INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

18 April 2018 (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. *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 11 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 A scientist collects the following data sample on the number of plants grown on newly fertilised plots of land.

Number of plants	*Frequency*
1	2
2	6
3	3
4	8
5	1

(i) Calculate the mean, median and mode of the sample. [3]

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

[Total 5]

2 Consider the following data, and the corresponding sums derived from the data:

$$x_i : 10.0 \quad 6.9 \quad 11.4 \quad 12.6 \quad 10.3 \quad 12.4 \quad 9.8$$

$$\sum x_i = 73.4; \sum x_i^2 = 792.22; \sum x_i^3 = 8{,}750.972.$$

(i) Determine the third moment about the mean for these data. [2]

(ii) (a) Write down the mathematical definition of the coefficient of skewness of a set of data. [1]

(b) Determine the coefficient of skewness for the data above. [2]

[Total 5]

3 The number of minutes late that students arrive at a lecture is a random variable following an exponential distribution with mean 5 minutes.

(i) Determine the probability that a student is more than 10 minutes late to the lecture. [1]

Twenty students arrive at the lecture independently of each other.

(ii) Determine the exact probability that fewer than two of the students are more than 10 minutes late. [4]

[Total 5]

4 Consider a random sample $X_1, \ldots, X_n$ from a normal distribution with expectation μ and variance σ^2. The sample size is $n = 9$. We define the statistics $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

Determine the following probabilities:

(i) $P[S^2 > \sigma^2]$ [2]

(ii) $P[\overline{X} > \mu \mid S^2 > \sigma^2]$ [2]

(iii) $P[\overline{X} - \mu > \sigma]$ [2]

(iv) $P[\overline{X} - \mu > S]$ [2]

[Total 8]

5 A random sample of size 49 from a normal distribution gives a 99% confidence interval for the population mean as (30, 50).

Determine a 90% confidence interval for the population mean based on this information. [5]

6 A sample is drawn from the normal distribution with mean μ and variance σ^2. Denote the sampling variance by S^2.

(i) Show that the expected value of S^2 is σ^2, using the sampling distribution of S^2. [2]

Due to a spreadsheet error a scientist accidentally uses the following to estimate the variance

$$G^2 = \frac{1}{n}\left(\Sigma x_i^2 - n\overline{x}^2\right)$$

(ii) (a) Determine the bias of G^2 as an estimator of σ^2. [2]

(b) Comment on how the bias behaves as n gets large. [1]

[Total 5]

7 We consider a random sample $X_1, \ldots, X_n$ from a normal distribution with expectation $E[X_i] = \frac{1}{2}\vartheta^2$ and variance $V(X_i) = \sigma^2$ for all i.

(i) Derive an estimator $\hat{\vartheta}$ for the unknown parameter ϑ using the method of moments. [3]

(ii) Estimate ϑ for a sample of size 200 for which $\sum_{i=1}^{200} x_i = 900$ using the estimator derived in part (i). [1]

(iii) Comment on the suitability of the estimator in part (i). In particular, consider the two cases $\sum_{i=1}^{n} X_i < 0$ and $\sum_{i=1}^{n} X_i \geq 0$. [1]
[Total 5]

8 Consider a one-way analysis of variance.

(i) List the assumptions required for an analysis of variance to be valid. [2]

A study involving three different types of treatment for a medical condition gave the following sums on a particular improvement score (y):

	n	Σy_i	Σy_i^2
Treatment 1	8	193.03	4,697.80
Treatment 2	9	259.49	7,508.30
Treatment 3	9	263.08	7,730.34

In the above table, n denotes the number of patients receiving each of the three types of treatment.

(ii) Test the hypothesis, using an analysis of variance, that the mean improvement score for each type of treatment is the same. [6]
[Total 8]

9 The random variables X and Y have joint probability density function (pdf)

$$f_{X,Y}(x,y) = \begin{cases} 24x^3y & \text{for } 0 < x < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(i) (a) Show that the marginal pdf of X is

$$f_X(x) = 12x^3(1 - x^2),\ 0 < x < 1.$$

(b) Show that the marginal pdf of Y is $f_Y(y) = 6y^5,\ 0 < y < 1$.

[2]

(ii) Determine the covariance cov (X, Y). [5]

(iii) Determine the conditional pdf $f_{X|Y}(x|y)$ together with the range of X for which it is defined. [2]

(iv) Determine the conditional probability $P\left(X > \frac{1}{3} \mid Y = \frac{1}{2}\right)$. [2]

(v) Determine the conditional expectation $E\left(X \mid Y = \frac{1}{4}\right)$. [3]

(vi) Verify that $E[E[X|Y]] = E[X]$ by evaluating each side of the equation. [3]

[Total 17]

10 A large pension scheme regularly investigates the lifestyle of its pensioners using surveys. In successive surveys it draws a random sample from all pensioners in the scheme and it obtains the following data on whether the pensioners smoke.

Survey 1: Of 124 pensioners surveyed, 36 were classed as smokers.
Survey 2: Of 136 pensioners surveyed, 25 were classed as smokers.

An actuary wants to investigate, using statistical testing at a 5% significance level, whether there have been significant changes in the proportion of pensioners, p, who smoke in the entire pension scheme.

(i) Perform a statistical test, without using a contingency table, to determine if the proportion p has changed from the first survey to the second. [7]

When a third survey is performed it is found that 26 out of the 141 surveyed pensioners, are smokers.

(ii) Perform a statistical test using a contingency table to determine if the proportion p is different among the three surveys. [7]

(iii) (a) Calculate the proportion of smokers in the third survey.

(b) Comment on your answers to parts (i) and (ii).

[2]

[Total 16]

11 A car insurance company wishes to investigate the relationship between the age of drivers and the average annual mileage. The company has asked drivers of specific ages about their annual mileage. The age of the drivers is denoted by x (where $x = 40, 45, \ldots, 75$), and the annual mileage (in 1,000 miles) is denoted by y. The company asked 100 drivers of each age.

The average annual mileage and the sample variance for the annual mileage for each age are shown in the following table, together with some relevant statistics.

									Sum	Sum of squares
age x	40	45	50	55	60	65	70	75	460	27,500
average mileage $\bar{y}$	15	14.5	14.1	13.4	13	12.1	11.8	11.4	105.3	1,398.23
sample variance	2.25	2.56	1.69	1.96	3.24	4.00	1.44	1.21		
$x \times \bar{y}$	600	652.5	705	737	780	786.5	826	855	5,942	

The second last column contains the sum of the eight other columns and the last column contains the sum of the squares of the eight other columns.

(i) Determine a 95% confidence interval for the average annual mileage of drivers aged 50 based on the sample of 100 drivers at this age, justifying any assumptions you make. [3]

(ii) Perform a test of the null hypothesis that the average annual mileage of drivers aged 40 is equal to the average annual mileage of drivers aged 50 based on the two samples of 100 drivers each. You should calculate an approximate p-value, make a test decision and justify your decision and any approximations. [4]

(iii) Determine the correlation coefficient between the observed average annual mileage $\bar{y}$ and the age x of the driver. [4]

Further studies show that the correlation coefficient between the actual annual mileage y for each individual driver and the age x of the driver based on the entire sample of 800 drivers is –0.63. You are not required to confirm this result.

(iv) Explain the difference between this correlation coefficient and the correlation coefficient calculated in part (iii). [2]

(v) State the circumstances under which the two correlation coefficients would be equal. [1]

(vi) Determine the parameters of the simple linear regression model with the actual annual mileage y for each individual driver being the response variable and age x the explanatory variable, including writing down the equation. [7]

[Total 21]

END OF PAPER