INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2010 examinations

Subject CT3 — Probability and Mathematical Statistics
Core Technical

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.

T J Birse
Chairman of the Board of Examiners

December 2010

Comments

The paper was answered well and overall performance was satisfactory. However, some questions were poorly attempted. A number of candidates could not answer Question 1 correctly and efficiently – the question required basic knowledge of data summaries. Question 8 was not answered very well – answers to questions that require candidates to "show" a particular statement, need to demonstrate intermediate steps clearly and accurately. The same applies to Question 10, where deriving specific results regarding maximum likelihood estimation was not performed accurately by many candidates.

1 Sample mean = $(1.1 \times 57.2) + 8 = 70.92$
Sample standard deviation = $1.1 \times 7.3 = 8.03$

2 (i) Sample median is not affected by the fact that the last two observations are censored.

It is therefore given by the 5.5^{th} ranked observation, i.e. $(355 + 379)/2 = 367$ days.

(ii) We know that the last two observations have minimum values 432 and 463.

Using these two values the sample mean would be equal to 3679/10 = 367.9.

So, the sample mean is at least equal to 367.9 days.

3 (i) Using the negative binomial distribution, or from first principles,

$P(5 \text{ policies required}) = \binom{5-1}{2-1}(0.2)^2(0.8)^3 = 0.0819$

(ii) Expected number = mean of negative binomial distribution = $\frac{2}{0.2} = 10$

4 Working in units of £1,000, sum of 100 claim amounts S has $E[S] = 100 \times 4 = 400$ and $V[S] = 100 \times 0.5^2 = 25$, and so $S \sim N(400, 5^2)$ approximately.

$P(S < 407.5) = P(Z < 1.5) = 0.933$

5 Sample proportion = 29/200 = 0.145

99% CI is given by $0.145 \pm 2.5758\sqrt{\frac{0.145 \times 0.855}{200}}$ i.e. 0.145 ± 0.064

i.e. (0.081, 0.209).

6 $E[X] = E[E(X \mid Y)] = E[Y] = e^{\mu+\sigma^2/2}$

$V[X] = E[V(X \mid Y)] + V[E(X \mid Y)] = E[Y] + V[Y] = e^{\mu+\sigma^2/2} + e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$

7 (i) Method

$0 < u \le 0.4 \Rightarrow x = 0$
$0.4 < u \le 0.7 \Rightarrow x = 1$
$0.7 < u \le 0.9 \Rightarrow x = 2$
$0.9 < u \le 1 \Rightarrow x = 3$

We get $x = 1, 2, 0$

(ii) Setting $u = \frac{1}{1-e^{-1}}\left(1 - e^{-x^2}\right) \Rightarrow e^{-x^2} = 1 - \left(1 - e^{-1}\right)u$

$\Rightarrow x = \left[-\log\left[1 - \left(1 - e^{-1}\right)u\right]\right]^{1/2}$

$u = 0.8149 \Rightarrow x = 0.851$

8 (i) (a) (1) Let X_i be a claim amount.

Mgf of X_i is $M_X(t) = \left(1 - \frac{t}{1.25}\right)^{-1}$

Mgf of $S = \sum_{i=1}^{10} X_i$ is $M_S(t) = [M_X(t)]^{10} = \left(1 - \frac{t}{1.25}\right)^{-10}$,

which is the mgf of a gamma(10, 1.25) variable.

(2) Mgf of $2.5S$ is $E[e^{t(2.5S)}] = E[e^{(2.5t)S}] = M_S(2.5t) = (1-2t)^{-10}$,

which is the mgf of a gamma(10, ½) variable, i.e. χ^2_{20}.

(b) $P(\text{total} > £10{,}000) = P(S > 10) = P(\chi^2_{20} > 25) = 1 - 0.7986 = 0.2014$

(ii) (a) S has mean $\frac{10}{1.25} = 8$ and variance $\frac{10}{1.25^2} = 6.4$. So $S \approx N(8, 6.4)$

$P(S > 10) \cong P(Z > \frac{10-8}{\sqrt{6.4}} = 0.791) = 1 - 0.786 = 0.214$

(b) n is not particularly large for the use of the CLT, but the approximation is still quite close to the true probability.

9 (i) $P(X = k+1) = e^{-\lambda}\dfrac{\lambda^{k+1}}{(k+1)!} = e^{-\lambda}\dfrac{\lambda^k}{k!}\dfrac{\lambda}{k+1} = \dfrac{\lambda}{k+1}P(X = k)$.

(ii) Using $P(X = 0) = e^{-1.186}$, $P(X = 8 \text{ or more}) = 1 - \sum_{i=0}^{7} P(X = i)$, and the recurrent formula, we obtain:

K	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8 or more*
$P(X = k)$	0.3054	0.3623	0.2148	0.0849	0.0252	0.0060	0.0012	0.0002	4×10^{-5}
Expected, e_k	305.4	362.3	214.8	84.9	25.2	6.0	1.2	0.2	0.0

(iii) Combining the last 4 categories to obtain expected frequencies greater than 5, we have:

k	*0*	*1*	*2*	*3*	*4*	*5 or more*
No. of policies, f_k	310	365	202	88	26	9
Expected, e_k	305.4	362.3	214.8	84.9	25.2	7.4

This gives

$$\chi^2 = \sum \frac{(f_k - e_k)^2}{e_k}$$

$$= 0.0693 + 0.0201 + 0.7628 + 0.1132 + 0.0254 + 0.3459 = 1.3367$$

DF = 6 – 1 – 1 = 4, and from statistical tables, $\chi^2_{0.05,4} = 9.488$.

Therefore, we do not have evidence against the hypothesis that the number of claims comes from a Poisson(1.186) distribution.

(Alternatively if we only combine the last 3 categories, the expected frequencies for 5 and 6 or more policies are 6 and 1.4, with observed frequencies 6 and 3 respectively. These give $\chi^2 = 2.819$ on 5 DF, and with $\chi^2_{0.05,5} = 11.071$ the conclusion is the same as before.)

10 (i) $P(\text{yes}) = P(5,6)P(\text{yes}\,|\,\text{main question}) + P(1,2,3,4)P(\text{yes}\,|\,\text{coin question})$

$$= \frac{2}{6}p + \frac{4}{6}\cdot\frac{1}{2} = \frac{1}{3}(p+1)$$

(ii) (a) $L(p) \propto [\frac{1}{3}(p+1)]^{56}[1-\frac{1}{3}(p+1)]^{100-56}$

$$\propto (p+1)^{56}(2-p)^{44}$$

(b) $\log L = 56\log(p+1) + 44\log(2-p) + const.$

$$\frac{d}{dp}\log L = \frac{56}{p+1} - \frac{44}{2-p}$$

$$= \frac{56(2-p)-44(p+1)}{(p+1)(2-p)} = \frac{68-100p}{(p+1)(2-p)}$$

Equate to zero $\Rightarrow$ $68-100p = 0$ $\therefore \hat{p} = \frac{68}{100} = 0.68$

(iii) Due to the invariance property of MLEs $\frac{1}{3}(\hat{p}+1) = \hat{\theta}$

$$\therefore \frac{1}{3}(\hat{p}+1) = \frac{56}{100} \quad \therefore \hat{p} = \frac{168}{100} - 1 = \frac{68}{100} = 0.68$$

(iv) (a) $$\frac{d^2}{dp^2}\log L = -\frac{56}{(p+1)^2} - \frac{44}{(2-p)^2}$$

at $\hat{p} = 0.68$, $\frac{d^2}{dp^2}\log L = -\frac{56}{1.68^2} - \frac{44}{1.32^2} = -45.0938$

(b) $CRlb = \frac{-1}{-45.0938} = 0.02218$ and $\hat{p} \approx N(p, 0.02218)$

(c) Approximate 95% CI for p is $\hat{p} \pm 1.96\sqrt{0.02218}$
giving 0.68 ± 0.292

(v) (a) Approximate 95% CI for p is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{100}}$

giving

$$0.68 \pm 1.96\sqrt{\frac{0.68(1-0.68)}{100}} \Rightarrow 0.68 \pm 1.96(0.0466) \Rightarrow 0.68 \pm 0.091$$

(b) Less accurate estimation is the penalty paid for using the randomised response method.

11 (i) We want to test H_0: $\mu_A = \mu_B$ against H_1: $\mu_A \neq \mu_B$.

Data give: $\bar{y}_A = 56.1/12 = 4.675$, $\bar{y}_B = 59.1/12 = 4.925$

$s_A^2 = (266.33 - 56.1^2/12)/11 = 0.36932$,
$s_B^2 = (297.03 - 59.1^2/12)/11 = 0.54205$

Assuming that the two samples come from normal distributions with the same variance,

we first compute the pooled variance as $s_p^2 = \dfrac{11s_A^2 + 11s_B^2}{22} = 0.455685$

which gives $t = \dfrac{\bar{y}_A - \bar{y}_B}{s_p\sqrt{2/12}} = -0.907$.

Critical values at 5% level are $t_{22}(0.025) = -2.074$ and $t_{22}(0.975) = 2.074$ so we don't have evidence against H_0 and conclude that the mean delay time is the same for claims associated with the two causes of illness.

(ii) Distribution of times can be skewed to the right, and we need a log transformation to normalise the data (for test to be valid).

(iii) (a) CI is given by $\left(\dfrac{s_A^2/s_B^2}{F_{11,11}(0.025)}, (s_A^2/s_B^2) * F_{11,11}(0.025) \right)$

$F_{11,11}(0.025) = 3.478$ (using interpolation in the tables)
giving CI as (0.68134/3.478, 0.68134*3.478) = (0.196, 2.370).

(b) The value "1" is included in the 95% CI, meaning that the assumption of common variance made for the test is valid.

(iv) $SS_T = 952.64 - 183^2/36 = 22.39$
$SS_B = (56.1^2 + 59.1^2 + 67.8^2)/12 - 183^2/36 = 6.155$
$\Rightarrow SS_R = 22.39 - 6.155 = 16.235$

Source of variation	*d.f.*	*SS*	*MSS*
Between	2	6.155	3.078
Residual	33	16.235	0.492
Total	35	22.390	

F = 3.078/0.492 = 6.256 on (2,33) degrees of freedom.

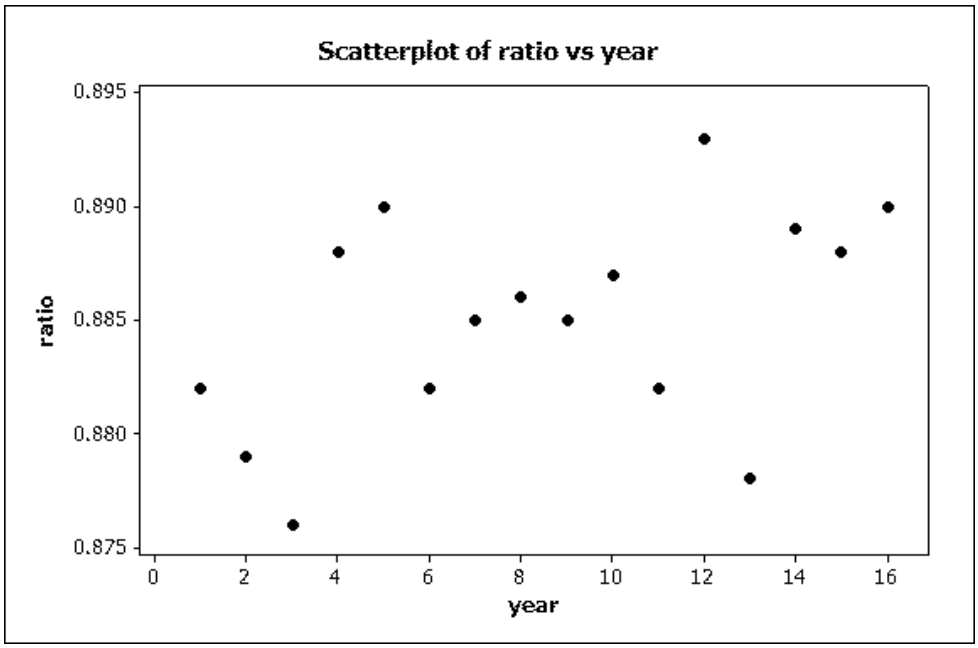
From tables, $F_{2,33}(0.05)$ is between 3.276 and 3.295, and $F_{2,33}(0.01)$ is between 5.289 and 5.336.

We have strong evidence against the null hypothesis and conclude that the three mean delay times are not equal.

(v) The assumptions are that the data come from normal populations with constant variance.

(vi) The plot suggests that the normality assumption is reasonable and that variance does not depend on cause. Test seems valid.

12 (i) Scatterplot with suitable axes and clearly labelled:



There does not appear to be much of a relationship, perhaps a slight increasing linear relationship but it is weak with quite a bit of scatter.

(ii) $n = 16$

$$S_{tt} = 1496 - \frac{136^2}{16} = 340$$

$$S_{yy} = 12.531946 - \frac{14.160^2}{16} = 0.000346$$

$$S_{ty} = 120.518 - \frac{(136)(14.160)}{16} = 0.158$$

$$\hat{\beta} = \frac{S_{ty}}{S_{tt}} = \frac{0.158}{340} = 0.0004647$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t} = \frac{14.160}{16} - (0.0004647)\frac{136}{16} = 0.88105$$

Fitted line is $y = 0.88105 + 0.000465t$

(iii) (a) s.e. $(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{tt}}}$ where $\hat{\sigma}^2 = \frac{1}{n-2}(S_{yy} - \frac{S_{ty}^2}{S_{tt}})$

$$\hat{\sigma}^2 = \frac{1}{14}(0.000346 - \frac{0.158^2}{340}) = 0.0000195$$

$$\therefore s.e.(\hat{\beta}) = \sqrt{\frac{0.0000195}{340}} = 0.000239$$

(b) Null hypothesis of "no linear relationship" is equivalent to H_0: $\beta = 0$

We use $t = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \sim t_{14}$ under H_0: $\beta = 0$

Observed $t = \frac{0.000465}{0.000239} = 1.95$ and $t_{0.025,14} = 2.145$

So we must accept H_0: no linear relationship at the 5% level.

(c) 95% CI is $0.000465 \pm 2.145 \times 0.000239$
giving 0.000465 ± 0.000513 or (–0.000048, 0.000978)

(iv) (a) Observed $t = \frac{0.000487}{0.000220} = 2.21$ – this is greater than $t_{0.025,14} = 2.145$

So we reject H_0: no linear relationship at the 5% level.

(b) 95% CI is $0.000487 \pm 2.145 \times 0.000220$
giving 0.000487 ± 0.000472 or (0.000015, 0.000959)

The two CIs overlap substantially, so there is no evidence to suggest that the slopes are different.

(c) Although the tests have different conclusions at the 5% level, the 100m observed t is only just inside the critical value of 2.145 and the 200m one is just outside. This in fact agrees with, rather than contradicts, the conclusion that the slopes are not different.

END OF EXAMINERS' REPORT