EXAMINATION

April 2006

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

June 2006

Comments

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

Question 6

This question was the "worst" one on the paper as regards quality of answers. The question linked the concept of a conditional distribution with the simulation of observations of normal random variables (Core Reading Unit 6, section 1.3 and Unit 4, section 5.2). There were few good answers. Many candidates simply did not submit answers, suggesting that they were not familiar with the basic approach to the simulation of observations — despite the fact that there were short questions on this topic in both immediately previous papers, for which solutions are readily available.

Question 7

The likelihood function in this question is not of "standard" form and expressing and graphing it correctly requires a good understanding of the likelihood concept. Many candidates did not think clearly about the range of values of θ for which the likelihood is positive and for which it is zero and so got the wrong graph.

Question 8

Many candidates ignored the fact that "only the first three claims in any one year are paid". Suppose Y denotes the numbers of claims which arise, then $Y \sim Poisson(0.8)$. Suppose X denotes the number of claims which are paid. Many candidates worked with the set of probabilities $P(X = i) = P(Y = i), i = 0,1,2,3$.

But these four probabilities do not sum to 1 and so do not provide a proper probability distribution for X. What is required is the set of four probabilities $P(X = i) = P(Y = i), i = 0,1,2$ together with $P(X = 3) = P(Y \geq 3)$.

Question 9

The wording of the question made it clear that candidates could assume the mgf of a Poisson random variable and, armed with this information, should use a "conditional expectation argument". Full marks were not awarded to candidates who jumped into the middle of the argument by assuming the mgf of a <u>compound</u> <u>Poisson</u> random variable.

Question 10

Candidates should be aware that when constructing a histogram with unequal group widths one must ensure that the areas (and not the heights) of the rectangles are proportional to the frequencies.

In part (ii), many candidates calculated a confidence interval for a different proportion to the one asked for.

Question 11

Many candidates were unsure of the definition of the coefficient of skewness (Core Reading Unit 3, section 3.4).

Question 12

In part (ii)(a), many candidates calculated S_{yy}, S_{xx}, and S_{xy} *but did not go on to calculate the regression and error sums of squares (SSREG, SSRES) as asked for. In part (iii)(a) many candidates failed to make any of the most pertinent possible comments (but credit was given for relevant comments other than those given here).*

1 $n = 40$, so the median is the 20.5^{th} observation, which is ½(7.7+7.8) = 7.75.
This represents €77,500.

2 If X is the number in the sample with group A, then X has a binomial (300, 0.45) distribution, so

$E[X] = 300 \times 0.45 = 135$ and $Var[X] = 300 \times 0.45 \times 0.55 = 74.25$.

Then, using the continuity correction,

$$P(X > 115) = P(X > 115.5) \approx 1 - \Phi\left(\frac{115.5 - 135}{\sqrt{74.25}}\right) = 1 - \Phi(-2.26) = \Phi(2.26) = 0.99.$$

3 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ so here $9S^2 \sim \chi^2_9$

$$P\left(S^2 > 1\right) = P\left(\chi^2_9 > 9\right)$$

$$= 1 - 0.5627 = 0.437 \quad \text{(tables p165)}$$

4
$$\hat{\mu} = \frac{1}{3}(276.7 + 254.6 + 296.3) = 275.87$$
$$\hat{\tau}_1 = 276.7 - 275.87 = 0.83$$
$$\hat{\tau}_2 = 254.6 - 275.87 = -21.27$$
$$\hat{\tau}_3 = 296.3 - 275.87 = 20.43$$

$$\hat{\sigma}^2 = \frac{SS_R}{27} = \frac{15508.6}{27} = 574.4$$

5 (i) Let X = number of policies with claims
So $X \sim$ binomial(25, p).
Poisson approximation is $X \approx$ Poisson(25p).

(a) using Poisson(2.5)
$P(X \leq 4) = 0.89118$ from tables [or evaluation]

(b) using Poisson(5)
$P(X \leq 4) = 0.44049$ again from tables

(ii) in (a) error is $0.8912 - 0.9020 = -0.0108$
in (b) error is $0.4405 - 0.4207 = 0.0198$
The approximation is valid for "small p", and, as p is smaller in (a), this gives a better approximation as noted with the smaller error.

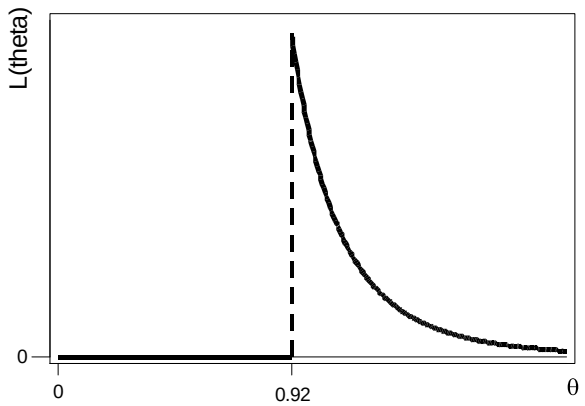
[Note: candidates may also comment on the fact that the sample size 25 is not "large" and so we would not expect the Poisson approximations to be very good anyway. In fact the key to the approximations is the small p and here the given approximations *are* quite good]

6 Solving $P(Z < z) = 0.5714 \Rightarrow z = 0.180 \quad \Rightarrow x = 200 + 10(0.180) = 201.80$
Solving $P(Z < z) = 0.8238 \Rightarrow z = 0.930 \quad \Rightarrow y = 201.80 + 0.930 = 202.73$
$\Rightarrow t = 201.80 \times 202.73 = 40911$
Solving $P(Z < z) = 0.3192 \Rightarrow z = -0.470 \quad \Rightarrow x = 200 + 10(-0.470) = 195.3$
Solving $P(Z < z) = 0.6844 \Rightarrow z = 0.480 \quad \Rightarrow y = 195.3 + 0.480 = 195.78$
$\Rightarrow t = 195.3 \times 195.78 = 38236$

7 (i) $L(\theta) = \left(\frac{1}{2\theta}\right)^n = c\left(\frac{1}{\theta}\right)^n$ for $-\theta < x_i < \theta$, $i = 1, 2, \ldots, n$ and $L(\theta) = 0$ otherwise

So, as θ increases from zero, $L(\theta)$ is zero until it reaches the largest observation in absolute value i.e. $\max |x_i|$, $i = 1, 2, \ldots, n$. For the data given, this value is 0.92.

It is then a decreasing function θ. Hence the graph is as below:



(ii) The maximum value of $L(\theta)$ is attained at the largest absolute value of the data. The ML estimate of θ is 0.92.

8 (i) By subtraction using entries in tables for Poisson(0.8), the probabilities for the Poisson distribution for 0, 1, 2 and ≥ 3 are: [or by evaluation]

0.44933, 0.35946, 0.14379 and (1 – 0.95258) = 0.04742

(ii) Let N = number of claims paid and let $X_1,\ldots, X_n$ be the claim amounts then $S = \Sigma X_i$ is the sum of the amounts.

$E[S] = E[N]E[X]$

Here $E[N] = 1(0.35946) + 2(0.14379) + 3(0.04742) = 0.7893$

and $E[X] = 2/1 = 2$ from gamma(2,1)

So $E[S] = (0.7893)(2) = 1.5786$ = £157.86

(iii) Given that $N > 0$, divide the probabilities in part (i) by (1 – 0.44933) = 0.55067 to give the probabilities for 1, 2 and 3 claims paid as:

0.6528, 0.2611 and 0.0861

$E[N] = 1(0.6528) + 2(0.2611) + 3(0.0861) = 1.4333$

So $E[S] = (1.4333)(2) = 2.8666$ = £286.66

9 (i) $M_S(t) = E[e^{tS}] = E[E[e^{tS}|N]]$

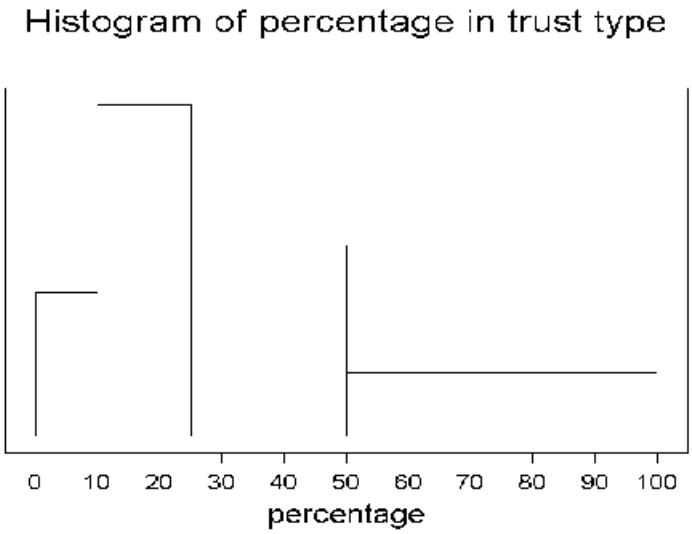
Now $E[e^{tS}|N = n] = E[\exp(tX_1 + \ldots + tX_n)] = \Pi\, E[\exp(tX_i)] = \{M_X(t)\}^n$
$\Rightarrow M_S(t) = E[\{M_X(t)\}^N] = E[\exp\{N\log M_X(t)\}] = M_N\{\log M_X(t)\}$
$= \exp[\lambda\{M_X(t) - 1\}]$ since $N \sim$ Poisson(λ)
$\therefore C_S(t) = \log M_S(t) = \lambda\{M_X(t) - 1\}$

(ii) $V[S] = C_S''(0) = \lambda\{M_X''(0)\} = \lambda E[X^2] = 20(10 + 20^2) = 8200$

OR $V[S] = E[N]V[X] + V[N]\{E[X]\}^2 = 20\times 10 + 20\times 20^2 = 8200$

10 (i) (a) The key feature of the histogram is that the areas of the four rectangles should be proportional to the frequencies.

See histogram below.



(b) Mean is calculated from the following frequency distribution:

x	5	17.5	37.5	75
f	22	76	73	49

$\Sigma f = 220,\ \Sigma fx = 7852.5 \qquad \therefore \bar{x} = \dfrac{7852.5}{220} = 35.7\%$

(ii) Estimated proportion is $\hat{p} = \dfrac{220}{650} = 0.338$ (or 33.8%)

95% confidence interval for underlying proportion is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{650}}$$

$$\Rightarrow \quad 0.338 \pm 1.96\sqrt{\frac{0.338(0.662)}{650}} \quad \Rightarrow \quad 0.338 \pm 0.036$$

as a percentage: 33.8% $\pm$ 3.6% or (30.2%, 37.4%)

(iii) Under the null hypothesis of no association between percentage in type of trust and satisfaction with current return, expected frequencies are

2.0	6.9	6.6	4.5	20
10.0	34.5	33.2	22.3	100
9.0	31.1	29.9	20.0	90
1.0	3.5	3.3	2.2	10
22	76	73	49	220

six are less than 5 which would invalidate a χ^2 test

(iv) expected frequencies (e) are

12.000	41.455	39.818	26.727
10.000	34.545	33.182	22.273

table of residuals ($o-e$) is

–3.000	–6.455	+3.182	+6.273
+3.000	+6.455	–3.182	–6.273

table of contributions to χ^2 is

0.750	1.005	0.254	1.472
0.900	1.206	0.305	1.767

giving $\chi^2 = 7.659$ on 3 d.f.

$\chi^2_3(5\%) = 7.815$ $\therefore$ must accept the null hypothesis that there is no relationship between percentage in type of trust and satisfaction with current return.

However this decision to accept is marginal at the 5% level and there is some evidence, but not strong, to suggest that satisfaction improves as the percentage increases.

11 (i) $\sigma^2 = E[X^2] - [E(X)]^2 = 12\theta^2 - (3\theta)^2 = 3\theta^2$

$\mu_3 = E[X^3] - 3\mu\, E[X^2] + 2\mu^3$

$= 60\theta^3 - 3(3\theta)(12\theta^2) + 2(3\theta)^3$

$= (60 - 108 + 54)\,\theta^3 = 6\theta^3$

coefficient of skewness = $\dfrac{\mu_3}{\sigma^3} = \dfrac{6\theta^3}{(\sqrt{3\theta^2})^3} = 1.155$

[OR: note that $X \sim$ gamma(3,1/θ) and use formulae in tables
so var = $3\theta^2$ and coef. of skew. = $\dfrac{2}{\sqrt{3}}$]

(ii) $L(\theta) = \prod_{i=1}^{n} \dfrac{x_i^2}{2\theta^3} \exp(-\dfrac{x_i}{\theta}) = \dfrac{\Pi x_i^2}{2^n \theta^{3n}} \exp\left(-\dfrac{\Sigma x_i}{\theta}\right)$

$\log L(\theta) = \log(\Pi x_i^2) - n\log 2 - 3n\log\theta - \dfrac{\Sigma x_i}{\theta}$

$\dfrac{\partial}{\partial\theta}\log L(\theta) = -\dfrac{3n}{\theta} + \dfrac{\Sigma x_i}{\theta^2}$

equate to zero: $\dfrac{3n}{\theta} = \dfrac{\Sigma x_i}{\theta^2} \quad \Rightarrow \quad \theta = \dfrac{\Sigma x_i}{3n}$

this clearly maximises $L(\theta)$ [or consider $\dfrac{\partial^2}{\partial\theta^2}\log L(\theta)$]

So MLE is $\hat{\theta} = \dfrac{\Sigma X_i}{3n} = \dfrac{\overline{X}}{3}$

$E\left[\hat{\theta}\right] = \dfrac{1}{3}E\left[\overline{X}\right] = \dfrac{1}{3}E\left[X\right] = \dfrac{1}{3}3\theta = \theta$ $\quad \therefore$ unbiased

(iii) (a) $\overline{x} = \dfrac{313.6}{50} = 6.272 \;\therefore\; \hat{\theta} = \dfrac{6.272}{3} = 2.091$

(b) $s^2 = \dfrac{1}{49}(2675.68 - \dfrac{313.6^2}{50}) = 14.465$

$\sigma^2 = 3\theta^2$ and $3\hat{\theta}^2 = 13.117$

s^2 is a bit larger but still quite close

(c) sample coefficient 1.149 is very close to the distribution value 1.155

(iv) (a) approximate 95% CI for μ is $\bar{x} \pm 1.96\sqrt{\frac{s^2}{n}}$

as $\mu = 3\theta$, divide by 3 for an approximate 95% CI for θ

$$\frac{1}{3}\left(\bar{x} \pm 1.96\sqrt{\frac{s^2}{n}}\right)$$

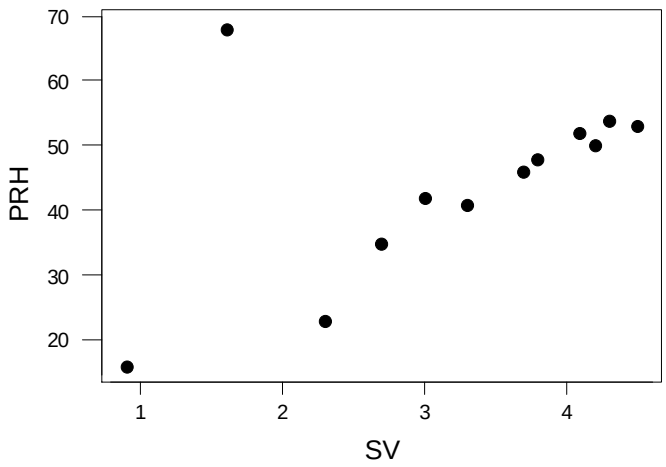
for data: $\frac{1}{3}\left(6.272 \pm 1.96\sqrt{\frac{14.465}{50}}\right)$

$\Rightarrow \quad 2.091 \pm 0.351 \quad$ or $\quad (1.740, 2.442)$

(b) $\sigma^2 = 3\theta^2 = 9.083$ at lower limit of 1.740
$= 17.890$ at upper limit of 2.442

$s^2 = 14.465$ is well within these values
confirming that s^2 is quite close to $3\hat{\theta}^2$.

12 (i)



(ii) (a) $SSTOT = S_{yy} = 25428 - 528^2/12 = 2196$

$S_{xx} = 137.16 - 38.4^2/12 = 14.28$, $S_{xy} = 1778.4 - (38.4 \times 528)/12 = 88.8$

$SSREG = 88.8^2/14.28 = 552.20$, $SSRES = 2196 - 552.20 = 1643.80$

(b) $R^2 = 552.2/2196 = 0.251$ (25.1%)

(c) $y = a + bx$: $\hat{b} = 88.8/14.28 = 6.2185$

$\hat{a} = 528/12 - (88.8/14.28)(38.4/12) = 24.101$

Fitted line is $y = 24.101 + 6.2185x$

(d) $s.e.(\hat{b}) = \left(\frac{1643.8/10}{14.28}\right)^{1/2} = 3.3928$

Observed $t = (6.2185 - 0)/3.3928 = 1.833 < t_{10}(0.025) = 2.228$
so we do not have evidence at the 5% level of testing to justify rejecting "$b = 0$" and concluding that the underlying slope is non-zero.

(iii) (a) Large change in slope (and intercept) of fitted line.

The total and error sums of squares are much reduced.

The fit of the linear regression model is much improved (R^2 is much increased — from 25.1% to 94.9%).

We have overwhelming evidence to justify concluding that the slope is non-zero.

(b) Fitted *PRH* value at *SV* = 3.5 is 3.757 + (11.377×3.5) = 43.577

$n = 11,\ \sum x = 38.4 - 1.6 = 36.8,\ \sum x^2 = 137.16 - 1.6^2 = 134.6$

$\Rightarrow S_{xx} = 11.4873$

s.e. of estimation = $\left[\left(\frac{1}{11} + \frac{(3.5 - 36.8/11)^2}{11.4873}\right)8.98\right]^{1/2} = 0.9138$

$t_9(0.025) = 2.262$

$\Rightarrow$ 95% CI for expected *PRH* is $43.577 \pm (2.262 \times 0.9138)$

i.e. 43.577 ± 2.067 or (41.51, 45.64)

END OF EXAMINERS' REPORT