INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2016 (with mark allocations)

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.

F Layton
Chair of the Board of Examiners
June 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, unless 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 only penalised once.

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. Candidates performed generally well, demonstrating good understanding and application of core topics in probability and mathematical statistics.

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

3. As in previous examination diets, a question involving the likelihood function of a non-standard model (censored observations in Question 7), was poorly answered. The idea of likelihood estimation is central in statistics, and candidates should be able to apply it also in situations that vary from basic simplified scenarios.

4. The paper included questions testing the understanding of fundamental concepts in probability and statistics (e.g. questions 6 part (ii) and 11 part (iv)). It is important that candidates at this level can demonstrate the ability to correctly and meaningfully interpret such concepts.

C. Pass Mark

The Pass Mark for this exam was 60%.

Solutions

Q1 (i) Range is [2, 35], or $35 - 2 = 33$. [1]

(ii) Median given as the observation with rank 8.5, i.e. 16. [1]

(iii) Mean is $266/16 = 16.625$. [1]

[TOTAL 3]

Well answered. In part (i) both answers shown above were given full credit.

Q2 (i) $\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}}$ [1]

(ii) $\mathrm{Cov}(X, aX + b) = aV(X)$ [1]

$V(Y) = V(aX + b) = a^2V(X)$ [1]

For $a < 0$ we obtain $\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \dfrac{aV(X)}{\sqrt{V(X)a^2V(X)}} = -1$ [1]

[TOTAL 4]

Part (i) was generally well answered. Some candidates gave the sample correlation coefficient, which is not what the question required.

Performance in part (ii) was not very competent, with many candidates failing to recognise the negative correlation.

Q3 (i) $z = \log y \Rightarrow y = e^z \therefore \dfrac{dy}{dz} = e^z$

$f(y) = \dfrac{\theta}{y^{\theta+1}}$

$\therefore f(z) = \theta(e^z)^{-(\theta+1)}\left|e^z\right| = \theta e^{-\theta z}$ [1½]

and since $y > 1 \Rightarrow z > 0$. [1½]

(ii) This is the probability density function of an exponential distribution with parameter θ. [1]

[OR: Obtain using the distribution function:

$F_Z(z) = \Pr(Z \le z) = \Pr(Y \le e^z) = F_Y(e^z)$

Also, $F_Y(y) = \int_1^y \frac{\theta}{u^{\theta+1}} du = -\left[u^{-\theta}\right]_1^y = 1 - y^{-\theta}$

Thus, $F_Z(z) = 1 - e^{-\theta z}$ implying an Exp(θ) distribution.]

[TOTAL 4]

Part (i) required some work involving probabilistic arguments and was not well answered. There were no problems with part (ii).

Q4 (i) Let A = Accident

Expected cost of an accident= Average Cost(Major)*P(Major|A)
+ Average Cost(Minor)*P(Minor|A)
= 0.2*1000 + 0.8*50 =£240 [1]

Expected total cost of accidents = £240 *1000 * 0.1 = £24,000 [1]

(ii) Let O = Office Staff &F = Factory Staff. Then $P(A|O) = 0.5\ P(A|F)$

$P(A) = P(A|O)P(O) + P(A|F)P(F) = P(A|O)*0.18 + 2*P(A|O)*0.82$ [2]

$\Rightarrow P(A|O) = 0.1 / (0.18+.82*2) = 0.0549$ [1]

(iii) $P(O|A) = \frac{P(A\,|\,O)P(O)}{P(A)} = \frac{0.0549*0.18}{0.1} = 0.099.$ [2]

[TOTAL 7]

Part (i) was well answered with most candidates giving a fully or partially correct answer. Answers in parts (ii) and (iii) were generally poor; some candidates were confused with the conditional probabilities and considered conditioning on the wrong event.

Q5 (i) $X_A \sim \text{Bin}(50, 0.5)$, $X_B \sim \text{Bin}(50, 0.5)$. [1]

$E(D) = 0$, $V(D) = V(X_A) + V(X_B) = 2*50*0.5*0.5 = 25$ [1]

From CLT: $D \sim N(0, 25)$ [1]

[Alternatively, using CLT $X_A, X_B \sim N(25,12.5)$, and therefore $D \sim N(0, 25)$.]

(ii) $Z = (X_A - X_B - 0) / 5 \sim N(0,1)$

$P(A \text{ wins}) = P(D \geq 5) = P(D > 4.5)$ with continuity correction [1]

$= P(Z > 4.5/5) = P(Z > 0.9) = 1 - 0.8159 = 0.1841$ [1]

[TOTAL 5]

Mixed performance in both parts. The use of the CLT is important in part (i), so reference to it must be made for full credit. A common error in part (ii) was not using or wrongly applying the continuity correction.

Q6 (i) $E[X] = (a + b) / 2 \Rightarrow b = 2E[X] - a$

$$\text{Var}(X) = (b-a)^2 / 12 = (2E[X] - 2a)^2 / 12 = (E[X] - a)^2 / 3.$$ [2]

$$\hat{a} = \bar{x} - \sqrt{3}s = 0.791 \Rightarrow \hat{b} = 2 * \bar{x} - \hat{a} = 18.129.$$ [2]

(ii) The largest observation is greater than our estimate of b in part (i). This would suggest the uniform distribution is not a good fit to this data, or the largest observation is a mistaken observation. This also highlights a potential weakness of the method of moments. [2]

[TOTAL 6]

Part (i) was very well answered. In part (ii) some candidates failed to recognise that there is zero probability of having a sample value outside the range given by the parameters.

Q7 $L(\lambda)=\left(\prod_{i=1}^{6}\lambda e^{-\lambda x_i}\right)\left(\prod_{i=7}^{10}P[X_i>3]\right)=\lambda^6 e^{-\lambda\sum_{i=1}^{6}x_i}(e^{-3\lambda})^4$ [1]

$l(\lambda)=6\log(\lambda)-\lambda\sum_{i=1}^{6}x_i-12\lambda=6\log(\lambda)-\lambda\left(\sum_{i=1}^{6}x_i+12\right)=6\log(\lambda)-24\lambda$ [1]

$l'(\lambda)=\frac{6}{\lambda}-24$ [1]

$\hat{\lambda}=\frac{6}{24}=0.25.$ [1]

[TOTAL 4]

> Mixed performance. Most candidates worked through the maximisation steps, but many could not figure out the part of the likelihood corresponding to the censored information.

Q8 (i) Test H_0 that means are the same, against H_1 that at least one pair different.

	SS	*df*	*MS*	*F*
SS_B	55.672	2	27.836	56.35
SS_R	13.332	27	0.494	
SS_T	69.004	29		

[3]

$F_{2,27}$ = 5.488 at 1% so reject H_0 that means are the same. [2]

(ii) (a) $\hat{\sigma}^2=0.494,\ t_{27;0.975}=2.052$ [2]

CI width = $2*\hat{\sigma}*t_{27,0.975}*\left(\frac{1}{10}+\frac{1}{10}\right)^{\frac{1}{2}}=2*0.703*\frac{2.052}{5^{\frac{1}{2}}}=1.290$ [2]

(b) From part (ii)(a), LSD = 1.290/2 = 0.645. Therefore, all strain means are significantly different from each other. [1]

(iii) The samples may not be independent of each other and, so, the ANOVA may not be valid. [2]

[TOTAL 12]

Part (i) was well answered. In part (ii)(a) there were mixed efforts with many partial answers providing intervals rather than the required width. Part (ii)(b) was poorly answered with some answers contradicting the findings in part (a). There seemed to be some misinterpretation in part (iii), where a number of candidates referred to a "smaller sample, affecting accuracy of test", rather than addressing the independence issue. Sensible comments along these lines were given credit.

Q9 (i) Total no. of claims $\sum f_i n_i = 871\times 0 + 117\times 1 + 5\times 2 + 5\times 3 + 2\times 4 = 150$ [1]

Rate = 150/1000 = 0.15 [1]

(ii)

x	*0*	*1*	*2*	*3*	*4*	*Total*
	1850	140	5	3	2	2000
$P(X\le x)$	0.86071	0.98981	0.9995	0.99998	1	
$P(X=x)$	0.86071	0.1291	0.00969	0.00048	2*E*–05	
e_x	1721.42	258.2	19.38	0.96	0.04	

[2]

So sums >5

x	*0*	*1*	*2*	*Total*
	1850	140	10	2000
$P(X\le x)$	0.86071	0.98981	1	
$P(X=x)$	0.86071	0.1291	0.01019	
e_x	1721.42	258.2	20.38	
$(a_x-e_x)^2/e_x$	9.60	54.11	5.29	69.00

[1]

So test statistic is 69.00 on 2d.f.Compared to χ_2^2 clearly significant at any reasonable level, so reject H_0 that number of claims follows Poisson(0.15). [3]

(iii) Let *Y*=claims from group with no claims in last 5 years, *Z*= claims from group with claim in last 5 years

$$\bar{y} = \frac{6,420,000}{70} = 91714,\ s_y^2 = \frac{8.76\times10^{11} - \frac{(6420000)^2}{70}}{69} = 4.162\times10^9 \qquad [2]$$

$$\bar{z} = \frac{9,220,000}{80} = 115250, s_z^2 = \frac{1.52\times10^{12} - \frac{(9220000)^2}{80}}{79} = 5.790\times10^9 \qquad [2]$$

Under assumption of equal variances, pooled variance is

$$s_p^2 = (69*4.162\times10^9 + 79*5.790\times10^9)/(69+79) = 5.031\times10^9 \qquad [1]$$

Test statistic = $(91714-115250)/\sqrt{5.031\times10^9*\left(\frac{1}{70}+\frac{1}{80}\right)} = -2.027$ [1]

$t_{148;0.975}$ is between 1.96 and 1.98 so reject H_0 that means are the same [2]

(iv) Overall number of claims for each category $= 2200*0.1*0.5 = 110$ [1]

Mean of total amount:

$$E(S) = E(S_1)+E(S_2) = E(N_1)\times E(X_1)+E(N_2)\times E(X_2)$$
$$= 110\times94,000+110\times120,000 = £23.54\text{m}. \qquad [2]$$

Variance of total amount:

$$V(S) = V(S_1)+V(S_2)$$
$$= E(N_1)\times V(X_1)+V(N_1)\times[E(X_1)]^2$$
$$+E(N_2)\times V(X_2)+V(N_2)\times[E(X_2)]^2$$
$$= 110(70,000^2+94,000^2)+110(70,000^2+120,000^2) = 3.63396\times10^{12} \qquad [2]$$

[TOTAL 21]

Parts (i) and (iii) were well answered. In part (ii) there was a wide range of answers, some involving errors in calculations. For full marks, categories with small expected frequencies must be combined. Some candidates performed a test of equality between the two rates, which is not what the question requires here.

Part (iv) was not answered competently. Most candidates gave a correct answer for the mean, but many did not attempt to calculate the variance at all – or gave a wrong answer. Note here that if a common variance is assumed, then $V(S) = 2*110*70,000^2 = £^2 1.078\times10^{12}$.

A small number of candidates worked along these lines, and full credit was given if this answer was provided.

Q10 (i) Mean: $\bar{x} = 3{,}852/20 = 192.6$ [1]

Median: (184+185)/2 = 184.5 [1]

SD: $s = \sqrt{(759{,}348 - 20 * 192.6^2)/19} = 30.31$ [1]

(ii) $\bar{x} \pm t_{0.025,19} \dfrac{s}{\sqrt{20}} = \left[192.6 - 2.093\dfrac{30.31}{\sqrt{20}},\ 192.6 + 2.093\dfrac{30.31}{\sqrt{20}}\right] = [178.4, 206.8]$

[2]

We need to assume that the claim size is normally distributed. [1]

(iii) $\left[\sqrt{\dfrac{19s^2}{\chi^2_{0.025,19}}}, \sqrt{\dfrac{19s^2}{\chi^2_{0.975,19}}}\right] = \left[30.31\sqrt{\dfrac{19}{32.85}}, 30.31\sqrt{\dfrac{19}{8.907}}\right] = [23.05, 44.27]$ [2]

(iv) The χ^2 distribution is not symmetric. [1]

(v) $\left[1.6 - 1.96\sqrt{\dfrac{1.6}{50}}, 1.6 + 1.96\sqrt{\dfrac{1.6}{50}}\right] = [1.25, 1.95]$ [2]

(vi) The length of the confidence interval is $2 * 1.96\sqrt{\dfrac{1.6}{n}} < 1/2$. Therefore,

$\dfrac{1.6}{n} < \left(\dfrac{1}{4 * 1.96}\right)^2$ giving $n > 1.6 * (4 * 1.96)^2 = 98.3$ [2]

So, $n \geq 99$. [1]

(vii) We are now having a compound Poisson distribution for total S.

Expected number of claims from 5,000 policies:

$E(N) = 1.6 * 5{,}000 = 8{,}000$

Expected value of total claim size:

$E(S) = E(N) * E(X) = 8{,}000 * 192.6 = 1{,}540{,}800$ [1]

$V(S) = E(N)V(X) + V(N)[E(X)]^2 = 1.6 * 5{,}000 * (30.31^2 + 192.6^2)$
$= 304{,}107{,}649$ [1]

SD = 17,438.68 [1]

[TOTAL 17]

Parts (i), (ii), (iii) and (v) were very well answered. Note that in part (ii) the assumptions must be stated for full marks.

In part (iv) a common error was not mentioning the asymmetry of the distribution, and part (vi) was tackled with mixed success. Part (vii) was not particularly well answered – many candidates failed to treat this as a compound distribution.

Q11 (i) (a) $S_{xx} = 15600 - \frac{376^2}{10} = 1462.4 \quad S_{yy} = 311.44 - \frac{(55.4)^2}{10} = 4.524$

$$S_{xy} = 2014.5 - \frac{(376)(55.4)}{10} = -68.54$$ [3]

$$r = \frac{-68.54}{\sqrt{1462.4 \times 4.524}} = -0.843$$ [1]

(b) There appears to be strong negative linear correlation between mileage and price. [1]

(ii) Least squares estimates:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{-68.54}{1462.4} = -0.0469$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{55.4}{10} + 0.0469\left(\frac{376}{10}\right) = 7.303$$ [2]

Line given as: $\hat{y} = 7.30 - 0.0469x$ [1]

(iii) (a) $$\hat{\sigma}^2 = \frac{\left(s_{yy} - \frac{s_{xy}{}^2}{s_{xx}}\right)}{n-2} = \frac{\left(4.524 - \frac{(-68.54)^2}{1462.4}\right)}{8} = 0.163957$$ [1]

$$se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0.163957}{1462.4}} = 0.01059$$ [1]

and $t_{8,0.025} = 2.306$ [1]

95% CI for $\hat{\beta}$ is given by $-0.0469 \pm 2.306(0.01059)$ [1]

i.e. -0.0469 ± 0.02442 or $(-0.071, -0.022)$ [1]

(b) Since the value zero is not included in the interval, the suggestion in the article seems valid. [2]

(iv) $\hat{\beta} \times 5 = -0.2345$

Estimated difference in price will be £234.34. [2]

[TOTAL 17]

Parts (i), (ii) and (iii) are typical questions on correlation and regression and were tackled without problems by most candidates. Note that in part (iii)(b) clear reasoning must be provided for full marks. Performance in part (iv) was mixed, with some answers suggesting inadequate understanding of the interpretation of regression coefficients.

END OF EXAMINERS' REPORT