Subject CT3 — Probability and Mathematical Statistics
Core Technical

September 2009 Examinations

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.

R D Muckart
Chairman of the Board of Examiners

December 2009

Comments for individual questions are given with the solutions that follow.

© Faculty of Actuaries
© Institute of Actuaries

Comments

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

1

(i) We have $60 + f_2 + f_4 = 100$ and $\frac{7 + 2f_2 + 60 + 4f_4 + 90 + 60 + 35}{100} = 4$. 1

These give $f_2 = 40 - f_4$ and $2f_2 + 4f_4 = 148$

from which we obtain $f_2 = 6$ and $f_4 = 34$. 1

(ii) Median is equal to the midpoint between the 50^{th} and 51^{st} ordered observations, i.e. median = 4. 1

We have mean = median, suggesting that the distribution of these data is roughly symmetric. 1

2

(i) With sample space $\{(i,j), i = 1, \ldots, 6, j = 1, \ldots, 6, j \neq i\}$

(that is, i is the number on the first ticket selected, j that on the second selected) there are 30 equally likely outcomes.

Favourable outcomes are (2,6), (3,5), (5,3), (6,2)

so probability = 4/30 = 2/15 = 0.133 2

(ii) Favourable outcomes are

(1,4), (4,1), (1,5), (5,1), (1,6), (6,1), (2,5), (5,2), (2,6), (6,2), (3,6), (6,3)

so probability = 12/30 = 0.4 2

OR: Use a sample space of size 15: $\{(i,j)\}$ where i is smaller number selected, j is larger.

Then event (i) has 2 favourable outcomes and event (ii) has 6.

3

(i) $M_Y(t) = E[e^{tY}] = E[e^{t(aX+b)}] = e^{tb}E[e^{atX}] = e^{bt}M_X(at)$

$\Rightarrow C_Y(t) = \log M_Y(t) = bt + \log M_X(at) = bt + C_X(at)$ 2

(ii) $C_Y(t) = 2t + \log(1 - 3t)^{-2} = 2t - 2\log(1 - 3t)$ 1

$C_Y'(t) = 2 + 6(1 - 3t)^{-1}$, $C_Y''(t) = 18(1 - 3t)^{-2}$, $C_Y'''(t) = 108(1 - 3t)^{-3}$ 2

$E[(Y - \mu_Y)^2] = C_Y''(0) = 18$, $E[(Y - \mu_Y)^3] = C_Y'''(0) = 108$

$\Rightarrow$ coefficient of skewness of $Y = 108/18^{3/2} = 2^{1/2} = 1.414$ 2

OR note that coefficient of skewness of Y = coefficient of skewness of X and just work with X (some candidates may recognise $X \sim$ Gamma(2,1) and comment on the formula for coefficient of skewness ($2/\sqrt{\square}$) given in the Yellow Book).

4

(i) $f_X(x) = \int_0^\infty f(x,y)dy = \int_0^\infty e^{-x-y}dy = e^{-x}\left[-e^{-y}\right]_0^\infty = e^{-x}$. 1

$f_Y(y) = \int_0^\infty f(x,y)dx = \int_0^\infty e^{-x-y}dx = e^{-y}\left[-e^{-x}\right]_0^\infty = e^{-y}$ [OR by symmetry]. 1

Since $f_{X,Y}(x,y) = e^{-x-y} = f_X(x)\,f_Y(y)$, X and Y are independent. 1

(ii) $F_{X,Y}(x,y) = \int_0^x \int_0^y e^{-u-v}dvdu$ 1

$\Rightarrow F_{X,Y}(x,y) = \int_0^x e^{-u}du \int_0^y e^{-v}dv = \left[-e^{-u}\right]_0^x \left[-e^{-v}\right]_0^y = 1-e^{-x} \quad 1-e^{-y}$ 1

5

$E[X] = \int_0^\theta xf(x)dx = \int_0^\theta 2x^2\theta^{-2}dx$ 1

$= \frac{2}{3}\theta^{-2}\left[x^3\right]_0^\theta = \frac{2}{3}\theta$. 1

Consider $Z = \frac{3}{2}X \Rightarrow E[Z] = \frac{3}{2}E[X] = \theta$.

Z is an unbiased estimator of θ. 2

6

(i) $L(\lambda) = \Pi\lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda\Sigma x_i}$ 1

$\log L(\lambda) = n\log\lambda - \lambda\Sigma x_i$ and $\frac{d}{d\lambda}\log L(\lambda) = \frac{n}{\lambda} - \Sigma x_i$ 1

Equate to zero for the MLE $\Rightarrow \hat{\lambda} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}$

(second derivative is clearly negative, so maximum) 1

(ii) $p = P(X > 4000) = \exp(-4000\lambda)$ for the exponential distribution 1

$\therefore \hat{p} = \exp(-4000\hat{\lambda})$ using the invariance property of MLE's 1

$\therefore \hat{p} = \exp(-4000(0.00124)) = 0.0070$ 1

7

(i) We need to calculate the basics sums $\sum x$, $\sum x^2$, $\sum y$, $\sum xy$

$n = 20$

$\sum x = 4(1) + 3(2) + 6(3) + 7(4) = 56$

$\sum x^2 = 4(1^2) + 3(2^2) + 6(3^2) + 7(4^2) = 182$

$\sum y = 4(18.6) + 3(21.7) + 6(23.2) + 7(27.1) = 468.4$

$\sum xy = 1(4)(18.6) + 2(3)(21.7) + 3(6)(23.2) + 4(7)(27.1) = 1381.0$

2

$\therefore S_{xy} = 1381.0 - \frac{1}{20}(56)(468.4) = 69.48$ and $S_{xx} = 182 - \frac{1}{20}(56)^2 = 25.2$ 1

$\therefore \hat{\beta} = \frac{69.48}{25.2} = 2.757$ 1

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{1}{20}[468.4 - (2.757)56] = 15.7$ 1

$\therefore \hat{y} = 15.7 + 2.757x$

(ii) (a) 95% CI for β is $\hat{\beta} \pm t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

So we need to calculate

$\hat{\sigma}^2 = \frac{1}{n-2}\sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ *or* $\frac{1}{n-2}(S_{yy} - \frac{S_{xy}^2}{S_{xx}})$. 1

Problem as we do not have the individual y_i values, only means of sets of them.

(b) We would need these individual y_i values (or the s.d. or $\sum y^2$ for each set). 2

8

$X|Y = 2$ takes values 0, 1, 2 with probabilities 1/8, 3/8, 4/8

(being in the ratios 1:3:4) 2

So $E[X|Y = 2] = 1(3/8) + 2(4/8) = 11/8 = 1.375$ 1

9

(i) $E(\text{amount}) = 0.8(1650) + 0.2(625) = 1{,}320 + 125 = £1{,}445$ 1

(ii) $E(\text{number of claims}) = 150000(0.15) = 22{,}500$ 1

$E(\text{total claim amount}) = 22500(1445) = £32{,}512{,}500$

10

(i) Number of sample members with the characteristic $X \sim bi(n, \theta)$ with mean $n\theta$ and variance $n\theta(1-\theta)$. $P = X/n$. 1

$E[P] = n\theta/n = \theta$ 1

$s.e.[P] = \{V[P]\}^{1/2} = \{V[X]/n^2\}^{1/2} = \{n\theta(1-\theta)/n^2\}^{1/2} = \{\theta(1-\theta)/n\}^{1/2}$ 1

(ii) $X \sim bi(200, 0.7)$ with mean 140 and variance 42 2

$$P(X \geq 150) = P\left(Z > \frac{149.5 - 140}{\sqrt{42}}\right) = P(Z > 1.466) = 0.071$$ 2

(iii) (a) Sample proportion $P \sim N(\theta, \theta(1-\theta)/200)$

Observed P = 146/200 = 0.73

Estimated standard error(P) = $(0.73 \times 0.27/200)^{1/2} = 0.03139$ 1

$$Pr\left(\frac{P - \theta}{e.s.e.\ P} > -1.645\right) = 0.95$$

$$\Rightarrow Pr\ \theta < P + 1.645 \times e.s.e.\ P = 0.95$$ 2

Upper 95% CI is given by (0, 0.73 + 1.645×0.03139)

i.e. (0, 0.782) 1

(b) By analogy with (i),

Lower 95% CI is given by (0.73 – 1.645×0.03139, 1)

i.e. (0.678, 1) 2

(c) The *P*–value indicates that the null hypothesis "$\theta = 0.7$" can stand and we do not have to conclude that $\theta > 0.7$. 1

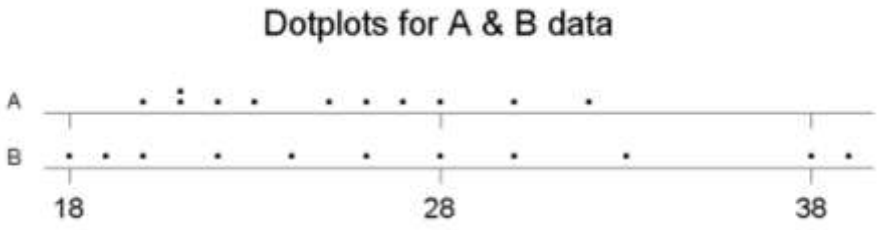
The CI in (iii)(b) includes values down to 0.678, so all such values, including 0.7, are consistent with the data when considering how low a value of θ is reasonable. 1

The two results complement each other. 1

11

(i) Dotplots on same scale are most suitable

[alternatively boxplots or histograms are acceptable]



2

The spread of the *B* data appears to be greater than that of the *A* data and so casts some doubt on the equal variance assumption. 1

(ii) (a) $s_A^2 = \frac{1}{10}\left(7033 - \frac{275^2}{11}\right) = 15.8$

$s_B^2 = \frac{1}{10}\left(8559 - \frac{297^2}{11}\right) = 54.0$ 1

$F = \frac{s_B^2}{s_A^2} = \frac{54.0}{15.8} = 3.418$ on 10, 10 df 1

For a two-sided test at the 5% level, critical value is

$F_{10,10}(2.5\%) = 3.717$ 1

So we accept $H_0 : \sigma_A^2 = \sigma_B^2$ at the 5% level. 1

(b) $F_{10,10}(2.5\%) = 3.717$ and $F_{10,10}(5\%) = 2.978$

So *P*-value is between 0.05 and 0.10 1

By interpolation: *P*-value is

$0.05 + \frac{3.717 - 3.418}{3.717 - 2.978}(0.10 - 0.05) = 0.05 + (0.405)(0.05) = 0.070$ 1

(iii) (a) If the samples have equal variances, then the absolute deviations will be similar in size for both samples; if one sample has a larger variance than the other, then the deviations will be more extreme such that the absolute deviations will be larger for that sample.

A two-sided two-sample *t*-test applied to these absolute deviations will test for a difference in the means of these absolute deviations and hence for a difference in the variances in the original samples. 2

(b) (1) $\bar{x}_A = \frac{275}{11} = 25$

So the deviations for sample A, i.e. $d_A = |x_A - \bar{x}_A|$, are

4 3 3 2 5 2 1 7 0 4 5 1

$$\bar{x}_B = \frac{297}{11} = 27$$

So the deviations for sample B, i.e. $d_B = |\, x_B - \bar{x}_B \,|$ are

8 9 11 6 3 12 5 7 1 1 3 1

(2) Calculations:

$$\Sigma d_A = 36, \quad \Sigma d_A^2 = 158 \text{ and } \Sigma d_B = 66, \quad \Sigma d_B^2 = 540$$

$$\bar{d}_A = \frac{36}{11} = 3.273 \quad \text{and} \quad s_{dA}^2 = \frac{1}{10}\left(158 - \frac{36^2}{11}\right) = 4.018$$

$$\bar{d}_B = \frac{66}{11} = 6.000 \quad \text{and} \quad s_{dB}^2 = \frac{1}{10}\left(540 - \frac{66^2}{11}\right) = 14.400$$

1

$$s_{dp}^2 = \frac{10(4.018) + 10(14.400)}{20} = 9.209 \quad \therefore s_{dp} = 3.035$$ 1

obs. $t = \dfrac{3.273 - 6.000}{3.035\sqrt{\dfrac{1}{11} + \dfrac{1}{11}}} = -\dfrac{2.727}{1.294} = -2.107$ on 20 df 1

For the two-sided test at the 5% level, critical value is $t_{20}(2.5\%) = 2.086$ 1

So we just reject $H_0 : \mu_{dA} = \mu_{dB}$ and hence $H_0 : \sigma_A^2 = \sigma_B^2$ at the 5% level. 1

(3) $t_{20}(2.5\%) = 2.086$ and $t_{20}(1\%) = 2.528$

So P-value is between 0.02 and 0.05 1

By interpolation: P-value is

$$0.02 + \frac{2.528 - 2.107}{2.528 - 2.086}(0.03) = 0.02 + (0.952)(0.03) = 0.049$$ 1

(iv) Tests in (ii) and (iii) give different results at the 5% level, but in fact have quite similar P-values.

Graphical approach in (i) casts doubt on H_0.

So all three are fairly consistent. 2

12

(i) (a) $y_{A\bullet} = 27$, $y_{B\bullet} = 14$, $y_{C\bullet} = 52$, $y_{\bullet\bullet} = 93$, $\Sigma y^2 = 865$

$SS_T = 865 - 93^2/15 = 288.4$

$SS_B = (27^2 + 14^2 + 52^2)/5 - 93^2/15 = 149.2$ 2

$\Rightarrow SS_R = 288.4 - 149.2 = 139.2$

Source of variation	*d.f.*	*SS*	*MSS*
Between	2	149.2	74.6
Residual	12	139.2	11.6
Total	14	288.4	

2

$F = 74.6/11.6 = 6.431$ on (2,12) degrees of freedom. 1

From yellow tables, $F_{2,12}(0.05) = 3.885$ and $F_{2,12}(0.01) = 6.927$. 1

We can reject the hypothesis of "no message effect" at the 5% significance level, but not at the 1% level. We have some evidence against the "no message effect" hypothesis and conclude that there is a message effect. 1

(b) $t_{12}(0.025) = 2.179$ 1

95% CI is $(5.4 - 10.4) \pm 2.179\left\{11.6\left(\frac{1}{5} + \frac{1}{5}\right)\right\}^{0.5}$ 2

i.e. -5 ± 4.694 or $(-9.694, -0.306)$ 1

(ii) (a) We have $\hat{y}_i = 10.4 - 5.0x_{i1} - 7.6x_{i2}$ and for classical music message we need $x_{i1} = x_{i2} = 0$.

This gives $\hat{y}_i = 10.4$ 2

(b) The *P*-value is 0.039. We have evidence to reject the hypothesis that $b_1 = 0$ at the 5% level of significance. 1

(c) For $x_{i1} = 1, x_{i2} = 0$ we have $\mu_A = a + b_1$

and $x_{i1} = x_{i2} = 0$ gives $\mu_C = a$ 1

$\therefore b_1 = \mu_A - \mu_C$ 1

The 95% CI for $\mu_A - \mu_C$ in (i)(b) can be used for testing H_0: $\mu_A - \mu_C = 0$ and equivalently H_0: $b_1 = 0$. The interval does not include the value 0, and thus we reject H_0 at the 5% level. 2

END OF EXAMINERS' REPORT