Queen Mary
University of London

B. Sc. Examination by course unit 2011

MTH4106 Introduction to Statistics

Duration: 2 hours

Date and time: 4 May 2011, 1430h–1630h

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.

Candidates should note that the Examination and Assessment Regulations state that possession of unauthorized materials by any candidate who is under examination conditions is an assessment offence. Please check your pockets now for any notes that you may have forgotten that are in your possession. If you have any, then please raise your hand and give them to an invigilator now.

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) Researchers at the Australian Institute of Sport measured the heights (in cm) of several female students who regularly took part in sport. They entered the heights into one column of a Minitab worksheet. The second column classifies the students according to their main sport, with 'Netball' coded as 1 and '400m running' coded as 2. The heights are summarized below.

Descriptive Statistics: height

Variable	sport	N	N*	Mean	SE Mean	StDev	Minimum	Q1
height	1	23	0	176.07	0.864	4.14	168.60	173.30
	2	11	0	169.34	1.67	5.55	162.00	163.00

Variable	sport	Median	Q3	Maximum
height	1	176.00	179.60	183.30
	2	170.80	174.10	177.00

(a) What type of data are these? [2]

(b) In general, which of these two sports has the taller women? Justify your answer in two ways, using the summary statistics. [4]

(c) Are the heights more spread for the netball players or for the 400m runners? Justify your answer in two ways, using the summary statistics. [6]

(d) How, if at all, would your conclusions to (b) and (c) change if the heights were measured in inches? Briefly justify your answer. [3]

Question 2 (5 marks) Explain the difference between an observational study and an experiment. Give an example of each. [5]

Question 3 (15 marks) Let X, Y and Z be random variables with a joint distribution.

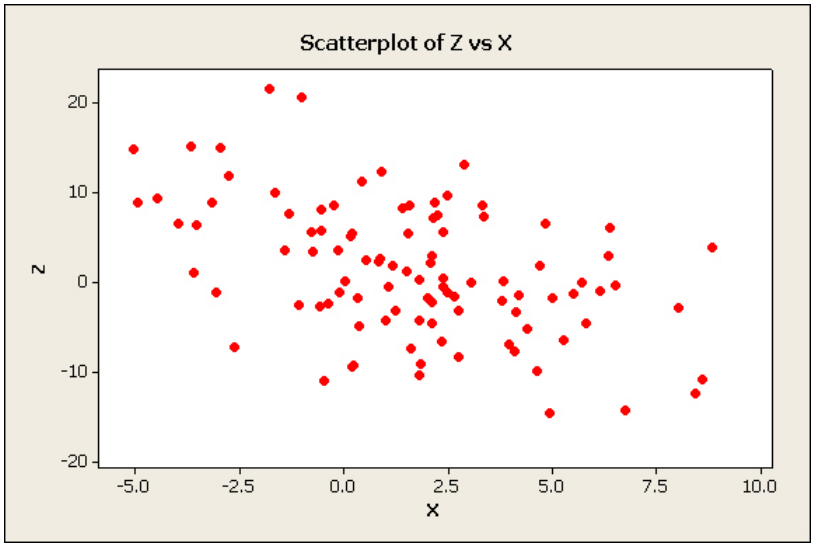
(a) Define the *covariance* $\mathrm{Cov}(X, Y)$. [2]

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

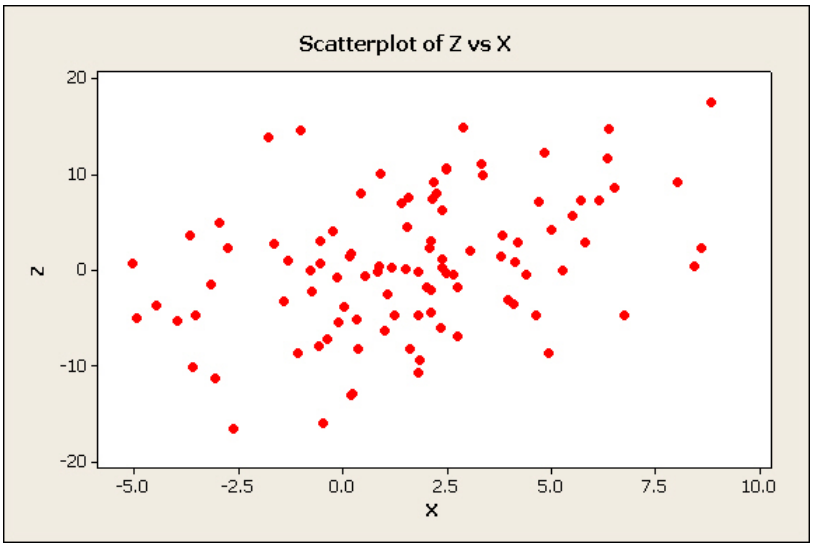
(c) Suppose that X and Y are independent, with $\mathrm{Var}(X) = 9$ and $\mathrm{Var}(Y) = 10$, and that $Z = X + 2Y$. Find the correlation between X and Z. [5]

(d) In a Minitab project, students are asked to simulate 100 rows of data from X and Y, where $X \sim N(2, 9)$ and $Y \sim N(0, 10)$, and X and Y are independent. Then they should calculate $Z = X + 2Y$, and produce a scatterplot of Z against X. Three scatterplots are shown on the next page: two of them are wrong. State which is correct. For each of the others, give a brief reason why it cannot be correct. [5]

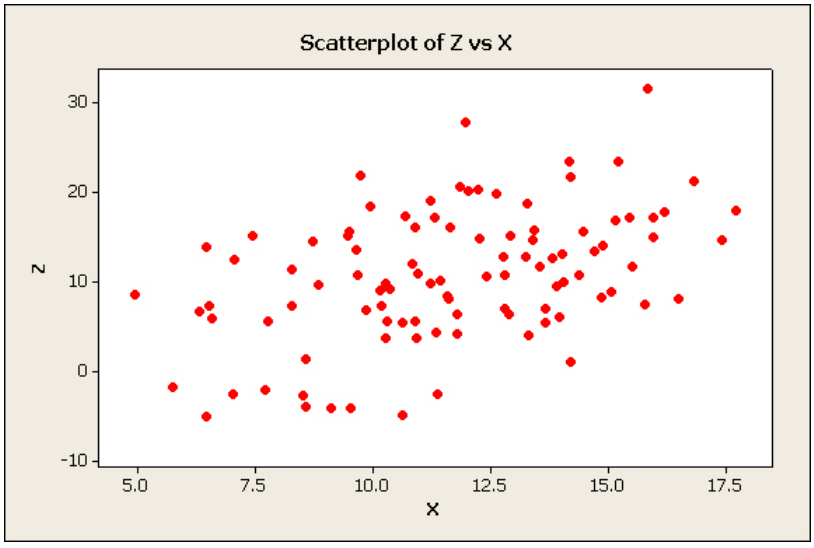
(i)



(ii)



(iii)



Question 4 (15 marks) Let X and Y be random variables all of whose values are non-negative integers.

(a) Define the *probability generating function* of X. [2]

(b) State a theorem linking the probability generating functions of X, Y and $X+Y$ under a suitable condition. [3]

(c) Suppose that $X \sim \text{Bin}(20, 0.4)$. Find the probability generating function of X. [5]

(d) Suppose that $X \sim \text{Bin}(20, 0.4)$, that $Y \sim \text{Bin}(30, 0.4)$ and that X and Y are independent of each other. Find the distribution of $X + Y$. [5]

Question 5 (10 marks) (a) Let T be an estimator for a quantity θ. Define the *bias* and *mean squared error* of T, and state a result which links them. [5]

(b) Let $X_1, \ldots, X_n$ be mutually independent random variables which all have the same distribution with expectation μ and variance σ^2. Let

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Find the bias and mean squared error of $\bar{X}$ if it is used as an estimator for μ. [5]

Question 6 (20 marks) Let p be the proportion of adults in the USA who have a laptop computer in their home. A survey company contacts a random sample of 3000 adults in the USA and asks them several questions about electronic gadgets. Let X be the number who say that they have a laptop computer in their home.

(a) State the distribution of the random variable X. Hence write down the expectation and variance of X. [4]

(b) Find $\mathbb{P}(X \geq 2425)$ if $p = 0.8$. [6]

(c) Suppose that exactly 2619 respondents in the sample say that they do have a laptop computer in their home.

(i) Find a 95% confidence interval for p. [8]

(ii) Your non-statistical friend asks what you mean by the confidence interval you have given. How do you answer? [2]

Question 7 (5 marks) The random variables X and Y are independent of each other, $X \sim N(0, 2)$ and $Y \sim N(-1, 7)$. Find the probability density function of $X + Y$. [5]

Question 8 (15 marks) A firm makes sprinkler systems for the prevention of fires in buildings. It is important that each sprinkler is activated quickly when a fire starts. The engineer in charge assumes that their activation times (in seconds) are approximately normally distributed, with unknown mean μ and unknown variance σ^2. He has designed the sprinklers to have $\mu \leq 25$. If $\mu > 25$ then he needs to change the design. He takes a random sample of 50 sprinklers, starts a fire near each one, and records their activation times x_i in seconds. He finds that

$$\sum_{i=1}^{50} x_i = 1396 \quad \text{and} \quad \sum_{i=1}^{50} x_i^2 = 45165.$$

(a) State the engineer's null and alternative hypotheses. [3]

(b) Find the sample mean and the sample standard deviation of the data. [4]

(c) Carry out the appropriate hypothesis test at the 10% significance level, and report the conclusion. [8]

End of Paper