EXAMINATION

April 2007

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 A Stocker
Chairman of the Board of Examiners

June 2007

Comments

Comments are given in the solutions that follow. Note that in some cases variations on the solutions given are possible — the examiners gave credit for all sensible comments and correct solutions.

The paper was well-answered overall and there are no particular topics that stand out as being poorly attempted. Similarly there were no particular misunderstandings evident widely, and no particular errors were made repeatedly and are worthy of comment.

In Question 13(iii) candidates were asked to plot a given set of residuals and comment. There was in fact a negative sign missing from the first residual (quoted as 1.6). The error was noted before marking commenced. No candidate was disadvantaged — all answers using 1.06 or -1.06 were accepted as being equally valid. There was no evidence in the scripts of any problem for candidates. The examiners wish to apologise for the minor error.

1 A Poisson random variable has mean = variance and this will be reflected in the sample mean and variance for a random sample.

Sample 2 has a very much higher variance than mean, whereas sample 1 has mean and variance approximately the same, so sample 1 is likely to be the one which comes from a Poisson distribution.

2 Approximate large sample confidence interval for the mean is given by

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

for 99% CI, $z_{\alpha/2} = 2.5758$

leading to $43.6 \pm 2.5758 \frac{82.2}{\sqrt{200}} \quad \Rightarrow \quad 43.6 \pm 15.0$

or (28.6, 58.6) or (£28,600, £58,600)

3 The mean number of claims per policy is $\bar{X} = \frac{123}{150} = 0.82$

Using the normal approximation to the Poisson distribution

approximate 95% confidence interval for λ is $\bar{X} \pm 1.96\sqrt{\frac{\bar{X}}{n}}$

$\rightarrow 0.82 \pm 1.96\sqrt{\frac{0.82}{150}} \quad \rightarrow 0.82 \pm 1.96(0.0739)$

$\rightarrow 0.82 \pm 0.145$ or $(0.675, 0.965)$

4 Answer = –0.982

(The relationship is now a negative one; the only change is the sign. An answer of +0.982 gets 1 mark.)

5 Let N = number of claims in the six months

Let X = a single claim size

Let S = sum of claim sizes for the six months

Then $N \sim$ Poisson(6)

$E(S) = E(N)E(X) = (6)(80) = £480$

$V(S) = E(N)V(X) + V(N)[E(X)]^2 = (6)(80^2) + (6)(80^2) = 76800$

$\therefore sd(S) = £277$

6 (i) $M_X(t) = E[e^{tx}]$

$$= \sum_{x=0}^{\infty} e^{tx} \frac{4}{5}\left(\frac{1}{5}\right)^x = \frac{4}{5}\sum_{x=0}^{\infty} \left(\frac{e^t}{5}\right)^x,$$

and for $e^t < 5$,

$$M_X(t) = \frac{4}{5}\frac{1}{1 - e^t/5} = 4\left(5 - e^t\right)^{-1}.$$

(ii) $M'(t) = 4e^t\left(5 - e^t\right)^{-2}$

Mean is given by $E(X) = M'(0)$

$$\therefore E[X] = 4e^0\left(5 - e^0\right)^{-2} = \frac{1}{4}.$$

[OR, by expansion as a power series.]

7 (i) Let X = the sum repaid for a single certificate.

$E(X) = 10(0.99) + 20(0.01) = 10.1$

$E(X^2) = 10^2(0.99) + 20^2(0.01) = 103$

$\therefore V(X) = 103 - 10.1^2 = 0.99 \quad \therefore sd(X) = 0.9950$

(ii) Let S = the sum repaid for 200 certificates.

$\therefore E(S) = 200(10.1) = 2020, \quad V(S) = 200(0.99) = 198 \quad \therefore sd(X) = 14.07$

$$P(S > 2040) = P(Z > \frac{2040 - 2020}{14.07} = 1.42)$$

$= 1 - 0.9222 = 0.0778$

(iii) $N \sim \text{binomial}(200, 0.01) \approx \text{Poisson}(2)$

$P(S > 2040) = P(N > 4)$

$= 1 - P(N \leq 4) = 1 - 0.94735 = 0.0527$

(iv) Clearly the Poisson approximation to the binomial is better than the Central Limit Theorem approximation.

OR:

Since S is discrete and increases in steps of 10, one can argue for the use of a continuity correction in (ii) above:

$$P\left(S > 2040\right) = P\left(Z \geq \frac{2045 - 2020}{14.07}\right) = P\left(Z > 1.78\right)$$

$= 1 - 0.96246 = 0.0375$

(Either approach is acceptable for the marks.)

8 (i) Mean $= \int_0^\infty x \frac{\alpha}{(1+x)^{\alpha+1}} dx$

$$= \int_0^\infty (1+x) \frac{\alpha}{(1+x)^{\alpha+1}} dx - \int_0^\infty 1 \frac{\alpha}{(1+x)^{\alpha+1}} dx$$

$$= \frac{\alpha}{\alpha-1} \int_0^\infty \frac{(\alpha-1)}{(1+x)^{(\alpha-1)+1}} dx - 1$$

$$= \frac{\alpha}{\alpha-1} - 1 = \frac{1}{\alpha-1}$$

(ii) Equate population mean to sample mean: $\frac{1}{\alpha-1} = \bar{x}$

Solve to get $\alpha = 1 + \frac{1}{\bar{x}}$, so MME = $1 + \frac{1}{\bar{X}}$

9 (i) $\text{Cov}(X, Y+Z) = \text{Cov}(X,Y) + \text{Cov}(X,Z) = 0$

[*Note:* The simple statement of the answer "0" is acceptable for the single mark available.]

(ii) $\text{Cov}(Z, 3X - 2Y) = 3\text{Cov}(Z,X) - 2\text{Cov}(Z,Y) = 0 - 2\rho_{YZ}\sigma^2$
$= -2 \times 0.5 \times 4 = -4$

(iii) $V[3X - 2Y + Z] = (9 + 4 + 1)\sigma^2 - 12\text{Cov}(X,Y) + 6\text{Cov}(X,Z) - 4\text{Cov}(Y,Z)$
$= 14(4) - 4(0.5)(4) = 48$

10 (i) $\hat{\mu} = \bar{Y}_{..} = \frac{871.9}{16} = 54.494$

$$\hat{\tau}_1 = \bar{Y}_{1.} - \bar{Y}_{..} = \frac{284.5}{5} - \frac{871.9}{16} = 2.406$$

$$\hat{\tau}_2 = \bar{Y}_{2.} - \bar{Y}_{..} = \frac{223.1}{4} - \frac{871.9}{16} = 1.281$$

$$\hat{\tau}_3 = \bar{Y}_{3.} - \bar{Y}_{..} = \frac{159.8}{3} - \frac{871.9}{16} = -1.227$$

$$\hat{\tau}_4 = \bar{Y}_{4.} - \bar{Y}_{..} = \frac{204.5}{4} - \frac{871.9}{16} = -3.369$$

(ii) $SS_T = \sum_i \sum_j y_{ij}^2 - \frac{Y_{..}^2}{n} = 120.430$

$$SS_B = \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 = (5 \times 2.406^2) + (4 \times 1.281^2) + (3 \times 1.227^2) + (4 \times 3.369^2) = 85.425$$

$$[\text{OR} : SS_B = \sum_i \left(\frac{Y_{i.}^2}{n_i} \right) - \frac{Y_{..}^2}{n} = 85.428]$$

$$SS_R = SS_T - SS_B = 35.002$$

The ANOVA table is:

Source	*DF*	*SS*	*MS*	*F*
Company (between treatments)	3	85.428	28.476	9.763
Residual	12	35.002	2.917	
Total	15	120.430		

At the 5% significance level, $F_{0.05,3,12} = 3.490$ (or $F_{0.01,3,12} = 5.953$)

Since $F = 9.763 > 3.490$, there is evidence against the null hypothesis, and we conclude that there are differences in the mean sums insured by the companies.

11 (i) $L(\underline{x}) = \dfrac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$

$\Rightarrow \ell(\lambda) = \log L(\lambda) = -n\lambda + \left(\sum x_i\right)\log\lambda + \text{constant}$

$\Rightarrow \dfrac{d\ell}{d\lambda} = -n + \dfrac{\sum x_i}{\lambda} = 0 \Rightarrow \hat{\lambda} = \dfrac{\sum X_i}{n} = \bar{X}$

(ii) $\dfrac{d^2\ell}{d\lambda^2} = -\dfrac{\sum x_i}{\lambda^2}$

$\Rightarrow -E\left[\dfrac{d^2\ell}{d\lambda^2}\right] = \dfrac{1}{\lambda^2}E\left[\sum X_i\right] = \dfrac{n\lambda}{\lambda^2} = \dfrac{n}{\lambda}$

$\Rightarrow \text{CRlb} = \dfrac{\lambda}{n}.$

(iii) (a) $E\left[\hat{\lambda}\right] = E\left[\bar{X}\right] = E\left[X\right] = \lambda$

$V\left[\hat{\lambda}\right] = V\left[\bar{X}\right] = \dfrac{V\left[X\right]}{n} = \dfrac{\lambda}{n}$,which is CRlb.

(b) The theory of asymptotic distributions of MLEs (and in this case the CLT) gives $\hat{\lambda} \sim N$ approximately, for large n so $\hat{\lambda} \sim N\left(\lambda, \dfrac{\lambda}{n}\right)$, approximately.

(iv) (a) Large sample approximate 95% CI for λ is given by

$\hat{\lambda} \pm \left(1.96 \times s.e.\left(\hat{\lambda}\right)\right)$ i.e. $\bar{x} \pm \left(1.96 \times s.e.\left(\bar{x}\right)\right)$

$s.e.\left(\hat{\lambda}\right) = \sqrt{\dfrac{\lambda}{n}}$ which we can estimate by using $\bar{x}$ to estimate λ, giving the estimated standard error $e.s.e.\left(\hat{\lambda}\right) = \sqrt{\dfrac{\bar{x}}{n}}$

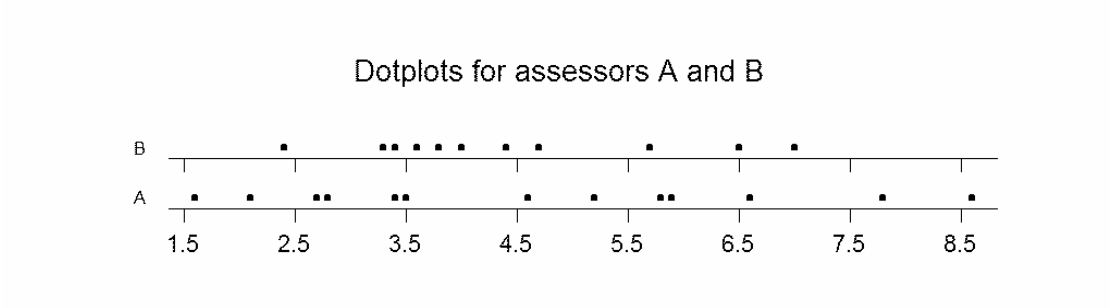
With $n = 100$, we get the 95% CI as

$\bar{x} \pm \left(1.96 \times \sqrt{\dfrac{\bar{x}}{100}}\right)$ *i.e.* $\bar{x} \pm 0.196\sqrt{\bar{x}}$.

(b) $\bar{x} = 215/100 = 2.15$

CI is $2.15 \pm 0.196(2.15)^{1/2}$ i.e. 2.15 ± 0.287 i.e. (1.86, 2.44)

12 (i)



dotplots on same scale are most suitable

[alternatively boxplots are acceptable]

(ii) (a) Let μ_A = mean initial estimate for this type of water damage for assessor *A* and μ_B = mean initial estimate for this type of water damage for assessor *B*.

$H_0 : \mu_A = \mu_B$ *v* $H_1 : \mu_A \neq \mu_B$

(b) dotplots show that normality assumption is reasonably valid
dotplots perhaps cast doubt on equal variances assumption

(c) test statistic is $t = \dfrac{\bar{x}_A - \bar{x}_B}{s_p\sqrt{\dfrac{1}{n_A}+\dfrac{1}{n_B}}} \sim t_{n_A+n_B-2}$ under H_0

From data: $\bar{x}_A = \dfrac{60.6}{13} = 4.662$, $s_A^2 = \dfrac{1}{12}(340.92 - \dfrac{60.6^2}{13}) = 4.8692$

$\bar{x}_B = \dfrac{48.8}{11} = 4.436$, $s_B^2 = \dfrac{1}{10}(236.80 - \dfrac{48.8^2}{11}) = 2.0305$

and $s_p^2 = \dfrac{12(4.8692)+10(2.0305)}{22} = 3.5789 \quad \therefore s_p = 1.8918$

observed $t = \dfrac{4.662 - 4.436}{1.8918\sqrt{\dfrac{1}{13}+\dfrac{1}{11}}} = \dfrac{0.226}{0.775} = 0.29$ on 22 df

Clearly *P*-value is very large, or noting that $t_{22}(40\%) = 0.2564$, then *P*-value is just a bit less than 0.8.

(d) So there is no evidence at all of any difference between assessors *A* and *B* as regards their mean initial estimates for this type of water damage.

(iii) (a) $H_0: \sigma_A^2 = \sigma_B^2$ v $H_1: \sigma_A^2 \neq \sigma_B^2$

(b) as in (i) dotplots show that normality assumption is reasonably valid

(c) test statistic is $F = \dfrac{s_A^2}{s_B^2} \sim F_{n_A-1, n_B-1}$ under H_0

observed $F = \dfrac{4.8692}{2.0305} = 2.40$ on 12,10 df

$F_{12,10}(10\%) = 2.284$ and $F_{12,10}(5\%) = 2.913$

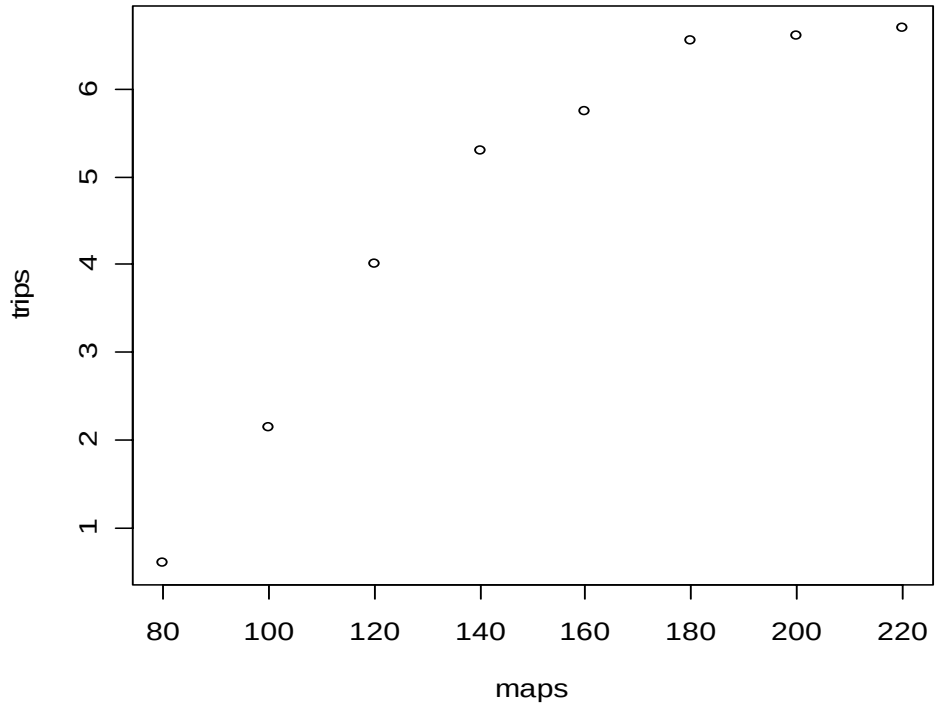
Thus *P*-value is between 0.10 and 0.20.

(d) So there is no real evidence of any difference between assessors *A* and *B* as regards the variances of their initial estimates for this type of water damage.

This validates the possibly doubtful assumption required in part (ii).

(iv) Overall there is no real evidence to distinguish any differences in the initial estimates for this type of water damage for the two assessors *A* and *B*.

13 (i) The scatterplot is shown below.



The plot suggests that there is a positive relationship between the increase in bus use and the number of maps distributed. The increase seems to be reasonably linear up to around 180000 maps, after which point it seems to level off (overall, relationship seems curved, possibly quadratic).

(ii) $S_{xx} = \Sigma x^2 - (\Sigma x)^2/n = 196800 - (1200)^2/8 = 16800$

$S_{yy} = \Sigma y^2 - (\Sigma y)^2/n = 213.4875 - (37.65)^2/8 = 36.29719$

$S_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)/n = 6378 - (1200)(37.65)/8 = 730.5$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{6}\left(36.29719 - \frac{(730.5)^2}{16800}\right) = 0.75558$$

$$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0.75558}{16800}} = 0.006706$$

To test $H_0 : \beta = 0$ v $H_1 : \beta \neq 0$, the test statistic is

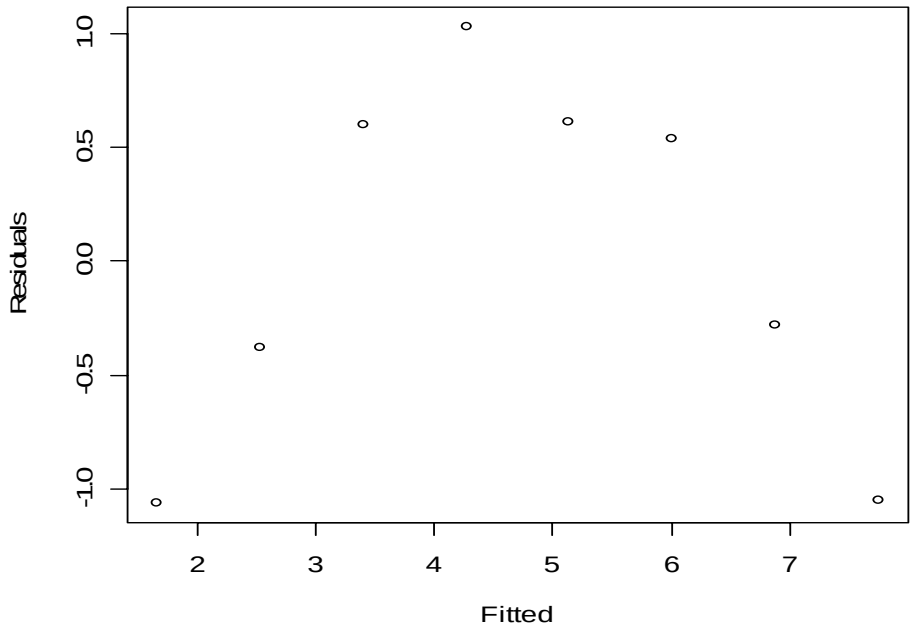
$$\frac{\hat{\beta} - 0}{\text{s.e.}(\hat{\beta})} = \frac{0.04348}{0.006706} = 6.484,$$

and under the assumption that the errors of the regression are *i.i.d.* $N(0, \sigma^2)$ random variables, it has a *t* distribution with $n - 2 = 6$ df.

From statistical tables we find $t_{6,0.025} = 2.447$ (or, $t_{6,0.005} = 3.707$).

Therefore, there is strong evidence against H_0. We conclude that a straight line representation of the relationship between the increase in bus use and the number of maps distributed would have a non-zero slope.

(iii) The plot is shown below.



Negative residuals are associated with the fitted values at the two ends of the data set, suggesting that the model is inadequate. Pattern suggests that a quadratic model might be appropriate.

(iv) Predicted value is $\hat{y} = -1.816 + 0.04348 \times 250 = 9.054$

This uses extrapolation on the fitted regression line. The prediction is probably not valid, especially as the linear model does not seem adequate.

END OF EXAMINERS' REPORT