INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 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.

Mike Hammer
Chair of the Board of Examiners
December 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, with most candidates demonstrating good understanding and application of core topics in probability and mathematical statistics.

2. Topics that were not particularly well answered in this paper include likelihood derivations (e.g. Q7) and linear regression methodology (e.g. Q9). Candidates are advised to revise all parts of the syllabus.

3. It is important that rigour mathematical and statistical notation is used when answering questions. In certain cases, e.g. Q7, poor or inaccurate notation shows inadequate understanding and may lead to loss of marks.

4. Question 9, parts (i) and (ii), contained methodological elements related to the derivation of the regression line. Performance in these parts was less satisfactory. Candidates are advised to give appropriate weight to all elements of the syllabus when preparing for the exam.

C. Pass Mark

The Pass Mark for this exam was 60.

Solutions

Q1

For original data: $\Sigma x_i = n\bar{x} = 20 \times 45 = 900$

Corrected data:

$\bar{x} = \frac{900+30-130}{20} = 40$ [1]

Original data:

$\Sigma x_i^2 = (n-1)s^2 + n\bar{x}^2 = 19 \times 25.4^2 + 20 \times 45^2 = 52758.04$ [1]

Corrected data:

$\Sigma x_i^2 = 52758.04 + 30^2 - 130^2 = 36758.04$ [1]

$s = \sqrt{\frac{36758.04-20*40^2}{19}} = 15.825$ [1]

[Total 4]

The question was very well answered by most candidates. A common mistake was using n instead of n-1 for the standard deviation.

Q2

(i) $P(X = 3) = 0.0256$ [1]

(ii) n=4 (since $P(X > 4) = 0$) and $P[X = 4] = 0.0016 > 0$ [1]

$P(X = 4) = p^4 = 0.0016$ [1]
$\Longrightarrow p = 0.2$ [1]

Therefore X ~ Binomial(4,0.2)

Alternative solutions, e.g. $P(X = 0) = (1-p)^4 = 0.4096 \Longrightarrow p = 0.2$

[The coefficient for t^2 in the question is incorrect (the correct coefficient value is 0.1536). If $E[X] = G_X'(1) = np$ is used then an answer of $p = 0.1996$ is obtained.]

[Total 4]

Generally well answered. In part (i), most candidates identified the polynomial coefficients correctly. However, in part (ii) some candidates attempted a complicated route relating to finding derivatives, leading to errors in many cases.
In cases where the incorrect coefficient for t^2 was used following the alternative solution with $E[X] = G_X'(1) = np$, full credit was given to candidates providing the slightly different answer under this approach.

Q3

(i)

(a) Needs to assume that each time the athlete tries she independently has the same probability *p* of passing the height, i.e. that attempts here are iid. [1]

(b) Given that the attempts are at the same event and on the same day, it is reasonable to assume that conditions are the same (independence) and that probability of success does not change. [2]

(ii) (a) If *X* is the corresponding random variable, we want:

$P(X > x + n \mid X > n)$ [0.5]

$= \frac{P(X>x+n)}{P(X>n)} = \frac{(1-p)^{x+n}}{(1-p)^n} = (1-p)^x = P(X > x)$ [1.5]

(b) The lack of success on the first *n* jumps is irrelevant – under this model the chances of success are not any better because there have been *n* attempts already. [1]

[Total 6]

Answers were mixed, with many candidates in part (i) failing to describe the assumptions or give justifications for them. All reasonable comments were given credit here. In part (ii), many candidates failed to express the conditional probability as required.

Q4

(i) $P[X_A \geq 2] = 1 - F_A(1) = 1 - 0.98248 = 0.01752$ using tables [2]

(ii) $P[X_A = 1]P[A] + P[X_B = 1]P[B] + P[X_C = 1]P[C]$ [1]

$= e^{-0.2} * 0.2 * 0.2 + e^{-0.1} * 0.1 * 0.2 + e^{-0.05} * 0.05 * 0.6 = 0.07938286$ = 0.0794 [2]

(iii) Let X^0 be the number of claims submitted last year

$P[A|X^0 = 1] = \frac{P[X_A=1]P[A]}{P[X^0=1]} = e^{-0.2} * 0.2 * \frac{0.2}{0.07938286} = 0.4125479$ = 0.4125 [2]

[Total 7]

Generally well answered. In part (i) some candidates mistakenly divided by the probability of being in group A, effectively conditioning on being in group A twice.

Q5

Observed

	Not overweight	Overweight	Obese	
Females	45	32	23	100
Males	33	41	26	100
	78	73	49	200

Expected

	Not overweight	Overweight	Obese	
Females	39	36.5	24.5	100
Males	39	36.5	24.5	100
	78	73	49	200

[1]

Squared Difference

	Not overweight	Overweight	Obese
Females	36	20.25	2.25
Males	36	20.25	2.25

[1]

Test statistic:

$2 \times \left[\frac{36}{39} + \frac{20.25}{36.5} + \frac{2.25}{24.5}\right] = 3.14$ [1]

2 degrees of freedom, Chi-Squared critical value is 5.991 at a 5% significance level. [1]

Do not reject null hypothesis that weight and gender are independent [1]

[Total 5]

Very well answered by most candidates.

Q6

(i)

If X denotes the "no" voters, under H_0 we have

$X \sim$ Binomial(1106, 0.5), or approximately $X \sim N(553, 276.5)$ [1]

Using a continuity correction, the z statistic is given as

$$z = \frac{607.5-553}{\sqrt{276.5}} = 3.28$$

[1.5]

Critical point for tables is $z_{0.05} = 1.6449$.

[0.5]

So we reject H_0at the 5% level in favour of H_1,
which means that we have evidence of a "no" vote. [1]

(ii) (a) 90% CI is given by $\hat{p} \pm 1.6449\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ [1]

This gives: $\frac{608}{1106} \pm 1.6449\sqrt{\frac{(\frac{608}{1106})(1-608/1106)}{1106}}$, i.e. (0.525, 0.574) [2]

(b) A larger sample would reduce the standard error of $\hat{p}$ and would therefore give a narrower interval. [1]

(iii) (a) The P-value is the probability, assuming $H_0$is true, of observing a test statistic at least as "extreme" as the value observed (or, it is the lowest significance level at which H_0 can be rejected).

[1]

(b) P-value $= P(X \geq 608) = P\left(Z > \frac{607.5-553}{\sqrt{276.5}}\right) = P(Z > 3.28) = 0.00052$ [2]

We have very strong evidence against H_0, which means that we have very strong evidence of a "no" vote. [1]

(c) Using a fixed level does not provide clear detailed information on the strength of the evidence against H_0, whereas using a P-value is more informative about the

strength of this evidence.

Here, using the *P*-value approach clearly tells us about how strong the evidence against H_0 is, which means we can put our conclusion in stronger terms. [2]

[Total 14]

Generally well answered. Notice that a continuity correction is needed in this question for full marks. In part (i) many candidates mistakenly used the sample estimate of p in the variance.

Q7

(i)

$$p_i(x) = \begin{cases} 0, & x < 0 \\ \frac{\lambda^x e^{-\lambda}}{x!}, & x \geq 0, i \neq July, August \\ \frac{(u\lambda)^x e^{-u\lambda}}{x!}, & x \geq 0, i = July, August \end{cases}$$

[1]

[1]

(ii)

$$L(\boldsymbol{x}; \lambda, u) = \prod_{i \neq Jul, Aug} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \prod_{i = Jul, Aug} \frac{(u\lambda)^{x_i} e^{-u\lambda}}{x_i!}$$

[1]

$$= \prod_i \frac{\lambda^{x_i}}{x_i!} \times \left(e^{-(10+2u)\lambda} \, u^{x_{Jul} + x_{Aug}} \right)$$

[1]

$$l(x; \lambda, u) = \sum_i x_i \ln \lambda - (10 + 2u)\lambda + \left(x_{Jul} + x_{Aug} \right) \ln u + C$$

[1]

(iii)

$$\frac{\delta l}{\delta u} = -2\lambda + \frac{x_{Jul} + x_{Aug}}{u} = 0$$

[1]

$$\Longrightarrow 2\lambda u = (x_{Jul} + x_{Aug})$$

[1]

$$\frac{\delta l}{\delta \lambda} = \frac{\Sigma x_i}{\lambda} - 10 - 2u = 0$$

$$\Longrightarrow \Sigma x_i - 10\lambda = 2\lambda u = (x_{Jul} + x_{Aug})$$

[1]

$$\Rightarrow \hat{\lambda} = \sum_{i \neq Jul,Aug} x_i /10$$

[1]

$$\Rightarrow \hat{u} = (x_{Jul} + x_{Aug})/2\hat{\lambda}$$
$$= (x_{Jul} + x_{Aug})/(2 \sum_{i \neq Jul,Aug} x_i /10)$$
$$= (x_{Jul} + x_{Aug})/(\sum_{i \neq Jul,Aug} x_i /5)$$

[1]
[Total 10]

There were mixed answers in parts (i) and (ii), often with poor notation for the likelihood. Part (iii) was well answered.

Q8

(i) $\text{C.I.} = 523 \pm \frac{t_{60;.975}s}{\sqrt{n}}$ [1]

$= 523 \pm 2.000 \times \frac{81}{\sqrt{61}}$ [1]

$= (502.3, 543.7)$ [1]

(ii) $\text{C.I.} = \left(\frac{(n-1)s^2}{X^2_{n-1;0.025}}, \frac{(n-1)s^2}{X^2_{n-1;0.975}}\right)$ [1]

$= \left(\frac{60 \times 81^2}{40.48}, \frac{60 \times 81^2}{83.30}\right)$ [1]

$= (4726, 9725)$ [1]

(iii) By the CLT and as λ is large $N \sim N(\lambda, \lambda) = N(250,250)$ [2]

$P(N > 270) = P(N > 270.5)$ continuity correction [1]

$= P\left(Z > \frac{270.5-250}{\sqrt{250}}\right) = P(Z > 1.297) = 1 - 0.903 = 0.097$ [2]

(iv) Now the rate of claims is $12 \times \lambda = 3000$ [1]

Mean $= \lambda_{new} * \mu_{claims} = 3000 * 523 = 1{,}569{,}000$ [1]

$$\text{Standard deviation} = \sqrt{\lambda_{annual}} * \sqrt{(\sigma^2_{claims} + \mu^2)}$$

$= \sqrt{3000} * \sqrt{81^2 + 523^2} = 28987$ [2]

(v) Want smallest n such that $P(\bar{X} < 503) \leq 0.05$ under H_0 [1]

i.e. $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{503-\mu}{\sigma/\sqrt{n}}\right) \leq 0.05$ [1]

$-1.6449 \geq \frac{503-523}{81/\sqrt{n}} = -\frac{20\sqrt{n}}{81}$ [1]

$\Longrightarrow n \geq \left(\frac{1.6449 \times 81}{20}\right)^2 = 44.38$ [1]

i.e. n is at least 45 claims. [1]

[Total 20]

Parts (i) and (ii) were very well answered. Part (iii) was generally well answered, although many candidates failed to justify the normal approximation and/or to apply a continuity correction. In part (iv) many candidates performed the calculations for the monthly amounts rather than annual.

Q9

(i) Using quantiles of the t_{50}-distribution as an approximation to the required t_{49}-distribution.

$\left[3.5 - 2.009\frac{2.3}{\sqrt{50}}, 3.5 + 2.009\frac{2.3}{\sqrt{50}}\right] = [2.8465, 4.1535]$

[2]

(ii) Total sample size: 50+65+60+35=210

(iii) [1]

Total units: $50 \times 3.5 + 65 \times 4.8 + 60 \times 5.1 + 35 \times 4.2 = 940$

[1]

Overall average: $\frac{940}{210} = 4.476$

[1]

(iv) ANOVA:

$$SS_B = 50 \times (3.5 - 4.476)^2 + 65 \times (4.8 - 4.476)^2 + 60 \times (5.1 - 4.476)^2 + 35 \times (4.2 - 4.476)^2 = 80.48$$

[3]

$SS_R = 49 \times 2.3^2 + 64 \times 1.8^2 + 59 \times 1.6^2 + 34 \times 1.1^2 = 658.75$

[2]

Test statistic: $F = \frac{80.48/3}{658.75/206} = 8.3891$ [1]

This compares to a 1% quantile of a $F_{3,206}$ distribution. [1]

This quantile is between3.782 and 3.949, and we therefore have sufficient evidence to reject the null hypothesis that the average number of units of alcohol per week is the same for all age groups. [1]

(v) Overall variance in sample:

$\frac{1}{209}SS_T = \frac{1}{209}(SS_R + SS_B) = \frac{1}{209}(658.75 + 80.48) = 3.54$ [1]

95% C.I.: $\left[4.476 - 1.96\sqrt{\frac{3.54}{210}}, 4.476 + 1.96\sqrt{\frac{3.54}{210}}\right] = [4.222, 4.73]$ [1]

[Alternative solution: $\hat{\sigma}^2 = SS_R/(n-k)$. Then CI is (4.234,4.718).]

(vi) The results in part (iii) indicate that age has an impact on drinking habits, and therefore, the overall average of units per week and the corresponding confidence interval in part (iv) might not be meaningful to describe the drinking habits of any specific individual. [2]

[Total 17]

Generally well answered. In part (ii) a few candidates calculated a simple mean instead of a weighted average. There were mixed answers in part (v) with many candidates failing to comment meaningfully on their results.

Q10

(i) $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right]^2$

[1]

(ii) Partially differentiate w.r.t. each parameter and equate to zero gives:

$2\sum_{i=1}^{n}[y_i - (\hat{\alpha} + \hat{\beta}x_i)] = 0 \Rightarrow \sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i$ [1]

$2\sum_{i=1}^{n} x_i[y_i - (\hat{\alpha} + \hat{\beta}x_i)] = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i = \hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2$ [1]

Eliminate $\hat{\alpha}$ from simultaneous equations:

$$\left(\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} y_i = n\hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$n\sum_{i=1}^{n} x_i y_i = n\hat{\alpha}\sum_{i=1}^{n} x_i + n\hat{\beta}\sum_{i=1}^{n} x_i^2$$

[1]

$\Rightarrow n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) = \hat{\beta}(n\sum_{i=1}^{n} x_i^2 - [\sum_{i=1}^{n} x_i]^2)$ [1]

$\Rightarrow \hat{\beta} = \left(n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)\right)/(n\sum_{i=1}^{n} x_i^2 - [\sum_{i=1}^{n} x_i]^2)$ [1]

(iii) $S_{xx} = 389{,}684 - (3{,}660)^2/44 = 85{,}238.55$ [1]

$S_{xy} = 13{,}609{,}918 - (3{,}660 \times 136{,}727)/44 = 2{,}236{,}718$ [1]

$\bar{x} = \frac{3{,}660}{44} = 83.1818, \bar{y} = 136{,}727/44 = 3{,}107.43$ [1]

$\hat{\beta} = S_{xy}/S_{xx} = 2{,}236{,}718/85{,}238.55 = 26.241$ [1]

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 3{,}107.43 - 26.241 \times 83.1818 = 924.68$ [1]

(iv) $S_{yy} = 500{,}813{,}951 - (136{,}727)^2/44 = 75{,}944{,}121$ [1]

$r = S_{xy}/\sqrt{S_{xx}S_{yy}} = 2{,}236{,}718/\sqrt{85{,}238.55 \times 75{,}944{,}121} = 0.879$ [1]

[Total 13]

Parts (i) and (ii) were more focussed on methodology compare to other questions and were not well answered. Parts (iii) and (iv) were very well answered.

END OF EXAMINERS' REPORT