INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 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 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. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

D C Bowie
Chairman of the Board of Examiners

December 2012

General comments on Subject CT3

For CT3 exams some questions admit alternative solutions or different ways in which the provided answer can be determined. All valid alternative solutions or answers received credit as appropriate. Rounding errors were not penalised, unless excessive rounding led to significantly different answers. In cases where the same error was carried forward to later parts of the answer, candidates were not penalised twice. In questions where comments were required, reasonable comments that were different from those provided in the solutions also received full credit.

Comments on the September 2012 paper

The overall performance was similar to recent sessions, but not as strong as in the last diet (April 2012). A good number of candidates achieved very high scores, although the high end of the mark distribution was negatively affected by the inability of most candidates to tackle certain questions.

As in past sessions, questions corresponding to parts of the syllabus that had not been recently examined were generally poorly answered (e.g. Q4). This highlights the need for candidates to cover the whole syllabus when they revise for the exam and not only rely on themes appearing in past papers. Problems were also recorded in questions where basic algebraic manipulations were required, such as in Q9(ii) and Q12(i).

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

1 $mean = \frac{1}{160}(54 + 2*58 + 3*28) = \frac{254}{160} = 1.5875$ 1

Median = value between 80^{th} and 81^{st} observation = 2 1

Mode = 2 1

Generally well answered. Note that the median is NOT the 80^{th} observation, as some candidates quoted.

2 (i) $Q_1 = \left(\frac{n+2}{4}\right)$th observation counting from below $= 5.5$th observation

$= \frac{335+368}{2} = 351.5$ 1

$Q_3 = \left(\frac{n+2}{4}\right)$th observation counting from above $= 5.5$th observation from above

$= \frac{807+686}{2} = 746.5$ 1

$IQR = Q_3 - Q_1 = 395$ 1

[With alternative definition:

$Q_1 = \left(\frac{n+1}{4}\right)$th observation counting from below $= 343.25$,

$Q_3 = \left(\frac{n+1}{4}\right)$th observation counting from above $= 776.75$, $IQR = 433.5$.]

(ii) The *length* of the interval containing the *central* half of the claim sizes is 395.
1

The vast majority of candidates calculated the quartiles correctly, although some were confused with their definition. Part (ii) was not very well answered.

3 $E[X] = 1\times0.3 + 2\times0.2 + 3\times0.1 = 1$

$\Rightarrow V[X] = (0-1)^2\times0.4 + (1-1)^2\times0.3 + (2-1)^2\times0.2 + (3-1)^2\times0.1$
$= 0.4 + 0.2 + 0.4 = 1$

(*OR* via $E[X^2] = 2$)

$V[Y] = 4V[X] = 4$

[*OR:* Directly from the distribution of Y, which is $Y = 10, 12, 14, 16$ with probabilities 0.4, 0.3, 0.2, 0.1 respectively.]

No particular problems encountered here. There are a variety of different methods for obtaining the correct answer.

4 Let $f_Z(z)$ be the density of $Z = U + X$.

$$f_Z(z) = \int_u f_U(u)\, f_X(z-u)\, du = \int_0^1 f_X(z-u)\, du$$

$$= \int_{z-1}^{z} f_X(x)\, dx = F_X(z) - F_X(z-1)$$

where we have used the substitution $u = z - x$, and where F_X is the distribution function of X.

This question was very poorly answered. A large number of candidates did not attempt it at all, while many others did not follow any reasonable approach. Note that this is based on standard bookwork, viz. Unit 6, Section 3 in the Core Reading.

5 (i) P(none of class A) = P(all 10 of class B or C) = $(0.8)^{10} = 0.1074$

(ii) (a) Let B = number of class B.

Note that $B \sim$ binomial (10, 0.5), so that $E(B) = (10)(0.5) = 5$

(b) $P(B > 5) = 1 - P(B \le 5) = 1 - 0.6230 = 0.3770$

[0.6230 is from tables; alternatively by evaluation]

This was generally very well answered. A common error in part (ii) (b) was to calculate P(B < or = 5) instead of P(B > 5).

6 (i) Population mean = 8θ

So MME is solution of $\bar{X} = 8\theta \Rightarrow$ MME $= \dfrac{\bar{X}}{8}$

(ii) $E\left(\dfrac{\bar{X}}{8}\right) = \dfrac{1}{8}E(\bar{X}) = \dfrac{1}{8}(8\theta) = \theta$

Bias = $E\left(\dfrac{\bar{X}}{8}\right) - \theta = 0$ (i.e. MME is unbiased for θ).

(iii) (a) Since MME is unbiased, $MSE\left(\frac{\bar{X}}{8}\right) = var\left(\frac{\bar{X}}{8}\right) = \frac{8\theta^2}{64n} = \frac{\theta^2}{8n}$

(b) MME gets more efficient (MSE gets smaller) as sample size increases.

There was a mix of quality in the answers, especially in parts (ii) and (iii). Attention to detail is required when determining the expected value and variance of functions of sample statistics (here the sample mean).

7 (i) With the larger sample of 100 claims the standard error of the sample mean will be smaller, giving a narrower confidence interval.

(ii) The replacement of the extreme value will give a smaller sample mean, which means that the interval will be shifted to the left.

The variance of the sample will also be smaller, which will again give a narrower interval.

Many candidates recognised the correct effect on the interval, without being able to justify it properly. Note that reasonably accurate wording is important in providing the comments and justification required here.

8 $E[N] = \sum n\, P(N = n) = 0.3 + 0.6 + 0.6 + 0.4 = 1.9$

$E[N^2] = \sum n^2\, P(N = n) = 0.3 + 1.2 + 1.8 + 1.6 = 4.9$

$V[N] = E[N^2] - (E[N])^2 = 1.29$

Also $E[Y] = \exp(\mu + \sigma^2/2) = e^{0.55} = 1.73325$

$V[Y] = (E[Y])^2(\exp(\sigma^2) - 1) = 1.73325^2 * (e^{0.1} - 1) = 0.31595$

Using known results

$E[S] = E[N]\ E[Y] = 1.9 * 1.73325 = 3.293$

$V[S] = E[N]\ V[Y] + V[N]\ (E[Y])^2 = 0.60031 + 3.87536 = 4.476$

Some frequent errors were due to mis-interpretation of the mean and variance of the log-normal distribution.

9 (i) (a) mean = $\dfrac{\alpha}{\lambda} = 4$ and s.d. = $\sqrt{\dfrac{\alpha}{\lambda^2}} = \sqrt{8} = 2.8$

(b) As claims are non-negative and the s.d. is quite large relative to the mean, then the distribution will be quite positively skewed.

(ii)
$$F(x) = \int_0^x \frac{1}{4} t e^{-\frac{1}{2}t}\,dt$$

$$= -\frac{1}{2}\int_0^x t\,d(e^{-\frac{1}{2}t})$$

$$= -\frac{1}{2}[te^{-\frac{1}{2}t}]_0^x + \frac{1}{2}\int_0^x e^{-\frac{1}{2}t}\,dt$$

$$= -\frac{1}{2}xe^{-\frac{1}{2}x} - [e^{-\frac{1}{2}t}]_0^x$$

$$= 1 - (1 + \frac{1}{2}x)e^{-\frac{1}{2}x}$$

(iii) (a) $F(x) = u$ i.e. $1 - (1 + \frac{1}{2}x)e^{-\frac{1}{2}x} = u$

(b) This equation would have to be solved numerically

(c) Using $u = 0.66$ on the vertical axis, we invert to get $x = 4.5$ on the horizontal axis.

In part (ii) many candidates failed to integrate correctly. A lot of problems were caused by not using the correct limits for the integral. In part (iii) a popular answer was to use "trial-and-error", which is not an appropriate approach here.

10 (i) $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

with ε_{ij} being i.i.d. $N(0,\sigma^2)$

In particular, it is assumed that the variance is the same in all groups.

(ii)

Source of variation	d.f.	SS	MSS
Between regions	3	4.4655	1.4885
Residual	16	8.892	0.55575

(iii) $H_0 : \tau_i = 0$ for all groups i

$F = 2.6784$ should be from F distribution with 3,16 d.f.

From the tables we know that this gives a *p*-value of 0.086 (with interpolation).

Reject at 10%, not at 5%, some but very weak evidence against H_0

Mainly well answered. Care is required in calculating the p-values correctly. Also, a number of candidates had difficulties in writing down a sensible form of the ANOVA model in part (i).

11 (i) This is an F distribution with 10, 8 degrees of freedom.

(ii) The interval is given by $\left(\frac{S_A^2 / S_B^2}{F_{10,8,0.025}}, \frac{S_A^2 / S_B^2}{F_{10,8,0.975}} \right)$

From tables $F_{10,8,0.025} = 4.295$ and $F_{10,8,0.975} = 1/F_{8,10,0.025} = 1/3.855$

giving $\left(\frac{0.692/0.813}{4.295}, \ (0.692/0.813)*3.855 \right)$ = (0.198, 3.281)

(iii) As the two samples are independent we have that

$$V\left(\bar{X}_A - \bar{X}_B\right) = \frac{V\left(X_A\right)}{11} + \frac{V\left(X_B\right)}{9} = \sigma^2(1/11 + 1/9)$$

Normality of the data then gives that $Z = \dfrac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sigma\sqrt{\frac{1}{11} + \frac{1}{9}}} \sim N(0,1)$

We are also given that $Y = \dfrac{18S_p^2}{\sigma^2} \sim \chi^2_{18}$ and with Z and Y being independent we can use that $\dfrac{Z}{\sqrt{Y/18}} \sim t_{18}$ to obtain $\dfrac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{S_p\sqrt{\dfrac{1}{11}+\dfrac{1}{9}}} \sim t_{18}$.

(iv) First compute $s_p^2 = \dfrac{10*0.692 + 8*0.813}{18} = 0.74577 \Rightarrow s_p = 0.864$

Then with $t_{18,0.025} = 2.101$ the interval is given by (4.05 – 4.36) ± 2.101 * 0.864 $(1/11 + 1/9)^{1/2}$ i.e. (– 1.126, 0.506).

(v) The interval includes the value 0, suggesting that there is no difference in the mean effectiveness of the two vaccines.

Part (iii) was problematic for many candidates. Many candidates struggled to provide a 'proof' that had sufficient rigour. There were errors also in determining the endpoints of the CI in part (ii), often due to using the wrong percentiles of the F distribution.

12 (i) Likelihood function

$$L(p) = p^6 (20p)^{114} (10p)^{62} (1-31p)^{18} = Cp^{182}(1-31p)^{18}$$

$$\log L(p) = \log C + 182\log p + 18\log(1-31p)$$

$$\frac{\partial}{\partial p}\log L(p) = \frac{182}{p} + \frac{18}{1-31p}(-31) = 0$$

$$\frac{182}{p} = \frac{558}{1-31p} \Rightarrow \frac{p}{182} = \frac{1-31p}{558} \Rightarrow \frac{p}{182} + \frac{31p}{558} = \frac{1}{558} \Rightarrow p\left(\frac{1}{182} + \frac{31}{558}\right) = 1/558$$

$$\hat{p} = 0.02935$$

(ii) H_0 : The proposed distribution is the true distribution of the data with non-specified parameter p (it is important to mention that the parameter itself is not part of the null hypothesis)

Under H_0 and using $\hat{p} = 0.02935$ from (i)(a) we obtain the following expected frequencies

Body-Mass-Index	< 18.5	18.5–25	25–30	>30
Expected frequency	5.87	117.4	58.7	18.03

Test-statistic is 0.286915

from a Chi-square distribution with 2 d.f.

The test statistic has a very small value, and there is no evidence against the null.

(iii) $P\left[BMI > 30\right]$

$$= P\left[BMI > 30 | single\right] P\left[single\right] + P\left[BMI > 30 | married\right] P\left[married\right]$$

$$= \frac{12}{158} * 0.5 + \frac{6}{42} * 0.5 = 0.1094$$

(iv) H_0 : Marital status is independent of BMI

Under H_0 we have:

Marital Status	Body-Mass-Index				Total
	< 18.5	18.5–25	25–30	>30	
Single	4.74	90.06	48.98	14.22	158
Married	1.26	23.94	13.02	3.78	42
Total	6	114	62	18	200

Use χ^2 test.

Test-statistic: $$C = \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{\left(f_{ij} - \frac{f_{i.} * f_{.j}}{n}\right)^2}{\frac{f_{i.} * f_{.j}}{n}} = 8.528399$$

C is χ^2-distributed with $(2-1)(4-1)=3$ degrees of freedom.

p-value: $P\left[C > 8.528399\right] < 1 - 0.9616 = 0.0384$

Therefore, we reject H_0 at 5% level, but not at the 1% level.

There were errors in part (i) caused by failure to differentiate correctly. In part (iv) alternative solutions involving merging of adjacent categories were given full redit where correct. However note that merging the first and last column is not correct in this question.

13 (i) (a) The scatter plot suggests a positive linear association between weight and stopping distance.

(b) $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.892$

(ii) We want to test $H_0\!: \rho = 0$ against $H_1\!: \rho > 0$.

Need to assume that data come from a bivariate normal distribution.

Fisher's (standardised) transformation statistic is given by

$$\frac{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)}{\sqrt{1/(n-3)}} = \frac{\sqrt{7}}{2}\log\left(\frac{1.892}{0.108}\right) = 3.79$$

and under H_0 this should be a value from the $N(0,1)$ distribution.

This gives P-value $= \Pr(Z \geq 3.79) \approx 0.0001$, so there is very strong evidence against H_0 and we conclude that motorcycle weight and stopping distance are positively correlated.

[Or by considering critical values of N(0,1) distribution.]

(iii) (a) $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{36.51}{3344.1} = 0.01092$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 14.17 - 0.01092 * 337.7 = 10.4823$

Fitted line is $\hat{y} = 10.48 + 0.01092x$

(b) $R^2 = \dfrac{S_{xy}^2}{S_{xx}S_{yy}} = \dfrac{36.51^2}{3344.1 * 0.501} = 0.7956$

This gives the proportion of total variation explained by the model.

(Note that R^2 can also be computed as r^2.)

(c) For every additional unit (kilogram) of weight the stopping distance is expected to increase by $\hat{\beta} = 0.01092$ metres. So, for 10 kilograms of weight the distance is expected to increase by 0.109 meters.

Generally adequately answered. Identifying the correct hypotheses in part (ii) was problematic in some cases, while many candidates failed to assume bivariate normality.

END OF EXAMINERS' REPORT