EXAMINATION

April 2005

Subject CT3 — Probability and Mathematical Statistics
Core Technical

EXAMINERS' REPORT

Introduction

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

M Flaherty
Chairman of the Board of Examiners

15 June 2005

1 $\Sigma x = 6212$, $\Sigma x^2 = 4186784$

$$\bar{x} = \frac{6212}{10} = £621.20$$

$$s = \sqrt{\frac{1}{9}\left(4186784 - \frac{6212^2}{10}\right)} = \sqrt{\frac{327889.6}{9}} = £190.87$$

2 (a) $P(A \cap B) = P(A) + P(B) - P(A \cup B) = \frac{1}{2} + \frac{1}{2} - \frac{3}{4} = \frac{1}{4}$

$P(A)P(B) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ so the events A and B are independent as $P(A \cap B) = P(A)P(B)$.

(b) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

$$= \frac{1}{2} + \frac{1}{2} + \frac{1}{3} - \frac{1}{4} - \frac{1}{6} - \frac{1}{6} + \frac{1}{12} = \frac{5}{6}$$

[OR $P(A \cup B \cup C) = P(A \cup B) + P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
$= \frac{3}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{6} + \frac{1}{12} = \frac{5}{6}$.]

[OR Use a Venn diagram]

3 $M(t) = (1 - 10t)^{-2}$
$M'(t) = (-2)(-10)(1-10t)^{-3} = 20(1 - 10t)^{-3}$
$M''(t) = (-60)(-10)(1-10t)^{-4} = 600(1-10t)^{-4}$ Putting $t = 0 \Rightarrow E[X^2] = 600$
$M'''(t) = (-2400)(-10)(1-10t)^{-5} = 24000(1-10t)^{-5}$ Putting $t = 0 \Rightarrow E[X^3] = 24000$

[OR use the power series expansion $M(t) = 1 + 20t + 600t^2/2! + 24000t^3/3! + \ldots$]
[OR use the result on $E[X^r]$ for a gamma(2,0.1) variable in the Yellow Book]

4 $\bar{X} \sim N(25, 0.25)$

$$P(\bar{X} > 26) = P\left(Z > \frac{26-25}{0.5}\right) = P(Z > 2) = 1 - 0.97725 = 0.02275$$

5 $$\frac{(n-1)S^2}{\sigma^2} = 5S^2 \sim \chi^2_{20}$$

$V[5S^2]$ = variance of χ^2_{20} = 40, so $V[S^2]$ = 40/25 = 1.6

6 $$\hat{p} = \frac{x}{n} = \frac{172}{200} = 0.86$$

95% CI is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\Rightarrow 0.86 \pm 1.96(0.0245) \Rightarrow 0.86 \pm 0.048$ or $(0.812, 0.908)$

7 Let $S = X_1 + X_2 + \ldots + X_N$ be the total claim amount.

Note that $E[N] = V[N] = 10$, $E[X] = 4/(1/5) = 20$, $V[X] = 4/(1/5)^2 = 100$

$E[S] = E[N]E[X] = 10(20) = 200$

$\therefore$ mean of total claim amount = £20,000.

$V[S] = E[N]V[X] + V[N]\{E[X]\}^2$
$= 10(100) + 10(20)^2 = 5000$

$\therefore$ s.d.$[S]$ = 70.71 $\therefore$ s.d. of the total claim amount = £7,071

8 (i) Width of 95% confidence interval: $\pm 1.96\frac{120}{\sqrt{n}}$ [or $2 \times 1.96\frac{120}{\sqrt{n}} = 3.92\frac{120}{\sqrt{n}}$]

$\therefore \pm 1.96\frac{120}{\sqrt{100}} = \pm 23.52$

[or $2 \times 1.96\frac{120}{\sqrt{100}} =$ £47.04]

(ii) For the width of a 95% confidence interval to be at most "±10" we require

$1.96\frac{120}{\sqrt{n}} \le 10$

$\sqrt{n} \ge \frac{1.96 \times 120}{10} = 23.52$, $n \ge 553.19$

i.e., take the sample size as 554.

(iii) The confidence interval in (ii) in narrower — to achieve this we require a much larger sample size.

9 (i) $\bar{X}$ approx ~ $N\left(\mu, \frac{3^2}{n}\right)$ for large n by the central limit theorem.

$0.05 \approx P(\text{reject } H_0|\mu = 0.8) = P(\bar{X} > k|\mu = 0.8)$

$$= 1-\Phi\left(\frac{k-0.8}{3/\sqrt{n}}\right)$$

$$\therefore \Phi\left(\frac{k-0.8}{3/\sqrt{n}}\right) \approx 0.95$$

and

$$0.1 \approx P(\text{do not reject } H_0|\mu = 1.2)$$

$$= P(\bar{X} < k \mid \mu = 1.2)$$

$$= \Phi\left(\frac{k-1.2}{3/\sqrt{n}}\right)$$

(ii) Significance level $\alpha = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = P(\bar{X} > k \mid \mu = 1)$

$$\approx 1-\Phi\left(\frac{1.025-1}{3/\sqrt{482}}\right) = 1-\Phi(0.18) = 1-0.57 = 0.43.$$

The significance level of the test is very high (43%).

10 (i) X takes positive values only so to have such a relatively high standard deviation the distribution must be positively skewed with sizeable probability associated with high values (i.e. the model embraces high claim sizes; the density has a long or heavy tail).

(ii) (a) Solving $r = F(x) \Rightarrow (1 + x/10) = (1-r)^{-0.2} \Rightarrow x = 10[(1-r)^{-0.2} - 1]$

(b) $R \sim U(0,1) \Rightarrow 1 - R \sim U(0, 1)$ so $(1 - r)$ is also a random number from (0, 1), so we can use $1 - r$ in place of r, giving the formula $x = 10\left[r^{-0.2} - 1\right]$

(c) $r = 0.0016 \Rightarrow$ claim = 262390
$r = 0.5154 \Rightarrow$ claim = 14175

11 (i) $SS_T = 77249 - 1203^2/20 = 4888.55$

$SS_B = 387^2/8 + 254^2/4 + 270^2/4 + 292^2/4 - 1203^2/20 = 2030.675$

$\therefore\ SS_R = 2857.875$

H_0: no treatment effects (i.e. population means are equal) v H_1: not H_0

Analysis of Variance

Source	DF	SS	MS	F
Factor	3	2031	677	3.79
Error	16	2858	179	
Total	19	4889		

$F_{3,16}(0.05) = 3.239$, $F_{3,16}(0.01) = 5.292$

P-value is lower than 0.05 (but higher than 0.01), so we can reject H_0 at least at the 5% level of testing (*Note:* actually *P-value* is 0.032). The data do indicate significant differences amongst the treatment means.

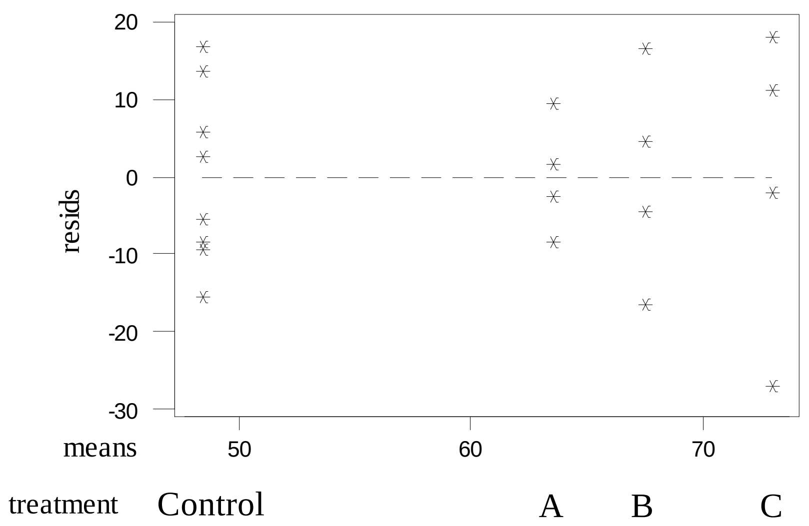
(ii) (a) Residual = observed value – treatment mean

Treatment means are: Control 48.375, A 63.5, B 67.5, C 73.0

Missing values are:

Control	***–5.4***	***–8.4***	16.6	2.6	–15.4	–9.4	5.6	13.6
Preparation A	9.5	–8.5	–2.5	1.5				
Preparation B	16.5	***–4.5***	–16.5	4.5				
Preparation C	***–27***	***18***	***11***	–2				

(b)



(c) Observations Y_{ij} (j^{th} value for treatment i) are <u>independent and normally</u> distributed with variance σ^2 which is <u>constant</u> across treatments.

(d) The assumptions seem reasonable — with the exception of the constant variance assumption, which is questionable — the data for preparation A appear to be less variable than the data for the other treatments.

(iii) The control mean <u>is lower</u> than <u>all</u> three treatment means (48.4 v 63.5, 67.5, 73.0) so there is prima facie evidence to support the suggestion.

One could perform a two-sample *t*-test of "control mean = treatment mean" by combining the data for the 3 preparations (and using samples of sizes 8 and 12).

12 (i) (a) The probability function for the zero-truncated Poisson distribution is given by

$$P(Y=y\,|\,Y>0)=\frac{P(Y=y \quad \text{and} \quad Y>0)}{P(Y>0)}$$

$$=\frac{\theta^y e^{-\theta}}{y!(1-P(Y=0))}$$

$$=\frac{\theta^y e^{-\theta}}{y!(1-e^{-\theta})} \quad (y=1,2,\ldots).$$

(b) Expectation of Y:

$$E[Y]=\sum_{y=1}^{\infty} y\frac{\theta^y e^{-\theta}}{y!(1-e^{-\theta})}$$

$$=\frac{\theta}{(1-e^{-\theta})}\left[\sum_{z=0}^{\infty}\frac{\theta^z e^{-\theta}}{z!}\right] \quad (z=y-1)$$

$$=\frac{\theta}{(1-e^{-\theta})}[1]=\frac{\theta}{(1-e^{-\theta})}.$$

(ii) (a) The log likelihood function for θ is:

$$\log L(\theta) = \sum_{i=1}^{n} y_i \log\theta - n\theta - n\log(1-e^{-\theta}) + \text{constant}$$

$$\frac{d\log L(\theta)}{d\theta} = \frac{n\overline{y}}{\theta} - n - n\frac{e^{-\theta}}{1-e^{-\theta}}$$

$\therefore$ the ML estimate is determined by the solution of the equation

$$\frac{d\log L(\theta)}{d\theta} = 0 \Rightarrow \overline{y} - \theta - \frac{\theta e^{-\theta}}{1-e^{-\theta}} = 0$$

As this equation may be rewritten as

$$\overline{y} = \frac{\theta}{1-e^{-\theta}} \text{ and } E[Y] = \frac{\theta}{1-e^{-\theta}}$$

the ML estimate is the same as the method of moments estimate.

(b) $$\frac{d^2}{d\theta^2}\log L(\theta) = -\frac{n\overline{y}}{\theta^2} + n\frac{e^{-\theta}}{(1-e^{-\theta})^2}$$

and since $E[\overline{Y}] = E[Y] = \dfrac{\theta}{(1-e^{-\theta})}$, the Cramer-Rao lower bound is given by,

$$CR\text{lb} = \frac{1}{-E\left[\dfrac{d^2}{d\theta^2}\log L(\theta)\right]} = \frac{1}{n\left(\dfrac{1}{\theta(1-e^{-\theta})} - \dfrac{e^{-\theta}}{(1-e^{-\theta})^2}\right)}$$

or $$\frac{\theta(1-e^{-\theta})^2}{n(1-e^{-\theta}-\theta e^{-\theta})}.$$

(iii) (a) The expected frequencies for the fitted zero-truncated Poisson model are given by

$$n\frac{\hat{\theta}^y e^{-\hat{\theta}}}{y!(1-e^{-\hat{\theta}})} \quad (y=1,2,\dots) \quad \text{where } \hat{\theta}=0.8925 \text{ and } n=2423$$

y	1	2	3	4	5	≥6	Total
e_i	1500.48	669.59	199.20	44.45	7.93	1.35	2423.00
f_i	1486	694	195	37	10	1	2423

$$\chi^2 = \sum\frac{(f_i-e_i)^2}{e_i} = \frac{(1486-1500.48)^2}{1500.48}+\cdots+\frac{(1-1.35)^2}{1.35} = 2.99$$

(on 4 df).

The Yellow Book gives that the probability value is greater than 50%, therefore there is no evidence to reject the null hypothesis, i.e. the model seems appropriate for the data.

[OR χ^2 = 2.68 on 3 df if ≥ 5 combined rather than ≥6.]

(b) As $\hat{\theta}$ approx. ~ $N(\theta, CR\text{lb})$ for large n, a 95% confidence interval for θ is given by
$\hat{\theta} \pm 1.96\sqrt{CR\text{lb}}$

$= 0.8925 \pm 1.96\sqrt{5.711574\times10^{-4}}$, since $CR\text{lb} = 5.711574 \times 10^{-4}$ at $\hat{\theta}$ = 0.8925,

= 0.8925 ± 1.96(0.0238989) = 0.8925 ± 0.04684
= (0.84566, 0.93934) 1

Then the 95% confidence interval for the mean of Y, $\dfrac{\theta}{1-e^{-\theta}}$, is given by

$$\left(\frac{0.84566}{1-e^{-0.84566}}, \frac{0.93934}{1-e^{-0.93934}}\right) = (1.48,\ 1.54).$$

13 (i) $S_{mm} = 129853.03 - (1136.1)^2/10 = 780.709$

$S_{ss} = 377700.62 - (1934.2)^2/10 = 3587.656$

$S_{ms} = 221022.58 - (1136.1)(1934.2)/10 = 1278.118$

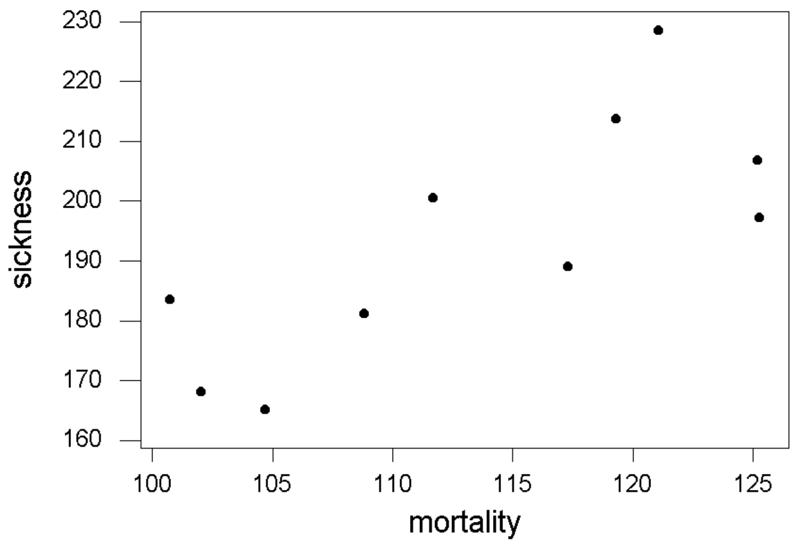
$$r = \frac{1278.118}{\sqrt{(780.709)(3587.656)}} = 0.764$$

H_0: $\rho = 0$ v. H_1: $\rho > 0$

$$t = \frac{r\sqrt{8}}{\sqrt{1-r^2}} = 3.35 \quad \Rightarrow \quad \text{Prob-value} = P(t_8 > 3.35) = 0.005 \text{ from tables.}$$

[OR use Fisher's transformation]

(ii) Given the issue of whether mortality can be used to predict sickness, we require a plot of sickness against mortality:



There seems to be an increasing linear relationship such that mortality could be used to predict sickness.

(iii) $\hat{\beta} = \dfrac{1278.118}{780.709} = 1.6371$ and $\hat{\alpha} = \dfrac{1}{10}[1934.2 - \hat{\beta}(1136.1)] = 7.426$

$$\hat{\sigma}^2 = \frac{1}{8}\{3587.656 - \frac{(1278.118)^2}{780.709}\} = 186.902$$

$$Var[\hat{\beta}] = \frac{\hat{\sigma}^2}{780.709} = 0.2394$$

Test H_0: $\beta = 2$ v. H_1: $\beta < 2$

$$t = \frac{1.6371 - 2}{\sqrt{0.2394}} = -0.74 \text{ on 8 df}$$

Prob-value large; no evidence to reject H_0: $\beta = 2$
So we can accept that the slope is as large as 2.

(iv) For a region with $m = 115$:

estimated expected $s = 7.426 + 1.6371(115) = 195.69$

with variance = $\hat{\sigma}^2\{\dfrac{1}{10} + \dfrac{(115 - 113.61)^2}{780.709}\} = 19.1528$

95% confidence limits are:

$195.69 \pm t_8(\text{s.e.})$

$\Rightarrow$ $195.69 \pm 2.306(4.376)$ $\Rightarrow$ 195.69 ± 10.09 or $(185.60, 205.78)$

END OF EXAMINERS' REPORT