EXAMINATIONS

April 2001

Subject 101 — Statistical Modelling

EXAMINERS' REPORT

1 Ordered data are:

198 200 203 208 209 210 210 212 215 215
216 220 221 223 224

$n = 15$ $Q_1 = 4\frac{1}{4}^{th}$ observation = £208.25, median = 8th observation = £212

$Q_3 = 11\frac{3}{4}^{th}$ observation = £219 $Q_3 - Q_1 = £10.75$

[OR: Using the alternative definitions of quartiles as 16/4th and 48/4th observations gives $Q_1 = 208$, $Q_3 = 220$, $Q_3 - Q_1 = 12$.]

2 Apply Bayes' theorem: $p = \frac{1}{1000}$ (Probability of having disease)

$$P(\text{has disease} \mid \text{positive result}) = \frac{\frac{98}{100}p}{\frac{98}{100}p + \frac{1}{100}(1-p)} = \frac{98}{1097} = 0.089$$

3 Let the number of events in a period of time of length t hours be X_t .

Then $X_1 \sim$ Poisson(μ) and $X_t \sim$ Poisson(μt).

Let the time between two consecutive events be T.

Then $P(T > t) = P$(no events in period of length t) $= P(X_t = 0) = \exp(-\mu t)$.

So $P(T < t) = 1 - \exp(-\mu t)$ [and so $f(t) = \mu \exp(-\mu t)$, $t > 0$]

Hence $T \sim$ exponential with mean 1/μ hours.

4 Binomial (10,000, 0.001) approximated by Poisson with mean of 10.

Approximate probability of no more than 5 claims:

$$e^{-10}\left(1 + 10 + \frac{10^2}{2!} + \frac{10^3}{3!} + \frac{10^4}{4!} + \frac{10^5}{5!}\right) = 0.0671.$$

[OR: 0.06709 using Green Book]
[OR: Use normal approximation with continuity correction]

5 PGF for binomial (n,p)

$$G_X(t) = E(t^X) = \sum_{x=0}^{n} t^x P(X = x)$$

$$= \sum t^x \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum \binom{n}{x} (pt)^x (1-p)^{n-x}$$

$$= (1 - p + pt)^n$$

MGF obtained by replacing t by e^t, i.e. $M_X(t) = (1 - p + pe^t)^n$.

6 $C(t) = \log M(t) = \log[\exp(\mu t + \sigma^2 t^2/2)]$ from Green Book
$= \mu t + \sigma^2 t^2/2$

κ_r is the coefficient of $t^r/r!$, $r = 2,3,\ldots$

$\therefore \kappa_2 = \sigma^2, \kappa_3 = \kappa_4 = 0$

7 Using formulae from the Green book

mgf of N is $\exp(10(e^t - 1))$ and mgf of X_i is $\exp(2(e^t - 1))$

mgf of $\sum_{i=1}^{N} X_i$ is $\exp\{10[\exp(2(e^t - 1)) - 1]\}$

8 (i) Let $X_A = 1$ if A dies and 0 if not. Similarly X_B for life B.

$E(X_A) = 0.1$, $V(X_A) = 0.1 \times 0.9 = 0.09$;

$E(X_B) = 0.2$, $V(X_B) = 0.2 \times 0.8 = 0.16$

Total losses $T = 5X_A + 3X_B$ (units of £10,000)

so $E(T) = (5 \times 0.1) + (3 \times 0.2) = 1.1$ i.e. £11,000

and $V(T) = (25 \times 0.09) + (9 \times 0.16) = 3.69$ so $SD(T) = 1.921$

i.e. £19,210

[OR: T takes values 0, 3, 5, and 8 with probabilities $0.9 \times 0.8 = 0.72$, $0.2 \times 0.9 = 0.18$, $0.1 \times 0.8 = 0.08$, and $0.1 \times 0.2 = 0.02$ respectively; hence find $E(T)$, $E(T^2)$, and $V(T)$.]

(ii) P(exactly one of A, B dies) = 0.18 + 0.08 = 0.26

P(A dies | exactly one dies) = P(A dies and B does not die) / P(exactly one dies)

= 0.08/0.26 = 4/13 ; so P(B dies | exactly one dies) = 9/13

So E(loss | exactly one dies) = 5(4/13) + 3(9/13) = 47/13 = 3.615

i.e. £36,150

9 (i) The change after a large number of time periods equals the sum of i.i.d. r.v.'s.

The Central Limit theorem leads to an approximate normal distribution.

(ii)

x :	+1	0	−1
p :	0.35	0.35	0.30

$E(X) = 0.35 - 0.30 = 0.05$

$E(X^2) = 0.35 + 0.30 = 0.65$

$\therefore \text{Var}(X) = 0.65 - (0.05)^2 = 0.6475 \quad \therefore \text{s.d.}(X) = 0.8047$

Let Y = price change after 500 periods

$\therefore Y$ has mean $500(0.05) = 25$ and s.d. $\sqrt{500}(0.8047) = 17.99$

$$P(Y > 20) \approx P(Z > \frac{20 - 25}{17.99} = -0.28) = 0.61$$

[Continuity correction can be used if desired]

10 (i) Using the basic binomial result

$$\hat{\theta} = \frac{x}{n} = \frac{5}{20} = 0.25$$

(ii) $\theta = P(X > 200)$ using $X \sim N(\mu\ ,\ 20^2)$

$$= P(Z > \frac{200 - \mu}{20}) = 1 - \Phi(\frac{200 - \mu}{20})$$

where $\Phi(.)$ is the cdf of $N(0,1)$.

Using the invariance property of MLE's, we get $\hat{\mu}$ from the equation

$$\hat{\theta} = 1 - \Phi(\frac{200 - \hat{\mu}}{20})$$

So with $\hat{\theta} = 0.25$, we get $\hat{\mu}$ from $\Phi(\frac{200 - \hat{\mu}}{20}) = 0.75$

$$\therefore \frac{200 - \hat{\mu}}{20} = 0.674 \text{ from tables}$$

$$\therefore \hat{\mu} = 200 - 20(0.674) = £186.52$$

11 (i) 95% confidence interval: $\bar{x} \pm 1.96\frac{75}{\sqrt{n}}$

Require $1.96\frac{75}{\sqrt{n}} \leq 10$

$$\Rightarrow \sqrt{n} \geq \frac{1.96 \times 75}{10} = 14.7$$

$n \geq 216.09$ $\therefore$ take $n = 217$.

(ii) 99% confidence interval: $\bar{x} \pm 2.576\frac{75}{\sqrt{n}}$

Require $2.576\frac{75}{\sqrt{n}} \leq 10$

$$\Rightarrow \sqrt{n} \geq \frac{2.576 \times 75}{10} = 19.32$$

$n \geq 373.3$ $\therefore$ take $n = 374$.

12 (i) $\hat{p} = 0.17$

95% confidence interval for the true proportion p is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{400}}$$

$$\Rightarrow 0.17 \pm 1.96\sqrt{\frac{(0.17)(0.83)}{400}} = 0.17 \pm 0.037 \quad \text{or} \quad (0.133, 0.207)$$

(ii) $\hat{p}_1 = 0.17$, $\hat{p}_2 = 0.2$

common $\hat{p} = \dfrac{68+80}{800} = 0.185$

$H_0 : p_2 = p_1$ v. $H_1 : p_2 > p_1$

$$z = \frac{0.2 - 0.17}{\sqrt{(0.185)(0.815)(\frac{1}{400} + \frac{1}{400})}} = \frac{0.03}{0.0275} = 1.09$$

P-value = $P(Z > 1.09) = 1 - 0.862 = 0.14$

There is <u>not</u> sufficient evidence to conclude that there is an increase in the true proportion. The observed difference could have occurred by chance. 1

13 (i) Estimate of $\mu = 4302/37 = 116.27$

Estimate of $\tau_1 = 115.13 - 116.27 = -1.14$;

estimate of $\tau_2 = 95.00 - 116.27 = -21.27$;

estimate of $\tau_3 = 135.42 - 116.27 = 19.15$

(ii) Total $SS = 521662 - 4302^2/37 = 21467.3$

Model (Explained) $SS = 15(115.13^2) + 10(95^2) + 12(135.42^2) - 4302^2/37$

$= 8942.0$ **

$\therefore$ Error $SS = 21467.3 - 8942.0 = 12525.3$ **

$F = (8942.0/2)/(12525.3/34) = 12.1$ on 2,34 degrees of freedom

Upper 1% point for $F_{2,34} \approx 5.3$ so here P-value << 0.01

Overwhelming evidence against hypothesis $H_0 : \tau_i = 0$, $i = 1,2,3$.

We conclude that real differences do exist : there *is* a company effect.

[** OR: Error $SS = (14 \times 196.41) + (9 \times 341.33) + (11 \times 609.36) = 12524.7$
$\therefore$ Model $SS = 21467.3 - 12524.7 = 8942.6$]

14 (i) The model greatly overestimates the number of policyholders who give rise to exactly one claim, and greatly underestimates the number who give rise to multiple claims.

(ii) $E(X) = 1 \times p(1-p) + 2 \times p(1-p)^2 + 3 \times p(1-p)^3 + \ldots$

$= p(1-p)[1 + 2(1-p) + 3(1-p)^2 + \ldots] = p(1-p)[1 - (1-p)]^{-2} = (1-p)/p$

Data mean is $(128 + 78 + 21)/1000 = 0.227$

so the method of moments estimate * of p is given by solving

$(1-p)/p = 0.227$, which gives estimate of $p = 0.81500$.

* it is also in fact the MLE

(iii) Estimates of probabilities for 0,1,2, and 3 claims are then 0.815,

$0.815 \times 0.185 = 0.15078$, $0.815 \times 0.185^2 = 0.02789$, and

$0.815 \times 0.185^3 = 0.00516$

So, for 4 or more claims, estimate is $1 - 0.99883 = 0.00117$.

Hence expected frequencies are as follows:

Number of claims per policyholder	0	1	2	3	≥4
Expected number of policyholders	815.0	150.8	27.9	5.2	1.2

(iv) (a) Using 5 cells:

$\chi^2 = (826-815)^2/815+(128-150.8)^2/150.8+(39-27.9)^2/27.9 +(7-5.2)^2/5.2$

$+ 1.2^2/1.2 = 9.84$ on 3 df

P-value is between 0.025 and 0.01

There is quite strong evidence against the null hypothesis (that the second statistician's model describes the data) and we conclude that the model does not provide a good fit to the data.

[Alternative solution: Using 4 cells (for 0 , 1 , 2 , and ≥ 3) the χ^2 value is 8.07 on 2df. *P*-value is again between 0.025 and 0.01.*]*

(b) The *P*-value exceeds 0.01 so we cannot reject the hypothesis.

The hypothesis that the model describes the data stands.

15 (i) (a)

(1) mean $\mu = \frac{4}{\lambda} \quad \therefore \bar{X} = \frac{4}{\lambda} \quad \Rightarrow \tilde{\lambda} = \frac{4}{\bar{X}}$

(2) $L(\lambda) = \text{const.}\lambda^{4n}(\Pi x_i)^3 e^{-\lambda \Sigma x_i}$

$\log L = const. + 4n \log \lambda + 3\log(\Pi x_i) - \lambda \Sigma x_i$

$\frac{\partial}{\partial \lambda} \log L = \frac{4n}{\lambda} - \Sigma x_i = 0 \quad \Rightarrow \hat{\lambda} = \frac{4n}{\Sigma X_i} = \frac{4}{\bar{X}}$

(b)

(1) $M_X(t) = E(e^{Xt}) = (1 - \frac{t}{\lambda})^{-4}$ from Green book

$M_{\Sigma X_i}(t) = E(e^{(\Sigma X_i)t}) = \prod_{i=1}^{n} E(e^{X_i t}) = (1 - \frac{t}{\lambda})^{-4n}$

$Y = 2n\lambda\bar{X} = 2\lambda\Sigma X_i$

$M_Y(t) = E(e^{(\Sigma X_i)2\lambda t}) = (1 - \frac{2\lambda t}{\lambda})^{-4n} = (1 - 2t)^{-4n}$

$\therefore 2n\lambda\bar{X} \sim \chi^2_{8n}$ being gamma($4n$, 1/2)

(2) $P(\chi^2_{0.975,8n} < 2n\lambda\bar{X} < \chi^2_{0.025,8n}) = 0.95$

95% confidence interval for λ is $(\frac{\chi^2_{0.975,8n}}{2n\bar{X}}, \frac{\chi^2_{0.025,8n}}{2n\bar{X}})$

(ii) (a) $n = 10, \bar{x} = 242 \quad \Rightarrow 2n\bar{x} = 4840$

$\chi^2_{0.975,80} = 57.15$, $\chi^2_{0.025,80} = 106.6$

95% confidence interval for λ is $(\frac{57.15}{4840}, \frac{106.6}{4840})$ = (0.0118, 0.0220)

(b)

(1) Approx. 95% confidence interval for μ is $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$

(2) $242 \pm 1.96\frac{112}{\sqrt{10}} = 242 \pm 69.4 = (172.6, 311.4)$

but $\mu = \frac{4}{\lambda}$

Conversion gives $(\frac{4}{311.4}, \frac{4}{172.6})$ = (0.0128, 0.0232)

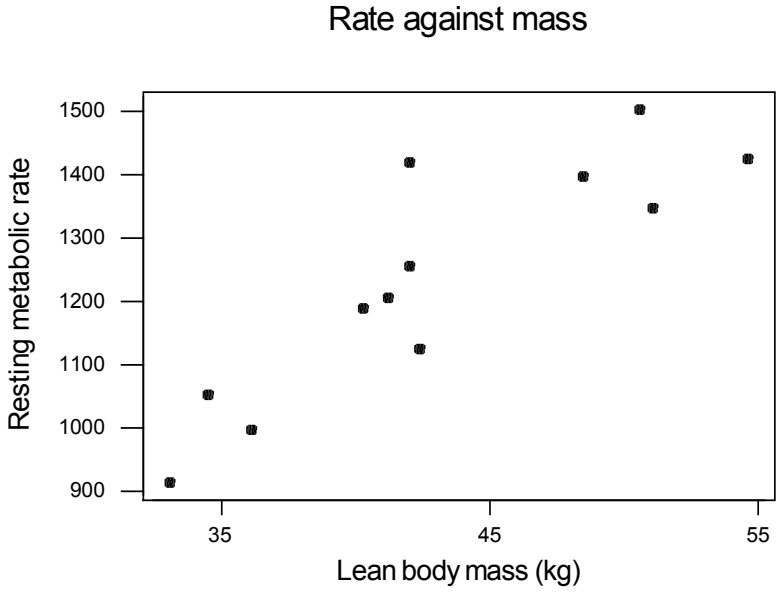
(3) This is quite close to the exact CI.

So the CLT approximation is quite good even though n=10 is not very large.

[Note: allow use of t in (ii)(b)

$242 \pm 2.262\frac{112}{\sqrt{10}} = (161.9, 322.1) \quad \Rightarrow (0.0124, 0.0247)$]

16 (i) x = lean body mass
y = resting metabolic rate



Linear relationship between x and y.

$$\sum x = 516.4\,, \quad \sum y = 14821$$

$$S_{xx} = \sum x^2 - \left(\sum x\right)^2 / n = 518.927$$
$$S_{yy} = \sum y^2 - \left(\sum y\right)^2 / n = 389954.92$$
$$S_{xy} = \sum xy - \left(\sum x\right)\left(\sum y\right) / n = 12467.767$$

(ii) Model: $y = \alpha + \beta x$

The least squares estimates of α and β:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{12467.8}{518.927} = 24.03$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{14821}{12} - (24.03)\frac{516.4}{12} = 201.2$$

$y = 201.2 + 24.03x$.

(iii) 95% confidence interval for β: $\hat{\beta} \pm t_{10}(2.5\%)\text{s.e.}\left(\hat{\beta}\right)$

$$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}, \hat{\sigma}^2 = \left(S_{yy} - \frac{S_{xy}{}^2}{S_{xx}}\right)/(n-2) = 9040.4$$

$\therefore \text{s.e.}(\hat{\beta}) = 4.17$

$24.03 \pm 2.228(4.17) = 24.03 \pm 9.299 = (14.7, \quad 33.3)$

Assuming normal errors with a constant variance.

(iv) 95% confidence intervals:

$$\hat{y} \pm t_{10}(2.5\%)\sqrt{\hat{\sigma}^2\left(\frac{1}{12} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}$$

(a) For $x = 50$ kg:

$$1402.5 \pm 2.228\sqrt{9040.4\left(\frac{1}{12} + \frac{(50-43.03)^2}{518.927}\right)}$$

$$1402.5 \pm 2.228 \times 40.0$$

$$= (1313, \quad 1492)$$

(b) For $x = 75$ kg:

$$2003.1 \pm 2.228\sqrt{9040.4\left(\frac{1}{12} + \frac{(75 - 43.03)^2}{518.927}\right)}$$

$$= 2003.1 \pm 2.228 \times 136.2$$

$$= (1700, \quad 2307)$$

(v) The confidence interval for $x = 50$ kg seems OK.

However, the confidence interval at $x = 75$ kg involves extrapolation. Care needed!