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 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)

Monday, 10 May 2004 : 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 of the answer book.

PLEASE TURN OVER

SECTION A

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

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

   i. If A and B are independent events, then the complementary events A^c and B^c are also independent events.

   ii. If, in a population of men of who are all 55 years old, there is a correlation coefficient of 0.9 between daily consumption of alcohol and expenditure on medical products, that must be because the alcohol causes illness.

   iii. If $(2.1, 3.2)$ is a 95% confidence interval for the standard deviation σ, then $(4.41, 10.24)$ is a 95% confidence interval for σ^2.

   iv. If T is an unbiased estimator of θ, then T^2 is not an unbiased estimator of θ^2.

   (8 marks)

   (b) Write brief notes on each of the topics below. Explain in what way each topic is part of statistics, and why it is important.

   i. Collinearity.

   ii. Uniform random variable.

   iii. Sample space.

   iv. Combinations.

   (6 marks)

   (c) On a given day out of 300 men visiting a public library, 219 borrowed or returned books, and that of 500 women, 430 borrowed or returned books.

   i. Do the observed proportions of men and women returning or borrowing books differ because of randomness, or is there sufficient evidence to show that those proportions differ?

   ii. Give a 95% Confidence Interval for the difference in proportions.

   (5 marks)

   (d) X is a random variable with expected value 2 and $P(X = 1) = 0.7$ and $P(X = 4) = 0.2$. X takes just one other value besides 1 and 4.

   i. What is the other value that X takes?

   ii. What is the variance of X?

   (4 marks)

(question continues on next page)

(e) i. *From your tables* find the probability that there are no more than 9 successes for the binomial distribution with 19 trials and probability of success 0.47.

ii. *From your tables* find the probability that $X \geq 11$ when X has a Poisson distribution with mean 5.55.

(4 marks)

(f) Two students, C and D consider applying for a bursary, that may or may not be awarded, depending on their exam results. C makes an application, and will get the bursary with probability 3/4, provided D does not apply. D will apply with probability p, and if D does apply, the probability that C will get the bursary is 1/3.

i. What is the probability (in terms of p) that C will get the bursary?

ii. If the probability that D does not apply given that C is awarded the bursary is 3/7, what is the probability p?

(5 marks)

(g) 40 observations are thought to be independent realisations x of a Poisson random variable X with mean $\mu = 7$. They are tabulated below:

$x \leq 6$	$6 < x \leq 10$	$10 < x$
16	20	4

Test the hypothesis that the observations on X are from a Poisson distribution with mean $\mu = 7$.

(8 marks)

SECTION B

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

2. (a) A box contains 5 green balls and 5 red balls. A ball is chosen at random by a student from the box. If it is a green ball then the student writes down the answer 'green'. If the ball chosen is red, then the student writes down at random one of either 'red' or 'green'. Suppose the student has written down 'green'. What is the probability that a red ball was chosen? Suppose the first ball chosen is replaced, and another student acts in the same way as the first student. What is the probability that at least one student chose a red ball if both students write down 'green'? **(12 marks)**

   (b) X is a random variable with density function

$$f_X(x) = 3(1+x)^2/19$$

   over the range $(1,2)$. Using calculus, find the mean and the expected value of $1/(1+X)$. What is $P(X > 1.5)$?

   (18 marks)

3. (a) Why is the sample variance s^2 said to be unbiased for the population variance σ^2? Should one always use s^2 to estimate σ^2?

   (5 marks)

   (b) Suppose that X is a random observation from an exponential distribution with mean $1/\lambda$, and that one wants to estimate $\theta = P(X > 2) = e^{-2\lambda}$. Find the mean squared error of the estimator T of θ, where

$$T = \begin{cases} 1 & X > 2 \\ 0 & X < 2 \end{cases}$$

   (12 marks)

   (c) The table below shows infant mortality per 100 live births for several countries in both 1935/39 and 1945/49.

	Years	
Country	1935/39	1945/49
Hungary	135	115
N. Ireland	77	53
Italy	103	84
Netherlands	38	42
Norway	40	33
Poland	136	109
Romania	181	166

(question continues on next page)

i. Find a 90% confidence interval for the difference between infant mortality for 1935/39 and 1945/49.

(9 marks)

ii. Test at the 5% level of significance the null hypothesis that 1945/49 has infant mortality exactly 20 lower than 1935/39.

(4 marks)

4. (a) Derive the sum of squares identity for a one-way analysis of variance, and describe the meaning of the terms in the identity.

(8 marks)

(b) The table below shows four cells in a two-way table. Each cell in the table contains the population mean for the cell. For each cell we now draw one random observation from a normal distribution with the mean given for the cell and variance 1. Are the four observations we will get, one for each of the four cells, suitable for a two-way analysis of variance?

2	3
4	6

(6 marks)

(c) The table below shows the average daily calorie intake of food as a percentage of basic needs in three groups of South American countries.

Group 1	Group 2	Group 3
128	94	101
87	92	123
108	94	93
113	98	108
108	84	107
118	96	
121	124	
	97	

i. Give a two-way analysis of variance table for these data.

(9 marks)

ii. Is there significant evidence of differences in life expectancy for the different groups of countries?

(4 marks)

iii. Give a 95% confidence interval for the difference between the population means for life expectancy for Groups 2 and 3.

(3 marks)

5. (a) i. Explain without using any algebra why the sum of the n squared residuals from a simple linear regression divided by $n-2$ might be expected to estimate the measurement error variance.

(6 marks)

ii. With the usual assumptions, show that the least squares estimator of the slope parameter in a simple linear regression is unbiased.

(6 marks)

(b) The table below shows the Average Weekly Wage in dollars, the number of Machines per Worker, and the annual Growth Rate of capital in cotton mills in several places in the period around 1910.

Country	Average Weekly Wage	Machines per worker	Growth Rate
New England	8.8	2.97	1.6
Canada	8.8	2.53	2.4
US (South)	6.5	2.65	9.4
Lancashire	5.0	2.04	0.9
Germany	3.8	1.28	3.1
France	3.7	1.11	1.7
Switzerland	3.7	1.40	-0.4

i. Fit a straight line to these data using Average Weekly Wage as the response variable and Machines per Worker as the explanatory variable.

(6 marks)

ii. Give an 90% prediction interval for a new observation on Average Weekly Wage for Machines per Worker of 2.30 of the regression.

(6 marks)

iii. The fitted regression model with both Machines per Worker, and Growth Rate included is

$$\text{Average Weekly Wage} = -0.04 + 3.06\text{Machines per Worker} - 0.119\text{Growth Rate}.$$

Interpret this model carefully, comparing the insight it gives you with that from the simpler model which just uses Machines per Worker.

(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 $$\binom{M}{x}\binom{N-M}{n-x}/\binom{N}{n}.$$ 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	$\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{[1/n_1+1/n_2]\{[(n_1-1)s_1^2+(n_2-1)s_2^2]/(n_1+n_2-2)\}}}$
Variances for differences of binomial proportions	
Pooled	$\left[\dfrac{(n_1p_1+n_2p_2)}{(n_1+n_2)}\right]\left[1-\dfrac{(n_1p_1+n_2p_2)}{(n_1+n_2)}\right]\left[\dfrac{1}{n_1}+\dfrac{1}{n_2}\right]$
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