INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

21 April 2017 (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 10 questions, beginning your answer to each question on a new page.*

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

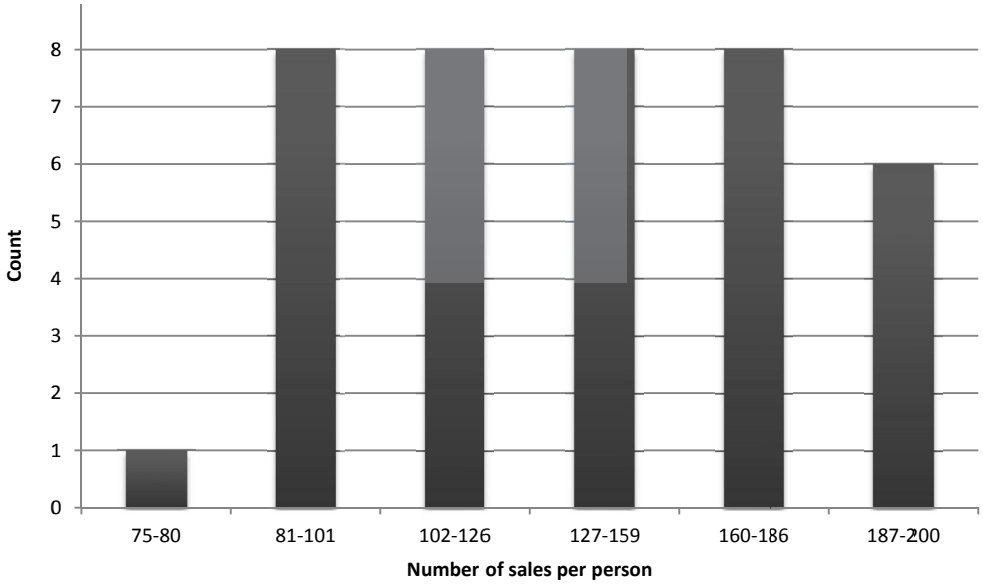
Graph paper is required for this paper.

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 A company is collecting data on its sales. It has 39 sales employees and records the number of sales that each one made in a month. The data are summarised by grouping the sales employees according to the number of sales they made and presented in the following chart:



Determine the mean and standard deviation of the number of sales made by a sales employee. [4]

2 Consider the random variable X having a distribution with probability density function:

$$f(x) = \nu\lambda x^{\nu-1}\exp(-\lambda x^{\nu}),\ 0 < x < \infty$$

where $\nu > 0$ and $\lambda > 0$ are the parameters of the distribution.

(i) Show that the cumulative distribution function of X is given by:

$$F(x) = \begin{cases} 1-\exp(-\lambda x^{\nu}), & x > 0 \\ 0, & x \le 0 \end{cases}$$ [2]

You are given a value $u = 0.671$ from the $U(0,1)$ distribution.

(ii) Determine by simulation a value of the random variable X when $\nu = 1.1$ and $\lambda = 0.2$. [2]

[Total 4]

3 Consider two random variables X and Y and assume that X and Y both follow a standard normal distribution but are not independent. Define the random variables:

$$Z^- = X - Y \text{ and } Z^+ = X + Y.$$

(i) Determine the covariance between Z^- and Z^+. [2]

(ii) Determine whether Z^- and Z^+ are uncorrelated based on your answer in part (i). [1]

[Total 3]

4 An insurance company calculates car insurance premiums based on the age of the policyholder according to three age groups: Group A consists of drivers younger than 22 years old; Group B consists of drivers 22–33 years old; and, Group C consists of drivers older than 33 years.

Its portfolio consists of 10% Group A policyholders, 38% Group B policyholders and 52% Group C policyholders.

The probability of a claim in any 12-month period for a policyholder belonging to Group A, B or C is 13%, 3% and 2%, respectively.

(i) Calculate the probability that a randomly chosen policyholder from this portfolio will make a claim during a 12-month period. [3]

One of the company's policyholders has just made a claim.

(ii) Calculate the probability that the policyholder is younger than 22 years. [2]

[Total 5]

5 Let $X_1, X_2, \ldots, X_n$ be a sequence of independent, identically distributed random variables with finite mean μ and finite (non-zero) variance σ^2.

(i) State the central limit theorem (CLT) in terms of the sum $\sum_{i=1}^{n} X_i$. [2]

Assume now that each X_i, $i = 1, 2, \ldots, 50$, follows an exponential distribution with parameter $\lambda = 2$ and let $Y = \sum_{i=1}^{50} X_i$.

(ii) Determine the approximate distribution of Y together with its parameters using the CLT. [2]

(iii) State the exact distribution of Y together with its parameters. [2]

(iv) Comment on the shape of the distribution of Y based on your answers to parts (ii) and (iii). [2]

[Total 8]

6 We consider the impact that different types of cars have on the amount spent on fuel per month. Three different types of cars are considered: small, medium and large. For each type of car a group of 15 drivers are asked about the amount of money (in £) spent on fuel per month. The results are summarised in the following table

Type of car	*Small*	*Medium*	*Large*
Sample mean	70	75	83
Sample standard deviation	16	19	16

For example, the 15 drivers of medium sized cars spent on average £75 per month with a sample standard deviation of £19.

(i) Perform a one-way analysis of variance to test the hypothesis that the type of car has no impact on the monthly amount spent on fuel. [6]

For some further investigation, only the difference between small and large cars is considered.

(ii) Determine a 95% confidence interval for the difference between the average amount spent on fuel for small cars and large cars, stating any assumptions you make. [4]

(iii) Test the null hypothesis that the average fuel costs for small and large cars are the same at a 5% significance level against the alternative that the fuel costs for small and large cars are different. [1]

[Total 11]

7 An investigation at a large airport focuses on the delay with which flights arrive. The delay time X, in minutes, is the difference between the actual time of arrival and the scheduled arrival time of delayed flights. Assume that X has an exponential distribution with parameter $\lambda > 0$.

(i) Derive the estimator $\hat{\lambda}$ for λ using the method of moments. [2]

The following table shows the observed values of X for a random sample of ten delayed flights.

45 20 120 90 60 30 45 90 60 150

(ii) Estimate the value of λ for this sample using the method of moments. [1]

To gain further insight into the distribution of flight delays, it is suggested that the time at which a flight is scheduled to arrive during a day has an impact on the delay. Therefore, assume now that X_i has an exponential distribution with a parameter λ that depends on the scheduled arrival time as follows:

$$X_i \sim \text{Exp}(\lambda_i) \text{ with } \lambda_i = \theta Z_i$$

where the random variable Z_i describes the scheduled arrival time (in minutes) after midnight on the day of arrival for the i^{th} randomly selected delayed flight and $\theta > 0$ is a parameter in this model.

(iii) Derive the maximum likelihood estimator $\hat{\theta}$ for the parameter θ. You should show that your solution is indeed a maximum. [5]

[Total 8]

8 An actuary models the number of claims X per year per policy as a discrete random variable with the following distribution

Number of claims	*0*	*1*	*2*	*3*	*More than 3*
Probability	*	p	$p/2$	$p/4$	$p/8$

where p is an unknown parameter.

(i) Show that $P[X=0]=\dfrac{8-15p}{8}$. [1]

(ii) Determine the range of possible values of p. [2]

In a sample of n independent policies there are N_0 policies with no claims during a year, N_1 policies with one claim, N_2 policies with two claims and N_3 policies with three claims. There are also some policies with more than three claims.

(iii) Show that the maximum likelihood estimator $\hat{p}$ for p based on observations of $N_0, \ldots, N_3$ in a sample of n independent claims is given by:

$$\hat{p}=\frac{8}{15}\frac{n-N_0}{n}.$$

You do not need to check that your solution is a maximum. [4]

(iv) Explain why the distribution of N_0 is a Binomial distribution specifying its parameters. [2]

(v) Verify that $\hat{p}$ is an unbiased estimator for p. [2]

Assume that in a sample of size $n = 300$ there were 100 policies with no claims during the previous year.

(vi) Determine the value of the variance of the estimator $\hat{p}$. [2]

The insurance company has now decided to limit the maximum number of claims per year to four per policy, but otherwise continue to use the distribution above. The claim amount of any individual claim is assumed to have a normal distribution with expectation 100 and standard deviation 20. Let S denote the total amount claimed in a portfolio of 300 independent policies during a year. We assume that claim amounts are independent of each other and independent of the number of claims.

Let X be the number of claims per policy per year and Y be the total number of claims per year.

(vii) (a) Show that $E(X) = 3.25p$ and $\text{Var}(X) = 7.25p - 10.5625p^2$.

Assume now that $p = 0.2$.

(b) Determine $E(Y)$ and $\text{Var}(Y)$.

(c) Determine the expected value and the standard deviation of S.

[7]

[Total 20]

9 A statistician is examining the survey methodology of a country's national statistics department. It conducts much of its data collection by telephoning individuals selected at random and asking them questions.

(i) Comment on whether this methodology will give a random sample. [2]

(ii) Comment on whether this methodology will give a representative random sample of the population. [2]

The department has been experimenting with surveying in person by visiting randomly selected individuals in their homes. To make this economical the department will only conduct surveys in a limited number of areas. It has asked the statistician to validate the effectiveness of its process.

For its first trial it conducts a small survey in two locations on the daily time spent accessing social media and gets the following results (in minutes).

	Number of interviews	*Mean*	*Standard deviation*
Area 1	25	50.0	20.2
Area 2	13	61.6	15.6

The statistician assumes that the underlying population is normally distributed.

(iii) (a) Determine a 95% confidence interval for the ratio of sample variances.

(b) Determine whether it is reasonable to assume that the variances are equal.

[4]

(iv) Perform a test at a 5% significance level to investigate whether the means are the same against a two sided alternative. [6]

The statistician then learns that there is an expectation that the mean of Area 2 is larger than the mean of Area 1.

(v) Perform a test to investigate whether the means are the same against an appropriate alternative at the same significance level as in part (iv). [3]

(vi) Comment on the results of parts (iv) and (v). [2]

[Total 19]

10 A geologist is trying to determine what causes sand granules to have different sizes. She measures the gradient of nine different beaches in degrees, g, and the diameter in mm of the granules of sand on each beach, d.

g	0.63	0.70	0.82	0.88	1.15	1.50	4.40	7.30	11.30
d	0.17	0.19	0.22	0.235	0.235	0.30	0.35	0.42	0.85

$$\sum g = 28.68, \sum g^2 = 206.2462, \sum d = 2.97, \sum d^2 = 1.33525, \sum gd = 15.55855$$

(i) Determine the linear regression equation of d on g. [5]

The geologist assumes that the error terms in the linear regression are normally distributed.

(ii) Perform a test to determine whether the slope coefficient is significantly different from zero. [4]

(iii) Determine a 95% confidence interval for the mean estimate of d on a beach with a slope of exactly 3 degrees. [5]

(iv) (a) Plot the data from the table above.

(b) Comment on the plot suggesting what the geologist might do to improve her analysis.

[4]

[Total 18]

END OF PAPER