INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

28 September 2012 (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. *Mark allocations are shown in brackets.*
4. *Attempt all 13 questions, beginning your answer to each question on a separate sheet.*
5. *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 Calculate the mean, the median and the mode for the data in the following frequency table.

Observation	0	1	2	3	4
Frequency	20	54	58	28	0

[3]

2 The following data are sizes of claims (ordered) for a random sample of 20 recent claims submitted to an insurance company:

174	214	264	298	335	368	381	395	402	442
487	490	564	644	686	807	1092	1328	1655	2272

(i) Calculate the interquartile range for this sample of claim sizes. [3]

(ii) Give a brief interpretation of the interquartile range calculated in part (i). [1]

[Total 4]

3 Let X be a discrete random variable with the following probability distribution:

X	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

Calculate the variance of Y, where $Y = 2X + 10$. [3]

4 Consider a random variable U that has a uniform distribution on (0,1) and a random variable X that has a standard normal distribution. Assume that U and X are independent.

Determine an expression for the probability density function of the random variable $Z = U + X$ in terms of the cumulative distribution function of X. [4]

5 A large portfolio consists of 20% class A policies, 50% class B policies and 30% class C policies. Ten policies are selected at random from the portfolio.

(i) Calculate the probability that there are no policies of class A among the randomly selected ten. [1]

(ii) (a) Calculate the expected number of class B policies among the randomly selected ten.

(b) Calculate the probability that there are more than five class B policies among the randomly selected ten.

[2]

[Total 3]

6 A random sample of size n is taken from a gamma distribution with parameters $\alpha = 8$ and $\lambda = 1/\theta$. The sample mean is $\bar{X}$ and θ is to be estimated.

(i) Determine the method of moments estimator (MME) of θ. [2]

(ii) Find the bias of the MME determined in part (i). [2]

(iii) (a) Determine the mean square error of the MME of θ.

(b) Comment on the efficiency of the MME of θ based on your answer in part (iii)(a).

[3]

[Total 7]

7 Analyst A collects a random sample of 30 claims from a large insurance portfolio and calculates a 95% confidence interval for the mean of the claim sizes in this portfolio. She then collects a different sample of 100 claims from the same portfolio and calculates a new 95% confidence interval for the mean claim size.

(i) Explain how the widths of the two confidence intervals will differ. [2]

Analyst B obtains a 95% confidence interval for the mean claim size of this portfolio based on a different sample of 30 claims. She subsequently realises that one of the claims in the sample has an extremely large value and can be considered as an outlier. She decides to replace this claim with a new randomly selected one, whose size is not an outlier, and obtains a new 95% confidence interval.

(ii) Explain how the two confidence intervals will differ in the case of Analyst B.

[3]

[Total 5]

8 The random variable S is given as $S = Y_1 + Y_2 + \ldots + Y_N$ (with $S = 0$ if $N = 0$) where the random variables Y_i are identically and independently distributed according to a lognormal distribution with parameters $\mu = 0.5$ and $\sigma^2 = 0.1$. N is also a random variable which is independent of Y_i, and its distribution given below.

N	0	1	2	3	4
$\Pr(N = n)$	0.1	0.3	0.3	0.2	0.1

Calculate the mean and the variance of the random variable S. [7]

9 An analyst is interested in using a gamma distribution with parameters $\alpha = 2$ and $\lambda = ½$, that is, with density function $f(x) = \frac{1}{4} x e^{-\frac{1}{2}x}$, $0 < x < \infty$.

(i) (a) State the mean and standard deviation of this distribution.

(b) Hence comment briefly on its shape.

[2]

(ii) Show that the cumulative distribution function is given by

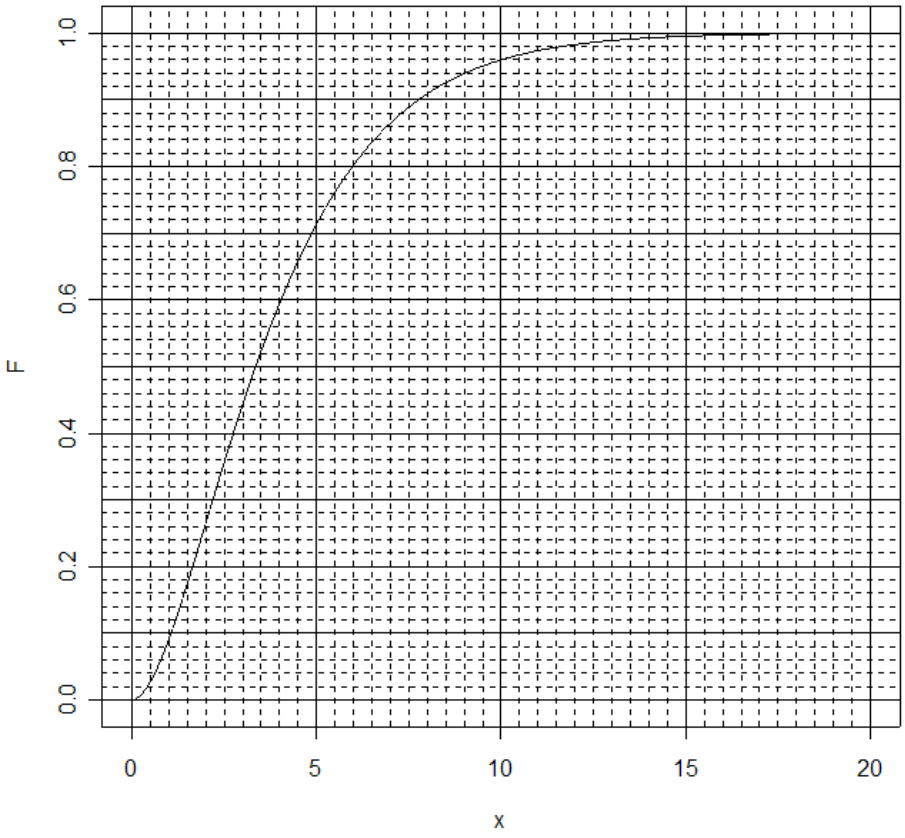
$$F(x) = 1 - (1 + \frac{1}{2}x)e^{-\frac{1}{2}x}, \quad 0 < x < \infty \quad \text{(zero otherwise).}$$ [3]

The analyst wishes to simulate values x from this gamma distribution and is able to generate random numbers u from a uniform distribution on (0,1).

(iii) (a) Specify an equation involving x and u, the solution of which will yield the simulated value x.

(b) Comment briefly on how this equation might be solved.

(c) The graph below gives $F(x)$ plotted against x. Use this graph to obtain the simulated value of x corresponding to the random number $u = 0.66$.



[3]
[Total 8]

10 The number of hours that people watch television per day is the subject of an empirical study that is carried out in four regions in a country. Five people are randomly selected in each of the regions and are asked about the average number of hours per day that they spent watching television during the last year. The results are shown in the following table, with the last column shows the average in each region.

						Average
Region 1	2.0	1.1	0.2	3.8	2.8	1.98
Region 2	1.2	1.0	0.9	1.1	1.6	1.16
Region 3	2.5	2.0	2.6	2.4	2.3	2.36
Region 4	1.2	1.7	1.0	1.8	1.3	1.40

Based on the above observations the following ANOVA table was obtained:

Source of variation	d.f.	SS	MSS
Between regions	...	4.4655	...
Residual	...	8.892	...

(i) State the mathematical model underlying the one-way analysis of variance together with all associated assumptions. [3]

(ii) Complete the ANOVA table. [1]

(iii) Carry out an analysis of variance to test the hypothesis that the region has no effect on the average time spent watching television. You should write down the null hypothesis, calculate the value of the test-statistic, state its distribution including any parameters, calculate the *p*-value approximately and state your conclusion. [4]

[Total 8]

11 In order to compare the effectiveness of two new vaccines, A and B, for a childhood disease, 11 infants were immunised with vaccine A and 9 infants were immunised with vaccine B. One month after immunisation the concentration of the disease antibodies in the blood of each infant was recorded in appropriate units. The sample mean and variance for each group is given below.

Vaccine A: $n_A = 11, \bar{x}_A = 4.05, s_A^2 = 0.692$

Vaccine B: $n_B = 9, \bar{x}_B = 4.36, s_B^2 = 0.813$

It is assumed that the distributions of the antibody concentration levels after immunisation with vaccine A and vaccine B are $N(\mu_A, \sigma_A^2)$ and $N(\mu_B, \sigma_B^2)$ respectively. You may assume that the samples are independent.

(i) State the distribution of the pivotal quantity $\dfrac{s_A^2 / \sigma_A^2}{s_B^2 / \sigma_B^2}$. [2]

(ii) Calculate an equal-tailed 95% confidence interval for the ratio $\dfrac{\sigma_A^2}{\sigma_B^2}$ using the pivotal quantity in part (i). (You are not required to show the derivation of the interval.) [4]

We now assume that $\sigma_A^2 = \sigma_B^2 = \sigma^2$. Under this assumption, you are given that the distribution of $\dfrac{18S_p^2}{\sigma^2}$ is χ_{18}^2, where S_p^2 is the pooled variance of the two samples and is independent from $\bar{x}_A$ and $\bar{x}_B$.

(iii) Explain why, under the above result, the sampling distribution of

$$\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{S_p\sqrt{\dfrac{1}{11} + \dfrac{1}{9}}}$$

is t_{18}. [4]

(iv) Calculate an equal-tailed 95% confidence interval for $\mu_A - \mu_B$ using the sampling distribution in part (iii). (You are not required to show the derivation of the interval.) [3]

(v) Comment on your results with regard to differences between vaccine A and vaccine B. [2]

[Total 15]

12 An insurer has collected data about the body mass index of 200 males between the age of 18 and 40. The results are shown in the following table.

Body mass index	< 18.5	18.5–25	25–30	>30
Observed frequency	6	114	62	18

A statistician suggests the following model for the distribution of the body mass index with an unknown parameter p.

Body mass index	< 18.5	18.5–25	25–30	>30
Relative frequency	p	$20p$	$10p$	$1-31p$

(i) Estimate the parameter p using the method of maximum likelihood. [4]

(ii) Perform a statistical test to decide whether the suggested distribution is appropriate for the observed data. You should state the null hypothesis for the test and your decision. [6]

To improve the description of the distribution of the body mass index, it is suggested that the marital status of the males in this study is also recorded. The results are shown in the following table.

Marital Status	Body mass index				Total
	< 18.5	18.5–25	25–30	>30	
Single	5	98	43	12	158
Married	1	16	19	6	42
Total	6	114	62	18	200

A life office has considered a sample of 10,000 men aged between 18 and 40 of which 50% are married and the other 50% are single.

(iii) Estimate the proportion of men with a body mass index of more than 30 in this sample, based on the data in the above table. [2]

(iv) Determine whether the body mass index is independent of the marital status or not, using an appropriate statistical test. You should state the null hypothesis for the test, calculate the value of the test statistic and the approximate p-value and state your decision. [8]

[Total 20]

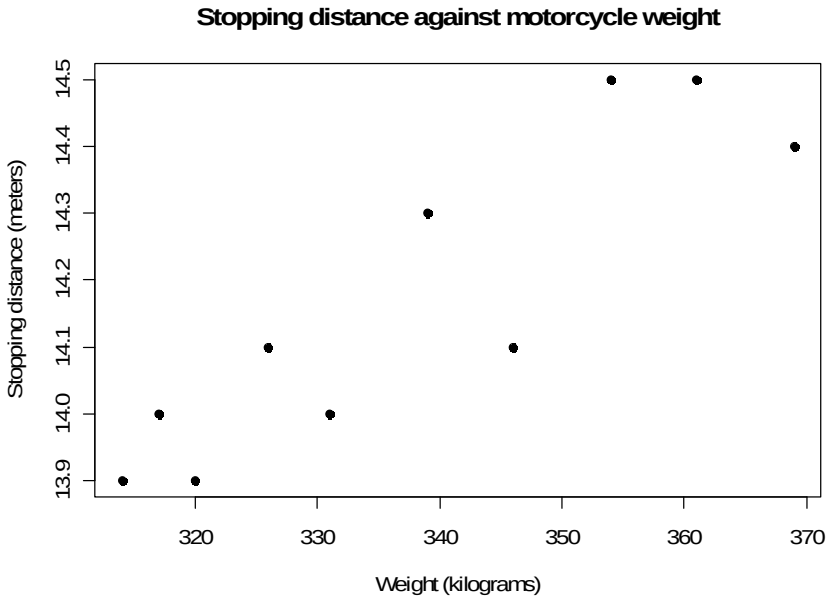
13 The following data give the weight, in kilograms, of a random sample of 10 different models of similar motorcycles and the distance, in metres, required to stop from a speed of 20 miles per hour.

Weight x	314	317	320	326	331	339	346	354	361	369
Distance y	13.9	14.0	13.9	14.1	14.0	14.3	14.1	14.5	14.5	14.4

For these data: $\sum x = 3{,}377$, $\sum x^2 = 1{,}143{,}757$, $\sum y = 141.7$, $\sum y^2 = 2{,}008.39$, $\sum xy = 47{,}888.6$

Also: $S_{xx} = 3{,}344.1$, $S_{yy} = 0.501$, $S_{xy} = 36.51$

A scatter plot of the data is shown below.



(i) (a) Comment briefly on the association between weight and stopping distance, based on the scatter plot.

(b) Calculate the correlation coefficient between the two variables. [2]

(ii) Investigate the hypothesis that there is positive correlation between the weight of the motorcycle and the stopping distance, using Fisher's transformation of the correlation coefficient. You should state clearly the hypotheses of your test and any assumption that you need to make for the test to be valid. [6]

(iii) (a) Fit a linear regression model to these data with stopping distance being the response variable and weight the explanatory variable.

(b) Calculate the coefficient of determination for this model and give its interpretation.

(c) Calculate the expected change in stopping distance for every additional 10 kilograms of motorcycle weight according to the model fitted in part (iii)(a).

[5]

[Total 13]

END OF PAPER