INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2016

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.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

Luke Hatter
Chair of the Board of Examiners
December 2016

A. General comments on the *aims of this subject and how it is marked*

1. The aim of the Probability and Mathematical Statistics subject is to provide a grounding in the aspects of statistics and in particular statistical modelling that are of relevance to actuarial work.

2. Some of the questions in this paper admit alternative solutions from these presented in this report, or different ways in which the provided answer can be determined. All mathematically correct and valid alternative solutions or answers received credit as appropriate.

3. Rounding errors were not penalised, but candidates lost marks where excessive rounding led to significantly different answers.

4. In cases where the same error was carried forward to later parts of the answer, candidates were given full credit for the later parts.

5. In questions where comments were required, reasonable comments that were different from those provided in the solutions also received full credit where appropriate.

B. General comments on *student performance in this diet of the examination*

1. Performance was generally good, and most candidates demonstrated very good understanding and application of core topics in probability and mathematical statistics.

2. The pass rate was in line with previous sessions and there were a number of excellent scripts achieving very high scores.

3. Questions involving maximum likelihood estimation, requiring calculus and algebra skills, are often not very competently answered (for example Question 9 part (iv) in this paper). Candidates are encouraged to practise these type of questions.

4. Candidates are also encouraged to work on questions that require deeper and more rounded understanding of concepts in the Core Reading – e.g. Question 9 parts (iii)–(v) in this paper.

C. Pass Mark

The Pass Mark for this exam was 60.

Solutions

Q1 (i) Mean = $413/20 = 20.65$ [1]

(ii) SD = $\sqrt{\frac{1}{19}\left[12{,}311 - 20*\left(\frac{413}{20}\right)^2\right]} = 14.11$ [2]

(iii) Median = 19.5 [1]

(iv) IQR = 29.5–10 = 19.5 [2]

[Total 6]

> Very well answered. Note that the Core Reading mentions two different ways for calculating the quartiles in part (iv). Both ways were given full credit when applied correctly.

Q2 (i) Mean [1]

(ii) Median – mode could be at the lowest point of the distribution and the mean could be raised by a long tail. [2]

[Total 3]

> Part (i) was answered correctly in most cases. Answers in part (ii) were mixed, with many candidates failing to justify their answers convincingly.

Q3 (i) Number of claims, $N \sim \text{Bin}(10000, 0.003)$

By CLT number of claims approximately $N(30, 29.91)$ [2]

Continuity correction applies

$$P(N > 40) = P(N > 40.5) = P\left(\frac{N-30}{\sqrt{29.91}} > \frac{40.5-30}{\sqrt{29.91}}\right)$$

$$= 1 - \Phi(1.920) = 1 - 0.973 = 0.027$$ [2]

(ii) 95% interval is $30 \pm Z_{0.975}\sigma = 30 \pm 1.96*5.469 = (19.28, 40.72)$ [2]

(iii) The probability that the result in any year will lie in the interval is 0.95 so there is a 5% probability that the company will see a result outside that range. [2]

[Total 8]

Performance here was generally good, but there were some common errors. In part (i) reference to the CLT was often omitted, as was the continuity correction (or it was applied in the wrong direction). In part (iii) many candidates did not identify the main point, about the 95% coverage of the interval.

Q4 Since the sample size for each portfolio is one, we have that $\hat{\beta}_A = 134$ and $\hat{\beta}_B = 91$. [1]

Using the normal approximation we find:

$$\hat{\beta}_A - \hat{\beta}_B = X_A - X_B \sim N(\beta_A - \beta_B, \beta_A + \beta_B) \quad [1]$$

The confidence interval is then given by

$$134 - 91 \pm 2.5758\sqrt{134+91} = 43 \pm 2.5758 \times 15 = [4.4, 81.6] \quad [2]$$

[Total 4]

Performance in this question was mixed. Many candidates seemed unsure as to what is the correct size of each sample (portfolio).

Q5 Company A: $n_1 = 150$, and $\hat{\theta}_1 = \frac{45}{150} = 0.3$

Company B: $n_2 = 150$, and $\hat{\theta}_2 = \frac{33}{150} = 0.22$ [1]

Combined: $n = 300$, and $\hat{\theta} = \frac{45+33}{300} = 0.26$

The test statistic is: $\dfrac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\dfrac{\hat{\theta}(1-\hat{\theta})}{n_1} + \dfrac{\hat{\theta}(1-\hat{\theta})}{n_2}}} = \dfrac{0.3 - 0.22}{\sqrt{0.26(1-0.26)\left(\dfrac{1}{150} + \dfrac{1}{150}\right)}} = 1.58$ [1]

The significance probability of the test of H_0: $\theta_1 = \theta_2$ against a two sided alternative is $2*P(Z > 1.58) = 2*(1 - 0.94295) = 0.114$. [1]

(Or, compare with the 97.5 percentile (1.96) of the normal distribution.)

Therefore there is insufficient evidence to reject the null hypothesis, and we conclude that there is no difference between the underlying proportions. [1]

[Total 4]

Mixed performance. The question was answered competently by well prepared candidates, but there were errors mainly concerning the use of a common θ estimate in the denominator. Also, note that this is a 2-sided test.

Q6 (i) $$\int_x\int_y f_{XY}(x,y)\,dydx = \int_0^1\int_x^1 kx^2y^2\,dydx = \int_0^1\left[\frac{k}{3}x^2y^3\right]_{y=x}^{y=1}dx$$ [2]

$$= \frac{k}{3}\int_0^1 x^2 - x^5\,dx = \frac{k}{3}\left[\frac{x^3}{3} - \frac{x^6}{6}\right]_0^1 = \frac{k}{3}\left(\frac{1}{3} - \frac{1}{6}\right) = \frac{k}{18}$$ [1]

Want integral equal to $1 \Rightarrow k = 18$ [1]

(ii) $$f_Y(y) = \int_x f_{XY}(x,y)\,dx = \int_0^y 18x^2y^2\,dx = \left[6x^3y^2\right]_{x=0}^{x=y} = 6y^5$$ [2]

(iii) $$P(X > 0.5|Y = 0.75) = \int_{0.5}^{0.75} f_{(x|Y=0.75)}(x)\,dx = \int_{0.5}^{0.75} f_{XY}(x,0.75)/f_Y(0.75)\,dx$$ [2]

$$= \int_{0.5}^{0.75} 18x^2 0.75^2/(6\times 0.75^5)dx = 3\times\left(\frac{4}{3}\right)^3\left[\frac{x^3}{3}\right]_{0.5}^{0.75} = 0.7037$$ [1]

[Total 9]

Parts (i) and (ii) were very well answered. Dealing with the conditional distribution in part (iii) was problematic.

Q7 (i) Observed frequencies:

	Large City	*Small City*	*Countryside*	*Total*
No claim	370	390	410	1,170
One claim	93	99	87	279
More than one claim	37	11	3	51
Total	500	500	500	1,500

Expected frequencies (under independence)

	Large City	*Small City*	*Countryside*	*Total*
No claim	390	390	390	1,170
One claim	93	93	93	279
More than one claim	17	17	17	51
Total	500	500	500	1,500

[2]

Values of $(e-f)^2 / e$:

	Large City	*Small City*	*Countryside*
No claim	1.025641	0	1.025641
One claim	0	0.387097	0.387097
More than one claim	23.52941	2.117647	11.52941

[2]

Test statistic:

$$\sum \frac{(e-f)^2}{e} = 2*1.026 + 2*0.387 + 23.53 + 2.118 + 11.53 = 40.004$$ [1]

This compares to a χ^2-distribution with (3–1)*(3–1) = 4 degrees of freedom. [1]

The value of the test statistic is clearly very high and the null hypothesis is rejected at all reasonable significance levels. We conclude that the number of claims depends on the place of living. [1]

(ii) (a) P [small city] = 0.25+0.06+0.02=0.33 [1]

(b) P [more than one claim]=0.04+0.02+0.01=0.07 [1]

(c) P [more than one claim | large city]

$$= \frac{0.04}{0.23+0.06+0.04} = \frac{0.04}{0.33} = 0.1212$$ [2]

(d) P [countryside | no claim or one claim]

$$= \frac{0.27+0.06}{0.23+0.25+0.27+3*0.06} = \frac{0.33}{0.93} = 0.3548$$ [2]

(e) P [small city or large city | one claim or more than one claim]

$$= \frac{2*0.06+0.04+0.02}{3*0.06+0.04+0.02+0.01} = \frac{0.18}{0.25} = 0.72 \qquad [2]$$

[Total 15]

Generally well answered. There were some calculation errors in part (i), while in part (ii) not all steps were always shown clearly.

Q8 (i) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{89.91}{81.15} = 1.108$ [1]

$\hat{a} = \bar{y} - \hat{\beta}\bar{x} = 7.15 - 1.108 \times 5.93 = 0.58$ [1]

Fitted model is: $\hat{y} = 0.58 + 1.108x$ [1]

(ii) A = 1, B = 8, C = 99.61/1 = 99.61, D = 21.63/8 = 2.704 [2]

(iii) $\hat{\sigma}^2 = 21.63/8 = 2.704$ [1]

(iv) (a) R^2 gives the proportion of the total variation of y that is explained by x. [1]

(b) $R^2 = 99.61/121.24 = 0.822$ [1]

(v) We want to test H_0: $\beta = 0$ v. H_1: $\beta \neq 0$ [1]

Under $H_0$the value $MS_{REG} / MS_{RES} = 99.61/2.704 = 36.84$ should be a value from the $F_{1,8}$ distribution. [2]

The 0.99 quantile of $F_{1,8}$is 11.26 [1]

We have strong evidence to reject $H_0$and we conclude that there is linear relationship between x and y. [1]

[Total 13]

Generally very well answered. However, many candidates struggled with the interpretation of the coefficient of determination in part (iv)(a) – this is an important (and very widely used) concept in regression analysis.

Q9 (i) For MME we want:

$$E(X)=\bar{x}\Rightarrow\frac{\tilde{\alpha}}{\tilde{\lambda}}=\bar{x}\text{ and }V(X)=s^2\Rightarrow\frac{\tilde{\alpha}}{\tilde{\lambda}^2}=s^2 \qquad [2]$$

These give

$$\tilde{\lambda}=\frac{\bar{x}}{s^2}=500/150^2=0.022 \quad \text{and} \quad \tilde{\alpha}=\bar{x}\tilde{\lambda}=\frac{\bar{x}^2}{s^2}=500^2/150^2=11.111 \qquad [2]$$

(ii) $E(S)=E(N)E(X)=14*500=7{,}000$ [1]

$V(S)=E(N)V(X)+V(N)E(X)^2=14*(150^2+500^2)=3{,}815{,}000$ [2]

(iii) (a) $$L(\lambda;x)=\prod_{i=1}^{n}\left(\frac{\lambda^{\alpha^*}}{\Gamma(\alpha^*)}x_i^{\alpha^*-1}e^{-\lambda x_i}\right)=\frac{\lambda^{n\alpha^*}}{\left\{\Gamma(\alpha^*)\right\}^n}e^{-\lambda\sum_1^n x_i}\prod_1^n x_i^{\alpha^*-1} \qquad [1]$$

$$l(\lambda)=n\alpha^*\log(\lambda)-\lambda\sum_1^n x_i+\text{constant} \qquad [1]$$

$$\frac{d}{d\lambda}l(\lambda)=0\Rightarrow\frac{n\alpha^*}{\hat{\lambda}}=\sum_1^n x_i\Rightarrow\hat{\lambda}=\frac{\alpha^*}{\bar{x}}=\frac{\alpha^*}{500} \qquad [1]$$

We can confirm that this is a maximum as $\frac{d^2}{d\lambda^2}l(\lambda)=-\frac{n\alpha^*}{\lambda^2}<0$. [1]

(b) CRLB $= -1/E\left(\frac{d^2}{d\lambda^2}l(\lambda)\right)=\lambda^2/(n\alpha^*)$ [1]

So, approximately, $\hat{\lambda}\sim N\left(\lambda,\ \lambda^2/5\alpha^*\right)$ [2]

(c) The approximation of the distribution relies on a very small sample ($n = 5$) and therefore may not be valid. [2]

(iv) (a) We now have

$$l(\alpha,\lambda)=n\alpha\log(\lambda)-n\log\left(\Gamma(\alpha)\right)+(\alpha-1)\sum_{i=1}^{n}\log(x_i)-\lambda\sum_{i=1}^{n}x_i \qquad [1]$$

From part (iii): $\hat{\lambda}=\frac{\alpha}{\bar{x}}$, and [1]

$$\frac{d}{d\alpha}l(\alpha,\lambda)=0 \Rightarrow n\log\left(\hat{\lambda}\right)-n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}+\sum_{i=1}^{n}\log(x_i)=0 \qquad [1]$$

$$\Rightarrow n\log\left(\frac{\hat{\alpha}}{\bar{x}}\right)-n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}+\sum_{i=1}^{n}\log(x_i)=0$$

$$\Rightarrow \log(\hat{\alpha})-\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}=\log(\bar{x})-\frac{\sum_{i=1}^{n}\log(x_i)}{n} \qquad [1]$$

(b) The equation for $\hat{\alpha}$ cannot be solved analytically, so a numerical solution is required. Then $\hat{\alpha}$ can be substituted in the equation for $\hat{\lambda}$.

[1]

[Total 21]

Parts (i), (ii) and (iii)(a) were very well answered, while performance in part (iii)(b) was mixed. In part (iii)(c) many candidates failed to comment on the effect of the small sample size on the approximation. Many answers to part (iv) were weaker. This is a maximum likelihood derivation, requiring moderate calculus and algebra skills.

Q10 (i) ANOVA table:

Source of variation	*df*	*SS*	*MSS*
Between groups	2	1.785	0.8925
Residual	27	6.579	0.2437
Total	29	8.364	

[2]

$$F=\frac{0.8925}{0.2437}=3.662 \text{ on } 2, 27 \text{ df} \qquad [1]$$

$F_{2,27}(5\%) = 3.354$, $F_{2,27}(1\%) = 5.488$, so reject the null hypothesis at the 5% level. [1]

There is evidence against the null hypothesis. We conclude that there are differences in the mean level of nervousness scores among the three groups.

[1]

(ii) Residuals given as:

$$r_{A1}=y_{A1}-\bar{y}_A=0.693-\frac{6.475}{10}=0.0455$$

$$r_{B1} = y_{B1} - \bar{y}_B = 0 - \frac{6.356}{10} = -0.6356 \qquad [2]$$

(iii) (a) Interval will be based on $\frac{SSR}{\sigma^2} \sim \chi^2_{27}$ [1]

and therefore a 95% CI is given as

$$\left(\chi^2_{27}(0.025) \le \frac{SSR}{\sigma^2} \le \chi^2_{27}(0.975) \right)$$

i.e. $\left(\frac{SSR}{\chi^2_{27}(0.975)} \le \sigma^2 \le \frac{SSR}{\chi^2_{27}(0.025)} \right)$ [1]

which gives $\left(\frac{6.579}{43.19} \le \sigma^2 \le \frac{6.579}{14.57} \right)$, i.e. (0.1523, 0.4515) [1]

(b) Interval now based on $\frac{SSB}{\sigma^2} \sim \chi^2_2$ [1]

Working similarly as above we obtain a 95% CI as:

$\left(\frac{1.785}{7.378} \le \sigma^2 \le \frac{1.785}{0.05064} \right)$, i.e. (0.2419, 35.2488) [2]

(iv) This interval is too wide. Notice that its validity depends on H_0 being true, which is rejected at 5% level. [2]

(v) We could perform a two-sample *t*-test of control mean = treatment mean by combining the data for the two treatment groups (and using samples of sizes 10 and 20). [2]

[Total 17]

Part (i) was well answered. However, performance in the remaining part of the question was not strong. Part (ii) involves basic understanding of the concept of residuals, and is not examined often. Parts (iii) and (iv) require deeper understanding of the construction and assumptions behind confidence intervals, while part (v) examines a more rounded knowledge and understanding of statistical testing.

END OF EXAMINERS' REPORT