INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2013 examinations

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.

D C Bowie
Chairman of the Board of Examiners

July 2013

General comments on Subject CT3

For CT3 exams some questions admit alternative solutions or different ways in which the provided answer can be determined. All valid alternative solutions or answers received credit as appropriate. Rounding errors were not penalised, unless excessive rounding led to significantly different answers. In cases where the same error was carried forward to later parts of the answer, candidates were not penalised twice. In questions where comments were required, reasonable comments that were different than those provided in the solutions also received full credit.

Comments on the April 2013 paper

Performance was generally good, but overall not as strong as in the previous examination diet. There was a wide distribution of marks, with well prepared candidates achieving very high scores. On the other hand, less well prepared candidates struggled with questions that did not appear in very similar form in recent examination papers. This is a recurring issue, and candidates are advised to take a wider and more inclusive approach in their preparation for the subject, rather than overly rely on questions appearing in past papers.

The comments on individual questions that follow concern specific parts that candidates answered poorly and important frequent errors.

1 Mean = $(7\times 0+9\times 1+3\times 2+3)/20=(9+6+3)/20=18/20=0.9$

Median = 1 (observation with rank 10.5)

Mode = 1

VAR = $\dfrac{7\times 0.9^2+9\times 0.1^2+3\times 1.1^2+2.1^2}{19}=\dfrac{5.67+0.09+3.63+2.1^2}{19}=\dfrac{13.80}{19}=0.7263$

STD = 0.8522

Well answered. Some working needs to be shown for full marks.

2 For any number x we get

$$P[X\le x]=P\left[F^{-1}(U)\le x\right]=P[U\le F(x)]=F(x)$$

which shows that F is the distribution function of the random variable X, which proves the result.

Very poorly answered. Most candidates did not attempt this question and very few completed it correctly.

3
$$P[X>10]=1-P[X\le 10]=1-F(10)=0.5$$
$$P[X<30]=P[X\le 20]=F(20)=0.7$$
$$P[X=40]=F(40)-F(30)=0.1$$
$$P[20<X<50]=F(40)-F(20)=0.25$$
$$P\left[\{X=20\}\cup\{X=40\}\right]=P[X=20]+P[X=40]$$
$$=\left[F(20)-F(10)\right]+\left[F(40)-F(30)\right]=0.2+0.1=0.3$$

Some problems were encountered here involving understanding and distinguishing the need (or not) for strict inequalities for discrete variables, e.g. P[X<30] = P[X≤20].

4 (i) The random variables $X_1,\ldots,X_n$ are independent and identically distributed with $X_i\sim N(\mu,\sigma^2)$

(ii) $\bar{X}$ and S^2 are independent

$$\bar{X}\sim N(\mu,\sigma^2/n)$$

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$$

(iii) $t_k = N(0,1)/\sqrt{\chi^2_k / k}$ where $N(0,1)$ and χ^2_k are independent

This result can be applied here, and we get $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

Mixed quality in the answers. Some candidates answered part (iii) in the process of answering part (ii) – this did not always show clear understanding, but full marks were given.

5 Under given assumptions $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$

and approximately $X_1 \sim N(\lambda_1, \lambda_1)$, $X_2 \sim N(\lambda_2, \lambda_2)$

giving $X_1 - X_2 \sim N(\lambda_1 - \lambda_2, \lambda_1 + \lambda_2)$, or $\dfrac{X_1 - X_2 - (\lambda_1 - \lambda_2)}{\sqrt{\lambda_1 + \lambda_2}} \sim N(0,1)$

Approximate 90% interval given as

$X_1 - X_2 \pm z_{0.05}\sqrt{\hat{\lambda}_1 + \hat{\lambda}_2} = X_1 - X_2 \pm z_{0.05}\sqrt{X_1 + X_2}$

$= 12 \pm 1.6449 \times (234)^{1/2} = 12 \pm 25.1621$ i.e. $(-13.162, 37.162)$

A common error here involved the normal approximation of the difference of the two variables – especially its variance.

6 (i) Using approximate normality, and with $\hat{p} = 0.3$ we can calculate the interval a

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = (0.21, 0.39)$$

(ii) Sample size is large (or np, or $np(1-p)$), so normal approximation is valid.

(iii) With larger sample size the standard error will be smaller, and therefore the interval will be narrower.

This was straightforward for most candidates. However, the explanation was often not clear or convincing.

7 (i) No of inspected policies ~ Negative binomial(5, 0.1).
Expected no of inspected policies = 5/0.1 =50

(ii) $\sum x = 479, \sum x^2 = 63705$
Mean = 479/5 = 95.8
Variance = $(63705 - 5*95.8^2)/4 = 4454.2$

(iii) $E[X] = \frac{\alpha}{\lambda} = 95.8,\ V[X] = \frac{\alpha}{\lambda^2} = 4454.2$

$$\Rightarrow \lambda = \frac{E[X]}{V[X]} = \frac{95.8}{4454.2} = 0.0215$$

$$\Rightarrow \alpha = \lambda E[X] = 0.0215 * 95.8 = 2.06$$

Generally very well answered with no particular issues.

8 (i) H_0: The means of the claims in the 3 regions are all equal; H_1: means are different for at least one pair.

F = 5.59 on 2 and 27 d.f. From tables the 1% critical point is 5.488.

Therefore, we have (strong) evidence against the null hypothesis, and conclude that there are differences in the means for the 3 regions.

(ii) (a) 95% CI for $\mu_A - \mu_B$ is given by

$$(\bar{y}_A - \bar{y}_B) \pm t_{0.025,27}\hat{\sigma}\sqrt{\frac{1}{10} + \frac{1}{10}} \text{ giving } (147.47 - 154.56) \pm 2.052\sqrt{396.8}\sqrt{\frac{1}{10} + \frac{1}{10}}$$

i.e -7.09 ± 18.28 or (–25.37, 11.19)

(b) The CI comfortably contains zero, suggesting no difference between the true means for regions A and B.

The significant result of the F test clearly comes from region C mean being much lower than the means for regions A and B.

Generally well answered. Some candidates failed to identify the connection between the conclusion of the ANOVA and that of the CI for regions A and B.

9 (i) (a) If X_i is the number of bananas for each monkey then $X_i \sim \text{Bin}(7, p)$

$$E(X_i) = \bar{x} \Rightarrow 7\hat{p} = \frac{33+37}{(6+11)} \Rightarrow \hat{p} = 0.588$$

(b) $$\hat{p}_A = \frac{33}{6*7} = 0.786, \hat{p}_B = \frac{37}{11*7} = 0.481$$

σ^2 = Variance of test statistic = 0.588 * (1 – .588) * (1/42 + 1/77)
= 0.00891

Test statistic = $\frac{\hat{p}_A - \hat{p}_B}{\sigma} = \frac{0.786 - 0.481}{\sqrt{0.00891}} = 3.23$

Test statistic has $N(0,1)$ distribution so p-value is 0.00124

i.e. reject $H_0 : p_A = p_B$

(ii) (a) Let n_i be the number of monkeys in group i and B_i be the total number of bananas taken by group i.

$$L(b;\theta) = (2\theta)^{B_A}(1-2\theta)^{7n_A - B_A}(\theta)^{B_B}(1-\theta)^{7n_B - B_B} \times \text{constant}$$

$$l(b;\theta) = \ln L(b;\theta)$$

$$= 33\ln(2\theta) + (42-33)\ln(1-2\theta) + 37\ln(\theta) + (77-37)\ln(1-\theta) + \text{constant}$$

$$= 33\ln(2\theta) + 9\ln(1-2\theta) + 37\ln(\theta) + 40\ln(1-\theta) + \text{constant}$$

(b) $$\frac{dl}{d\theta} = \frac{66}{2\theta} - \frac{18}{1-2\theta} + \frac{37}{\theta} - \frac{40}{1-\theta} = \frac{70}{\theta} - \frac{18}{1-2\theta} - \frac{40}{1-\theta}$$

Set equal to zero and solve

$$\frac{70(1-2\theta)(1-\theta) - 18\theta(1-\theta) - 40\theta(1-2\theta)}{\theta(1-2\theta)(1-\theta)} = 0$$

$$\Rightarrow 70 - 210\theta + 140\theta^2 - 18\theta + 18\theta^2 - 40\theta + 80\theta^2 = 0$$

$$\Rightarrow 238\theta^2 - 268\theta + 70 = 0$$

$$\Rightarrow \theta = 0.412 \text{ or } 0.714$$

As $\theta < 0.5$, $\hat{\theta} = 0.412$.

(iii) (a) Expected values under 2 models are:

	A	*B*
Model in (i)	42 * 0.588 = 24.7	77 * 0.588 = 45.3
Model in (ii)	42 * 2 * 0.412 = 34.6	77 * 0.412 = 31.7
Observed	33	37

Model in (ii) seems to provide a better fit as expected values are closer to observed.

(b) In part (i)(b) we rejected $p_A = p_B$ which suggests a model with a common value of p would not be appropriate. The comparison above suggests that an improved model can be used.

There were some common errors here, mainly involving part (i)(b) where many candidates failed to identify an appropriate test to perform. There were also basic errors with algebraic and calculus operations.

We note that in part (i)(b)an alternative solution can be given, using a chi-square test with 1 d.f. in a 2x2 table (4 cells). This is exactly equivalent to the test presented here and full credit was given when completed correctly.

10 (i) Y_i has a compound distribution, so

$$E(Y_i) = E(N_i)E(X_{ij}) = \lambda\mu$$

$$V(Y_i) = E(N_i)V(X_{ij}) + V(N_i)E(X_{ij})^2 = \lambda\sigma^2 + \lambda\mu^2$$

(ii) S also has a compound distribution.

$$E(S) = E(M)E(Y_i) = \kappa\lambda\mu$$

$$V(S) = E(M)V(Y_i) + V(M)E(Y_i)^2 = \kappa\lambda(\sigma^2 + \mu^2) + \kappa\lambda^2\mu^2 = \kappa\lambda(\sigma^2 + \mu^2 + \lambda\mu^2)$$

(iii) $$P(X_{ij} \le x + C | X_{ij} > C) = \frac{P(C < X_{ij} \le x + C)}{P(X_{ij} > C)}$$

$$= \frac{\left(1 - e^{-x-C} - 1 + e^{-C}\right)}{e^{-C}} = 1 - e^{-x}$$

$$= P(X_{ij} \le x)$$

(iv) (a) $\lambda^* = 1000 \times P\left(X_{ij} > 2\right) = 1000e^{-2} = 135.3$

(b) From definition of new variable and part (iii) we have that

$$P\left(X_{ij}^* \le x\right) = P(X_{ij} - 2 \le x | X_{ij} > 2) = P(X_{ij} \le x + 2 | X_{ij} > 2) = P(X_{ij} \le x)$$

meaning that X_{ij}^* has the same distribution as X_{ij}, i.e. Exp(1).

(c) $E\left(S_R\right) = \kappa\lambda^*\mu = 4 \times 135.3 \times 1 = 541.2$

$$V\left(S_R\right) = \kappa\lambda^*(\sigma^2 + \mu^2 + \lambda^*\mu^2) = 4 \times 135.3 \times \left(1 + 1 + 135.3\right) = 74306.8$$

Most candidates found this question challenging. Answers to the memoryless property of the exponential distribution (amply discussed in the CR) in part (iii) were often disappointing, and the relevant application in part (iv) was poorly attempted. These shortcomings highlight the issue of being prepared to tackle questions that deviate from the form that appears in past papers.

11 (i) Notation: n_i = number in group i; r_i = number with disease in group i.

$$\hat{\theta} = \frac{\sum r_i}{\sum n_i} = \frac{43}{100} = 0.43$$

(ii) (a) Expected frequencies (in brackets) are given assuming constant probability of disease for all groups, independently of age:

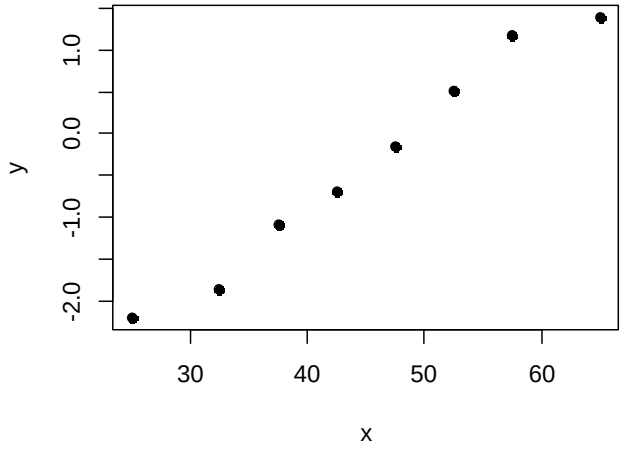
	Disease		
Age group	*Yes*	*No*	*Total*
20–29	1	9	10
	(4.30)	(5.70)	
30–34	2	13	15
	(6.45)	(8.55)	
35–39	3	9	12
	(5.16)	(6.84)	
40–44	5	10	15
	(6.45)	(8.55)	
45–49	6	7	13
	(5.59)	(7.41)	
50–54	5	3	8
	(3.44)	(4.56)	
55–59	13	4	17
	(7.31)	(9.69)	
60–69	8	2	10
	(4.30)	(5.70)	
Total	43	57	100

(b) $\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = \frac{(1-4.3)^2}{4.3} + \ldots + \frac{(2-5.7)^2}{5.7} = 26.6$ on 7 d.f.

From tables, $\chi^2_{7,0.01} = 18.48$

We have (strong) evidence against the hypothesis of no differences in probability of disease among age groups.

(iii) (a) Plot given below. Linear model seems appropriate for middle ages, but perhaps not for younger and older ages.



(b) $S_{xx} = 17437.5 - \frac{360^2}{8} = 1237.5$

$S_{yy} = 13.615 - \frac{(-2.9392)^2}{8} = 12.535$

$S_{xy} = -9.0429 - \frac{(360)(-2.9392)}{8} = 123.22$

Least squares estimates:

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{123.22}{1237.5} = 0.09957$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -0.3674 - 0.09957(45) = -4.85$

Fitted line: $\hat{y} = -4.85 + 0.09957x$

(c) $$\hat{\sigma}^2 = \frac{\left(S_{yy} - \frac{S_{xy}{}^2}{S_{xx}}\right)}{n-2} = \frac{\left(12.535 - \frac{123.22^2}{1237.5}\right)}{6} = 0.04430$$

$$se\left(\hat{\beta}\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{0.210}{\sqrt{1237.5}} = 0.0060$$

and $t_{6,0.005} = 3.707$

99% CI for $\hat{\beta}$ is given by $0.09957 \pm 3.707(0.0060)$

i.e. 0.09957 ± 0.0222 or $(0.0774,\ 0.1218)$

(d) In (ii) it was found that the probability of having the disease is different for different age groups. In part (iii)(c) it was also found that the probability of disease depends on age, as zero was not included in the interval for the slope parameter.

The quality of the answers was mixed, with some common errors appearing in part (ii) where many candidates failed to produce an appropriate 8×2 table (both "yes" and "no" columns) and perform the correct chi-square test (with 7 d.f.).

It is noted that in part (ii)(b) the chi-square test can alternatively be performed by combining some of the age groups (both columns) to achieve expected frequencies greater than 5, with no change in the conclusion of the test. Although this is not strictly required in this case, full credit was given to candidates that combined groups sensibly and completed the question correctly.

END OF EXAMINERS' REPORT