INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

25 April 2013 (pm)

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. *Mark allocations are shown in brackets.*

4. *Attempt all 11 questions, beginning your answer to each question on a separate sheet.*

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

Graph paper is 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 The following data represent the number of claims for twenty policyholders made during a year.

0	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	2	2	2	3

Determine the sample mean, median, mode and standard deviation of these data. [5]

2 Consider a random variable U that has a uniform distribution on $[0,1]$ and let F be the cumulative distribution function of the standard normal distribution.

Show that the random variable $X = F^{-1}(U)$ has a standard normal distribution. [3]

3 A discrete random variable X has a cumulative distribution function (CDF) with the following values:

Observation	10	20	30	40	50
CDF	0.5	0.7	0.85	0.95	1

Calculate the probability that X takes a value:

(i) larger than 10. [1]
(ii) less than 30. [1]
(iii) exactly 40. [1]
(iv) larger than 20 but less than 50. [2]
(v) exactly 20 or exactly 40. [2]
[Total 7]

4 Consider a random sample, $X_1, \ldots, X_n$, from a normal $N(\mu, \sigma^2)$ distribution, with sample mean $\bar{X}$ and sample variance S^2.

(i) Define carefully what it means to say that $X_1, \ldots, X_n$ is a random sample from a normal distribution. [2]

(ii) State what is known about the distributions of $\bar{X}$ and S^2 in this case, including the dependencies between the two statistics. [3]

(iii) Define the t-distribution and explain its relationship with $\bar{X}$ and S^2. [2]
[Total 7]

5 Bank robberies in various countries are assumed to occur according to Poisson processes with rates that vary from year to year. It was reported that the number of robberies in a particular country in a specific year was 123. The number of robberies in a different country in the same year was 111. It can be assumed that each robbery is an independent event and that robberies occur independently in the two countries.

Determine an approximate 90% confidence interval for the difference between the true yearly robbery rates in the two countries. [6]

6 A survey is undertaken to investigate the proportion p of an adult population that support a certain government policy. A random sample of 100 adults is taken and contains 30 who support the policy.

(i) Calculate an approximate 95% confidence interval for p. [2]

(ii) Comment on the validity of the interval obtained in part (i). [1]

A different sample of 1,000 adults is taken and it contains 300 who support the policy.

(iii) Explain how the width of a 95% confidence interval for p in this case will compare to the width of the interval in part (i), without performing any calculations. [1]

[Total 4]

7 A regulator wishes to inspect a sample of an insurer's claims. The insurer estimates that 10% of policies have had one claim in the last year and no policies had more than one claim. All policies are assumed to be independent.

(i) Determine the number of policies that the regulator would expect to examine before finding 5 claims. [1]

On inspecting the sample claims, the regulator finds that actual payments exceeded initial estimates by the following amounts:

£35 £120 £48 £200 £76

(ii) Find the mean and variance of these extra amounts. [3]

It is assumed that these amounts follow a gamma distribution with parameters α and λ.

(iii) Estimate these parameters using the method of moments. [3]

[Total 7]

8 A random sample of 10 independent claim amounts was taken from each of three different regions and an analysis of variance was performed to compare the mean level of claims in these regions. The resulting ANOVA table is given below.

Source	*d.f.*	*SS*	*MSS*
Between regions	2	4,439.7	2,219.9
Residual	27	10,713.5	396.8
Total	29	15,153.2	

(i) Perform the appropriate F test to determine whether there are significant differences between the mean claim amounts for the three regions. You should state clearly the hypotheses of the test. [4]

The three sample means were:

Region	*A*	*B*	*C*
Sample mean	147.47	154.56	125.95

It was of particular interest to compare regions A and B.

(ii) (a) Calculate a 95% confidence interval for the difference between the means for regions A and B.

(b) Comment on your answer in part (ii)(a) given the result of the F test performed in part (i). [4]

[Total 8]

9 A behavioural scientist is observing a troop of monkeys and is investigating whether social status affects the amount of food that an individual takes. The monkeys are divided into two groups of different social rank and the scientist counts the number of bananas each individual takes. Each monkey can take a maximum of 7 bananas.

Social rank	A	B
Number of monkeys	6	11
Total bananas taken	33	37

(i) It is first suggested that the number of bananas taken by each individual of each group follows the same binomial distribution with common parameter p and n=7.

(a) Use the method of moments to estimate the parameter p.

(b) The scientist is unsure whether a common parameter is appropriate and wishes to compare p_A and p_B, the probability that a banana is taken by an individual in groups A and B respectively.

Test the hypothesis that $p_A = p_B$.

[7]

(ii) A statistician suggests an alternative model. The number of bananas taken by an individual still follows a binomial distribution with n=7, but for group A the parameter is 2θ and for group B the parameter is θ, where $\theta < 0.5$.

(a) Show that the log likelihood for θ is given by:

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

(b) Hence calculate the maximum likelihood estimate of θ.

[6]

(iii) (a) Compare the fit of the two suggested models in parts (i) (with common parameter p) and (ii) by considering the expected number of bananas taken in groups A and B under the two models. You are not required to perform a formal test.

(b) Comment on the above comparison in relation to your answer in part (i)(b).

[4]
[Total 17]

10 The random variable S represents the annual aggregate claims for an insurer from policies covering damage due to windstorms. S is modelled as follows:

$$S = \sum_{i=1}^{M} Y_i$$

where:

M denotes the number of windstorms each year and has a Poisson distribution with mean κ

Y_i denotes the aggregate claims from the ith windstorm and is modelled as

$$Y_i = \sum_{j=1}^{N_i} X_{ij}$$

where:

N_i denotes the number of claims from the ith windstorm.

$N_1, N_2, \ldots, N_M$ are independent and identically distributed random variables, each with a Poisson distribution with rate λ.

X_{ij} denotes the amount of the jth claim from the ith windstorm.

X_{ij}, $i = 1, \ldots, M, j = 1, \ldots, N_i$ is a sequence of independent and identically distributed random variables, each with mean μ and variance σ^2.

It is assumed that the random variables M, N_i and X_{ij} are independent of each other.

(i) Derive expressions for the mean and the variance of Y_i in terms of λ, μ and σ. [2]

(ii) Derive expressions for the mean and the variance of S in terms of κ, λ, μ and σ. [3]

Now suppose that X_{ij} has an exponential distribution with mean 1.

(iii) Show that for any positive numbers x and C

$$P(X_{ij} \le x + C | X_{ij} > C) = P(X_{ij} \le x).$$ [3]

Consider the new random variable S_R given as:

$$S_R = \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}^*$$

where: $X_{ij}^* = \begin{cases} X_{ij} - 2 & \text{if } X_{ij} \geq 2 \\ 0 & \text{otherwise} \end{cases}$.

Let N_i^* be the number of non-zero X_{ij}^* amounts, i.e. the number of claim amounts from the *i*th windstorm that are greater than 2.

Also assume that $N_1^*, N_2^*, \ldots, N_M^*$ are independent and identically distributed Poisson random variables, with parameter λ^*.

Let $\kappa = 4$, $\lambda = 1{,}000$.

(iv) (a) Show that $\lambda^* = 135.3$.

(b) Explain why the distribution of X_{ij}^* is exponential with mean 1.

(c) Calculate the mean and variance of S_R. [7]

[Total15]

11 The table below gives the frequency of a critical illness disease by age group in a certain study. The table also gives the age midpoint (x), the number of people in each group (n), and $y = \log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right)$, where $\hat{\theta}$ denotes the proportion in an age group with the disease.

		Contracted disease?			
Age group	*x*	*Yes*	*No*	*n*	*y*
20–29	25	1	9	10	−2.19722
30–34	32.5	2	13	15	−1.87180
35–39	37.5	3	9	12	−1.09861
40–44	42.5	5	10	15	−0.69315
45–49	47.5	6	7	13	−0.15415
50–54	52.5	5	3	8	0.51083
55–59	57.5	13	4	17	1.17865
60–69	65	8	2	10	1.38629

$\sum x = 360$; $\sum x^2 = 17437.5$; $\sum y = -2.9392$; $\sum y^2 = 13.615$; $\sum xy = -9.0429$

(i) Calculate an estimate of the probability of having the disease under the assumption that the probability is the same for all age groups. [1]

Consider the hypothesis that there are no differences in the probability of having the disease for the different age groups.

(ii) (a) Construct an 8×2 contingency table which includes the expected frequencies under this hypothesis.

(b) Conduct a χ^2 test to investigate the hypothesis.

[6]

Consider the linear regression model $y = \alpha + \beta x + \varepsilon$, where the error terms (ε) are independent and identically distributed following a $N(0, \sigma^2)$ distribution.

(iii) (a) Draw a scatterplot of y against x and comment on the appropriateness of the considered model.

(b) Calculate the fitted regression line of y on x.

(c) Calculate a 99% confidence interval for the slope parameter.

(d) Interpret the result obtained in part (ii) with reference to the confidence interval obtained in part (iii)(c).

[14]
[Total 21]

END OF PAPER