Queen Mary
University of London

B. Sc. Examination by course unit 2010

MTH4106 Introduction to Statistics

Duration: 2 hours

Date and time: 6 May 2010, 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 (20 marks) Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be data collected on the same n items.

(a) Define the *sample covariance* s_{xy} and the *sample correlation coefficient* r. [5]

(b) There were 17 entrants in a beauty contest in a small town in the United States. Their weights and heights were recorded: x_i is the height of contestant i in inches, and y_i is the weight of contestant i in pounds. For these data,

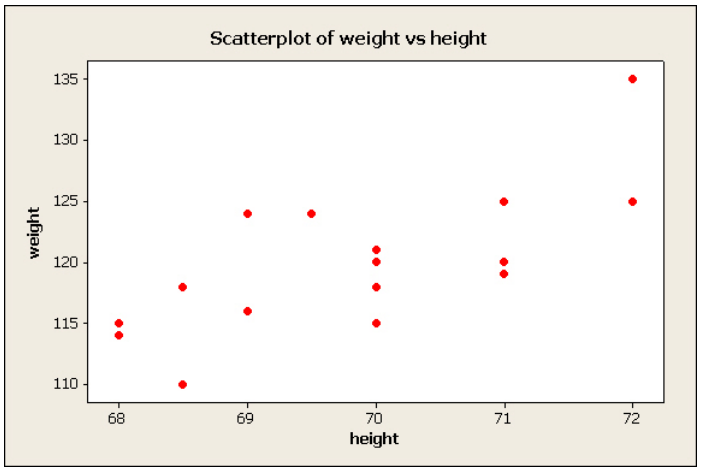
$$\sum_{i=1}^{17} x_i = 1135.5, \quad \sum_{i=1}^{17} y_i = 2035,$$

$$\sum_{i=1}^{17} x_i^2 = 75870.75, \quad \sum_{i=1}^{17} y_i^2 = 244135 \quad \text{and} \quad \sum_{i=1}^{17} x_i y_i = 136007.$$

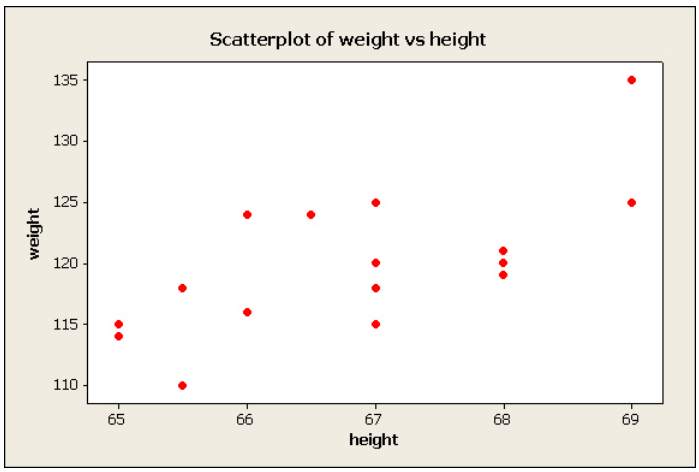
Calculate the sample correlation coefficient. [6]

(c) The data for weight and height are plotted on one of the following three scatterplots. State which scatterplot is the correct one. For each of the other two scatterplots, explain briefy why it cannot be correct. [5]

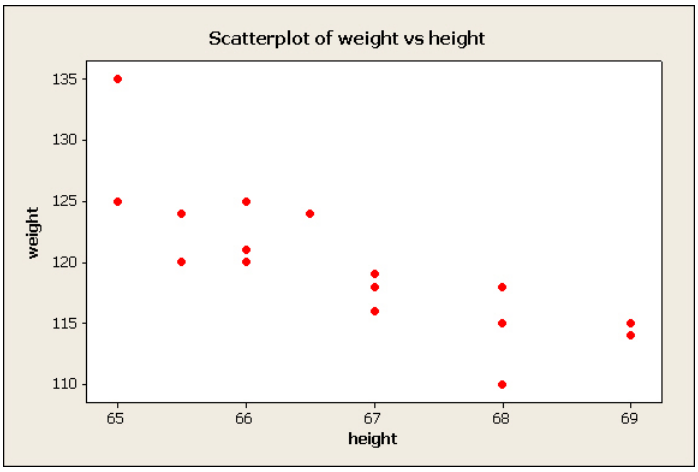
(i)



(ii)



(iii)



(d) How would the sample correlation coefficient change if the weights had been measured in kilograms rather than pounds? Justify your answer. [4]

Question 2 (15 marks) Let X be a discrete random variable all of whose values are non-negative integers. Let $G(x)$ be the probability generating function of X.

(a) Prove that

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

and

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

[8]

(b) If $0 < p < 1$ and $X \sim \mathrm{Geom}(p)$ then

$$G(x) = \frac{px}{1-qx},$$

where $q = 1 - p$. Find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$. [7]

Question 3 (10 marks) Using Minitab, a student enters the numbers 1–200 in order into column `C1`. He simulates 200 values from the distribution Exp(0.25) and puts them in column `C2`. Call these values $x_1, \ldots, x_{200}$. He uses the calculator to find `Partial Sum(C2)` and store it in column `C3`, and then to calculate `C3/C1` and store it in column `C4`. Finally he creates a scatterplot of `C4` against `C1`.

(a) Write down the entries in row 4 of columns `C1`, `C2`, `C3` and `C4`. [4]

(b) Describe the important features of the scatterplot, either in words or with a labelled sketch. [4]

(c) What theorem does this scatterplot illustrate? [2]

Question 4 (20 marks) Consider the population of all people who are ordinarily resident in the London borough of Tower Hamlets. Let p be the proportion of these who have ever been prescribed glasses or contact lenses. An optician wants to estimate p, so she chooses a random sample of 512 residents and asks each one if they have ever been prescribed glasses or contact lenses. Let X be the number who answer "yes".

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

(b) Find $\mathbb{P}(X \leq 200)$ if $p = 1/3$. [6]

(c) Let $Y = X/512$. Show that Y is an unbiased estimator of p. [4]

(d) Find the mean squared error of Y as a function of p. [4]

(e) Suppose that 157 residents in the sample answer "yes". Estimate the value of p, giving your answer to two decimal places. [2]

Question 5 (10 marks) Let $X_1, \ldots, X_n$ mutually independent random variables which all have the same distribution with expectation μ and variance σ^2. Let

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

(a) Prove that $\mathbb{E}(\bar{X}) = \mu$. [3]

(b) Find the variance of $\bar{X}$. [4]

(c) State the distribution of $\bar{X}$ when $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. [3]

Question 6 (15 marks) A manufacturer of soft drinks sells them in bottles which are labelled as containing 350 ml each. A new manager thinks that the firm is putting slightly too much drink into each bottle, and therefore spending too much money. He assumes that the distribution of the volume (in ml) of drink in the bottles is $N(\mu, 4)$, and thinks that if $\mu > 355$ then the firm should adjust the machines that fill the bottles.

He takes a random sample of twelve filled bottles, and measures the volume of drink they contain, as follows (all data are in ml).

353	357	356	357	360	357
356	359	355	358	356	358

(a) State the manager'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]

Question 7 (10 marks) (a) Explain what is meant by a 99% *confidence interval*. [3]

(b) A forester is studying the growth rate of red pine trees at an early stage. He measures the heights x_i (in cm) of 40 red pine seedlings one year after they were planted. He finds that the sample mean is 1.715 and the sample variance is 0.2252.

He assumes that he can regard these as a random sample of all such seedlings, and that the population heights are normally distributed, with unknown mean μ and unknown variance σ^2.

Find a 99% confidence interval for μ. [7]

End of Paper