This paper is not to be removed from the Examination Halls

UNIVERSITY OF LONDON

279 004b ZA
990 004b ZA
996 D04b ZA

BSc degrees and Diplomas for Graduates in Economics, Management, Finance and the Social Sciences, the Diploma in Economics and Access Route for Students in the External Programme

Statistics 2 (half unit)

Tuesday, 10 May 2005 : 2.30pm to 4.30pm

Candidates should answer **THREE** of the following **FIVE** questions: **QUESTION 1** of Section A (40 marks) and **TWO** questions from Section B (30 marks each). **Candidates are strongly advised to divide their time accordingly.**

A list of formulae is given at the end of the paper.

Graph paper is provided. If used, it must be securely fastened inside the answer book.

New Cambridge Statistical Tables (second edition) are provided.

A hand held non-programmable calculator may be used when answering questions on this paper. The make and type of machine must be stated clearly on the front cover of the answer book.

PLEASE TURN OVER

SECTION A

Answer all **seven** parts of question 1 (40 marks).

1. (a) For each of i to iv below, say whether the statement is true or false and briefly give your reasons:

   i. If X is a random variable with a normal distribution, then $P(X = 0) = 0$.

   ii. If, in a regression model, the variance of the measurement error ε is zero, then the Coefficient of Determination R^2 is 100%.

   iii. If we have a random sample from a normal distribution with unknown mean μ and known variance σ^2, then $(-\infty, \bar{x} + z_{0.05}\sigma/\sqrt{n})$ is a 95% confidence interval for μ.

   iv. If the bias of the estimator T of θ is smaller than the bias of the estimator S of θ, then T is a better estimator than S.

(8 marks)

(b) Write brief notes on each of the topics below. Explain the rôle of each in statistics.

   i. Sample space.

   ii. Row effects.

   iii. Diagnostic plots.

   iv. Type I error.

(6 marks)

(c) Out of a random sample of 381 births to mothers aged twenty-two, 200 were male children. Out of a random sample of 444 births to mothers aged twenty-three there were 227 male children.

   i. Do the observed proportions of male children differ with the age of the mother because of randomness, or is there sufficient evidence to show that the population proportions differ?

   ii. Give a 96.8% Confidence Interval for the difference in proportions.

(5 marks)

(d) X is a random variable with $E(X^2) = 3.6$ and $P(X = 2) = 0.6$ and $P(X = 3) = 0.1$. The random variable X takes just one other value besides 2 and 3.

   i. What is the other value that X takes?

   ii. What is the variance of X?

(4 marks)

(e) i. What is the probability that there are fewer than 10 successes for the binomial distribution with 15 trials and probability of success 0.69?

(question continues on next page)

ii. What is the probability that $X > 15$ when X has a Poisson distribution with mean 18.8?

(4 marks)

(f) There are six tickets, three of which are numbered 1,2,3 and the other three are labelled 0. If three tickets are drawn at random without replacement, what is the probability of drawing a total of three for the labels? **(5 marks)**

(g) X is thought to be a Binomial random variable with number of trials $n = 3$, and probability of success π, where π is unknown. The results of 50 observations of X are tabulated below:

$X = 0$	$X = 1$	$X = 2$	$X = 3$
6	14	23	7

Estimate π, and test the hypothesis that X has a binomial distribution with $n = 3$. **(8 marks)**

SECTION B

Answer **two** questions from this section (30 marks each).

2. (a) i. State and prove Bayes' Theorem. **(4 marks)**

ii. First year students in statistics take a test of mathematical aptitude. Students taking the test are either well-prepared, or less well-prepared. Of the well-prepared students 95% will pass the test, whereas of the less well-prepared only 10% will pass. The pass rate for the test is 75%.
What is the proportion of students who are well-prepared?
Given that a student has passed the test, what is the probability that the student is not well-prepared?

(10 marks)

(b) X is a random variable with density function

$$f_X(x) = a + bx$$

over the range $(0, 1)$, where a, b are constants. The mean of X is 0.5.

i. Find a, b.
ii. Find the cumulative distribution function of X.
iii. Find the variance of X.

(16 marks)

3. (a) What is a minimum variance unbiased estimator? **(4 marks)**

(b) Suppose that for a random sample of size n an estimator W is unbiased for the population variance σ^2, and that it has has variance $2\sigma^4/(n-1)$. Find for what value of the constant a the estimator of σ^2 of the form $T = aW$ has minimum mean squared error. **(8 marks)**

(c) The table below shows heights in cm of small Thuja orientalis trees that have been subject to two different growing regimes.

Regime	
1	2
51.1	56.4
57.4	53.6
58.5	54.9
53.9	60.5
56.1	50.9
50.2	54.9
53.9	58.3
50.7	57.9
57.5	57.1
59.6	

(question continues on next page)

i. Find a 94% confidence interval for the difference between mean heights of trees grown under the two regimes. **(13 marks)**

ii. Test at the 4% level of significance the null hypothesis that the mean heights under Regime 2 are 4cm greater than under Regime 1. **(5 marks)**

4. (a) Derive using algebra the sum of squares identity for a two-way analysis of variance, and describe the meaning of the terms in the identity. **(4 marks)**

(b) Describe the statistical model corresponding to a two-way analysis of variance, giving all the assumptions. **(5 marks)**

(c) The table below shows the logarithms of forecasts of the exchange rates per US$ for the local currencies in four countries over several years.

	Years		
Countries	2004	2005	2006
Albania	2.02	2.01	2.02
Algeria	1.86	1.84	1.86
Angola	1.92	1.94	2.00
Argentina	0.46	0.48	0.46

i. Give a two-way analysis of variance table for these data. **(11 marks)**

ii. Is there significant evidence of differences in log forecast exchange rates for the different years? **(5 marks)**

iii. Give a 95% simultaneous confidence intervals for the differences between the log forecast exchange rates for different years. **(5 marks)**

5. (a) We are given three pairs of observations(x_i, y_i) for $i = 1,2,3$, where $x_1 = 2, x_2 = 1, x_3 = 1$. Suppose that

$$\begin{aligned} y_1 &= x_1\beta + \varepsilon_1 \\ y_2 &= \alpha + x_2\beta + \varepsilon_2 \\ y_3 &= 2\alpha + x_3\beta + \varepsilon_3 \end{aligned}$$

where α, β are unknown parameters, and the ε_i are random variables with mean 0 and variance σ^2. Find the least squares estimators of α, β. Show that the least squares estimator of β is unbiased. **(8 marks)**

(b) The table below shows the yields in hkg/ha of Soft Wheat and Spelt, of Rye and Maslin and Barley over several years in the Czech republic.

PLEASE TURN OVER

(question continues on next page)

Year	Soft Wheat and Spelt	Rye and Maslin	Barley
1996	46.5	31.9	37.5
1997	44.1	34.3	38.4
1998	42.1	36.3	36.2
1999	46.5	36.7	39.4
2000	42.1	34.2	32.9
2001	48.5	37.2	39.7
2002	45.6	33.8	36.7
2003	40.7	38.0	37.6
2004	58.8	53.0	50.5

i. Fit a straight line to these data using yield of Soft Wheat and Spelt as the response variable and yield of Rye and Maslin as the explanatory variable. **(7 marks)**

ii. Give a 95% confidence interval for the mean response when the yield of Rye and Maslin is 39.0 hkg/ha. **(9 marks)**

iii. The fitted regression model with both Rye and Maslin, and yield of Barley included as explanatory variables is

$$\text{soft wheat} = 4.96 - 0.154\ \text{rye} + 1.21\ \text{barley}\ .$$

Interpret this model carefully, comparing the insight it gives you with that from the simpler model which just uses Rye and Maslin. **(6 marks)**

Formulae for Statistics

Discrete Distributions

Binomial Distribution

The probability of x successes in n trials is

$$\binom{n}{x} \pi^x (1-\pi)^{n-x}$$

for $x = 0, 1, \ldots, n$ The mean number of successes is $n\pi$ and the variance is $n\pi(1-\pi)$.

Poisson Distribution

The probability of x is

$$e^{-\mu}\frac{\mu^x}{x!}.$$

The mean number of successes is μ and the variance is μ.

Hypergeometric Distribution

The probability of x successes in a sample of size n from a population of size N with M successes is

$$\binom{M}{x}\binom{N-M}{n-x} \Big/ \binom{N}{n}.$$

The mean number of successes is nM/N and the variance is $n(M/N)(1-M/N)(N-n)/(N-1)$.

Sample Quantities

Sample Variance	$s^2 = \sum(x_i - \bar{x})^2/(n-1) = (\sum x_i^2 - n\bar{x}^2)/(n-1)$
Sample Covariance	$\sum(x_i - \bar{x})(y_i - \bar{y})/(n-1) = (\sum x_i y_i - n\bar{x}\bar{y})/(n-1)$
Sample Correlation	$(\sum x_i y_i - n\bar{x}\bar{y})/\sqrt{(\sum y_i^2 - n\bar{y}^2)(\sum x_i^2 - n\bar{x}^2)}$

Inference

Variance of Sample Mean σ^2/n

One-sample t statistic $\sqrt{n}(\bar{x}-\mu)/s$ with $(n-1)$ degrees of freedom

Two-sample t statistic
$$\frac{(\bar{x_1}-\bar{x_2})-(\mu_1-\mu_2)}{\sqrt{[1/n_1+1/n_2]\{[(n_1-1)s_1^2+(n_2-1)s_2^2]/(n_1+n_2-2)\}}}$$

Variances for differences of binomial proportions

Pooled
$$\left[\frac{(n_1p_1+n_2p_2)}{(n_1+n_2)}\right]\left[1-\frac{(n_1p_1+n_2p_2)}{(n_1+n_2)}\right]\left[\frac{1}{n_1}+\frac{1}{n_2}\right]$$

Separate
$$p_1(1-p_1)/n_1+p_2(1-p_2)/n_2$$

Least Squares

Estimates for $y=\alpha+\beta x$ fitted to (y_i,x_i) for $i=1,2,\ldots,n$ are $a=\bar{y}-b\bar{x}$ and
$$b=\sum(x_i-\bar{x})(y_i-\bar{y})/\sum(x_i-\bar{x})^2.$$
The estimate of variance is
$$[\sum(y_i-\bar{y})^2-b^2\sum(x_i-\bar{x})^2]/(n-2).$$
The variance of b is $\sigma^2/\sum(x_i-\bar{x})^2$

Chi-squared Statistic $\sum(\text{Observed}-\text{Expected})^2/\text{Expected}$, with degrees of freedom depending on the hypothesis tested.

END OF PAPER