# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINATION

5 October 2016 (am)

## Subject CT3 – Probability and Mathematical Statistics
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *You have 15 minutes of planning and reading time before the start of this examination. You may make separate notes or write on the exam paper but not in your answer booklet. Calculators are not to be used during the reading time. You will then have three hours to complete the paper.*

4. *Mark allocations are shown in brackets.*

5. *Attempt all 10 questions, beginning your answer to each question on a new page.*

6. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** Consider the following sample with 20 observations $x_i$:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 5 | 7 | 9 | 11 | 11 | 14 | 14 | 19 |
| 20 | 21 | 23 | 28 | 28 | 31 | 39 | 41 | 43 | 47 |

$$\sum_{i=1}^{20} x_i = 413 \text{ and } \sum_{i=1}^{20} x_i^2 = 12,311$$

(i) Calculate the mean of this sample. [1]

(ii) Calculate the standard deviation of this sample. [2]

(iii) Calculate the median of this sample. [1]

(iv) Calculate the interquartile range of this sample. [2]

[Total 6]

**2** A sample of data has a distribution that has a single mode and is strongly positively skewed. An analyst computes three measures of location for these data: the mean, the median and the mode.

(i) State the largest of these three location measures. [1]

(ii) Suggest, with a reason, which would be the best measure of location. [2]

[Total 3]

**3** An insurance company has a portfolio of 10,000 policies. Based on past data the company estimates that the probability of a claim on any one policy in a year is 0.003. It assumes no policy will generate more than one claim in a year.

(i) Determine the approximate probability of more than 40 claims from the portfolio of 10,000 policies in a year. [4]

(ii) Determine an approximate equal-tailed interval into which the number of claims per year will fall with probability 0.95. [2]

In practice 42 claims were received in a particular year. A Director of the company complains about the range of estimates in part (ii) being wrong.

(iii) Comment on the Director's complaint. [2]

[Total 8]

**4** Consider two portfolios, A and B, of insurance policies and denote by $X_A$ the number of claims received in portfolio A and by $X_B$ the number of claims received in portfolio B during a calendar year. The observed numbers of claims received during the last calendar year are 134 for portfolio A and 91 for portfolio B. $X_A$ and $X_B$ are assumed to be independent and to have Poisson distributions with unknown parameters $\beta_A$ and $\beta_B$.

Determine an approximate 99% confidence interval for the difference $\beta_A - \beta_B$. You may use an appropriate normal distribution. [4]

**5** For each of two life insurance companies, A and B, a random sample of 150 policies is examined. The number of policies which have given rise to claims in the past year is 45 for company A and 33 for company B.

Test the null hypothesis that the underlying proportions of policies which have given rise to claims in the past year are equal for the two companies. [4]

**6** Let $X$ and $Y$ be random variables with joint probability distribution:

$$f_{XY}(x,y) = \begin{cases} kx^2y^2, & 0 < x < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

where $k$ is a constant.

(i) Show that $k = 18$. [4]

(ii) Determine $f_Y(y)$, the marginal density function of $Y$. [2]

(iii) Determine $P(X > 0.5 | Y = 0.75)$. [3]

[Total 9]

**7** An analyst is investigating the number of car insurance claims made by policyholders living in different parts of the country. Denote by $X$ the number of claims made by policyholders living in large cities, $Y$ the number of claims made by policyholders living in small cities, and $Z$ the number of claims made by policyholders living in the countryside. For each of the three groups of policyholders consider a random sample of size 500 and count the number of claims made during the last calendar year.

The following table shows the results for the three groups of policyholders:

| | *Large City* | *Small City* | *Countryside* | *Total* |
|---|---|---|---|---|
| No claim | 370 | 390 | 410 | 1,170 |
| One claim | 93 | 99 | 87 | 279 |
| More than one claim | 37 | 11 | 3 | 51 |
| Total | 500 | 500 | 500 | 1,500 |

For example, 390 of the 500 policyholders living in small cities had no claim during the last year, and 3 of the 500 policyholders living in the countryside had more than one claim during the same year.

(i) Perform a $\chi^2$-test to test the null hypothesis that the number of claims per policy is independent of the place of living. [7]

After some further analysis, an actuary has estimated that the joint distribution of the number of claims last year and the place of living is given by the following table:

| | *Large City* | *Small City* | *Countryside* |
|---|---|---|---|
| No claim | 0.23 | 0.25 | 0.27 |
| One claim | 0.06 | 0.06 | 0.06 |
| More than one claim | 0.04 | 0.02 | 0.01 |

For example, the probability that a randomly selected policyholder lives in the countryside and made no claim last year is 0.27.

(ii) Determine the probability that a randomly selected policyholder:

(a) lives in a small city.

(b) has submitted more than one claim last year.

(c) has submitted more than one claim last year given that the policyholder lives in a large city.

(d) lives in the countryside given the policyholder submitted fewer than two claims last year.

(e) lives in a city (small or large) given the policyholder made at least one claim last year.

[8]
[Total 15]

**8** Ten pairs of data on a predictor variable ($x$) and a response variable ($y$) are available with the following summary statistics:

$$\bar{x} = 5.93 \quad \bar{y} = 7.15 \quad \sum_{i=1}^{10}(x_i - \bar{x})^2 = 81.15 \quad \sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 89.91 .$$

A linear model of the form $y = \alpha + \beta x + \varepsilon$ is fitted to the data, where the error terms ($\varepsilon$) independently follow a $N(0, \sigma^2)$ distribution with $\sigma^2$ being an unknown parameter.

(i) Determine the fitted line of the regression model. [3]

A partially completed ANOVA table for this regression analysis is given below.

| *Source of variation* | *Degrees of freedom* | *Sums of squares* | *Mean squares* |
|---|---|---|---|
| Regression | A | 99.61 | C |
| Residual | B | 21.63 | D |
| Total | 9 | 121.24 | |

(ii) Determine the missing values A, B, C and D in the table. [2]

(iii) Determine an estimate of the variance $\sigma^2$ based on the above table. [1]

(iv) (a) Give the interpretation of the coefficient of determination, $R^2$, in a linear regression model.

(b) Determine the value of $R^2$ for the regression model fitted here, using the above table.

[2]

(v) Perform an $F$ test to test the null hypothesis that there is no linear relationship between $x$ and $y$, based on the above table. [5]

[Total 13]

**9** A statistical model is used to describe the total loss, $S$ (in pounds), experienced in a certain portfolio of an insurance company over a period of one year. The total loss is given by:

$$S = X_1 + X_2 + \cdots + X_N$$

where $X_i$ gives the size of the loss from claim $i$ =1,…, $N$. $N$ is a random variable representing the number of claims per year and follows a Poisson distribution. The $X_i$ s are independent, identically distributed according to a gamma distribution with parameters α and λ, and are also independent of $N$.

Data from previous years show that the average number of claims per year was 14, while the average size of claims was £500 and their standard deviation was £150.

(i) Estimate the parameters α and λ using the method of moments. [4]

(ii) Estimate the mean and the variance of the total loss $S$ using the information from the data above. [3]

Now suppose that the value of parameter α is known to be equal to α* and $n$ = 5 claims have been made in a particular year with average size again £500.

(iii) (a) Derive an expression for the maximum likelihood (ML) estimate of the parameter λ in terms of α*. You should verify that your answer corresponds to a maximum.

(b) Derive the asymptotic distribution of the ML estimator of the parameter λ in terms of α*.

(c) Comment on the validity of the distribution in part (iii)(b).

[9]

Now suppose that the values of both parameters α and λ are unknown and $n$ claims have been made in a particular year.

(iv) (a) Show that the ML estimate, $\hat{\alpha}$ of the parameter α needs to satisfy the equation:

$$\log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \log(\bar{x}) - \frac{\sum_{i=1}^{n} \log(x_i)}{n}$$

where $\Gamma'(\hat{\alpha})$ denotes the first derivative of $\Gamma(\hat{\alpha})$ with respect to $\hat{\alpha}$.

(b) Comment on how the ML estimates of the parameters α and λ can be obtained in this case.

[5]
[Total 21]

**10** A randomised clinical trial was conducted with the aim of investigating the effectiveness of two drugs (A and B) for tackling stage-fright in musicians before a performance. Ten musicians were allocated to each of two groups: group A (taking drug A) and group B (taking drug B). A further 10 musicians were allocated to a third group C, which served as a control group with the musicians not taking any drug. All group allocations were random and the musicians did not know which treatment they received.

At the end of the performance all 30 musicians were asked to give a score indicating their stage-fright, on a scale 1–5, with a score of 1 implying "not nervous at all" and a score of 5 implying "extremely nervous". The scores were then transformed by taking their logarithm (values denoted by $y$), and the results are shown below:

| *Group* | | | | | | | | | | | $\sum y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.693 | 1.099 | 0 | 1.099 | 0.693 | 0.693 | 0 | 1.099 | 1.099 | 0 | 6.475 |
| B | 0 | 0.693 | 0.693 | 1.386 | 0 | 0.693 | 1.099 | 0.693 | 0 | 1.099 | 6.356 |
| C | 0 | 1.609 | 1.099 | 1.386 | 1.099 | 1.609 | 0.693 | 1.099 | 1.609 | 1.386 | 11.589 |

For these data: $SS_T = 8.364$, $SS_B = 1.785$, $SS_R = 6.579$ (as defined in page 26 of the Formulae and Tables).

(i) Perform an analysis of variance to test the null hypothesis that the mean level of stage-fright is the same among the three groups. [5]

(ii) Determine the residuals for the first musician in groups A and B using the fitted model in part (i). [2]

(iii) Determine a 95% confidence interval for the variance of the scores (on the logarithmic scale), based on:

(a) the residual sum of squares.
(b) the sum of squares between treatments.

[6]

(iv) Comment on the interval obtained in part (iii)(b). [2]

It is suggested that any differences in the scores could be explained by the difference between the scores of the control group and the groups receiving a stage-fright drug.

(v) Suggest how this effect can be formally tested. You should not carry out any test. [2]

[Total 17]

## END OF PAPER