INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2018

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
June 2018

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 satisfactory and most candidates demonstrated good understanding and application of core topics in probability and mathematical statistics.

2. Topics that were not particularly well answered in this paper include sampling distributions (e.g. Q4) and conditional expectation involving joint distributions (e.g. Q9). Candidates are advised to revise all parts of the syllabus.

3. Answers requiring calculus elements (e.g. integration in part Q9) contained a considerable number of mathematical errors. 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 55.

Solutions

Q1 (i) $\text{mean} = \dfrac{\Sigma f_i x_i}{\Sigma f_i} = \dfrac{60}{20} = 3$ [1]

20 observations so median is 10.5^{th} value = 3 [1]

mode = 4 [1]

(ii) $\Sigma f_i x_i^2 = 206$ [1]

$$s = \sqrt{\frac{206 - 20*3^2}{19}} = 1.170$$ [1]

[Total 5]

This question was very well answered by most candidates. In part (ii), a common mistake was using n instead of n-1 in the denominator.

Q2 (i) $$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^3 - 3x_i^2\bar{x} + 3x_i\bar{x}^2 - \bar{x}^3\right)$$

$$= \frac{1}{n}\left(\sum x_i^3 - 3\bar{x}\sum x_i^2 + 3\bar{x}^2\sum x_i - n\bar{x}^3\right)$$ [1]

$$= \frac{1}{7}\left[8750.972 - 3\frac{73.4}{7}792.22 + 3\left(\frac{73.4}{7}\right)^2 73.4 - 7\left(\frac{73.4}{7}\right)^3\right]$$

$$= -\frac{29.316}{7} = -4.188$$ [1]

(ii) (a) Coefficient of skewness = $\dfrac{\sum\left(x_i - \bar{x}\right)^3 / n}{\left(\sum\left(x_i - \bar{x}\right)^2 / n\right)^{3/2}}$ [1]

(b) $\sum\left(x_i - \bar{x}\right)^2 / n = \left(\sum x_i^2 - \left(\sum x_i\right)^2 / n\right) / n = (792.22 - 73.4^2/7)/7 = 3.224$

[1]

Coefficient of skewness = $-4.188 / 3.224^{1.5} = -0.723$ [1]

[Total 5]

Answers in part (i) were mixed, with many candidates struggling with the formula and failing to arrive at the correct answer. Parts (ii) and (iii) were answered well. Some candidates used n instead of n-1 in the formula for the variance.

Q3 (i) If X is the time (in minutes) arriving late, we have $X \sim \text{Exp}(0.2)$ [0.5]

$$P(X > 10) = \int_{10}^{\infty} 0.2e^{-0.2t}dt = \left[-e^{-0.2t}\right]_{10}^{\infty} = 0 + e^{-2} = 0.1353$$ [0.5]

(Or use CDF from tables.)

(ii) Let Y be the number of students arriving more than 10 minutes late.

$Y \sim \text{Bin}(20, 0.1353)$ [1]

$$P(Y < 2) = P(Y = 0) + P(Y = 1)$$ [1]

$$= (1 - 0.1353)^{20} + \binom{20}{1}(0.1353)(1 - 0.1353)^{19} = 0.2255$$ [2]

[Total 5]

Part (i) was very well answered in general. Part (ii) was not answered correctly by a number of candidates who were not able to identify the binomial distribution.

Q4 (i) $$P\left[S^2 > \sigma^2\right] = P\left[\frac{S^2}{\sigma^2} > 1\right] = P\left[\frac{(n-1)S^2}{\sigma^2} > n-1\right] = P\left[\chi_8^2 > 8\right] = 0.4335$$ [2]

(ii) Since $\bar{X}$ and S^2 are independent: [1]

$$P[\bar{X} > \mu \mid S^2 > \sigma^2] = P[\bar{X} > \mu] = 0.5$$ [1]

(iii) $$P[\bar{X} - \mu > \sigma] = P\left[\frac{\bar{X} - \mu}{\sigma} > 1\right] = P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{9}} > 3\right] = P[Z > 3] = 0.00135$$

[2]

(iv) $P\left[\bar{X}-\mu > S\right] = P\left[\dfrac{\bar{X}-\mu}{S/\sqrt{9}} > 3\right] = P[t_8 > 3]$ is between 0.005 and 0.01 (see tables) [2]

A linear interpolation yields: $1-\dfrac{1}{2}\times\dfrac{3-2.896}{3.355-2.896} = 0.8867$

[Total 8]

The key to this question was to come up with a suitable expression of the variables concerned and identify the correct distribution in each case. In part (i) many candidates correctly identified the chi-square distribution, but failed to produce the correct answer. Part (ii) was poorly answered. Many candidates failed to establish independence between and the mean and variance and thus could not arrive at the correct expression. Parts (iii) and (iv) were also poorly attempted with a common error being the failure to apply the correct standard deviation.

Q5 From the 99% CI we know that

$\bar{x} - 2.576\dfrac{s}{7} = 30$ and $\bar{x} + 2.576\dfrac{s}{7} = 50$. [1]

Solving these two equations we obtain

$\bar{x} = 40$ and $s = \dfrac{70}{2.576} = 27.17391$ [2]

So 90% CI is

$40 \pm 1.645\dfrac{27.17391}{7}$ i.e. (33.614, 46.386). [2]

[Total 5]

This question was very well answered in general. A common mistake in wrong answers was using incorrect critical values.

Q6 (i) $E\left[\dfrac{(n-1)S^2}{\sigma^2}\right] = n-1$ (since $\sim X^2_{n-1}$) [1]

$\Rightarrow E\left[S^2\right] = \dfrac{\sigma^2}{n-1}\times(n-1) = \sigma^2$ [1]

(ii) (a) $G^2 = \dfrac{n-1}{n}S^2 \Rightarrow E\left[G^2\right] = \dfrac{n-1}{n}\sigma^2$ [1]

$\Rightarrow \text{bias} = \dfrac{n-1}{n}\sigma^2 - \sigma^2 = -\dfrac{\sigma^2}{n}$ [1]

(b) As *n* gets large the bias tends to zero. [1]

[Total 5]

Answers in part (i) were mixed, with some candidates attempting to answer it without using the sampling distribution of the variance. Part (ii) was better answered, while responses in part (iii) were mixed as many candidates did not relate the bias arrived in part (ii)(a) to their answers in this part.

Q7 (i) We have $\bar{X} = \dfrac{1}{2}\vartheta^2$ [1]

and solving for ϑ we obtain the estimator $\hat{\vartheta} = \pm\sqrt{2\bar{X}}$ [1]

only if $\sum x_i \geq 0$. [1]

(ii) For the given sample we obtain $\bar{x} = \frac{9}{2}$ and $\hat{\vartheta} = 3$ or $\hat{\vartheta} = -3$. [1]

(iii) Since we consider a normal distribution it is possible that $\sum_{i=1}^{n} X_i < 0$ in which case the estimator in part (i) is not defined. [1]

[Total 5]

Part (i) was answered well by most candidates, but only a relatively small number achieved full marks. A common mistake was not including the negative solution, while only a small number of

candidates noted that the solution was only valid for a non-negative sum. Parts (ii) and (iii) were answered well in general.

Q8 (i) The variances of the errors are common to all treatments. [1]

The errors are independently distributed [0.5]
with a N(0,σ^2) distribution. [0.5]

(ii) $y_{..} = 193.03 + 259.49 + 263.08 = 715.6$

$$SST = 4697.8 + 7508.3 + 7730.34 - \frac{715.6^2}{26} = 240.93$$ [1]

$$SSB = \left(\frac{193.03^2}{8} + \frac{259.49^2}{9} + \frac{263.08^2}{9} \right) - \frac{715.6^2}{26} = 133.85$$ [1]

	df	*SS*	*MSS*	*F stat*
Between	2	133.85	66.93	14.36
Residual	23	107.08	4.66	
Treatment	25	240.93		

[2]

$F_{2,23;0.95} = 3.422 <$ *F* stat [1]

Therefore reject H_0 and conclude that the means are not the same at 5% level [1]

[Total 8]

Generally very well answered. There were some calculation errors in part (ii).

Q9 (i) (a) $f_X(x) = \int_x^1 24x^3 y\,dy = 12x^3\left(1 - x^2\right), \quad 0 < x < 1$ [1]

(b) $f_Y(y) = \int_0^y 24x^3 y\,dx = 6y^5, \quad 0 < y < 1$ [1]

(ii) $E[X] = \int_0^1 x f_X(x)\, dx = \int_0^1 x 12x^3(1-x^2)\, dx = 12\left[\frac{x^5}{5} - \frac{x^7}{7}\right]_0^1 = 24/35$ [1]

$E[Y] = \int_0^1 y 6y^5\, dy = 6/7$ [1]

$E[XY] = \int_0^1 \int_0^y xy\, 24x^3 y\, dx\, dy = \frac{24}{5}\int_0^1 y^7 dy = 3/5$ [1]

$\text{cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{3}{5} - \frac{24}{35} \times \frac{6}{7} = \frac{3}{245}$ [2]

(iii) $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{24x^3 y}{6y^5} = 4x^3 y^{-4}$ for $0 < x < y$, or 0 otherwise.

[2]

(iv) $P\left(X > \frac{1}{3} | Y = \frac{1}{2}\right) = \int_{1/3}^{1/2} f_{X|Y}\left(x \mid \frac{1}{2}\right) dx = \int_{1/3}^{1/2} 4x^3 2^4\, dx = 16\left[x^4\right]_{1/3}^{1/2} = \frac{65}{81}$ [2]

(v) $E(X|Y=y) = \int_0^y x f_{X|Y}(x \mid y)\, dx = \int_0^y x\, 4x^3 y^{-4} dx = \frac{4}{5y^4}\left[x^5\right]_0^y = \frac{4}{5}y$ [2]

So, $E(X|Y=1/4) = 1/5$. [1]

(vi) From (v) we have $E[X|Y] = \frac{4}{5}Y$.

Therefore, $E\left[E[X|Y]\right] = E\left[\frac{4}{5}Y\right]$ and using (ii):

$E\left[E[X|Y]\right] = \frac{4}{5}\frac{6}{7} = 24/35$.

This is the same as $E[X]$ from (ii). [3]

[Total 17]

Part (i) was answered correctly by the majority of candidates. Part (ii) was well answered, although many candidates used incorrect limits of integration. Part (iii) was generally answered well. Parts (iv)-(vi) were not answered particularly well, while a number of candidates did not attempt them at all.

In part (iv) many candidates did not use the correct expression, and in part (v) candidates struggled with the calculation and the correct limits of integration. In part (vi) some candidates provided a general proof of the equation, which is not what was required here.

Q10 (i) $\hat{p}_1 = \frac{36}{124} = 0.290,\ \hat{p}_2 = \frac{25}{136} = 0.184$ [2]

common $\hat{p} = (36+25)/(124+136) = 0.235$ [1]

As the sample is large we can use a normal approximation.

$$\text{Test statistic} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})/124 + \hat{p}(1-\hat{p})/136}}$$

$$= \frac{(0.290 - 0.184)}{\sqrt{0.235(1-0.235)/124 + 0.235(1-0.235)/136}}$$

$$= \frac{0.106}{\sqrt{0.00145 + 0.00132}}$$

$$= 2.014$$

[2]

As $Z_{0.975} = 1.96$ is less than the test statistic we reject $H_0 : p_1 = p_2$ at a 5% significance level. [2]

(ii) H_0: Proportions are the same.

Calculate totals

Survey	*1*	*2*	*3*	
Y	36	25	26	87
N	88	111	115	314
Total	124	136	141	401

[1]

Contingency table

Survey	*1*	*2*	*3*	
Y	26.90	29.51	30.59	87

N	97.10	106.49	110.41	314
Total	124.00	136.00	141.00	401

[2]

Test statistic

$$= \sum \frac{(e_i - a_i)^2}{e_i} = 3.078 + 0.688 + 0.689 + 0.852 + 0.191 + 0.191 = 5.69 \quad [1]$$

d.f. = (3 – 1)(2 – 1) = 2 [1]

The 95% point of $X_2^2 = 5.991$. As test statistic is lower, do not reject that the proportion of smokers is equal. [2]

(iii) (a) $\hat{p}_3 = 26/141 = 0.184$ [1]

(b) In the first case the test rejected that the proportions were the same, but in the second it did not reject that they were, as the proportion in the third survey is almost identical to that in the second. [1]

[Total 16]

The answers in part (i) were generally good. Common errors included not calculating a common proportion p and calculation mistakes when computing the test statistic. Part (ii) was also well answered in general, although there were some calculation errors. Part (iii) (a) was well answered by most candidates. Answers in (iii)(b) were mixed, with most candidates making the correct observation. However, many candidates failed to identify appropriate reasoning.

Q11 (i) There are 100 observations for age 50 and we can therefore use the normal distribution: [1]

$$\left[14.1 - 1.96\frac{\sqrt{1.69}}{\sqrt{100}}, 14.1 + 1.96\frac{\sqrt{1.69}}{\sqrt{100}}\right] = [13.8452, 14.3548] \quad [2]$$

(Alternative solution: using the t_{99} distribution and interpolation to obtain the critical value 1.987, gives confidence interval of (13.84, 14.36).)

(ii) $H_0 : \mu_{40} = \mu_{50}$

Test statistic: $z = 10\frac{15 - 14.1}{\sqrt{2.25 + 1.69}} = 4.534$ [1]

which is approximately standard normal under H_0 due to the large sample size (100 drivers per age). [1]

The p -value is therefore very close to 0, [1]

and the null hypothesis is rejected. We conclude that the average annual mileage at age 50 is not the same as the average annual mileage at age 40. [1]

(iii) $S_{xx} = 27,500 - \frac{460^2}{8} = 1,050$ [1]

$S_{yy} = 1,398.23 - \frac{105.3^2}{8} = 12.21875$ [1]

$S_{xy} = 5,942 - 460\frac{105.3}{8} = -112.75$ [1]

$r = \frac{-112.75}{\sqrt{1,050 \times 12.21875}} = -0.99542$ [1]

(iv) For the correlation coefficient in part (iii) the variation amongst drivers of the same age is ignored. [2]

(Therefore there seems to be a stronger linear relationship between age and annual mileage than for the case where variations amongst drivers of the same age is considered (part (iv)).

(v) Only if the variance in each group is zero will the two coefficients coincide. [1]

(vi) We have

$$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n = 100 \times 5,942 - \frac{(100 \times 460)(100 \times 105.3)}{800}$$

$$= -11,275$$ [2]

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 100 \times 27,500 - \frac{(100 \times 460)^2}{800}$$

$$= 2,750,000 - 2,645,000 = 105,000$$ [2]

Therefore,

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -\frac{11,275}{105,000} = -0.10738 \qquad [1]$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{105.3}{8} + 0.10738 \times \frac{460}{8} = 19.337 \qquad [1]$$

$$\hat{y} = 19.337 - 0.10738\,x \qquad [1]$$

[Total 21]

Part (i) was very well answered. Answers in part (ii) were mostly correct. Some candidates used a t99 distribution with pooled variance, in which case an assumption of equal variances needs to be made and explicitly mentioned. Part (iii) was well answered in general. Answers in parts (iv)-(v) were mixed, with many students failing to state in (iv) that the variation was ignored. Part (vi) was not particularly well answered, with many candidates using the Sxy and Sxx sums from a previous part without further explanation regarding why they would be the same here.

END OF EXAMINERS' REPORT