UNIVERSITY COLLEGE LONDON

University of London

EXAMINATION FOR INTERNAL STUDENTS

For The Following Qualifications:-

B.Sc. *B.Sc.(Econ)M.Sci.*

Mathematics M252: Probability and Statistics

COURSE CODE	:	**MATHM252**
UNIT VALUE	:	**0.50**
DATE	:	**19-MAY-06**
TIME	:	**14.30**
TIME ALLOWED	:	**2 Hours**

TURN OVER

All questions may be attempted but only marks obtained on the best **four** *solutions will count.*
The use of an electronic calculator **is** *permitted in this examination.*

New Cambridge Statistical Tables are provided.

1. (a) Consider an experiment with sample space Ω. Let P be a function that allocates probabilities to subsets of Ω. State the three conditions that P must satisfy (i.e. the three axioms of probability).

   (b) In the context of an experiment with sample space Ω, give definitions of the following:

   (i) Random variable.

   (ii) Continuous random variable.

   You should explain clearly any additional notation or concepts that are needed in your definitions.

   (c) Suppose that X is a random variable with distribution function $F(\cdot)$. Use the axioms of probability to show that if $a < b$, $P(a < X \leqslant b) = F(b) - F(a)$. Deduce that $F(\cdot)$ is a non-decreasing function.

   (d) Show that if X is a continuous random variable, $P(X = x) = 0$ for all x.

2. (a) Write down the probability density function for a uniform distribution with parameters 0 and 1. Find an expression for the corresponding distribution function.

   (b) Suppose U has a uniform distribution with parameters 0 and 1, and let $X = -\log(U^2)$. Show that X has an exponential distribution, and give the parameter of this distribution. Derive the moment generating function (MGF) of X.

   (c) Suppose now that $U_1, \ldots, U_n$ are independent random variables, each distributed as $U(0,1)$. Write down an expression for the MGF of $S = -\sum_{i=1}^{n} \log(U_i^2)$.

   (d) It can be shown that the MGF of a chi-squared distribution with m degrees of freedom is $M(t) = (1-2t)^{-m/2}$. Deduce that in part (c), S has a chi-squared distribution. What are the degrees of freedom of this distribution?

   (e) Use the tables provided to evaluate $P\left(\prod_{i=1}^{6} U_i > 0.1\right)$, where $U_1, \ldots, U_6$ are independent random variables each distributed as $U(0,1)$.

   (You may use, without proof, the result that the MGF of a sum of independent random variables is the product of their individual MGFs).

3. (a) X and Y are independent Poisson random variables, each with mean λ.

   (i) Write down an expression for $P(X < 2)$, in terms of λ.

   (ii) Show that $P(Y < X | X < 2) = \lambda e^{-\lambda}/(1+\lambda)$.

   (iii) Without carrying out any calculations, explain why $P(Y < X) = P(X < Y)$. Deduce that $P(Y < X) = \frac{1}{2}[1 - P(X = Y)]$.

   (iv) Show that

$$P(X = Y) = \sum_{k=0}^{\infty} \frac{\lambda^{2k} e^{-2\lambda}}{(k!)^2} .$$

   When $\lambda = 1$, the value of this expression is 0.309 to three decimal places (do *not* attempt to verify this). What is $P(Y < X)$ in this case?

   (v) Show that

$$P(Y < X | X \geqslant 2) = \frac{P(Y < X) - P(Y < X | X < 2) P(X < 2)}{P(X \geqslant 2)} ,$$

   and evaluate this expression when $\lambda = 1$.

   (b) In a certain country, accidents caused by speeding motorists have in the past occurred in a Poisson process at a rate of 1 per month in any given location. At the end of last December the government, anxious to reduce the number of accidents, installed speed cameras at every location where there had been 2 or more accidents during the month. It was subsequently found that at about 80% of these speed camera locations, the number of accidents in January was less than that in December. The government was delighted by this, and concluded that speed cameras are an extremely effective way to prevent accidents.

   Is the government's conclusion justified? Explain your answer clearly.

4. The germination of plant seeds can be encouraged by soaking them in a solution of Gibberellic Acid-3 (GA-3). The strength of solution required depends on the plant species. In this question, consider a hypothetical species whose seeds will always germinate if treated with GA-3 solution in a concentration of between 500 and 550 parts per million (ppm), but will never germinate otherwise.

   (a) A botanist regularly makes up solutions of GA-3 by mixing powder with water. The concentration of GA-3 in each batch of solution is normally distributed, independently in each batch, with a mean of 525ppm and a standard deviation of 15ppm.

   (i) What is the probability that a batch of solution will enable seeds to germinate?

   (ii) Out of 4 batches of solution, what is the probability that all of the batches will allow seeds to germinate? What is the probability that exactly 3 batches will allow seeds to germinate?

(b) To increase the probability of enabling seeds to germinate, the botanist could try to reduce the standard deviation of the GA-3 concentration. What standard deviation would be required so that 99% of batches enable germination?

(c) The botanist signs up to a special offer from a horticultural supplier, as a result of which she receives 20 free bottles of GA-3 solution. The GA-3 concentrations in the bottles are independently normally distributed, with a mean of 500ppm and standard deviation of 20ppm. Out of 4 of these bottles, what is the probability that exactly 3 will enable seeds to germinate?

(d) After using 10 of the free bottles, the botanist makes up 10 batches of her own GA-3 solution (with a standard deviation of 15ppm) and stores these in the empty bottles, on a different shelf from the remaining free bottles. She then goes on holiday. By the time she comes back, she has forgotten which shelf contains the free bottles, and which contains the new solution. She takes four bottles from one of the shelves, and uses each bottle to treat a tray of seeds. Germination takes place in three of the four seed trays. Given this information, what is the probability that she is using bottles containing the new solution?

5. (a) n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ have been generated from the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1, \ldots, n),$$

where the (ε_i) are independent normally distributed random variables, each with mean zero and variance σ^2. Show that the least squares estimates of β_1 and β_0 are $\hat{\beta}_1 = C_{xy}/C_{xx}$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ respectively, in a notation which you should define.

(b) In a study to determine the effect of temperature upon the amount of converted sugar in a certain process, the following data were obtained:

Temperature (°C)	60	61	62	63	64	65	66	67	68	69	70
Converted sugar (g)	8.1	7.8	8.5	9.8	9.5	8.9	8.6	10.2	9.3	9.2	10.5

(i) Estimate the parameters β_0 and β_1 in a linear regression model to predict converted sugar from temperature.

(ii) It can be shown that for such a linear regression model, the standard error of the estimator $\hat{\beta}_1$ is $\sigma^2/\sqrt{C_{xx}}$. For these data, the variance σ^2 in the linear regression model is estimated as $\hat{\sigma}^2 = 0.400$.
Is there any evidence here that temperature has a genuine effect on the amount of converted sugar?

6. In February 2006, an overworked university lecturer realised at rather short notice that he needed to prepare an assessment for a course that he was teaching. As he didn't have time to think of a complete set of new questions for the assessment, he decided to re-use a question paper that had been set for the same course in 2002. The assessment was marked out of 100 on both occasions, yielding percentage marks for each student. A summary of the results for each year is as follows:

	2002	**2006**
Number of students:	16	44
Mean percentage mark:	65.25	62.18
Variance of percentage marks:	332.2	714.8

(a) Give the standard formulae used to calculate sample means and variances such as those in the table above. Clearly define any notation that you use.

(b) Using the tables provided, show that the lower 2.5% point of an $F_{15,43}$ distribution is between 0.370 and 0.418.

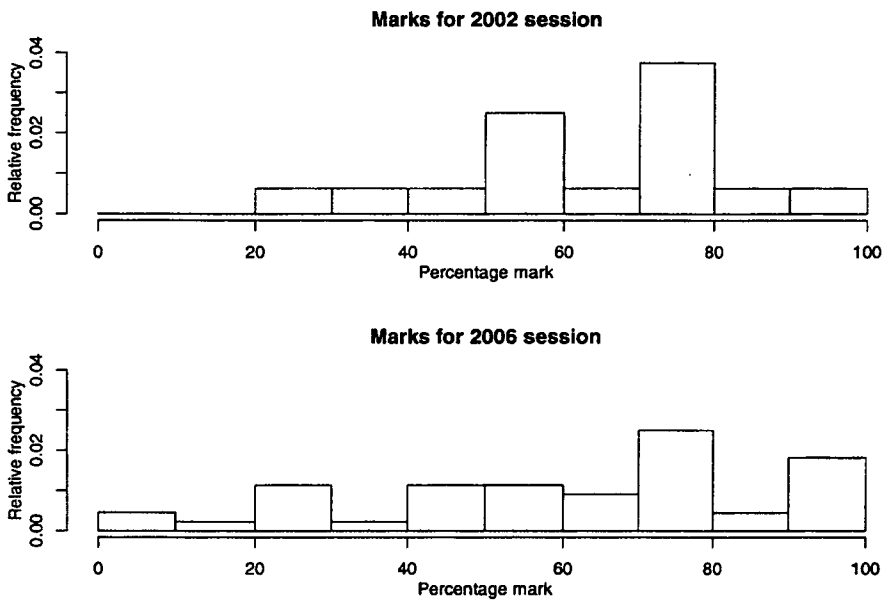
(c) Assume that the two sets of marks can be regarded as independent samples from populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, respectively.

(i) Test, at the 95% level and using a 2-tailed test, the hypothesis that $\sigma_1^2 = \sigma_2^2$.

(ii) Assuming that $\sigma_1^2 = \sigma_2^2$, calculate a 95% confidence interval for $\mu_1 - \mu_2$.

(d) State any additional assumptions made during the analyses in part (b). If these assumptions are satisfied, what do the analyses tell you about student performance in 2006 compared to that in 2002?

(e) Histograms of the marks obtained in each year are as follows:



Do these tell you anything about the validity of the assumptions made during the previous analyses?

MATHM252 END OF PAPER