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 2009, 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.

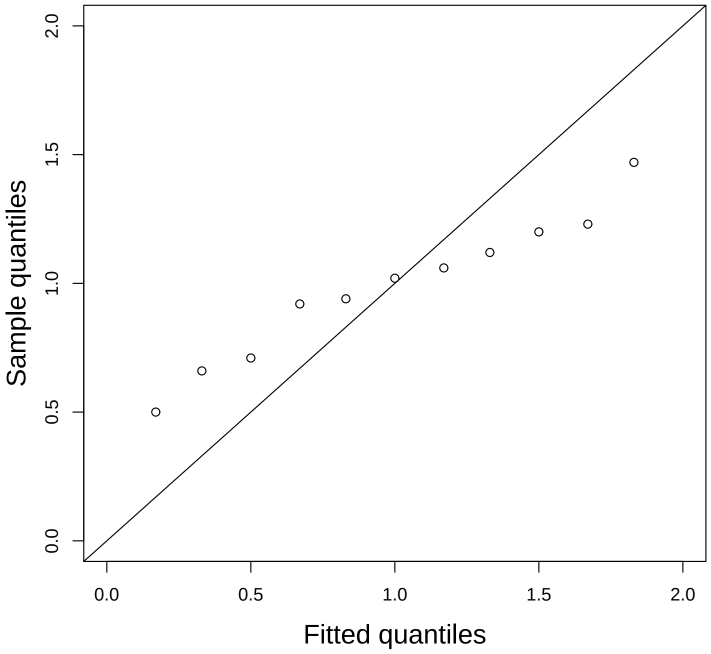
A.1 (**8 marks**)

The following (ordered) data is thought to come from a distribution with distribution function $F_X(x) = x/2$ for $0 < x < 2$ (and $F_X(x) = 0$ for $x \leq 0$ and $F_X(x) = 1$ for $x \geq 2$).

0.50	0.66	0.71	0.92	0.94	1.02	1.06	1.12	1.20	1.23	1.47

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

(ii) Consider the Q-Q plot of the sample quantiles against the fitted quantiles of F_X shown below. Compute the values of the coordinates of the point corresponding to the smallest observation, clearly indicating your method. Say in one or two sentences what the plot may indicate about the difference between the tails of the fitted distribution and the tails of the actual distribution from which the sample was drawn.



A.2 (**8 marks**)

Let $x_1, x_2, \ldots, x_n$ be the observed values of a random sample of size n from the Normal $N(\mu, \sigma^2)$ distribution and let $\hat{\mu}$ and $\hat{\sigma}^2$ be the corresponding method of moments estimates.

(i) Write down two equations which jointly define $\hat{\mu}$ and $\hat{\sigma}^2$ for this distribution.

(ii) Hence find expressions for $\hat{\mu}$ and $\hat{\sigma}^2$ in terms of $x_1, x_2, \ldots, x_n$.

A.3 (**8 marks**)

A supermarket issues scratch-cards to customers with every purchase. Each card is chosen at random from a large stock in which 60% have value 5 points, 30% have value 10 points, and 10% have value 20 points.

(i) Let X denote the points value of a single card. Compute $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

(ii) Let T denote the total points value of 50 cards collected by a given customer. Using the central limit theorem, write down the approximate distribution of T.

(iii) A second customer has also collected 50 cards. Let U denote their total points value. Using a similar approximation for the distribution of U, compute $P(U - T > 20)$.

[No marks will be lost for not using a continuity correction.]

A.4 (**8 marks**)

(i) Assume U and V are independent random variables where U has the Normal $N(0, 1)$ distribution and V has the Chi-square distribution with r degrees of freedom.

State the distribution of $U/\sqrt{V/r}$.

(ii) Let $X_1, \ldots, X_n$ be a random sample of size n from the Normal distribution with mean μ and variance σ^2. Let $\bar{X} = \sum_{i=1}^{n} X_i/n$ and let $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$.

Outline briefly how the result in (i) can be used to identify the distribution of $\sqrt{n}(\bar{X} - \mu)/S$, stating clearly any results you need concerning the distribution of $\bar{X}$ and S^2.

A.5 (**8 marks**)

Let $X_1, \ldots, X_n$ be a random sample of size n from the Normal $N(\mu, \sigma^2)$ distribution and let $\bar{X} = \sum_{i=1}^{n} X_i/n$. For each i, let x_i be the observed value of X_i and let $\bar{x} = \sum_{i=1}^{n} x_i/n$.

(i) Assume σ has known value σ_0. State the distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma_0$. Write down expressions in terms of $\bar{x}, \sigma_0, \alpha$ and n for the end points of a $100(1 - \alpha)\%$ confidence interval for μ. Compute the end points when $\bar{x} = 5, \sigma_0 = 3, \alpha = 0.10, n = 10$.

(ii) Say, with brief reasons, how the length of this confidence interval might compare with that of the appropriate confidence interval constructed in each of the following three separate cases: (a) if you had taken more observations, (b) if the required confidence level was greater, (c) if the value of σ^2 was unknown.

B.1 (**30 marks**)

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

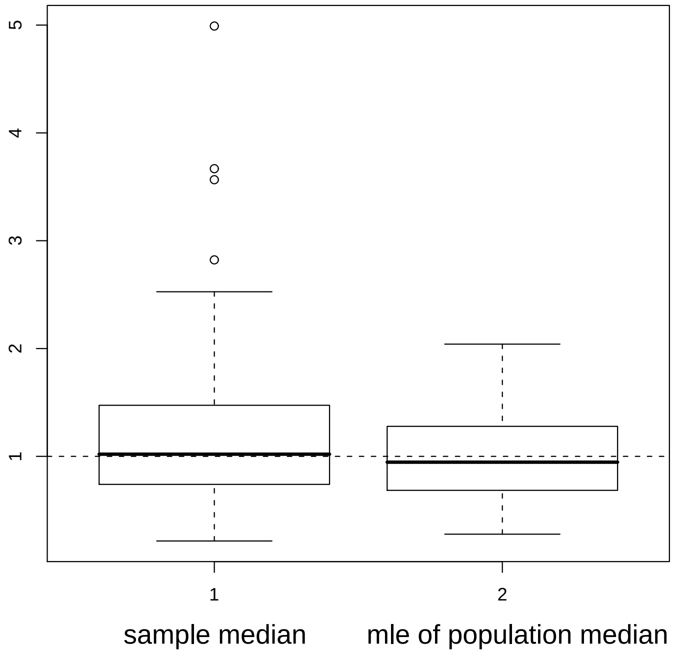
$$f(x;\theta) = \begin{cases} \dfrac{\theta}{(1+x)^{\theta+1}} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

(a) (**10 marks**)
Derive the likelihood equation for this distribution and hence show that the maximum likelihood estimate of θ is given by $\hat{\theta}_{\text{mle}} = n / \sum_{i=1}^{n} \log(1 + x_i)$.

(b) (**8 marks**)
Define the bias and mean square error of an estimator $\hat{\theta}$ of θ. Write down an expression for the mean square error of the estimator in terms of its bias and variance. Use the fact that $\mathrm{E}(\hat{\theta}_{\text{mle}}) = n\theta/(n-1)$ to compute the bias of $\hat{\theta}_{\text{mle}}$ as an estimator of θ.

(c) (**6 marks**)
You are given that the population median τ for this distribution satisfies the equation $(1+\tau)^{\theta} = 2$. Let $\hat{\tau}_{\text{mle}}$ denote the maximum likelihood estimate for τ. Find an expression for $\hat{\tau}_{\text{mle}}$ in terms of the observed values $x_1, \ldots, x_n$.

(d) (**6 marks**)
The simplest estimate of the population median is just the sample median m. Estimates m and $\hat{\tau}_{\text{mle}}$ were calculated for each of 1000 independent simple random samples, each of size 10, generated from a Pareto distribution for which $\tau = 1$. Boxplots of the values for each estimator are shown below, with a dashed line added at the true value $\tau = 1$. On the basis of these plots, compare the performance of m and $\hat{\tau}_{\text{mle}}$ in terms of their bias and mean square error as estimators of τ.



B.2 (**30 marks**)

(a) (**8 marks**)
In the context of a testing a simple null hypothesis, say $H_0 : \mu = \mu_0$, against a simple alternative hypothesis, say $H_1 : \mu = \mu_1$, define the following terms: (i) Type I error, (ii) Type II error, (iii) the significance level of the test, (iv) the power of the test.

(b) (**15 marks**)
As part of a strategy to help students reduce their body fat, a college randomly selected ten students to attend a trial course of weekly lunchtime exercise classes. The table below shows the percentage of body fat for each student, measured before (x_i) and after (y_i) the course, together with the individual differences ($d_i = x_i - y_i$).

Before (x_i)	26.3	19.6	31.7	18.9	22.9	27.5	27.4	30.1	31.8	21.0
After (y_i)	25.0	17.3	26.7	17.7	23.8	23.2	25.7	27.6	32.9	20.2
Difference ($x_i - y_i$)	1.3	2.3	5.0	1.2	-0.9	4.3	1.7	2.5	-1.1	0.8

The summary statistics for the differences are: $\sum_{i=1}^{10} d_i = 17.1$, $\sum_{i=1}^{10} d_i^2 = 63.71$.

Test at the 5% level the hypothesis that the attending the course does not affect the percentage of body fat against the alternative that it reduces the percentage of body fat. 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, a formal statement of the hypotheses being tested, the value of the test statistic, the p-value of your test statistic and a clear summary of your conclusions.

(c) (**7 marks**)
Now assume the ten differences are observed values of a random sample $D_1, \ldots, D_{10}$ from the Normal distribution with unknown mean μ and known variance $\sigma_0^2 = 4$. You are given that a test of $H_0 : \mu = 0$ against $H_1 : \mu = 1$ at the 5% level has test statistic $T = \sqrt{10}\bar{D}/2$ and critical region $C = \{T > 1.645\}$, where $\bar{D} = \sum_{i=1}^{10} D_i/10$.
Find the power of this test at $\mu = 1$.

B.3 (**30 marks**)

In a response-time study, an ambulance crew responded to simulated emergency calls from a representative sample of ten local addresses. For each address, let x_i denote the time (in minutes) from the call being made to arrival at the address and let y_i denote the address distance (in miles) from the ambulance base. Summary statistics for the data are: $\sum_{i=1}^{10} x_i = 30$; $\sum_{i=1}^{10} y_i = 73.62$; $\sum_{i=1}^{10} x_i^2 = 110$; $\sum_{i=1}^{10} y_i^2 = 599.8738$; $\sum_{i=1}^{10} y_i x_i = 254.3$

Assume the response-time data satisfies the standard Normal linear regression model under which $x_1, \ldots, x_n$ are given values of a predictor variable and each y_i is the observed value of a response variable $Y_i = \alpha + \beta x_i + e_i$, where $e_1, \ldots, e_n$ are independent random variables each with the $N(0, \sigma^2)$ distribution. Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimator for α and β and let $\hat{\sigma}^2$ be the corresponding estimator for σ^2.

(a) (**10 marks**)
Derive the two *normal equations* satisfied by $\hat{\alpha}$ and $\hat{\beta}$. Hence show $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and derive an expression for $\hat{\beta}$ in terms of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.

Calculate $\hat{\alpha}$ and $\hat{\beta}$ for the response-time data, and hence estimate the time the ambulance would take to reach an address 2.5 miles from the base.

(b) (**10 marks**)
You are given that $(\hat{\alpha} - \alpha)/s_{\hat{\alpha}}$ has the t-distribution with $n - 2$ degrees of freedom. Compute the endpoints of a 95% confidence interval for α for the response-time dataset, explaining your method briefly but clearly.

Note: to obtain $s_{\hat{\alpha}}$, you may use the following output taken from the **R** summary of a linear regression analysis of the dataset, but with the least squares estimates removed. Alternatively, you may use the formula $s_{\hat{\alpha}}^2 = \hat{\sigma}^2(1/n + \bar{x}^2/(\sum_{j=1}^n x_j^2 - n\bar{x}^2))$.

```
Coefficients:
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept)     --        0.3682     6.372    0.000215 ***
x               --        0.1110    15.062    3.73e-07 ***

Residual standard error: 0.4964 on 8 degrees of freedom
Multiple R-Squared: 0.9659,     Adjusted R-squared: 0.9617
F-statistic: 226.9 on 1 and 8 DF,  p-value: 3.732e-07
```

(c) (**10 marks**)
For fixed x, you are given that $\hat{\alpha} + \hat{\beta}x$ can be expressed in the form

$$\hat{\alpha} + \hat{\beta}x = \sum_{i=1}^{n} Y_i(1/n + b_i(x - \bar{x})) \quad \text{where} \quad b_i = (x_i - \bar{x})/\sum_{j=1}^{n}(x_j - \bar{x})^2.$$

Show that $\sum_{i=1}^n b_i = 0$ and $\sum_{i=1}^n b_i^2 = 1/\sum_{j=1}^n (x_j - \bar{x})^2$. Hence derive an expression for $\mathrm{Var}(\hat{\alpha} + \hat{\beta}x)$ in terms of n, σ^2, x and $x_1, \ldots, x_n$.

End of examination.