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 2008, 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**)

The following data set gives the resting pulse rate of each of nine patients.

68 72 60 96 56 84 59 64 89

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

(ii) Sketch a boxplot of the data, naming the main features and indicating their value.

A.2 (**8 marks**)

On average 50% of guests at a particular hotel ask for 2 rounds of toast for breakfast, 30% of guests ask for 1 round of toast, and 20% of guests do not want any toast. The hotel has 100 guests independently taking breakfast one morning.

(i) Find the mean and variance of the number of rounds of toast required by a randomly chosen guest. Using the central limit theorem, write down the approximate distribution of the total number of rounds of toast required by all 100 guests.

(ii) The hotel has prepared 150 rounds of toast for the morning. Using an appropriate continuity correction, find the probability this will be enough to satisfy all the guests.

A.3 (**8 marks**)

Let U, V and W be independent random variables, where U has the $N(0,1)$ distribution, V has the Chi-square distribution with r degrees of freedom, and W has the Chi-square distribution with s degrees of freedom. You are given that, for $t < 1/2$, V has moment generating functions $\mathcal{M}_V(t) = (1-2t)^{-r/2}$.

(i) State the distribution of $\dfrac{U}{\sqrt{V/r}}$.

(ii) Show that $V + W$ has the Chi-square distribution with $r + s$ degrees of freedom.

A.4 (**8 marks**)

Let $x_1, \ldots, x_{25}$ denote the observed values of a simple random sample of size $n = 25$ from the $N(\mu, \sigma_0^2)$ distribution, with $\sum_{i=1}^{25} x_i = 48$. You are given that $\bar{X} \sim N(\mu, \sigma_0^2/n)$, where μ is unknown but σ_0^2 takes the known value $\sigma_0^2 = 16$.

(i) State the formulae for the upper and lower endpoints of a 90% confidence interval for μ and compute the values of the endpoints based on this sample.

(ii) Say, with brief reasons, how you would expect the length of the confidence interval to differ if: (a) you had taken fewer observations, (b) the required confidence level was smaller.

A.5 (**8 marks**)

Let $X_1, X_2, \ldots, X_n$ be n independent observations from a 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:

(i) Type I error and Type II error.

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

B.1 (**30 marks**)

(a) Let $x_1, \ldots, x_n$ be the observed values of a random sample of size n from the Exponential distribution with unknown parameter $\theta > 0$ and probability density function

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

Derive the likelihood equation for this distribution and hence show that the maximum likelihood estimate of θ is given by $\hat{\theta}_{mle} = n/\sum_{i=1}^{n} x_i$.

(b) Now let $x_1, \ldots, x_n$ be the observed values of a random sample of size n from a double Exponential distribution with unknown parameters α and β and probability density function

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{\beta}{2} e^{-\beta|x-\alpha|} & -\infty < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

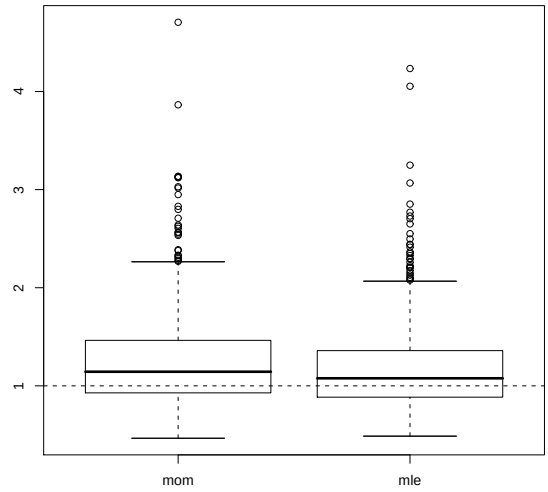
You are given that for this distribution

$$\mathrm{E}(X;\alpha,\beta) = \alpha, \quad \mathrm{E}(X^2;\alpha,\beta) = \alpha^2 + 2/\beta^2.$$

State the equations jointly satisfied by the method of moments estimates $\hat{\alpha}$ and $\hat{\beta}$ and hence find explicit expressions for $\hat{\alpha}$ and $\hat{\beta}$ in terms of the sample moments $m_1 = \sum_{i=1}^{n} x_i/n$ and $m_2 = \sum_{i=1}^{n} x_i^2/n$.

(c) Let $\hat{\beta}_{mom}$ denote the method of moments estimate for β and let $\hat{\beta}_{mle}$ denote the maximum likelikhood estimate for β for the model in part (b) above. To compare the properties of $\hat{\beta}_{mom}$ and $\hat{\beta}_{mle}$, 1000 independent simple random samples, each of size 11, were generated from a double Exponential distribution with $\alpha = 5$ and $\beta = 1$. The estimates $\hat{\beta}_{mom}$ and $\hat{\beta}_{mle}$ were calculated for each sample and boxplots of their values are shown below, with a dotted line added at the true value $\beta = 1$.

Compare the performance of the two estimators, commenting on what the plots indicate about their bias, spread and mean square error.



B.2 (**30 marks**)

A large public sector employer ran a competition, just for its own employees, where successful entrants had their names put forward for promoted positions as vacancies arose in the future. Out of 22 employees that entered the competition, 10 were successful and 12 were unsuccessful. The competition aroused interest because some of the unsuccessful entrants challenged the result in court, alleging that they had been unfairly discriminated against on grounds of age.

Let $x_1, \ldots, x_{12}$ be the observed ages of the unsuccessful entrants and $y_1, \ldots, y_{10}$ be the observed ages of the successful entrants. The summary statistics for the data are:
$\sum_1^{12} x_i = 597$; $\sum_1^{12} x_i^2 = 30315$. $\sum_1^{10} y_i = 451$; $\sum_1^{10} y_i^2 = 20763$.

(a) Assume that the data are random samples from distributions with the same variance, say σ^2. Compute a pooled estimate $\hat{\sigma}^2$ for σ^2.

Hence test at the 5% level the hypothesis that there is no difference between the mean age of unsuccessful and successful entrants, against the alternative that the mean age of unsuccessful entrants is greater than that of unsuccessful entrants. Your answer should contain a brief but clear description of your working at each stage of the test procedure, including details of any model assumptions, the hypotheses being tested, the value of the test statistic, the critical region of the test and a clear summary of your conclusions.

(b) Compute the p-value of your test statistic and say, with reasons, whether or not this value is consistent with your conclusions above.

(c) Now consider the more realistic case where you can not assume that the populations from which the samples were taken have the same variance. State an appropriate test statistic for a test of the hypotheses in (a) above, explaining how this test statistic differs from the one used above (you do not have to state the formula for the degrees of freedom for its distribution).

B.3 (**30 marks**)

The data in the following table was collected from $n = 12$ randomly selected households as part of a pilot study into the likely demand for an additional compostable garden waste collection in that area. For the study week, the households were offered the opportunity to put out compostable garden waste as a separate addition to their normal household waste. In each data pair, x_i denotes the weight of household waste (in kg) collected from the ith household and y_i denotes the weight of compostable waste (in kg).

You may assume each y_i is the observed value of a response variable Y_i satisfying the standard simple Normal linear regression model $Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n$, where $e_1, \ldots, e_n$ are independent $N(0, \sigma^2)$ random variables.

Household (x)	6.9	8.6	13.9	16.3	17.7	18.1	20.0	22.5	23.7	23.9	25.2	28.9
Compostable (y)	3.0	2.5	4.8	4.1	3.6	5.8	6.8	6.2	7.6	5.1	8.2	7.8

Summary statistics for the data set are:
$\sum_1^{12} x_i = 225.7$; $\sum_1^{12} y_i = 65.5$; $\sum_1^{12} x_i^2 = 4730.77$; $\sum_1^{12} y_i^2 = 398.23$; $\sum_1^{12} y_i x_i = 1354.02$

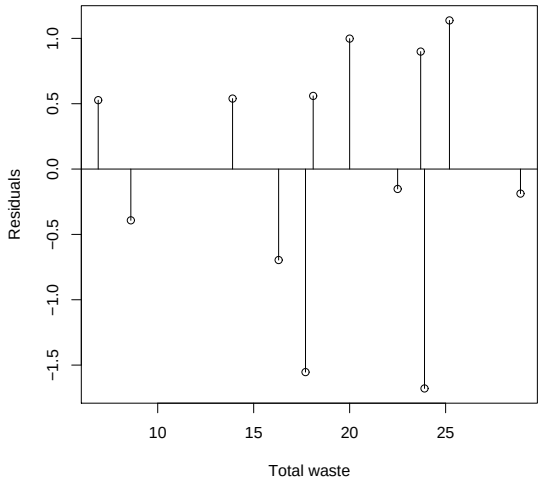
(a) Calculate the least squares estimates of α and β and the corresponding estimate of σ^2 for this dataset, in each case displaying your working briefly but clearly.

Currently a typical household in that area is known to put out 22kg of household waste per week. Based on your least squares estimates, estimate the amount of compostable waste such a household is likely to put out each week.

(b) The diagram below plots the residual values $\hat{e}_i$ against the predictor values x_i for the model above, which assumes that the e_i are independent, have mean zero and common variance σ^2 and are normally distributed. Say briefly how a residual plot like this may be used to assess the fit and the validity of the model, and comment on the model assumptions for this particular dataset.

(c) Calculate the end points of a 90% confidence interval for the slope of the regression line.

[Hint: You are given that $(\hat{\beta} - \beta)/\sqrt{\hat{\sigma}^2/ss_{xx}} \sim t_{n-2}$, where $\hat{\beta}$ and $\hat{\sigma}^2$ are the respective estimators of β and σ^2, and $ss_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.]



End of examination.