EXAMINATION

September 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

December 2007

Comments

The paper was answered quite 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 $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{25(120.2)+18(142.7)}{25+18} = \dfrac{5573.6}{43} = 129.6$

Using the fact that $\Sigma x^2 = (n-1)s^2 + n\bar{x}^2$, then for the combined set

$$\Sigma x^2 = [24(58.1)^2 + 25(120.2)^2] + [17(62.2)^2 + 18(142.7)^2] = 874525.14$$

$$\therefore s^2 = \frac{874525.14 - (5573.6)^2/43}{42} = 3621.02 \quad \therefore s = 60.2$$

2 Method: set uniform(0,1) random number $r = F(x) = 1 - 1/x^2$

$\Rightarrow$ simulated observation $x = [1/(1-r)]^{1/2}$

Here we get $x = 1.528$, 2.684, 1.198.

Note: We can do away with the step of subtracting r from 1 and use $x = (1/r)^{1/2}$.

This gives $x = 1.322$, 1.078, 1.817.

3 Let N be the number who have another type of account.

$\therefore N \sim \text{binomial}(250, 0.24) \approx N(60, 45.6) = N(60, 6.753^2)$

$P(N < 50) \quad \to P(N < 49.5)$ with continuity correction

$$= P(Z < \frac{49.5-60}{6.753} = -1.55) = 1 - 0.93943 = 0.061$$

4 Sample proportion $P = 68/200 = 0.34$

99% CI is $0.34 \pm 2.576\sqrt{\dfrac{0.34 \times 0.66}{200}}$ i.e. 0.34 ± 0.086 i.e. $(0.254, 0.426)$

5 (i) $\bar{x} = \dfrac{56.7}{8} = 7.0875.$

$$s^2 = \frac{1}{7}\left(403.95 - \frac{56.7^2}{8}\right) = 0.298 \Rightarrow s = 0.546\,.$$

90% CI for the true mean is given by:

$$\bar{x} \pm t_{7,0.05}\frac{s}{\sqrt{n}} = 7.0875 \pm 1.895\frac{0.546}{\sqrt{8}} = 7.0875 \pm 0.3658$$

i.e. the 90% CI is (6.722, 7.453), or (£6722, £7453).

(ii) The value £6500 is not included in the CI above, and therefore we conclude that the data are not consistent with the expert's assessment at the 10% significance level.

6 Use the result $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$ under H_0.

From tables $t_{0.01,13} = 2.650$

so critical value is solution of $\dfrac{r}{\sqrt{1-r^2}}\sqrt{13} = 2.65$

Solving gives $r = \sqrt{\dfrac{\dfrac{2.65^2}{13}}{1+\dfrac{2.65^2}{13}}} = 0.592$

7 From the Yellow Book:

Mean: $E(S) = E(N)E(X) = (20)(10000) = £200{,}000$

Variance: $\mathrm{var}(S) = E(N)\,\mathrm{var}(X) + \mathrm{var}(N)[E(X)]^2$

$= 20(2000^2) + 20(10000^2) = 2080 \times 10^6$

$\therefore$ Standard deviation = £45607

[***OR***, using compound Poisson results (in Yellow Book)

$E(S) = \lambda m_1$ and $\mathrm{var}(S) = \lambda m_2$ where $\lambda = E(N)$ and $m_r = E(X^r)$]

8 $X \sim N$ with mean $\mu = 30$ and $\sigma = 4$ (working in units of £1000)

(i) (a) $P(X > 35) = P\left(Z > \dfrac{35-30}{4}\right) = P(Z > 1.25) = 1 - 0.89435 = 0.10565$

(b) $P(X > 36) = P\left(Z > \dfrac{36-30}{4}\right) = P(Z > 1.5) = 1 - 0.93319 = 0.06681$

(ii) $P(X > 36 \mid X > 35) = P(X > 36 \text{ and } X > 35) / P(X > 35)$

$= P(X > 36) / P(X > 35)$

$= P(Z > 1.5)/P(Z > 1.25) = 0.06681/0.10565 = 0.632$

(iii) $\dbinom{5}{2} \times 0.1056^2 \times 0.8944^3 = 0.0798$

9 (i) If X is the random variable denoting the number of policies giving a claim, then $X \sim \text{binomial}(250, 0.5)$.

Using the normal approximation (CLT), $X \approx N(125, 62.5)$.

Using the appropriate continuity correction we have:

$P(X \geq 139) = P(X > 138.5)$

$$= P\left(Z > \frac{138.5 - 125}{\sqrt{62.5}}\right) = 1 - \Phi(1.7076) = 0.044.$$

(ii) This is a one-sided test of $H_0 : p = 0.5$ v $H_1 : p > 0.5$.

P-value of the test is 0.044 from part (i).

The evidence against the hypothesis that $p = 0.5$ (and in favour of $p > 0.5$) is not strong enough to justify rejecting it at the 1% level of testing — we cannot conclude "$p > 0.5$".

10 (i) Chi-square statistic is doubled and has value 9.722

OR work it out

(ii) *P*-value is given by $P\left(\chi_2^2 > 9.722\right) = 0.0077$

Note: answer = 0.008 is acceptable for the mark

(iii) Comment: With the first table we do not have strong enough evidence to justify rejecting the hypothesis of no association. In the second table, we have the same proportions in the columns, but based on more data, and now we <u>do</u> have strong enough evidence (*P*-value < 1%) to justify rejecting the hypothesis of no association.

11 (i) (a) $Y = X^{1/3} \Rightarrow X = Y^3$, and range of Y is $(0, \infty)$.

The cdf is given by

$$F_Y(y) = P(Y \le y) = P(X \le y^3) = F_X(y^3)$$

$$\therefore F_Y(y) = \begin{cases} 1 - \exp(-\lambda y^3), & y \ge 0 \\ 0, & y < 0 \end{cases}$$

(using formulae or by integration).

Then, the pdf of Y can be derived as

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 3\lambda y^2 \exp\left(-\lambda y^3\right).$$

[***OR***, directly as

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \lambda e^{-\lambda y^3} 3y^2$$

$$\Rightarrow f_Y(y) = 3\lambda y^2 \exp\left(-\lambda y^3\right),$$

OR, from formulae, identifying the cdf as that of a Weibull distribution with $c = \lambda, \gamma = 3$.]

(b) First simulate $X \sim \exp(\lambda)$ as

$$u = 1 - e^{-\lambda x} \Rightarrow x = -\frac{1}{\lambda}\log(1-u),$$

then set $y = x^{1/3}$.

[***OR***, use cdf of Y directly, i.e. $u = 1 - e^{-\lambda y^3} \Rightarrow y = \left\{-\frac{1}{\lambda}\log(1-u)\right\}^{1/3}$]

(ii)

(a) $$L(\lambda) = \prod_{i=1}^{n} f(y_i;\lambda) = \prod_{i=1}^{n}\left\{3\lambda y_i^2 \exp\left(-\lambda y_i^3\right)\right\} = 3^n \lambda^n \prod_i y_i^2 \exp\left(-\lambda \sum_i y_i^3\right)$$

$$\ell(\lambda) = \log L(\lambda) = n\log(\lambda) - \lambda\sum_i y_i^3 + \text{constant}$$

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_i y_i^3$$

$$\ell'(\lambda) = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_i y_i^3}$$

[Check that $\ell''(\lambda) = -\frac{n}{\lambda^2} < 0.$]

(b) For the given data we have $\sum_i y_i^3 = 16.3952$

$$\therefore \hat{\lambda} = \frac{n}{\sum_i y_i^3} = \frac{8}{16.3952} = 0.488.$$

(iii) (a) For $X \sim \exp(\lambda)$ we have

$$h(x) = \frac{f(x)}{S(x)} = \frac{\lambda e^{-\lambda x}}{1-\left(1-e^{-\lambda x}\right)} = \lambda$$

For Y (using pdf and cdf derived above):

$$h(y) = \frac{f(y)}{S(y)} = \frac{3\lambda y^2 \exp\left(-\lambda y^3\right)}{1-\left(1-\exp\left(-\lambda y^3\right)\right)} = 3\lambda y^2.$$

(b) X has a constant hazard rate $h(x) = \lambda$, and therefore should only be used when the force of mortality can be assumed constant, e.g. over a one-year period of time in mortality studies. For longer periods of lifetime the r.v. Y is more suitable, as it gives an increasing hazard function with time.

12 (i) Let X be a reading and M be the number of readings which are less than 1

(a) Since $X \sim U(0, \theta)$, $P(X < 1)$ = length of $[0,1]$/length of $[0, \theta] = 1/\theta$

(b) $$L(\theta) \propto \left(\frac{1}{\theta}\right)^m \left(1-\frac{1}{\theta}\right)^{n-m} \Rightarrow \ell(\theta) = -m\log\theta + (n-m)\log\left(\frac{\theta-1}{\theta}\right)$$

$$\Rightarrow \ell(\theta) = (n-m)\log(\theta-1) - n\log\theta$$

$$\Rightarrow \frac{\partial \ell}{\partial \theta} = \frac{n-m}{\theta-1} - \frac{n}{\theta} \quad \text{set to zero} \Rightarrow \hat{\theta} = n/m$$

OR Since $M \sim \text{bi}(n, 1/\theta)$, MLE of $1/\theta$ is the sample proportion of readings which are <1, namely m/n, so

$$\widehat{(1/\theta)} = m/n \quad \Rightarrow 1/\hat{\theta} = m/n \quad \Rightarrow \hat{\theta} = n/m$$

(c) $$\frac{\partial \ell}{\partial \theta} = \frac{n-m}{\theta-1} - \frac{n}{\theta} \quad \Rightarrow -\frac{\partial^2 \ell}{\partial \theta^2} = \frac{(n-m)}{(\theta-1)^2} - \frac{n}{\theta^2}$$

$$\therefore E\left[-\frac{\partial^2 \ell}{\partial \theta^2}\right] = E\left[\frac{n-M}{(\theta-1)^2}\right] - \frac{n}{\theta^2} = \frac{n-n/\theta}{(\theta-1)^2} - \frac{n}{\theta^2} = \frac{n}{\theta^2(\theta-1)}$$

$$\therefore \text{CRlb} = \frac{\theta^2(\theta-1)}{n}$$

Large sample distribution of $\hat{\theta}$ is $\hat{\theta} \sim N\left(\theta, \frac{\theta^2(\theta-1)}{n}\right)$

(ii) (a) $n = 100$, $m = 45$, $\hat{\theta} = 100/45 = 2.222$

Estimate of standard error of

$\hat{\theta} = [(100/45)^2(100/45 - 1)/100]^{1/2} = 0.2457$

$\Rightarrow$ approximate 95% CI for θ is given by $2.222 \pm 1.96 \times 0.2457$

i.e. 2.222 ± 0.482 i.e. $(1.74, 2.70)$.

(b) Under H_0: $\hat{\theta} \sim N(3, 0.18)$

$$P\text{-value} = P\left(\hat{\theta} < 2.222\right) = P\left(Z < \frac{2.222-3}{0.4243}\right) = P(Z < -1.834)$$

= 0.033 (by interpolation in the table)

We can reject H_0 (at levels of testing down to 3.3%) and conclude that $\theta < 3$.

[OR note that $P(Z < -1.834)$ is less than 0.05, so "reject H_0 at 5% level"]

13 (i) (a) $SS_T = 48196 - 872^2/16 = 48196 - 47524 = 672$
$SS_B = (186^2 + 236^2 + 243^2 + 207^2)/4 - 47524 = 523.5$
$SS_R = 672 - 523.5 = 148.5$

$$F = \frac{523.5/3}{148.5/12} = \frac{174.5}{12.375} = 14.10$$

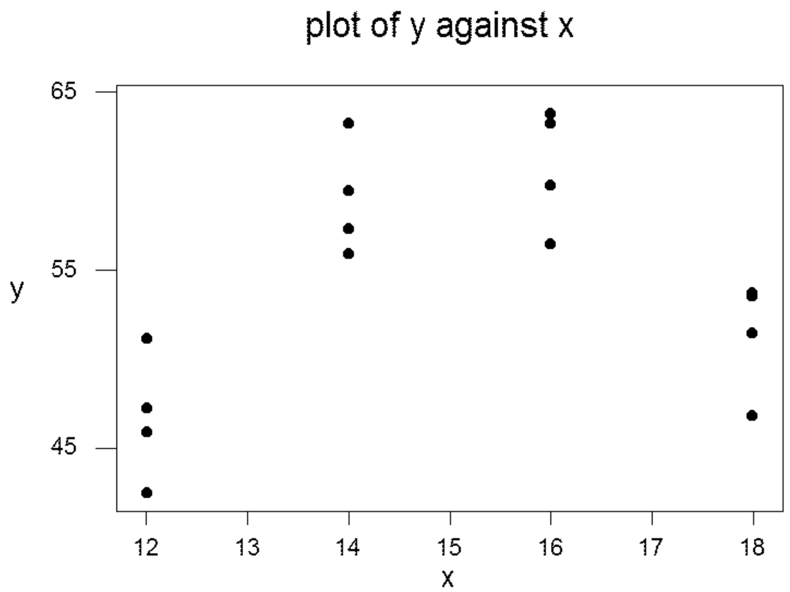
[or construct an ANOVA table]

(b) $F_{3,12}(1\%) = 5.953$ from tables

since $14.10 >> 5.953$, P-value $<< 0.01$

(c) Statistician A could plot either the individual y values or the four means of y against x to see what "shape" the effect might take.

Here is a plot of the individual y values:



The shape of the effect seems to be curved, initially increasing, then decreasing.

(d) The implications are simply that there is nothing to invalidate the assumptions required for the analysis.

(ii) (a) $S_{xx} = 3680 - \frac{240^2}{16} = 80$

$S_{yy} = 48196 - \frac{872^2}{16} = 672$

[or could state it is the same as SS_T from (i)]

$S_{xy} = 13150 - \frac{(240)(872)}{16} = 70$

$\hat{\beta} = \frac{70}{80} = 0.875$ as required

$\hat{\alpha} = \frac{1}{16}(872 - 0.875(240)) = 41.375$ as required.

(b) $\hat{\sigma}^2 = \frac{1}{14}(672 - \frac{70^2}{80}) = 43.625$

$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{43.625}{80}} = 0.7385$

$t = \frac{0.875 - 0}{0.7385} = 1.185$ on 14 d.f.

P-value = 2 × P(t_{14} >1.185)

As $P(t_{14} > 1.345) = 0.10$ from tables, P-value > 2(0.10), i.e. > 0.20

This implies that there is no evidence against H_0, and hence that there is no <u>linear</u> relationship between x and y.

[not that there is no "relationship"]

(c) Residual plot suggests that there may be a curved, rather than linear, relationship between x and y.

(d) Statistician B could try a quadratic regression (or some other curved form) of y on x.

END OF EXAMINERS' REPORT