Examination papers and Examiners' reports

Statistics 2

2790**04b**, 9900**04b**, 996D**04b**

2003, 2004

Undergraduate study in
Economics, Management, Finance and the Social Sciences

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UNIVERSITY OF LONDON 18 58 EXTERNAL PROGRAMME

Examiner's report 2004

Zone A

General Remarks

This is the second year for Statistics 2. The pattern of the paper was the same as that for 2003. Candidates could make life easier for examiners by writing on the right hand pages only, and by starting each question on a new page. Intertwining parts of different questions should be avoided if possible.

Section A, for 40% of the marks, is a compulsory question with several parts. It is designed to test general knowledge and understanding of the whole syllabus.Here you should give reasoned answers, with some explanation, avoiding one word responses, which will never be given any marks.

Section B has four questions of which two, and **no more than** two, should be presented for marking. They are meant to test a greater depth of knowledge on parts of the syllabus. You may here use your calculator as much as you wish, but usually few marks will be awarded for a part of a question that can be done easily with a built-in calculator routine.

It is hard to emphasise enough that memorising answers to past questions is not the best way to study for this paper. It is important to understand the material you write down, and to be able to develop it all from scratch as you write it. There are often several different ways to obtain good marks for a question. The examiners are not looking for 'the' correct answer because there are several correct answers.

Candidates can make a more pleasing impression on examiners by avoiding probabilities outside the range 0 to 1, and negative sums of squares. In fact, if you notice that any of your answers is wrong, it's best to note that on your script for the examiner to read.

Specific comments on questions

Question 1

(a)
It is important to give reasons here, not just the answer 'true' or 'false'. Quite often, the judgment as to whether a statement is true or false depends on how it is interpreted, so the examiner must see reasons to assess the worth of an answer. In i, if the events A, B are independent, so are the complementary events A^c and B^c. This can be shown by working out $P(A^c \cap B^c)$ assuming that A and B are independent events.

For ii one must remember that correlation and causation are distinct ideas. Correlation between alcohol consumption and expenditure on medical products might be entirely due, for instance, to variation in the wealth of individuals. Richer people might drink more alcohol and spend more on medicine.

The answer in iii depends on what you assume the question means. If (T_l, T_u) cover the parameter σ with probability 0.95, both T_l and T_u being greater than 0, then it is clear that (T_l^2, T_u^2) will cover the parameter σ^2 with probability 0.95. So the latter will be a valid 95% confidence interval for σ^2. The important thing here is to mention the coverage probabilities.

For iv note that if T is unbiased for θ, then $ET = \theta$, whereas for T^2 unbiased for θ^2 we must have $E(T^2) = \theta^2$. Since $E(T^2) = \mathrm{var}(T) + [E(T)]^2$, we know that T^2 is not unbiased for θ^2 unless $\mathrm{var}(T) = 0$. This is not likely to be the case in any practical statistical estimator T, for it implies that we can estimate θ without error.

(b)
Only four or five lines for each topic are necessary here. Collinearity is a topic in regression.

(c)
This is a test and confidence interval for a difference between two proportions. One uses results for the binomial distribution, or for part i a chi-squared test.

(d)
To answer this question one needs only to know that probabilities sum to 1, and how to calculate the mean and variance of a discrete random variable.

(e)
For this one needs familiarity with the statistical tables used in the examination room. It is best to use them when revising for the exam. Typical calculators in use by candidates this year can't give a correct answer as easily as the tables, which the question requires you to use. 'No more than 9 successes' means '9 successes or fewer.' The probability that $X \geq 11$ is $1 - P(X \leq 10)$.

(f)
It is easy to use the formula for total probability to obtain

$$P(C) = 3/4 - 5p/12.$$

Then

$$3/7 = P(D^c|C) = P(D^c \cap C)/P(C) = [3/4 \times (1-p)]/[3/4 - 5p/12]$$

Solving for p gives $p = 3/4$.

(g)
This is a goodness-of-fit test. The expected values are found by multiplying by 40 the probabilities obtained from the Poisson tables for $\mu = 7$. Remember to state the null and alternative hypotheses, the significance level used, the critical value and the outcome of the test.

Question 2

(a)
This uses Bayes' Theorem. The probability that the student writes down 'green' is

$$\frac{1}{2} \times \frac{5}{10} + \frac{5}{10} = \frac{3}{4}.$$

So, the chance that the ball is red given that the student writes down 'red' is

$$\frac{\frac{1}{2} \times \frac{5}{10}}{\frac{3}{4}} = \frac{1}{3}.$$

When there are 2 students, the probability that both write down 'green' is $\left(\frac{3}{4}\right)^2 = \frac{9}{16}$. So, the chance that at least one student chose red is

$$\frac{\frac{1}{4}^2 + 2\frac{1}{4}\frac{1}{2}}{\frac{9}{16}} = \frac{5}{9}.$$

(b)
This question requires the use of the formulae

$$EX = \int_1^2 x3(1+x)^2 dx/19,$$

The expected value of $1/(1+X)$ is

$$\int_1^2 3(1+x)dx/19.$$

$P(X > 1.5)$ is

$$\int_{1.5}^2 3(1+x)^2/19dx.$$

Question 3

(a)
The sample variance is unbiased for σ^2 because the expected value of s^2 over repeated independent random samples from a population with variance σ^2 is σ^2. If one chooses to estimate σ^2 with minimum mean squared error, then it is better to use a divisor n or $n+1$ rather than the $(n-1)$used for s^2.

(b)
The estimator T has mean

$$E(T) = 0 \times P(X < 2) + 1 \times P(X > 2) = P(X > 2) = e^{-2\lambda} = \theta.$$

So T is unbiased for θ. Also

$$E(T^2) = 0^2 \times P(X < 2) + 1^2 \times P(X > 2) = P(X > 2) = e^{-2\lambda} = \theta$$

so the variance of T is $ET^2 - (E(T))^2 = \theta - \theta^2$. So the mean squared error of T is also $\theta - \theta^2 = \theta(1 - \theta)$.

(c)
The data shows repeated observations in each of 7 countries, so one should use paired differences. Built-in calculator routines can be used to find the standard deviation and mean for the 7 differences. In part ii the Null Hypothesis is that the infant mortality is exactly 20 lower in 1945/49, and the Alternative Hypothesis is that in 1945/49 it is not exactly 20 lower. This is a two sided alternative, so we carry out a two-sided test. Remember to state the significance level that you use.

Question 4

(a)
This is book-work, but you should say why cross-product terms sum to 0. It is best to define any symbols that you use, such as $\bar{X}_i$. The Within Groups sum of squares measures the dispersion of the observations due to measurement error. The Between Groups sum of squares measures that part of the dispersion of the observations due to differences between group means. The Total sum of squares measures the dispersion in the whole set of observations.

(b)
The cells are not suitable for a two-way analysis, because there is an interaction between rows and columns. The difference between the columns depends on which row one looks at. In the first row the difference is 1, and in the second row it is 2. The question gives population means, so there is no doubt about this conclusion.

(c)
One can use built-in calculator routines for the sums of squares in the analysis of variance table. Since the given table is really more suitable for one-way analysis of variance than for two-way analysis, full marks were given for many different answers. It is important to state Null and Alternative hypotheses, and significance levels, and to state clearly conclusions from tests. In part iii there is no need to use simultaneous confidence intervals, as only one interval is asked for.

Question 5

(a)
For part i the fitted regression line is an estimate of the population regression line, and the residuals estimate the measurement errors ϵ_i. So if we take some average of the squares of the residuals, we should get an estimate of the variance of the measurement errors, which have mean zero. The divisor $(n-2)$ is a minor change in the more intuitive n to get unbiasedness. Part ii is bookwork.

(b)
For part i, the intercept and slope can be taken from built-in calculator routines. Part ii is book-work. In part iii, the interpretation is that wages go up when there are more machines per worker, whether ignoring Growth Rate or holding it fixed. It can be seen that for a fixed number of machines per worker, an increased Growth Rate seems to lower the wage. This may be because where industry is expanding rapidly the workers are fairly unskilled, and can't produce so much as in a developed business.

Examination paper for 2005

The form of the paper will be the same, though the proportion of marks for the more routine parts of questions will be about 15% greater.