INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2017

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
July 2017

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. Answers requiring algebraic manipulations and elements of calculus contained a considerable number of mathematical errors (e.g. Question 7(iii)). Candidates are encouraged to revise relevant core mathematical topics and practise their skills as part of their preparation for the CT3 examination.

C. Pass Mark

The Pass Mark for this exam was 60.

Solutions

Q1

Mid point	77.5	91	114	143	173	193.5
Count	1	8	8	8	8	6

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{5406.5}{39} = 138.63$$ [2]

$$\text{Variance} = \frac{\sum fx^2 - n\bar{x}^2}{n-1} = \frac{803899.8 - 39 \times 138.63^2}{38} = 1431.24$$ [1½]

Standard dev = 37.83 [½]

[Total 4]

This was generally very well answered. There were some minor mistakes in the calculations, typically taking the wrong midpoints for each range of values.

Q2

(i) $$F(x) = \int_0^x \nu\lambda u^{\nu-1} \exp(-\lambda u^{\nu}) du = \left[-\exp(-\lambda u^{\nu}) \right]_0^x = 1 - \exp(-\lambda x^{\nu})$$ [2]

(ii) Using the inverse CDF method

$$u = 1 - \exp(-\lambda x^{\nu}) \Rightarrow x = \left(-\frac{\log(1-u)}{\lambda} \right)^{\frac{1}{\nu}}$$ [1]

and with $\nu = 1.1$, $\lambda = 0.2$ and $u = 0.671$ we have $x = 4.756$. [1]

[Total 4]

The question was reasonably well answered, with some errors in part (i) where many candidates used infinity as the upper limit in the integration.

Q3

(i) $$\text{Cov}(X - Y, X + Y) = \text{Var}(X) - \text{Cov}(Y, X) + \text{Cov}(X, Y) - \text{Var}(Y) = 1 - 1 = 0$$ [2]

Alternative solution:

$$E(Z^-) = E(X - Y) = E(X) - E(Y) = 0$$

$$E(Z^+) = E(X + Y) = E(X) + E(Y) = 0 + 0 = 0$$

$$E\left(Z^{-}Z^{+}\right)=E\left\{\left(X-Y\right)\left(X+Y\right)\right\}=E\left(X^{2}-Y^{2}\right)=E\left(X^{2}\right)-E(Y^{2})=0$$

So: $Cov\left(Z^{-},Z^{+}\right)=E\left(Z^{-}Z^{+}\right)-E\left(Z^{-}\right)E\left(Z^{+}\right)=0$

(ii) Since cor(Z^-, Z^+) = cov(Z^-, Z^+)/{sd(Z^-) sd(Z^+)} it follows that Z^- and Z^+ are uncorrelated. [1]

[Total 3]

Generally well answered. A typical mistake in part (i) was not substituting for Var(*X*) and Var(*Y*). Note that in part (ii) the answer needs to be justified, e.g. by connecting correlation and covariance using the definition.

Q4 Denote by A, B, C the event that policyholder belongs to the corresponding group. Also let *F* be the event that a policyholder makes a claim.

(i) $P(F) = P(F|\text{A})P(\text{A}) + P(F|\text{B})P(\text{B}) + P(F|\text{C})P(\text{C})$

$= 0.13*0.1 + 0.03*0.38 + 0.02*0.52 = 0.0348$ [3]

(ii) $P\left(A|F\right)=\dfrac{P(\text{F}\mid\text{A})P\left(\text{A}\right)}{P\left(F\right)}=\dfrac{0.13*0.1}{0.0348}=0.374$ [2]

[Total 5]

Generally very well answered, with no particular issues.

Q5 (i) The CLT states that as $n\to\infty$, approximately, $\sum_{i=1}^{n}X_i$ approaches the $N(n\mu,\ n\sigma^2)$ distribution. [2]

(ii) The mean of $X_i$is 0.5 and its variance is 0.25.

Therefore, from CLT, $Y=\sum_{i=1}^{50}X_i \sim N(25, 12.5)$ approximately. [2]

(iii) Exact distribution is $Y=\sum_{i=1}^{50}X_i \sim$ Gamma(50, 2). [2]

(iv) Generally the gamma distribution is an asymmetric distribution. Here, as n is large, the CLT suggests that the distribution of Y is approximately normal, and therefore symmetric. [2]

[Total 8]

The performance in parts (i)–(iii) was generally good. In part (i) the approximation needs to be clearly indicated in the answer. In part (iv) a typical issue was not giving a direct conclusion on the shape based on the approximation.

Q6 (i) $SS_R = 14\left(16^2 + 19^2 + 16^2\right) = 12,222$ [1]

$$\bar{Y} = \frac{70+75+83}{3} = 76 \quad [1]$$

$$SS_B = 15\left((70-76)^2 + (75-76)^2 + (83-76)^2\right) = 1,290 \quad [2]$$

$$F_{2,42} = \frac{\frac{SS_B}{2}}{\frac{SS_R}{42}} = \frac{1290}{2}\frac{42}{12222} = 2.216 \quad [1]$$

This is clearly a rather small value compared to the 5% point from a $F_{2,42}$ distribution which is 3.22 (from Tables, using interpolation), so the null hypothesis is not rejected. We conclude that there is no evidence that the type of car has an impact on the monthly amount of money spent on petrol. [1]

(ii) We need to assume equal variances and also that observations are independent and normally distributed. [1]

Pooled variance: $\frac{14\times16^2 + 14*16^2}{28} = 16^2 = 256$ [1]

$$\bar{x}_L - \bar{x}_S \pm t_{0.025,28}\times16\sqrt{\frac{2}{15}} = 13 \pm 2.048\times16\times\sqrt{\frac{2}{15}} = [1.035, 24.965] \quad [2]$$

(iii) $0 \notin [1.035, 24.965]$, therefore, we would reject the null hypothesis of equal amounts spent on petrol for large cars and small cars. [1]

[Total 11]

Generally well answered. In part (ii) some candidates did not give the assumptions of the model. In part (iii) a number of candidates performed a

full t-test – this was not required but was given full credit where performed correctly. Also note that there are alternative ways for calculating the sums in part (i) and these also received full credit when performed correctly.

Q7 (i) $\bar{X} = \widehat{E[X]} = 1/\hat{\lambda}$ [1]

$\hat{\lambda} = 1/\bar{X}$ [1]

(ii) $\hat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{71} = 0.014085$ [1]

(iii) $L(\theta) = L(\theta) = \theta^n \prod_{i=1}^{n} Z_i \exp(-\theta Z_i X_i)$ [1]

$l(\theta) = n \log\theta + \sum_{i=1}^{n} \log Z_i - \theta \sum_{i=1}^{n} Z_i X_i$ [1]

$l'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} Z_i X_i = 0$ [1]

$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} Z_i X_i}$ [1]

This is indeed a maximum since the second derivative of $l(\theta)$ is $-n\theta^{-2} < 0$ for $\theta = \hat{\theta} > 0$. [1]

[Total 8]

Parts (i) and (ii) were very well answered. Performance in part (iii) was mixed. Most candidates exhibited a soundunderstanding of how to approach the question; many encountered problems related to writing the sums and the products and doing the required maths.

Q8 (i) $P[X=0]=1-p-\frac{p}{2}-\frac{p}{4}-\frac{p}{8}=1-\frac{8+4+2+1}{8}p=1-\frac{15}{8}p=\frac{8-15p}{8}$ [1]

(ii) To ensure that $0 \le P[X=k] \le 1$ for all k we only need to check this condition for $k=0,1$, and we need that $p \in \left[0, \frac{8}{15}\right]$. All other probabilities will then also be between 0 and 1. [2]

(iii) Let $N_4 = n - N_0 - N_1 - N_2 - N_3$, the number of policies with more than three claims.

MLE:

$$L(p) = \left(\frac{8-15p}{8}\right)^{N_0} p^{N_1} \left(\frac{p}{2}\right)^{N_2} \left(\frac{p}{4}\right)^{N_3} \left(\frac{p}{8}\right)^{N_4}$$

$$\log L(p) = N_0 \log(8-15p) + (N_1 + N_2 + N_3 + N_4) \log p + C$$

where C is a constant which does not depend on p. [1]

First derivative:

$$l'(p) = \frac{-15N_0}{8-15p} + \frac{N_1 + N_2 + N_3 + N_4}{p} = \frac{-15N_0}{8-15p} + \frac{n-N_0}{p} = 0$$ [1]

Solving this equation:

$$\frac{15N_0}{8-15p} = \frac{n-N_0}{p} \Leftrightarrow \frac{8-15p}{15N_0} = \frac{p}{n-N_0} \Rightarrow \hat{p} = \frac{8}{15}\frac{n-N_0}{n}$$ [2]

(iv) N_0 has a binomial distribution since it counts the outcome "no claim" in n independent trials [1]

The distribution of N_0 is $B\left(n, \frac{8-15p}{8}\right)$. [1]

(v) $E[\hat{p}] = \frac{8}{15}\left(1 - \frac{E[N_0]}{n}\right)$ [1]

And therefore with $E(N_0) = n\frac{8-15p}{8}$ we obtain $E[\hat{p}] = p$, so $\hat{p}$ is unbiased. [1]

(vi) $\text{Var}\left(\hat{p}\right)=\left(\frac{8}{15n}\right)^2\text{Var}\left(N_0\right)=\left(\frac{8}{15}\right)^2\frac{1}{n}\left(1-\frac{15}{8}p\right)\frac{15}{8}p$ [1]

$\hat{p}=\frac{8}{15}\frac{300-100}{300}=\frac{8}{15}\times\frac{2}{3}=\frac{16}{45}=0.35555$ [½]

Estimated variance of $\hat{p}$:

$\left(\frac{8}{15}\right)^2\frac{1}{n}\left(1-\frac{15}{8}\times\frac{16}{45}\right)\frac{15}{8}\times\frac{16}{45}=\left(\frac{8}{15}\right)^2\frac{1}{300}\times\frac{1}{3}\times\frac{2}{3}=0.0002107$ [½]

(vii) (a) $E\left[X\right]=p+2\frac{p}{2}+3\frac{p}{4}+4\frac{p}{8}=3.25p$ [1]

$E\left[X^2\right]=p+\frac{4}{2}p+\frac{9}{4}p+\frac{16}{8}p=7.25p$ [1]

$\text{Var}\left(X\right)=E\left[X^2\right]-E\left[X\right]^2=7.25p-10.5625p^2$ [1]

(b) Let $Y=\sum_{i=1}^{300}X_i$ be the total number of claims.

Expected total number of claims: $E\left[Y\right]=300\times3.25p=975p=195$ [½]

$\text{Var}(Y)=300\left(7.25p-10.5625p^2\right)=308.25$ [½]

(c) $E\left[S\right]=195\times100=19,500$ [1]

$\text{Var}\left(S\right)=195\times20^2+308.25\times100^2=3,160,500$ [1½]

$\text{Std}\left(S\right)=\sqrt{3,160,500}=1,777.78$ [½]

[Total 20]

Parts (i) and (ii) were very well answered. In part (iii) most candidates followed the correct approach, but started with a wrong likelihood function. Answers in part (iv) often did not provide a reasonable justification of why this was a binomial distribution. In part (v) most candidates gave answers that exhibited understanding of what bias is but failed to arrive at the final result. Performance in parts (vi) and (vii) was mixed, with many candidates using wrong formulas and missing the variance of the linear combination.

Q9 (i) A random sample should be independent and identically distributed. As people are chosen at random the methodology should give a random sample. [2]

(ii) While the sample chosen will be independent, they will not necessarily be representative of the population as a whole. In many places phone ownership may be restricted by economic, cultural or geographic limitations so some parts of the population may be excluded. [2]

(iii) (a) $C.I. = \left(\frac{s_1^2}{s_2^2} \frac{1}{F_{24,12;0.975}}, \frac{s_1^2}{s_2^2} F_{12,24;0.975} \right)$ [1]

$= \left(\frac{20.2^2}{15.6^2} \frac{1}{3.019}, \frac{20.2^2}{15.6^2} 2.541 \right) = (0.555, 4.260)$ [2]

Alternative solution:

CI for $\frac{\sigma_2^2}{\sigma_1^2}$ is (0.234, 1.802)

(b) As 1 lies in the confidence interval it is reasonable to assume the standard deviations are the same. [1]

(iv) $H_0 : \mu_1 = \mu_2$ *vs.* $H_1 : \mu_1 \neq \mu_2$ [1]

Pooled variance $s_P^2 = \left((n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 \right) / (n_1 + n_2 - 2)$

$= \left(24 \times 20.2^2 + 12 \times 15.6^2 \right) / (25 + 13 - 2) = 353.15$ [1]

Test statistic

$= (\mu_1 - \mu_2) / s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (61.6 - 50) / \sqrt{353.15 \left(\frac{1}{25} + \frac{1}{13} \right)} = 1.805$ [2]

$t_{36;0.975} = 2.028 >$ test statistic [1]

So do not reject H_0 at a 5% significance level [1]

(v) Now $H_1 : \mu_1 < \mu_2$ [1]

Test statistic is the same, but now use $t_{36;0.95} = 1.688$ [1]

This time we reject H_0 [1]

(vi) The results of the tests in parts (iv) and (v) were different. The additional information allowed us to choose a more appropriate alternative hypothesis.

[2]

[Total 19]

The performance in this question was mixed. In parts (i) and (ii) many candidates failed to demonstrate that they can distinguish between a sample being random and being representative. A typical error in part (iii) was the use of wrong critical values. Note that a 2-sided test is required in part (iv) – some candidates used a 1-sided test instead. Parts (v) and (vi) were generally well answered.

Q10 (i) $S_{gg} = \left(206.2462 - \frac{28.68^2}{9} \right) = 114.8526$ [1]

$S_{gd} = 15.55855 - \frac{2.97 * 28.68}{9} = 6.09415$ [1]

$\hat{\beta} = \frac{S_{gd}}{S_{gg}} = \frac{6.09415}{114.8526} = 0.05306$ [1]

$\hat{\alpha} = \bar{d} - \beta \bar{g} = \frac{2.97 - 0.05306 * 28.68}{9} = 0.1609$ [1]

So $d = 0.1609 + 0.05306g$ [1]

(ii) $S_{dd} = 1.33525 - \frac{2.97^2}{9} = 0.35515$ [1]

$\widehat{\sigma^2} = \frac{1}{7}\left(S_{dd} - \frac{S_{dg}^2}{S_{gg}} \right) = \frac{1}{7}\left(0.35515 - \frac{6.09415^2}{114.8526} \right) = 0.004542$ [1]

$\text{test statistic} = \hat{\beta} / \sqrt{\frac{\widehat{\sigma^2}}{S_{gg}}} = 0.05306 / \sqrt{\frac{0.004542}{114.8526}} = 8.438$ [1]

$t_{7;0.975} = 2.365$ (two sided) so reject $H_0 : \beta = 0$ [1]

(iii) If g_0=3 then $\hat{d}_0 = 0.1609 + 0.05306 * 3 = 0.32008$ [1]

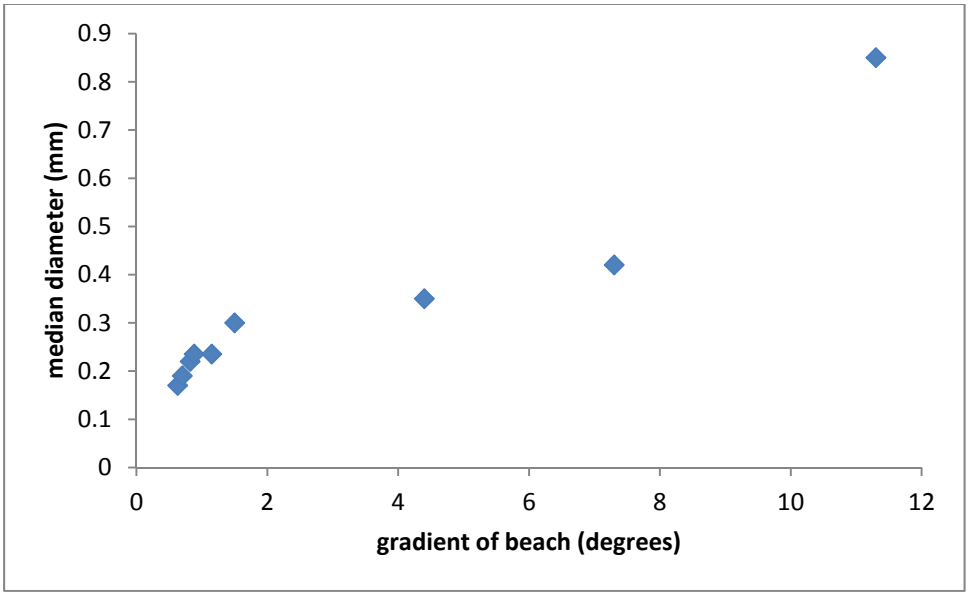
$$\text{Var}\left(\hat{d}_0\right) = \left\{\frac{1}{n} + \frac{\left(g_0 - \bar{g}\right)^2}{S_{gg}}\right\}\hat{\sigma}^2 = \left\{\frac{1}{9} + \frac{\left(3 - 3.187\right)^2}{114.8526}\right\} * 0.004542$$

$$= 5.060 \times 10^{-4} \quad [2]$$

$$\text{C.I.} = \hat{d}_0 \pm t_{7;0.975} * \text{Var}\left(\hat{d}_0\right)^{\frac{1}{2}} = 0.32008 \pm 2.365 * \left(5.060 \times 10^{-4}\right)^{\frac{1}{2}}$$

$$= (0.267, 0.373) \quad [2]$$

(iv) (a)



[2]

(b) With only three observations for g>1.5, the slope is determined by a small amount of data. Getting more observations in that range would give a better analysis. [2]

Alternatives (up to 2 marks total): could try a data transformation (e.g. logarithmic); other non-linear regression; more data.

[Total 18]

Parts (i) and (ii) were very well answered. However, a small number of candidates performed the regression using the wrong response and explanatory variables (g on d). Also in part (ii) some candidates attempted a test using the correlation coefficient. For full marks the equivalence of the two tests should be explicitly mentioned. In part (iii) there were some computational errors, while there were a few problems with the plot in part (iv) with inappropriate scales, missing axes labels etc.

END OF EXAMINERS' REPORT