INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2012 examinations

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Introduction

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and who are using past papers as a revision aid, and also those who have previously failed the subject. The Examiners are charged by Council with examining the published syllabus. Although Examiners have access to the Core Reading, which is designed to interpret the syllabus, the Examiners are not required to examine the content of Core Reading. Notwithstanding that, the questions set, and the following comments, will generally be based on Core Reading.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report. Other valid approaches are always given appropriate credit; where there is a commonly used alternative approach, this is also noted in the report. For essay-style questions, and particularly the open-ended questions in the later subjects, this report contains all the points for which the Examiners awarded marks. This is much more than a model solution – it would be impossible to write down all the points in the report in the time allowed for the question.

T J Birse
Chairman of the Board of Examiners

July 2012

General comments on Subject CT3

All valid alternative solutions receive credit as appropriate. Rounding errors are not penalised, unless if excessive rounding has led to significantly different answers. In cases where the same error is carried forward to later parts of the question, candidates are not penalised twice. In questions where comments are required, reasonable comments that are different than those provided in the solutions also receive credit.

Comments on the April 2012 paper

The performance of candidates was overall better than in the last session (September 2011), but generally not as strong as in previous diets. There were some excellent scripts achieving very high scores, but a few poor efforts were also recorded at the other end.

In general, answers were not as satisfactory as expected when questions deviated from the usual context, although dealing with commonly examined concepts – e.g. Q1, Q7, Q10. Also, many problems were encountered with straightforward algebraic manipulations, such as differentiation in Q11.

The comments on individual questions that follow concern specific parts that candidates answered poorly and important frequent errors.

1 We want the *first quartile* of the data.

$$Q_1 = \left(\frac{n+2}{4}\right)th \text{ observation counting from below} = 6.5\text{th observation}$$

$$= \frac{1.7+1.8}{2} = 1.75 \text{ hours.}$$

[With alternative definition:

$$Q_1 = \left(\frac{n+1}{4}\right)th \text{ observation counting from below} = 1.725\,]$$

Most answers were quite poor. Many candidates tried to work with a normal or t distribution, when this was not justified (or required). Only a small number of candidates realised that quartiles were required – but then a large proportion of them used the wrong quartile.

2 (i) From plot median = 60.5 days, *IQR* = 112.5 – 26 = 86.5 days.

(ii) The distribution is skewed to the right and a number of values appear to be outliers.

Well answered.

3 P(1st selected is male and 2nd selected is female) = $\frac{64}{100}.\frac{36}{99}$

P(1st selected is female and 2nd selected is male) = $\frac{36}{100}.\frac{64}{99}$

$\Rightarrow$ P(selected students are of different genders) = $2.\frac{64}{100}.\frac{36}{99} = \frac{128}{275} = 0.465$

[*OR* P(selected students are of different genders) = $\frac{\binom{64}{1}\binom{36}{1}}{\binom{100}{2}} = \frac{64\times 36\times 2}{100\times 99} = 0.465\,]$

Very well managed by most candidates. Some tried to calculate the probabilities with replacement.

4 Claim amount ~ $N(\mu, 35^2)$ $\Rightarrow$ difference between 2 claim amounts $D \sim N(0, 2\times35^2)$

i.e. $D \sim N(0, 2450)$

$\Rightarrow P(|D| > 100) = P(|Z| > 100/2450^{1/2}) = P(|Z| > 2.020) = 2 * 0.0217 = 0.043$

Performance was mixed here. There were some problems with specifying the correct variance, and a number of answers gave a one-sided probability.

5 (i) number of claims in a month $X \sim$ Poisson(2)

from tables: $P(X=0) = 0.1353$

[*alternative:* $P(X=0) = e^{-2}$]

(ii) number of claims in a year $X \sim$ Poisson(24)

from tables: $P(X > 30) = 1 - P(X \le 30) = 1 - 0.9042 = 0.0958$

[*alternative:* use normal approximation with continuity correction which gives 0.0923]

Well answered by majority of candidates.

6 (i) Use the normal approximation $\hat{\theta} \sim N\left(\theta, \frac{\theta(1-\theta)}{200}\right)$

to give the 95% confidence interval $\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{200}}$.

With $\hat{\theta} = 70/200 = 0.35$ we obtain 0.35 ± 0.066, that is (0.284, 0.416).

(ii) If we take a large number of samples from this population, we expect 95% of the resulting CIs to include the true value of θ.

There were no problems with the first part. However, many candidates struggled with providing a reasonable interpretation in part (ii).

7 (i) "$X_i \geq x$" ≡ "no heads in first x–1 tosses" so $P(X_i \geq x) = (1-p)^{x-1}$, $x = 1,2,3, \ldots$

[*OR* Recognise (as geometric) and sum the probabilities

$(1-p)^{x-1}p + (1-p)^{x}p + (1-p)^{x+1}p + \ldots = p(1-p)^{x-1}\{1-(1-p)\}^{-1}$]

(ii) (a) "$Y \geq y$" ≡ "all X_i's are $\geq y$"

so $P(Y \geq y) = P(X_1 \geq y, \ldots, X_n \geq y) = P(X_1 \geq y) \ldots P(X_n \geq y)$ (independent)

$= ((1-p)^{y-1})^n = ((1-p)^n)^{y-1}$

(b) The probability in part (a) implies that Y has the same distribution as X, but with $1-(1-p)^n$ in place of p

i.e. $P(Y = y) = r(1-r)^{y-1}$, $y = 1, 2, 3, \ldots$ where $r = 1-(1-p)^n$.

[*OR* $P(Y = y) = P(Y \geq y) - P(Y \geq y+1) = ((1-p)^n)^{y-1} - ((1-p)^n)^y$

$= (1-p)^{n(y-1)}\{1-(1-p)^n\}$ as above]

Most candidates had problems with part (ii). Carefully expressed probability statements are required here. A common error was to try to differentiate the CDF, despite this being a discrete distribution.

8 (i) (a) $SS_R = SS_T - SS_B = 673.5 - 148.3 = 525.2$

$$\hat{\sigma}^2 = \frac{SS_R}{n-k} = \frac{525.2}{36} = 14.59$$

(b) Associated d.f. 36

(ii) Alternatively, an unbiased estimator could be given using only part of the data, e.g. responses from treatment i: $S_i^2 = \dfrac{\sum_j (Y_{ij} - \bar{Y}_{i.})^2}{n_i - 1}$

(iii) The estimator used in part (i) should be preferred as it is based on all data and is therefore more accurate.

Part (i) was well answered, although the df were wrongly given in many answers. Answers in part (ii) were very poor – this question required good understanding of ANOVA concepts and critical thinking.

9 (i) The variables Y_i are independent between them and have a Bernoulli(p) distribution with mean p and variance $p(1-p)$.

Therefore $E(Y) = E(Y_1 + \cdots + Y_{200}) = E(Y_1) + \cdots + E(Y_{200}) = 200p$

$V(Y) = V(Y_1 + \cdots + Y_{200}) = V(Y_1) + \cdots + V(Y_{200}) = 200p(1-p)$

(ii) Again, with Y_i being iid Bernoulli(p) random variables and n being sufficiently large,

the central limit theorem implies that $Y = \sum_{i=1}^{200} Y_i$ follows approximately a normal distribution with mean and variance given by the mean and variance of Y as derived in (i), i.e.

$Y \sim N(200p, 200p(1-p))$

(iii) From (ii) $\hat{P} = Y/200 \sim N(p, p(1-p)/200)$ approximately, which gives a 90% confidence interval of the form $\hat{p} \pm z_{0.05}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{200}}$

$\hat{p} = 0.19$ giving $0.19 \pm 1.6445 \times 0.02774$

i.e. (0.144, 0.236).

Generally well tackled. Some students failed to work with the indicator variables (Bernoulli), which was key to this question.

10 (i) (a) S takes values 0, 1, 2, 3, 4 and we have

$$P(S=0) = 0.4$$
$$P(S=1) = 0.4 \times 0.7 = 0.28$$
$$P(S=2) = 0.4 \times 0.3 + 0.2 \times 0.7^2 = 0.218$$
$$P(S=3) = 0.2 \times 2 \times 0.7 \times 0.3 = 0.084$$
$$P(S=4) = 0.2 \times 0.3^2 = 0.018$$

(b) Hence

$$E(S) = 0.28 + 2 \times 0.218 + 3 \times 0.084 + 4 \times 0.018 = 1.04$$

(ii) (a) $E(S|N=0) = 0, \quad E(S|N=1) = E(X) = 0.7 + 2 \times 0.3 = 1.3$
$E(S|N=2) = E(2X) = 2 \times 1.3 = 2.6$

(b) Hence, $E(S) = 1.3 \times 0.4 + 2.6 \times 0.2 = 1.04$ as before.

Most candidates encountered problems here, as they failed to work out the probability function from first principles in part (i). Also, many did not recognise this as a compound distribution type of question.

11 (i) (a) $L(\theta) = k\theta^{n_A}\left(\theta^2\right)^{n_B}\left(1-\theta-\theta^2\right)^{n_C}$

$$\ell(\theta) = (n_A + 2n_B)\log\theta + n_C \log\left(1-\theta-\theta^2\right) + c$$

$$U(\theta) = \frac{n_A + 2n_B}{\theta} - \frac{n_C(1+2\theta)}{1-\theta-\theta^2}$$

(b) Setting $U(\theta) = 0 \Rightarrow (n_A + 2n_B)(1 - \theta - \theta^2) = n_C\theta(1 + 2\theta)$

$\Rightarrow \hat{\theta}$ satisfies

$$(n_A + 2n_B + 2n_C)\theta^2 + (n_A + 2n_B + n_C)\theta - (n_A + 2n_B) = 0$$

(ii) (a) $$\frac{\partial U(\theta)}{\partial\theta} = -\frac{n_A + 2n_B}{\theta^2} - n_C \frac{2\left(1-\theta-\theta^2\right) - (1+2\theta)(-1-2\theta)}{\left(1-\theta-\theta^2\right)^2}$$

$$= -\frac{n_A + 2n_B}{\theta^2} - \frac{n_C\left(3+2\theta+2\theta^2\right)}{\left(1-\theta-\theta^2\right)^2}$$

(b) $$E\left[-\frac{\partial U(\theta)}{\partial\theta}\right] = \frac{n\theta + 2n\theta^2}{\theta^2} + \frac{n\left(1-\theta-\theta^2\right)\left(3+2\theta+2\theta^2\right)}{\left(1-\theta-\theta^2\right)^2}$$

$$= \frac{n\left(1+4\theta-\theta^2\right)}{\theta\left(1-\theta-\theta^2\right)}$$

(iii) (a) $\hat{\theta}$ satisfies $149\theta^2 + 116\theta - 83 = 0 \Rightarrow \hat{\theta} = 0.4525$

(b) Using the Cramer-Rao lower bound, estimate of asymptotic standard error is

$$\left[\frac{\hat{\theta}(1-\hat{\theta}-\hat{\theta}^2)}{100\left(1+4\hat{\theta}-\hat{\theta}^2\right)}\right]^{1/2} = 0.0244$$

(c) 95% CI for θ is 0.4525 ± (1.96×0.0244) i.e. 0.4525 ± 0.0478 i.e. (0.405, 0.500)

Part (i) was very well answered. The differentiation in part (ii) was problematic. Also, many candidates could not identify the random variable for which expectation was required in (ii)(b).

12 (i) $\hat{p} = \frac{\bar{X}}{n} = \frac{220+2*90}{400*2} = 0.5$

We obtain the following table to test H_0:

Possible realisation of X_i	0	1	2
Number of observations	90	220	90
expected frequency under H_0	100	200	100
$(f_j - e_j)^2$	100	400	100
$(f_j - e_j)^2 / e_j$	1	2	1

The test-statistic is $= \sum_{j=0}^{2} (f_j - e_j)^2 / e_j$. For the given data the value of C is $c = 4$.

C is χ^2-distributed with 3−1−1 = 1 degree of freedom.

H_0 is rejected since the $(1-\alpha)$-quantile (α = 0.05) of the χ^2-distribution with one degree of freedom is 3.841 < 4.

(iii) Since the estimated value is 0.5, any reasonable test will not reject that value, since the value 0.5 will always be in the acceptance region of the test. In other words, 0.5 will always be in any confidence interval around the estimate 0.5.

(iv) We now have: $H_0 : X_i \sim Bin(2,0.5)$ and

$H_1 : X_i$ does not follow $Bin(2,0.5)$ (emphasis on both Bin, $p = 0.5$)

(v) The value of the test-statistic is still $c = 4$ but the distribution of C is now a χ^2-distribution with 3−1 = 2 degrees of freedom.

Now H_0 is NOT rejected at a 5%-level since the $(1-\alpha)$-quantile ($\alpha = 0.05$) of the χ^2-distribution with two degrees of freedom is $5.991 > 4$.

(vi) The result in part (ii) states that a binomial distribution does not fit the data well and is rejected. However, in part (iii) we found that, under the assumption of a binomial distribution, $p_0 = 0.5$ cannot be rejected. A specific binomial distribution with parameter $p = 0.5$ is not rejected in part (v) for the same data. The reason is that the additional degree of freedom in part (v) allows for a larger value of the test-statistic under the null.

Most candidates answered very well the parts of this question that concerned "knowledge" and "application" aspects of the tests. However, there were problems with the comments and reasoning.

13 (i) $S_{xx} = 24,\ S_{yy} = 2500,\ S_{xy} = 195$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\ S_{yy}}} = 0.796084$$

(ii) $W = \frac{1}{2}\log\frac{1+r}{1-r}$ is normally distributed with mean $\frac{1}{2}\log\frac{1+\rho}{1-\rho}$ and standard deviation $1/\sqrt{n-3}$

observed value of W is $w = 1.087828$

Under H_0 the mean of W is 1.098612

And the standard deviation is 0.447214

p-value is $P[W < 1.087828] = P[Z < (1.087828 - 1.098612)/\ 0.447214]$
$= P[Z < -0.024113527] = 1 - F(0.024113527) > 0.49$

No evidence against the null hypothesis.

(iii) $Y_i = a + bX_i + \varepsilon_i$

For b we obtain: $\hat{b} = \left\{n\sum x_i y_i - \sum x_i \sum y_i\right\}\left\{n\sum x_i^2 - (\sum x_i)^2\right\}^{-1}$

And therefore: $\hat{b} = \frac{8*10695 - 56*1500}{8*416 - 56^2} = 8.125$

$$\text{And } \hat{a} = \frac{1}{8}\left(\sum y_i - \hat{b}\sum x_i\right) = 130.625$$

(iv) $R^2 = 0.796084^2 = 0.634$

(v) Any increase in school quality by 1 index-point, leads to an increase of 8.125 in the house price index.

Mostly very well answered.

END OF EXAMINERS' REPORT