SINGAPORE INSTITUTE OF MANAGEMENT

PRELIMINARY EXAM 2012

PROGRAMME(S) : University of London Degree and Diploma Programmes
(Lead College: London School of Economics & Political Science)

SUBJECT : **04A STATISTICS 1**

DATE : Tuesday, 28 February 2012

DURATION : 2 hours

INSTRUCTIONS:-

DO NOT TURN OVER THIS QUESTION PAPER UNTIL YOU ARE TOLD TO DO SO.

Candidates should answer **THREE** of the following **FOUR** questions: **QUESTION 1** of Section A (50 marks) and **TWO** questions from Section B (25 marks each).

Graph paper is provided. When used, it should be fastened securely **inside** the answer book.

A formula sheet and statistical tables are attached.

A handheld calculator may be used when answering questions on this paper, but it must not be pre-programmed or able to display graphics, text, or algebraic equations. The make and type of the machine must be clearly stated on the front cover of the answer book.

Candidates are strongly advised to divide their time accordingly.

Total number of pages: 6 (including this page)

SECTION A

Answer **all** parts of Question 1 (50 marks in total).

1. (a) Display the following data using a stem-and-leaf plot:

298	394	312	284	305	326	271	287	279	302	293	277	324	333
344	297	281	295	291	307	315	303	340	272	311	318	312	337

Use the plot to find the median of the data.

(6 marks)

(b) Give the mean, the median, the range and the interquartile range of the data below:

11	14	22	46	34	21	32	75	57	28	16	85

(4 marks)

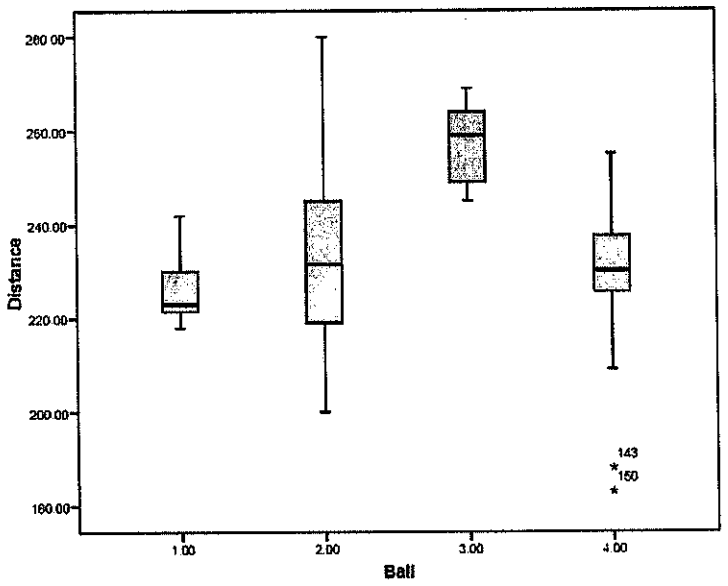
(c) If the line $y = -2x - 3$ is the least squares regression line of y on x, what value of y would you predict for

i. a value of $x = 1$

ii. a value of $x = -3$.

(2 marks)

(d) A sports magazine hired a golf professional to drive four different brands of golf ball. Each ball was driven fifty times and the distance achieved recorded. The outcomes are represented in a box plot produced by a statistical software package:



Distances have been measured in yards and the four types of ball labelled 1, 2, 3 and 4. Give at least three prominent features of the data which will aid interpretation.

(6 marks)

(e) A 95% confidence interval for the mean examination marks in a certain university is found to be (25%, 80%). Say in one sentence what this means. Would a 99% confidence interval be better than a 95% one? (Say why.)

(5 marks)

(f) In an examination the scores of students who attend schools of type A are normally distributed about a mean of 61 with a standard deviation of 5. The scores of students who attend type B schools are normally distributed about a mean of 64 with a standard deviation of 4. Which type of school would have a higher proportion of students with marks above 70?

(5 marks)

(g) x_i has the following values: 1, 4, 6, 2 and 3 when i is 1, 2, 3, 4 and 5, respectively. Find

i. $\sum_{i=1}^{i=3}(x_i - 1)^2$ ii. $\sum_{i=1}^{i=4}(x_i + 1)$ iii. $\sum_{i=2}^{i=5} 3x_i$

(6 marks)

(h) Give an example of a 2×2 contingency table in which there is:

i. No association.

ii. Strong association.

(4 marks)

(i) State whether the following are possible or not. Give a brief explanation. (*Note that no marks will be awarded for a simple possible/not possible reply.*)

i. Quota sampling gives biased estimators for parameters.

ii. A census always gives more accurate results than a reasonable size random sample survey.

iii. If the probability that it will rain tomorrow is 1/4 and the probability that you will take your umbrella with you is given, then the probability that it rains and you take your umbrella is 5/16.

iv. A positive regression line can have an associated negative correlation coefficient.

(8 marks)

(j) In a large lecture, 60% of the students are female and 40% are male. Records show that 15% of the female students and 20% of the male students are part-time students.

i. If a student is chosen at random from the lecture, what is the probability that the student is a part-time student?

ii. If a randomly chosen student studies part-time, what is the probability the student is male?

(4 marks)

SECTION B

Answer **two** questions from this section (25 marks each).

2. (a) The conventional treatment for a disease has been shown to be effective in 60% of all cases. A new drug is being promoted by a pharmaceutical company; the Department of Health wishes to test whether the new treatment is more effective than the conventional treatment.

   i. Write down the null and alternative hypotheses for this problem.

   ii. A simple random sample of $n = 400$ patients suffering from the disease were given the new drug; the treatment was effective for 320 of them. What decision would you reach about the new drug? Give reasons for your decision.

   iii. Say what Type I and Type II errors mean in this case, and what their implications would be.

   (13 marks)

   (b) A charity believes that when it puts out an appeal for charitable donations, the donations it receives will be normally distributed with a mean of £48 and standard deviation £4, and it is assumed that donations will be independent of each other.

   i. Find the probability that the first donation it receives will be greater than £40.

   ii. Find the probability that it will be between £56 and £60.

   iii. Find the value x such that 5% of donations are more than £x.

   iv. Find the probability that the first donation is at least £3 more than the second donation.

   (12 marks)

3. (a) i. In a large country each district is permitted to have its own policy on the death penalty. Some districts choose to have it, others choose not to. The table below shows the relationship between having the death penalty (No, Yes) and the crime rate (Low, High) for a sample of 200 districts.

Death penalty	**Crime rate** Low	High	Total
No	30	70	100
Yes	60	40	100
Total	90	110	200

Calculate the value of the chi-squared statistic for the table and say what you would conclude.

ii. If you look separately at the relationship for poor districts and rich districts you get the following two tables (poor districts = left table, rich districts = right table):

Death penalty	**Crime rate** Low	High	Total
No	16	63	79
Yes	8	16	24
Total	24	79	103

Death penalty	**Crime rate** Low	High	Total
No	14	7	21
Yes	52	24	76
Total	66	31	97

For the poor districts, the value of the chi-squared statistic is 1.76. For the rich districts, the value of the chi-squared statistic is 0.023. What would you conclude for each table, and overall?

(15 marks)

(b) i. Explain the strengths and weaknesses of using interviews to collect information in a survey rather than asking chosen respondents to fill in questionnaires themselves.

ii. You have been asked to design a random survey for a government department. Accuracy is very important and you have been given a realistic budget. The aim is to find out the attitude of old people to the care they receive in care homes or from their relatives. Would you use interviews or ask the old people to fill in forms? Explain and outline the main aspects of your survey design.

(10 marks)

4. (a) For each of 12 regions the following table relates the divorce rate (number of divorces per 100 married couples over a year) to an affluence measure (average house value).

Note: Summary statistics for these data are shown below the data table.

House price (£000s), x	100	130	180	210	220	250
Divorce rate, y	1.50	2.14	1.98	2.80	2.43	3.01

House price (£000s), x	250	253	290	295	350	360
Divorce rate, y	4.02	2.50	3.80	4.30	4.15	4.40

Summary statistics for these data:

$$\sum x = 2888,\ \sum x^2 = 764034,\ \sum y = 37.03,\ \sum y^2 = 125.4779,\ \sum xy = 9704.2$$

i. Using the graph paper provided, draw a scatter diagram for these data.

ii. Calculate the correlation coefficient. Is this a high value?

iii. Comment on the relationship between 'House price' and 'Divorce rate'.

iv. Compute the best fitting straight line for these data and draw this line on your scatter diagram. Label it carefully.

v. What would you expect the divorce rate to be in a region with a house price of £150,000?

(12 marks)

(b) A researcher was investigating computer usage among students at a particular university. Three hundred undergraduates and one hundred postgraduates were chosen at random and asked if they owned a laptop. It was found that 150 of the undergraduates and 80 of the postgraduates owned a laptop.

i. Find a 95% confidence interval for the difference in the proportion of undergraduates and postgraduates who own a laptop. On the basis of this interval, do you believe that postgraduates and undergraduates are equally likely to own a laptop?

ii. Notice that 57.5% of the students interviewed (230/400) owned a laptop. Explain, with reasons, whether the figure of 57.5% will be a good estimate of the proportion of all students who own a laptop.

(13 marks)

END OF PAPER

ST104a Statistics 1
Examination Formula Sheet

Expected value of a discrete random variable:

$$\mu = E[X] = \sum_{i=1}^{N} p_i x_i$$

Standard deviation of a discrete random variable:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} p_i (x_i - \mu)^2}$$

The transformation formula:

$$Z = \frac{X - \mu}{\sigma}$$

Finding Z for the sampling distribution of the sample mean:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Finding Z for the sampling distribution of the sample proportion:

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Confidence interval endpoints for a single mean (σ known):

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Confidence interval endpoints for a single mean (σ unknown):

$$\bar{x} \pm t_{n-1} \frac{s}{\sqrt{n}}$$

Confidence interval endpoints for a single proportion:

$$p \pm z \sqrt{\frac{p(1-p)}{n}}$$

Sample size determination for a mean:

$$n \geq \frac{Z^2 \sigma^2}{e^2}$$

Sample size determination for a proportion:

$$n \geq \frac{Z^2 p(1-p)}{e^2}$$

Z-test of hypothesis for a single mean (σ known):

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

t-test of hypothesis for a single mean (σ unknown):

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Z-test of hypothesis for a single proportion:

$$Z \cong \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

t-test for the difference between two means (variances unknown):

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Pooled variance estimator:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Confidence interval endpoints for the difference in means in paired samples:

$$\bar{x}_d \pm t_{n-1} \frac{s_d}{\sqrt{n}}$$

Pooled proportion estimator:

$$P = \frac{R_1 + R_2}{n_1 + n_2}$$

χ^2 test of association:

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Spearman rank correlation:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Z-test for the difference between two means (variances known):

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Confidence interval endpoints for the difference between two means:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

t-test for the difference in means in paired samples:

$$t = \frac{\bar{X}_d - \mu_d}{S_d / \sqrt{n}}$$

Z-test for the difference between two proportions:

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{P(1 - P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Confidence interval endpoints for the difference between two proportions:

$$(p_1 - p_2) \pm z \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

Sample correlation coefficient:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Simple linear regression line estimates:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$