma d^2 = 705$ so $\bar{d} = 7.3$, $s^2 = 19.122$

$t_9(0.05) = 1.833$

$\therefore$ 95% CI (one-sided) for α is $(7.3 - 1.833(19.122/10)^{½}, \infty)$ i.e. $(4.77, \infty)$

The early part of this question (on comparing models) looked hard, but, pleasingly, was generally well-attempted.

15 (i) see plot



there seems to be an increasing and linear relationship.

(ii) $$S_{xx} = 224.8554 - \frac{47.12^2}{10} = 2.82596$$

$$S_{yy} = 253.5796 - \frac{50.02^2}{10} = 3.37956$$

$$S_{xy} = 238.3676 - \frac{(47.12)(50.02)}{10} = 2.67336$$

$$\hat{\beta} = \frac{2.67336}{2.82596} = 0.946001$$

$$\hat{\alpha} = \frac{50.02}{10} - (0.946001)\frac{47.12}{10} = 0.544$$

$y = 0.544 + 0.9460x$

(iii) $R^2 = \dfrac{(2.67336)^2}{(2.82596)(3.37956)} = 0.748$ or 74.8%

quite high, showing agreement with a linear relationship.

(iv) $\hat{\sigma}^2 = \dfrac{1}{8}(3.37956 - \dfrac{2.67336^2}{2.82596}) = 0.1063$

For confidence interval use $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$

$$\Rightarrow \left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2}(0.025)}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2}(0.975)} \right)$$

$$= \left(\frac{8(0.1063)}{17.53}, \frac{8(0.1063)}{2.180} \right) = (0.0485, 0.3902)$$

(v) $\hat{\beta} = 0.9460$

its standard error is $\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\dfrac{0.1063}{2.82596}} = 0.1939$

95% confidence interval is $\hat{\beta} \pm t_8(0.025) \times s.e.$

= 0.9460 ± 2.306(0.1939) = 0.946 ± 0.447 or (0.499, 1.393)

(vi) estimate is $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}(5.0) = 0.544 + 0.9460(5.0) = 5.274$

$$s.e.(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{(5.0 - \bar{x})^2}{S_{xx}})}$$

$$= \sqrt{0.1063(\frac{1}{10} + \frac{(5.0 - 4.712)^2}{2.82596})} = 0.1173$$

95% confidence limits are $\pm 2.306(0.1173) = \pm 0.270$ or (5.004, 5.544)