Subject CT3 — Probability and Mathematical Statistics Core Technical

EXAMINERS' REPORT

September 2008

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

November 2008

Comments

The paper was answered well overall and there are no particular topics that stand out as being 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 $n = 30, \bar{x} = 5200$

$n_1 = 6, \bar{x}_1 = 8000$

$n_2 = 10, \bar{x}_2 = 3100$

$n_3 = 14$

$$\bar{x} = \frac{\sum x}{n} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$$

$$\Rightarrow \bar{x}_3 = \frac{n\bar{x} - n_1\bar{x}_1 - n_2\bar{x}_2}{n_3} = \frac{30\times 5200 - 6\times 8000 - 10\times 3100}{14} = \frac{77000}{14} = 5500$$

2 data: $\Sigma f = 540$, $\Sigma fx = 1469$, $\Sigma fx^2 = 5081$

mean = $\frac{1469}{540} = 2.72$ years

variance = $\frac{1}{539}(5081 - \frac{1469^2}{540}) = 2.0126$ $\quad\therefore$ s.d. = 1.42 years

3 (i) $M_Y(t) = E(e^{tY}) = E(e^{t(X_1+X_2)})$

$= E(e^{tX_1})E(e^{tX_2}) = M_{X_1}(t)M_{X_2}(t)$

(ii) $M_{X_i}(t) = (1-\frac{t}{\lambda})^{-\alpha_i}$ $\quad\therefore M_Y(t) = (1-\frac{t}{\lambda})^{-(\alpha_1+\alpha_2)}$

so that Y is a gamma r.v. with parameters $(\alpha_1 + \alpha_2, \lambda)$.

4 $t_{14}(0.005) = 2.977$

99% CI is $94.2 \pm 2.977\sqrt{\frac{24.86}{15}}$ i.e. 94.2 ± 3.83 i.e. $(90.37, 98.03)$

5 $E(X) = E_Y\left\{E_X(X|Y)\right\} = E(Y) = \frac{a}{b}$.

$E(X^2) = \text{var}(X) + E^2(X)$

with

$$\text{var}(X) = E_Y\left\{\text{var}_X(X|Y)\right\} + \text{var}_Y\left\{E_X(X|Y)\right\}$$
$$= E(Y) + \text{var}(Y) = \frac{a}{b} + \frac{a}{b^2}$$

giving

$E(X^2) = \frac{a}{b} + \frac{a}{b^2} + \frac{a^2}{b^2}$.

[*OR* $E(X^2) = E_Y\left\{E_X(X^2|Y)\right\}$
$= E_Y\left\{\text{var}_X(X|Y) + E_X^2(X|Y)\right\} = E(Y) + E(Y^2)$
$= E(Y) + \text{var}(Y) + E^2(Y)$
$= \frac{a}{b} + \frac{a}{b^2} + \frac{a^2}{b^2}$.]

6 (i) $T \sim \text{Exp}(0.5)$ and therefore $P(T > 1) = e^{-0.5\times 1} = 0.6065$.

(ii) The median, *M*, is such that $\int_0^M f(t)dt = 0.5 \Rightarrow \int_0^M 0.5e^{-0.5t}dt = 0.5$

which gives

$1 - e^{-0.5M} = 0.5 \Rightarrow M = -2\log(0.5)$, or $M = 2\log(2) = 1.386$.

(Note: the cdf is available from the Yellow Book, p11.)

(iii) From CLT, $Y = \sum_{i=1}^{30} T_i \sim N(30\times 2,\ 30\times 4)$, i.e. $N(60,120)$, approximately.

Then,

$$P(Y > 45) = P\left(Z > \frac{45-60}{\sqrt{120}}\right) = P(Z > -1.3693) = P(Z < 1.3693) = 0.915.$$

[OR $Y \sim$ gamma(30,1/2), that is $Y \sim \chi^2_{60}$, from which we can then use the normal approximation as above, or get $P(Y > 45) = 0.922$ (approximately) by interpolating in tables of percentage points of χ^2_{60} (Yellow Book p168).]

7 (i) $E(S) = E(N)E(X) = (60)(500) = £30,000$

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

$= (60)(400^2) + (60)(500^2) = 24,600,000 \quad \therefore sd(S) = £4,960$

(ii) As S is the sum of a large number of i.i.d. variables, then the central limit theorem gives an approximate normal distribution for S.

$$P(S > 40000) = P(Z > \frac{40000 - 30000}{4960} = 2.016)$$

$= 1 - 0.9781 = 0.0219$

[*Note*: 2.02 leading to 0.0217 is also acceptable.]

8 (i) From Yellow Book Table

$P(Z < 1.43) = 0.9236$ giving x value (10*1.43) + 200 = 214.3
$P(Z < -0.65) = 0.2578$ giving x value (10*(–0.65)) + 200 = 193.5

(ii) Setting $r = P(Y < y) = 1 - \exp(-y/100) \Rightarrow y = -100*\log(1 - r)$

$r = 0.3287 \Rightarrow y = -100\log(0.6713) = 39.85$

$r = 0.9142 \Rightarrow y = -100\log(0.0858) = 245.6$

Note: We can do away with the step of subtracting r from 1 and use.

$y = -100*\log(r)$. This gives $y = 111.3, 8.971$.

9 (i) $SS_R = SS_T - SS_B = 420.05 - 337.32 = 82.73$.

The degrees of freedom are 3 – 1 = 2 for the treatment (company) *SS*, and

12 – 1 – 2 = 9 for the residual *SS*.

These give $F = \dfrac{337.32/2}{82.73/9} = 18.348.$

From tables, $F_{0.01,2,9} = 8.022$, and therefore we have strong evidence against the hypothesis that the means of the insured sums are equal for the 3 companies.

(ii) To perform the ANOVA we assume that the data follow normal distributions and that their variance is constant.

(iii) The variance of the residuals seems to depend on the company from which the data come. This violates the assumption of constant variance in the response variable, and therefore the analysis may not be valid.

10 (i) P(rejected at 1st) = 1 – P(cleared at 1st) = 1 – θ

P(rejected at 2nd) = P(cleared at 1st)P(rejected at 2nd | cleared at 1st)
= θ (1 – θ)

P(progressing after two) = P(cleared at 1st) P(cleared at 2nd) = θ^2

(ii) (a) $L(\theta) = [(1-\theta)]^{x_1}[\theta(1-\theta)]^{x_2}[\theta^2]^{x_3}$

$= \theta^{x_2+2x_3}(1-\theta)^{x_1+x_2}$

$\therefore \log L(\theta) = (x_2 + 2x_3)\log\theta + (x_1 + x_2)\log(1-\theta)$

$\therefore \dfrac{\partial}{\partial\theta}\log L(\theta) = \dfrac{x_2 + 2x_3}{\theta} - \dfrac{x_1 + x_2}{1-\theta}$

(b) equate to zero for MLE

$\therefore \theta(x_1 + x_2) = (1-\theta)(x_2 + 2x_3)$

$\therefore \theta(x_1 + 2x_2 + 2x_3) = x_2 + 2x_3$

$\therefore \hat{\theta} = \dfrac{x_2 + 2x_3}{x_1 + 2x_2 + 2x_3}$

(iii) (a) $\dfrac{\partial^2}{\partial\theta^2}\log L(\theta) = -\dfrac{x_2 + 2x_3}{\theta^2} - \dfrac{x_1 + x_2}{(1-\theta)^2}$

$E\{\dfrac{\partial^2}{\partial\theta^2}\log L(\theta)\} = -\dfrac{n\theta(1-\theta) + 2n\theta^2}{\theta^2} - \dfrac{n(1-\theta) + n\theta(1-\theta)}{(1-\theta)^2}$

$$= -\frac{n}{\theta}(1+\theta) - \frac{n}{(1-\theta)}(1+\theta) = -n(1+\theta)(\frac{1}{\theta} + \frac{1}{1-\theta}) = -\frac{n(1+\theta)}{\theta(1-\theta)}$$

$$\text{CRlb} = \frac{1}{-E[\frac{\partial^2}{\partial\theta^2}\log L(\theta)]} = \frac{\theta(1-\theta)}{n(1+\theta)}$$

(b) $\hat{\theta} \approx N(\theta, CRlb)$ for large n

using CRlb = $\frac{\hat{\theta}(1-\hat{\theta})}{n(1+\hat{\theta})}$, then $\hat{\theta} \approx N(\theta, \frac{\hat{\theta}(1-\hat{\theta})}{n(1+\hat{\theta})})$

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

(iv) $\hat{\theta} = \frac{96+2(794)}{110+2(96)+2(794)} = \frac{1684}{1890} = 0.8910$

$$\text{CRlb} \approx \frac{0.8910(1-0.8910)}{1000(1+0.8910)} = 0.0000514 \quad \therefore \sqrt{CRlb} = 0.00717$$

$\therefore$ 95% CI is $0.8910 \pm 1.96(0.00717)$

$\Rightarrow 0.891 \pm 0.014$ *or* $(0.877, 0.905)$

11 (i) (a) Males: $n_1 = 40$ $\bar{x}_1 = 215/40 = 5.375$

Females: $n_2 = 35$ $\bar{x}_2 = 168/35 = 4.8$

95% CI:

$$\bar{x}_1 - \bar{x}_2 \pm z_{0.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= 5.375 - 4.8 \pm 1.96\sqrt{\frac{3^2}{40} + \frac{2.5^2}{35}}$$

$$= 0.575 \pm (1.96)(0.6353)$$

$$= 0.575 \pm 1.245 \quad \text{or} \quad (-0.67, 1.82)$$

(b) As this CI includes the value 0 we would not eliminate the possibility that the males and females have the same expected length of stay.

(ii) (a)

$s_1^2 = (1481 - 215^2/40)/39 = 8.34295$

$s_2^2 = (1026 - 168^2/35)/34 = 6.45882$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(39)(8.34295)+(34)(6.45882)}{40+35-2} = 7.46541$$

Two sample *t*-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{5.375-4.8}{\sqrt{7.46541\left(\frac{1}{40}+\frac{1}{35}\right)}} = 0.909$$

$t_{73}(0.025) = 1.996$ (by interpolation of 2.000 and 1.980 for 60 df and 120 df)

[OR just quote the $N(0,1)$ value 1.96 in place of the t_{73} value.]

Therefore there is no evidence to reject the null hypothesis that the means for males and females do not differ at the 5% significance level, and we conclude that the mean lengths of stay are the same.

(b) $$\frac{s_1^2}{s_2^2} = \frac{8.34295}{6.45882} = 1.29$$

Comparing this to an $F_{39,34}$ distribution, which has a 5% critical point between 2.075 and 1.717 (two-sided test), there is no evidence that the population variances differ.

The assumption of common variance was made when conducting the test in (ii)(a), and this seems valid given the result of the test in (ii)(b).

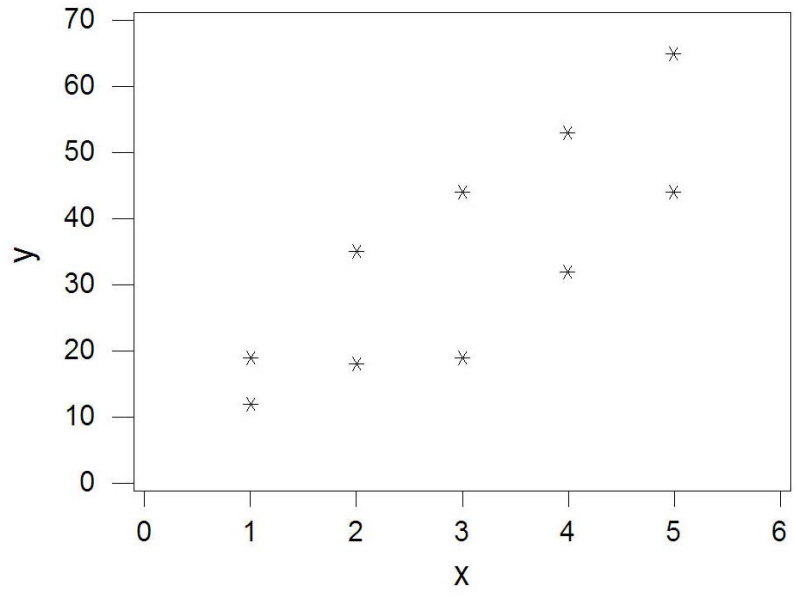
(c) $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

$$= \frac{5.375-4.8}{\sqrt{\frac{8.34295}{40}+\frac{6.45882}{35}}}$$

$= 0.917$

Compare with $N(0,1)$, e.g. 1.96 for 5% level test. Therefore we reach exactly the same conclusion (as in (ii)(a) but without making the assumptions of equal variances and normal distributions – we have large samples and can rely on CLT).

12 (i) (a)



(b) SSTOT = $S_{yy} = 14345 - 341^2/10 = 2716.9$

$\Sigma x = 30$, $\Sigma x^2 = 110$ so $S_{xx} = 110 - 30^2/10 = 20$

$S_{xy} = 1211 - 30*341/10 = 188$

$\therefore$ SSREG = $188^2/20 = 1767.2$

SSRES = 2716.9 – 1767.2 = 949.7

$R^2 = 1767.2/2716.9 = 0.650$ (65.0%)

(c) $y = a + bx$: $\hat{b} = 188/20 = 9.4$

$\hat{a} = 341/10 - 9.4 \times (30/10) = 5.9$

Fitted line is $y = 5.9 + 9.4x$

(d) $s.e.\left(\hat{b}\right) = \left(\frac{949.7/8}{20}\right)^{1/2} = 2.4363$

$t_8(0.025) = 2.306$

95% confidence interval for b is given by $9.4 \pm 2.306 \times 2.4363$

i.e. 9.4 ± 5.62 i.e. (3.78, 15.02)

(ii) When we replace the pair of responses by their mean:

the equation of the fitted line remains the same
but otherwise the analyses do not produce "equivalent results"
the "fit" of the line is very much better [the goodness-of-fit measure R^2 increases to a very high value – from 65% to 98.5%]

plus, for example:

the estimate of the slope has a much lower standard error (2.436 drops to 0.6733)
the SSTOT drops hugely (from 2716.9 on 9df to 897.2 on 4df)
the residual error (SSRES) drops hugely [from 949.7 on 8df (error variance estimate 118.7) to 13.60 on 3 df (error variance estimate 4.53)]
BUT we lose all information on the variation of the response for a given value of the explanatory variable

Note: these and other relevant comments will receive credit.

END OF EXAMINERS' REPORT