EXAMINATION

September 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 A Stocker
Chairman of the Board of Examiners

November 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.

The most common problems noted by the examiners are summarised below.

Some candidates were unsure of basic concepts in probability (such as the independence of two events) and gave poor answers to Questions 2 and 3.

Question 5

Many candidates used $\hat{\pi} = 0.34$ *(wrongly) rather than* $\pi = 0.4$ *(correctly) in the expression for the standard error of the estimate (the sample proportion) under* H_0*. However, it makes little difference numerically, and the examiners were generous on this point when marking.*

Question 7

was poorly attempted, with many candidates failing to realise that the distribution of the total waiting time can be approximated by a normal distribution, by virtue of the central limit theorem.

Question 8

Some candidates did not know the result on the variance of the mean of a random sample of size n, namely $Var\left[\bar{X}\right] = \frac{\sigma^2}{n}$.

Question 9

Some candidates displayed a lack of familiarity with the use of conditional expectations, and in particular with the application of the result

$$Var[Y] = Var[E(Y|X)] + E[Var(Y|X)]$$

Question 10

Some candidates did not know that the asymptotic standard error of a maximum likelihood estimator is found from evaluating $\sqrt{1/I}$ *, where*

$$I = E\left[-\frac{d^2\ell}{d\lambda^2}\right] \text{ and } \ell(\lambda) \text{ is the log-likelihood.}$$

Question 11

In the part on equality of variances (part (iii)(a)) some candidates who worked with $\frac{s_1^2}{s_2^2}$ *(= 0.607) did not know how to find the lower 2.5% point of* $F_{7,11}$ *(which is the reciprocal of the upper 2.5% point of* $F_{11,7}$ *, and is approximately 1/4.71 = 0.212).*

1 (i) P(second ball drawn is B) = P(first ball drawn is B) = 8/14 = 0.571

OR $P(1^{st}\ B \text{ and } 2^{nd}\ B) + P(1^{st}\ W \text{ and } 2^{nd}\ B)$
$= (8/14)\times(7/13) + (6/14)\times(8/13) = 8/14$

(ii) $P(1^{st}\ \text{W} \mid 2^{nd}\ B) = P(1^{st}\ W \text{ and } 2^{nd}\ B)/P(2^{nd}\ B) = (6/14)\times(8/13)/(8/14)$
$= 6/13 = 0.462$

2 Since A and B are independent, $P(A) = P(A \mid B) = P(A \mid \bar{B})$

Noting that $\overline{(\bar{B})} = B$, it follows immediately that $P(A) = P(A \mid \bar{B}) = P\left(A \mid \overline{(\bar{B})}\right)$ and so A and $\bar{B}$ are independent.

[OR

Since A and B are independent $P(A \cap B) = P(A)P(B)$.

Thus,

$P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)\{1 - P(B)\} = P(A)P(\bar{B})$

$\therefore$ A and $\bar{B}$ are independent.

3 P(no claims on 6 policies) = 0.5314 (from tables p186 — or using 0.9^6)

P(1 claim on 6 policies) = 0.8857 – 0.5314 = 0.3543 (or using $6(0.1)(0.9^5)$)

So required probability = 0.5314 × 0.3543 = 0.188.

4 Let X be the number in force for more than five years
then $X \sim$ binomial(500,0.65)

Using a normal approximation, $X \approx N(325, 10.665^2)$

$P(X < 300)$ becomes $P(X < 299.5)$ using continuity correction

$$\simeq P(Z < \frac{299.5 - 325}{10.665}) \quad \text{where } Z \sim N(0,1)$$

$$= P(Z < -2.39) = 1 - 0.99158 = 0.0084$$

5 Under H_0: sample proportion P is approximately normally distributed with mean 0.4 and standard error $(0.4\times0.6/200)^{1/2} = 0.03464$

$\therefore$ *P-value* of observed proportion (68/200 = 0.34)

$$= P\left(Z < \frac{0.34 - 0.4}{0.03464} \right) = P\left(Z < -1.732 \right) = 0.042$$

We reject H_0 at the 5% level of testing and conclude that the proportion of policyholders who are female is less than 0.4.

[OR This is actually better - working with the *number* of female policyholders (observed = 68), the *P-value* is

$$P\left(Z < \frac{68.5 - 80}{\sqrt{200(0.4)(0.6)}} = -1.660 \right) = 0.048 \quad]$$

Note: We can word the conclusion: we reject H_0 at levels of testing down to 4.2% (or 4.8%) and conclude …

6 (i) P(no claims) = $P(X = 0)$ where $X \sim$ Poisson(0.5)
= 0.60653 from tables [or evaluation]

(ii) Let Y = number of years with a claim
then $Y \sim$ binomial(3,0.3935) [or just directly as below]

$P(Y = 1) = 3(0.3935)(0.6065)^2 = 0.434$

(iii) Let T = time until next claim
then $T \sim$ exp(0.5)

$P(T > 2) = e^{-0.5(2)}$ [or by integration]
$= e^{-1} = 0.368$

[OR: answer = $\{P(\text{no claim})\}^2 = 0.60653^2 = 0.368$]

[OR: claim rate for period of 2 years = 1, so P(no claim in 2 years) $= e^{-1} = 0.368$]

7 (i) As stated in the question, if X_i is the waiting time on day i, then X_i has an exponential distribution with parameter $\frac{1}{15}$ so $E(X_i) = 15$, $Var(X_i) = 15^2 = 225$.

If X is the total waiting time over the 100 days, $X = \sum_{i=1}^{100} X_i$,

so $E[X] = 1500$ and $Var[X] = 22500$ and by the CLT

X has approximately an $N(1500, 22500)$ distribution,
so $P(X > 1620) \approx 1 - \Phi\left(\frac{1620-1500}{150}\right) = 1 - \Phi(0.8) = 0.2119.$

(ii) If Y_j is the waiting time on day j of the extra 99 days, then $E(Y_j) = 10$ and $Var(Y_j) = 100$ so that if $Y = \sum_{j=1}^{99} Y_j$ is the total waiting time over the 99 days, then Y is approximately $N(990,9900)$ by CLT.

If $Z = X + Y$ (so that Z is the total waiting time over the whole 199 days), then since X and Y are independent, Z is approximately $N(1500+990, 22500+9900)$, i.e. $N(2490, 32400)$.

Hence $P(Z > 2400) \approx 1 - \Phi\left(\frac{2400-2490}{180}\right) = 1 - \Phi(-0.5) = \Phi(0.5) = 0.6915.$

8 (i) $E(W) = E(\alpha\bar{X}_1 + (1-\alpha)\bar{X}_2)$

$= \alpha E(\bar{X}_1) + (1-\alpha)E(\bar{X}_2) \ = \alpha\mu + (1-\alpha)\mu = \mu$

Therefore W is unbiased.

(ii) $\text{MSE}(W) = \text{var}(W) + \{\text{bias}(W)\}^2$

W is unbiased

$\therefore \text{MSE}(W) = \text{var}(W)$

$= \text{var}(\alpha\bar{X}_1 + (1-\alpha)\bar{X}_2)$

$= \alpha^2 \text{var}(\bar{X}_1) + (1-\alpha)^2 \text{var}(\bar{X}_2)$ (independent samples)

$= \alpha^2 \frac{\sigma_1^2}{n} + (1-\alpha)^2 \frac{\sigma_2^2}{n}$

(iii) $\frac{d\text{MSE}}{d\alpha} = 2\alpha\frac{\sigma_1^2}{n} - 2(1-\alpha)\frac{\sigma_2^2}{n}$

$\frac{d\text{MSE}}{d\alpha} = 0 \Rightarrow (\sigma_1^2 + \sigma_2^2)\alpha = \sigma_2^2$

$$\therefore\ \alpha = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

$$\frac{d^2\text{MSE}}{d\alpha^2} = 2\frac{\sigma_1^2}{n} + 2\frac{\sigma_2^2}{n} > 0 \quad \therefore \text{ minimum}$$

(iv) The maximum likelihood estimator of μ in the special case with $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is

$$\hat{\mu} = \frac{\text{sum of observations}}{\text{number of observations}} = \frac{n\bar{X}_1 + n\bar{X}_2}{2n}$$

$$= \frac{1}{2}\bar{X}_1 + \frac{1}{2}\bar{X}_2$$

This is the same as W since

$$\alpha = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{\sigma^2}{\sigma^2 + \sigma^2} = \frac{1}{2} \Rightarrow W = \frac{1}{2}\bar{X}_1 + \frac{1}{2}\bar{X}_2.$$

9 (i) $E[E(Y|X)] = \int E[Y|X = x]\, f(x)dx$
$= \int\int yf(y|x)dy\, f(x)dx$
$= \int\int y\, f(y|x)\, f(x)dydx$

but $f(y|x)\, f(x) = f(x,y)$, the joint pdf of X and Y, so

$E[E(Y|X)] = \int\int y\, f(x,y)dxdy = E[Y]$

(ii) $E[Y] = E[E(Y|X)] = E[X^2 + 1]$
$= V[X] + \{E[X]\}^2 + 1 = 1 + 0 + 1 = 2$

$Var[Y] = Var[E(Y|X)] + E[Var(Y|X)] = Var[X^2 + 1] + E[X^2 + 1]$
$= Var[X^2] + E[X^2] + 1$

but $Z = X^2$ is χ_1^2 so has variance 2 and expectation 1

Thus $Var[Y] = 2 + 1 + 1 = 4$

10 (i) (a) Mgf of X_i is $(1 - t/\lambda)^{-4.5}$ so mgf of $\sum_{i=1}^{n} X_i$ is

$$\prod_{i=1}^{n}\left(1-t/\lambda\right)^{-4.5} = \left(1-t/\lambda\right)^{-4.5n}$$

Hence mgf of $2\lambda\sum_{i=1}^{n} X_i = 2\lambda n\bar{X}$ is $\left(1-2\lambda t/\lambda\right)^{-4.5n} = \left(1-2t\right)^{-4.5n}$

This is the mgf of a χ^2 variable — with $9n$ degrees of freedom.

(b) $P\left(a < 2\lambda n\bar{X} < b\right) = 0.95 \Rightarrow P\left(\frac{a}{2n\bar{X}} < \lambda < \frac{b}{2n\bar{X}}\right) = 0.95$

where a and b are such that

$P\left(\chi^2_{9n} < a\right) = 0.025$ and $P\left(\chi^2_{9n} > b\right) = 0.025.$

so a 95% CI for λ is given by $\left(\frac{a}{2n\bar{X}}, \frac{b}{2n\bar{X}}\right)$.

(c) $9n = 90$, and from tables of χ^2 with 90df we have $a = 65.65$, $b = 118.1$

CI is $\left(\frac{65.65}{2\times 21.47}, \frac{118.1}{2\times 21.47}\right) = (1.53\ ,\ 2.75).$

(ii) (a) $L\left(\lambda\right) \propto \lambda^{4.5n}\exp\left(-\lambda\sum x_i\right)$ so

$\ell\left(\lambda\right) = \left(4.5n\right)\log\lambda - \lambda\sum x_i + \text{constant}$

$\Rightarrow \frac{d\ell}{d\lambda} = 4.5n/\lambda - \sum x_i$ Setting $\frac{d\ell}{d\lambda} = 0 \Rightarrow \hat{\lambda} = \frac{4.5n}{\sum X_i} = \frac{4.5}{\bar{X}}$

(b) $-\frac{d^2\ell}{d\lambda^2} = 4.5n/\lambda^2$ so $s.e.(\hat{\lambda}) \cong \frac{\hat{\lambda}}{\left(4.5n\right)^{1/2}}$

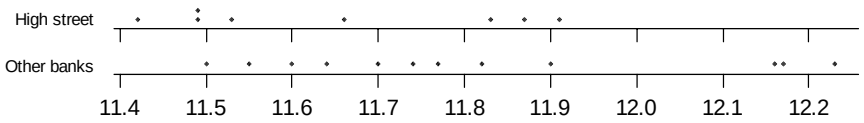
(c) 95% CI is $\hat{\lambda} \pm \left\{1.96 \times s.e.\left(\hat{\lambda}\right)\right\}$

In the case $n = 100$, $\Sigma x = 225.3$,

$$\hat{\lambda} = 4.5/2.253 = 1.9973 \text{ and } s.e.(\hat{\lambda}) \cong \frac{4.5/2.253}{(450)^{1/2}} = 0.0942$$

so CI is $1.9973 \pm (1.96 \times 0.0942)$ i.e. (1.81 , 2.18).

11 (i) Maturity values for high street banks and other banks



(ii) x_1: maturity value for high street bank
x_2: maturity value for other bank

$$\bar{x}_1 = \frac{93.20}{8} = 11.650$$

$$\bar{x}_2 = \frac{141.78}{12} = 11.815$$

$$s_1^2 = \frac{1}{7}\left(1086.0470 - \frac{93.20^2}{8}\right) = 0.038143$$

$$s_2^2 = \frac{1}{11}\left(1675.8224 - \frac{141.78^2}{12}\right) = 0.062882$$

Pooled estimate of σ:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{7(0.038143) + 11(0.062882)}{18} = 0.053261$$

$$\therefore \quad s_p = 0.2308$$

95% confidence interval for $\mu_1 - \mu_2$ is

$$11.650 - 11.815 \pm t_{18}\,(2½\%)\, s_p \sqrt{\frac{1}{8} + \frac{1}{12}}$$

$$= -0.165 \pm (2.101)(0.2308)\sqrt{\frac{1}{8} + \frac{1}{12}}$$

i.e. -0.165 ± 0.2213

i.e. $(-0.386, 0.056)$

i.e. the confidence interval for the difference between the means for high street banks and other banks ($\mu_1 - \mu_2$) is −£386 to £56.

As zero is within the confidence interval, there is insufficient evidence, at 5% level, to reject the null hypothesis that the mean maturity values do not differ for the accounts offered by high street banks and other banks.

(iii) (a) $\dfrac{S_2^2}{S_1^2} \sim F_{11,7}$

under the assumption that the variances are equal for high street and other banks,

i.e. H_0: $\sigma_1^2 = \sigma_2^2$

$$\frac{s_2^2}{s_1^2} = \frac{0.062882}{0.038143} = 1.65$$

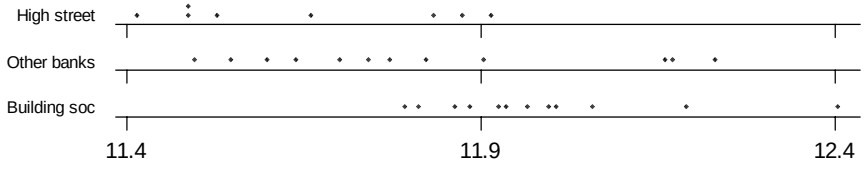
We cannot reject the null hypothesis at the 5% level as the two-sided critical value of a 5% level test is approximately 4.71 (by interpolation using 2½% one-sided F table in Yellow Book).

[OR probability value is $p > 0.20$ as a two-sided 20% level test has a critical value of approximately 2.69.]

(b) The plot in (i) indicates that the assumption of a normal distribution for maturity values is reasonable (but small samples) for both high street and other banks. The assumption of equal variance also seems valid as the test in (iii)(a) is not significant (and the plot above supports this).

(iv) Adding points for building societies to previous plot in (i).

Maturity values



(v) $\Sigma x = 93.20 + 141.78 + 143.83 = 378.81$

$\Sigma x^2 = 1086.0470 + 1675.8224 + 1724.2449 = 4486.114$

$$SS_T = 4486.114 - \frac{378.81^2}{32} = 1.832$$

$$SS_B = \frac{93.20^2}{8} + \frac{141.78^2}{12} + \frac{143.83^2}{12} - \frac{378.81^2}{32} = 0.551$$

$\therefore \; SS_R = SS_T - SS_B = 1.832 - 0.551 = 1.281$

Analysis of variance table

Source of variation	df	SS	MSS
Financial institution types	2	0.551	0.276
Residual	29	1.281	0.044
Total	31	1.832	

$$F = \frac{0.276}{0.044} = 6.27 \text{ on (2, 29) degrees of freedom}$$

$F_{2,29}$ (5%) = 3.328 and $F_{2,29}$ (1%) = 5.42

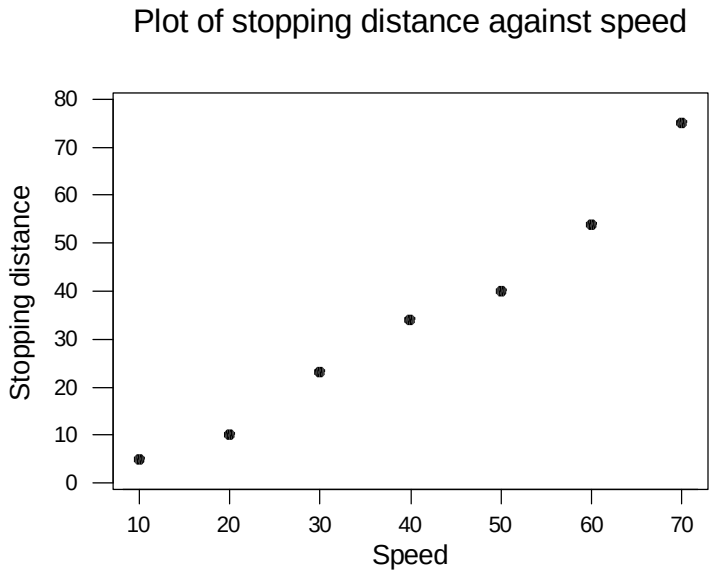
Reject H_0: $\mu_1 = \mu_2 = \mu_3$ (population means are equal) at 1% level.

Strong evidence of differences between the 3 financial institutions.

The plot shows nothing strong enough to invalidate the assumptions of normality and equal variances, even though the variability for the building societies is a bit smaller than for the banks.

(vi) Part (ii) indicates that there are no differences between the mean maturity values of the two types of bank, but (v) indicates that there are differences between the mean maturity values of the 3 types of financial institution. Therefore, in conclusion, it seems that the mean maturity value for building societies is not equal to the mean maturity values of the banks. Also, the plot in (iv) suggests that the maturity value for building societies is higher than the mean maturity values for the banks.

12 (i)



There is a suggestion of a curve but linear regression might still be reasonable.

(ii) $n = 7$

$S_{xx} = \Sigma x^2 - (\Sigma x)^2/n = 14000 - (280)^2/7 = 2800$

$S_{yy} = \Sigma y^2 - (\Sigma y)^2/n = 11951 - (241)^2/7 = 3653.714$

$S_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)/n = 12790 - (280)(241)/7 = 3150$

Model: $E[Y] = \alpha + \beta x$

Slope: $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{3150}{2800} = 1.125$

Intercept: $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 241/7 - (1.125)(280/7) = -10.571$

The equation of the least-squares fitted regression line is:

Distance = $-10.571 + 1.125$ Speed

(iii) $$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{1}{5}\left(3653.714 - \frac{(3150)^2}{2800} \right) = 21.99$$

$$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{21.99}{2800}} = 0.0886$$

95% confidence interval for slope:

$\hat{\beta} \pm t_{n-2}(0.025)$ s.e.($\hat{\beta}$) (df = n – 2 = 5)

= 1.125 ± (2.571)(0.0886) = 1.125 ± 0.228 or (0.897, 1.353)

$\beta = 1$ is within this 95% confidence interval, therefore we would not reject the null hypothesis $\beta = 1$ at the 5% significance level.

(iv) When $x = 50$: $y = -10.571 + 1.125(50) = 45.7$ m

When $x = 100$: $y = -10.571 + 1.125(100) = 101.9$ m

The stopping distance of 45.7 m when the speed is 50 mph can be regarded as a reliable estimate as $x = 50$ is well within the range of the x data values.

However, the stopping distance for a speed of 100 mph may be unreliable as $x = 100$ is outside the range of the data and involves extrapolation.

END OF EXAMINERS' REPORT