UNIVERSITY OF BRISTOL

Examination for the Degrees of B.Sc. and M.Sci. (Level 1)

STATISTICS 1

MATH 11400

(Paper Code MATH-43C)

May/June 2010, 1 hour 30 minutes

This paper contains two sections, Section A and Section B.
Answer each section in a separate answer book.

Section A contains **five** *questions,* **ALL** *of which will be used for assessment.*
This section is worth 40% of the marks for the paper.
Section B contains **three** *questions. A candidate's* **TWO** *best answers will be used for assessment. This section is worth 60% of the marks for the paper.*

Calculators of the approved type are permitted in this examination.
Statistical tables will be provided.

Do not turn over until instructed.

A.1 (**8 marks**)

According to a scientific hypothesis, the following nine readings (which have been ordered) should be drawn from a Uniform$(0,1)$ distribution, that is with distribution function

$$F(x) = \begin{cases} 0 & x < 0, \\ x & 0 \le x \le 1, \quad \text{and} \\ 1 & x > 1. \end{cases}$$

0.19	0.33	0.49	0.65	0.72	0.75	0.75	0.82	0.85

(a) Construct a stem-and-leaf plot of these data.

(b) Draw a Q-Q plot of the sample quantiles against the theoretical quantiles of the Uniform$(0,1)$ distribution.

(c) In one sentence, comment on the credibility of the stated hypothesis.

A.2 (**8 marks**)

Let $x_1, x_2, \ldots, x_n$ be the observed values of a simple random sample from the Gamma(α, λ) distribution, with expectation α/λ and variance α/λ^2, where both α and λ are unknown.

(a) Write down two equations relating the method of moments estimates $\hat{\alpha}$ and $\hat{\lambda}$ to the sample moments m_1 and m_2.

(b) Find explicit expressions for $\hat{\alpha}$ and $\hat{\lambda}$ in terms of the sample moments m_1 and m_2.

A.3 (**8 marks**)

For the Poisson distribution with parameter θ, the derivative of the logarithm of the probability mass function is

$$\frac{\partial}{\partial \theta} \log p_X(x; \theta) = -1 + \frac{x}{\theta}.$$

(a) Show that the maximum likelihood estimate of θ given a simple random sample $x_1, x_2, \ldots, x_n$ from the Poisson distribution is $\hat{\theta} = \overline{x} = \sum_1^n x_i/n$.

(b) Write down the maximum likelihood estimate for $p_X(0; \theta) = \exp(-\theta)$, the probability that $X = 0$.

A.4 (**8 marks**)

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Let $\overline{X} = (1/n) \sum_{i=1}^n X_i$ and $S^2 = (1/(n-1)) \sum_{i=1}^n (X_i - \overline{X})^2$.

(a) State the distributions of the random variables $(\overline{X} - \mu)/\sigma$ and $(n-1)S^2/\sigma^2$.

(b) Are these two random variables dependent or independent?

(c) State the distribution of the random variable $(\overline{X} - \mu)/(S/\sqrt{n})$.

Continued over...

A.5 (**8 marks**)

Suppose we are given observations from a specified distribution with unknown parameter θ. Consider a test of the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$. Define the following terms

(a) Type I error and type II error.

(b) The significance level and the power of the test.

Continued over...

B.1 (**30 marks**)

Let $x_1, x_2, \ldots, x_n$ be observed values of a random sample from a Rayleigh distribution, given by the probability density function

$$f(x;\theta) = \begin{cases} \theta x e^{-\theta x^2/2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) (15 marks) Derive the likelihood equation for this model and hence show that the maximum likelihood estimate of θ is given by $\hat{\theta}_{\text{mle}} = 2n/\sum_{i=1}^{n} x_i^2$.

(b) (6 marks) Show that the distribution function for this Rayleigh distribution is given by

$$F(x;\theta) = \begin{cases} 1 - e^{-\theta x^2/2} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

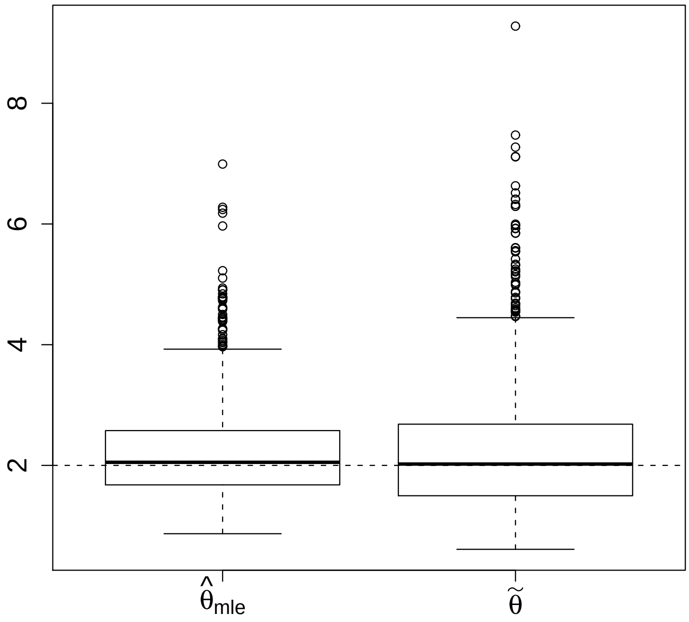
and hence show that the median of the distribution is $\sqrt{(2\log 2)/\theta}$.

(c) (3 marks) An alternative estimate of θ, which we will denote by $\tilde{\theta}$, can be obtained by equating this expression for the median to the sample median of the data, M say, and solving for θ. Show that

$$\tilde{\theta} = \frac{2\log 2}{M^2}$$

.

(d) (6 marks) To compare the performance of the two estimators $\hat{\theta}_{\text{mle}}$ and $\tilde{\theta}$, a simulation experiment was performed. 1000 samples of size $n = 10$ were generated from this Rayleigh distribution, with $\theta = 2$. A boxplot of the 1000 values for each estimate is shown below, with a dashed line added at the correct value of θ. Comment on the relative quality of the two estimators on the basis of this evidence.



Continued over...

B.2 (**30 marks**)

(a) (5 marks) You are given two samples of data and are asked to perform a test of the hypothesis that the means of the populations from which they are drawn are equal, using a test based on the Normal distribution. Explain briefly under what circumstances you would use a paired-comparison test, and when a two-sample test. What assumptions would need to be made to ensure the validity of the test in each case?

(b) (20 marks) Independent random samples of $n = 20$ men and $m = 25$ women from a certain population are given a questionnaire designed to elicit their confidence about their personal financial prospects over the next 12 months. The total scores each subject obtained on the questionnaire, $x_1, x_2, \ldots, x_{20}$ for the men and $y_1, y_2, \ldots, y_{25}$ for the women, can be assumed to be drawn from Normal distributions, and the data are summarised as follows: $\sum_1^{20} x_i = 1219$, $\sum_1^{20} x_i^2 = 74675$, $\sum_1^{25} y_j = 1416$ and $\sum_1^{25} y_j^2 = 80768$.

Assume that the data are random samples from distributions with the same variance σ^2. Compute a pooled estimate S_p^2 for σ^2.

Hence test the hypothesis that there is no difference in mean financial confidence between men and women in this population. State clearly what alternative hypothesis you think is appropriate in this case. Compute the p-value of the test, and interpret the result of the test carefully.

(c) (5 marks) Finally, suppose that in a different study, the questionnaire was given to n married couples, the man and woman answering independently. Briefly (in one sentence), what difference if any would this make to the analysis that would be appropriate, to test for differences in financial confidence between husbands and wives?

Continued over...

B.3 (**30 marks**)

Let $x_1, x_2, \ldots, x_n$ be n given values of a predictor variable, and for each $i = 1, 2, \ldots, n$ assume that the response variables Y_i satisfy the simple Normal linear regression model

$$Y_i = \alpha + \beta x_i + e_i$$

where $e_1, e_2, \ldots, e_n$ are independent random variables each with the $\mathrm{N}(0, \sigma^2)$ distribution. Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimators for α and β, $\hat{\sigma}^2$ the corresponding estimator for σ^2, and let $ss_{xx} = \sum_1^n (x_i - \overline{x})^2$.

(a) (10 marks) State the distribution of $\hat{\beta}$ and the distribution of $(n-2)\hat{\sigma}^2/\sigma^2$. Quoting the relevant standard result, indicate briefly why it follows that $(\hat{\beta} - \beta)/\sqrt{\hat{\sigma}^2/ss_{xx}}$ has the t-distribution with $n - 2$ degrees of freedom.

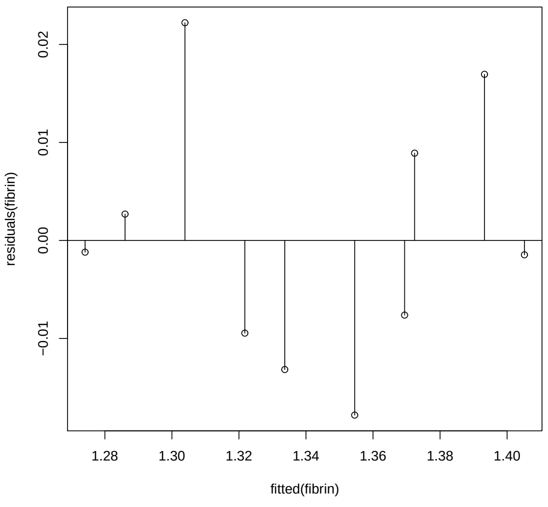
(b) The following data arises from the monitoring of a paient with kidney failure, and shows how the level of fibrin in the blood varies with time after injection of a substance into the bloodstream.

x, Time after injection	8	12	19	20	25
y, log(fibrin level)	1.4037	1.4102	1.3813	1.3618	1.3367
x, Time after injection	32	36	42	48	52
y, log(fibrin level)	1.3205	1.3123	1.3261	1.2887	1.2729

Summary statistics for these data are: $\sum x_i = 294$, $\sum y_i = 13.4142$, $\sum x_i^2 = 10686$, $\sum x_i y_i = 388.2936$ and $\sum y_i^2 = 18.01371$.

i. (15 marks) Assuming a linear regression model is appropriate for these data, calculate ss_{xx}, $\hat{\beta}$ and $\hat{\sigma}^2$ for these data, and hence find the endpoints of a 95% confidence interval for the true value of β, explaining your method briefly but clearly.

ii. (5 marks) The figure below shows a plot of the residuals from a linear regression of y on x, plotted against the fitted values. Comment on the features of this plot, and on whether you think that linear regression is really an adequate model for these data.



End of examination.