Queen Mary
University of London

B. Sc. Examination by course unit 2012

MTH4106 Introduction to Statistics

Duration: 2 hours

Date and time: 1 May 2012, 1000h–1200h

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

You should attempt all questions. Marks awarded are shown next to the questions.

Calculators ARE permitted in this examination. The unauthorized use of material stored in pre-programmable memory constitutes an examination offence. Please state on your answer book the name and type of machine used.

Statistical functions provided by the calculator may be used provided that you state clearly where you have used them.

The New Cambridge Statistical Tables are provided.

Complete all rough workings in the answer book and cross through any work which is not to be assessed.

Important note: the Academic Regulations state that possession of unauthorized material at any time by a student who is under examination conditions is an assessment offence and can lead to expulsion from QMUL.

Please check now to ensure you do not have any notes, mobile phones or unauthorised electronic devices on your person. If you have any, then please raise your hand and give them to an invigilator immediately. Please be aware that if you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. Disruption caused by mobile phones is also an examination offence.

Exam papers must not be removed from the examination room.

Examiner(s): R. A. Bailey and H. Maruri-Aguilar

 TURN OVER

Question 1 (15 marks) Suppose that we have data $x_1, \ldots, x_n$.

(a) Name two measures of the location of the data, and explain how they are calculated.

Explain which of these is more appropriate if the data consist of the annual income of n people. [5]

(b) Name two measures of the spread of the data, and explain how they are calculated. [4]

(c) For $i = 1 \ldots, n$, let $y_i = ax_i + b$, where a and b are constants and $a > 0$. Let m_x be the median of $x_1, \ldots, x_n$, and let m_y be the median of $y_1, \ldots, y_n$. State and prove a result giving m_y in terms of m_x. [6]

Question 2 (15 marks) Biologists from Queen Mary are investigating fish in the River Lea. They trap them in nets, measure them, then release them again. For each fish, they record its species (minnow, bullhead, stickleback, etc.), its weight in grams and its length in millimetres.

(a) State the type of each of these three variables. [5]

(b) Explain briefly, using a sketch, how to display these three variables on a single diagram. [4]

(c) Suppose that species have been entered into column `C1` of Minitab (coded as minnow = 1, bullhead = 2, etc.), weights have been entered into column `C2` and lengths have been entered into column `C3`. Explain briefly how to use Minitab to produce the diagram in (b). [4]

(d) If the correlation between weight and length is 0.82, what can you predict about the appearance of the diagram in (b)? [2]

Question 3 (15 marks) Let X be a discrete random variable all of whose values are non-negative integers. Let $G(t)$ be the probability generating function of X.

(a) Prove that

$$\left.\frac{\mathrm{d}G(t)}{\mathrm{d}t}\right|_{t=1} = \quad \mathbb{E}(X)$$

and

$$\left.\frac{\mathrm{d}^2G(t)}{\mathrm{d}t^2}\right|_{t=1} = \quad \mathbb{E}(X^2) - \mathbb{E}(X).$$

[8]

(b) Suppose that λ is a positive constant and that $X \sim \text{Poisson}(\lambda)$. Then $G(t) = \mathrm{e}^{\lambda(t-1)}$. Use the results of (a) to find $\mathbb{E}(X)$ and $\text{Var}(X)$. [7]

Question 4 (15 marks) I have a 50 pence coin, and I want to find out if it is fair. Let p be the probability that it shows heads when I toss it. My null hypothesis is

$$\mathrm{H}_0 : p = 0.5$$

and my alternative hypothesis is

$$\mathrm{H}_1 : p \neq 0.5.$$

I plan to test the null hypothesis by tossing the coin 16 times and recording the number X of times that it shows heads. I shall use X as the test statistic and the set $\{0, 1, 2, 3\} \cup \{13, 14, 15, 16\}$ as the rejection region.

(a) Define *Type I error* and *Type II error*. [2]

(b) Find the significance level of this test. [3]

(c) Find the power of this test when $p = 0.4$. [4]

(d) I toss the 50 pence coin 16 times and find that it falls heads exactly twice. As a statistician, what do you conclude, and how do you report it? [6]

Question 5 (10 marks) In a class of 100 students, let M be the unknown number whose blood type is rhesus negative. A random sample of 20 students is taken without replacement, and their blood is tested. Let X be the number of students in the sample whose blood is rhesus negative.

(a) State the distribution of the random variable X. Hence write down $\mathbb{E}(X)$. [4]

(b) Let $Y = 5X$. Show that Y is an unbiased estimator of M. [4]

(c) Suppose that exactly three students in the sample have rhesus negative blood. Estimate the value of M. [2]

Question 6 (15 marks) Let W, X and Y be random variables which have a joint distribution.

(a) Define the *covariance* $\mathrm{Cov}(X, Y)$. [2]

(b) Prove that $\mathrm{Cov}(W, X + Y) = \mathrm{Cov}(W, X) + \mathrm{Cov}(W, Y)$. [3]

(c) Suppose that X and Y are independent of each other, that $X \sim N(-2, 1)$ and $Y \sim N(1, 8)$. Let $T = X + Y$.

(i) State the distribution of T. [3]

(ii) Find the correlation between X and T. [3]

(iii) Find $\mathbb{P}(T \leq -2)$, giving the answer correct to four decimal places. [4]

Question 7 (5 marks) State the Central Limit Theorem. [5]

Question 8 (10 marks) Environmental scientists are concerned about the effect on Lake Michigan of salt being put onto neighbouring city streets throughout the winter. One winter, they take 32 samples of water from the lake and measure the amount x_i of sodium in the i-th sample, for $i = 1, \ldots, 32$, giving the result in parts per million (ppm), recorded to one decimal place. Thus $x_1 = 18.5$, $x_2 = 16.8$, and so on.

They assume that they can regard these as a random sample, and that the distribution of sodium (in ppm) in the lake is normal with unknown mean μ and unknown variance σ^2. They find that

$$\sum_{i=1}^{32} x_i = 610.2 \quad \text{and} \quad \sum_{i=1}^{32} x_i^2 = 11964.30.$$

(a) Find the sample mean and the sample standard deviation of the data. [3]

(b) Find a 95% confidence interval for μ. [7]

End of Paper