SIM GLOBAL EDUCATION

SINGAPORE INSTITUTE OF MANAGEMENT

PRELIMINARY EXAM 2013

PROGRAMME(S) :	University of London Degree and Diploma Programmes (Lead College: London School of Economics & Political Science)
SUBJECT :	**ST104A STATISTICS 1 (Half Unit)**
DATE :	Wednesday, 27 February 2013
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. If used, it must be detached and 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) State whether each of these statements is **true** or **false**, giving brief reasons why this is so. (*Note that no marks will be awarded if there is no explanation.*)

   i. When using a large random sample, we cannot assume that its mean forms part of a normal distribution.

   ii. The least squares regression line minimises the sum of absolute deviations.

   iii. The power of a test is the probability of a Type II error.

   iv. If two variables are correlated, then they must have a linear relationship.

   (8 marks)

   (b) Suppose $x_1 = 2$, $x_2 = 3$, $x_3 = 1$, $x_4 = 7$ and $x_5 = 9$. Find

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

   (6 marks)

   (c) The summary statistics for two datasets are as follows:

	Sample size	Sample mean
x-data	19	7.0
y-data	25	5.1

   Compute the mean of the combined datasets.

   (2 marks)

   (d) i. Two fair dice are thrown. What is the probability of at least one odd number? What is the probability of this if four fair dice are thrown?

   ii. In a manufacturing plant, machine A produces 10% of a certain product and machine B produces 90% of this product. Of the production by machine A, 10% are defective; for machine B the defective rate is 5%. If a product is selected at random from one of the machines, what is the probability that it is defective?
   If a company inspector, sampling production at random, discovers a defective product, what is the chance that it came from machine B?

   (8 marks)

   (e) i. What conclusions would you draw from a test which is significant at the 1% level?

   ii. What conclusions would you draw from a test which is significant at the 10% level, but not the 5% level?

   iii. An accounting firm wishes to test the claim that no more than 1% of a large number of transactions contains errors. In order to test this claim, they examine a random sample of 144 transactions and find that exactly 3 of these are in error. What conclusion should the firm draw?

   (8 marks)

(f) For a certain variable, the standard deviation in a large population is equal to 8.5. How big a sample is needed to be 90% sure that the sample mean is within 1.5 units of the population mean?

(4 marks)

(g) Assume that the number of weekly study hours for students at a certain university is approximately normally distributed with a mean of 22 and a standard deviation of 6.

  i. Find the probability that a randomly chosen student studies less than 12 hours.

  ii. Estimate the percentage of students that study more than 37 hours.

  iii. A certain lecture group consists of 225 students. You may assume that this class group forms a simple random sample from the students in the university. Find the probability that the average number of study hours for this group is between 21 and 23 hours.

(6 marks)

(h) Define quota sampling. In what circumstances would you use it?

(4 marks)

(i) In what circumstances would you use stratified random sampling? Give two ways in which it differs from quota sampling.

(4 marks)

SECTION B

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

2. (a) The table below shows the relationship between gender and party identification in a US state.

Gender	Party identification			Total
	Democrat	Independent	Republican	
Male	279	73	225	577
Female	165	47	191	403
Total	444	120	416	980

   i. Test for association between gender and party affiliation at two appropriate levels and comment on your results.

   ii. Does any one party appear to have a higher proportion of women identifying with it? Test for significance at the 5% level and comment.

   (13 marks)

   (b) Following the results above, political researchers wish to look at the relationship between gender, social and employment status and political affiliation in your country. Your market research company is asked to suggest a possible interviewer survey. Outline your proposed random sample design, giving an appropriate sampling frame and at least three possible stratification and/or clustering factors. How would you minimise the possibility of non-response?

   (12 marks)

3. (a) At a time of economic growth but political uncertainty, a sample of 40 economists (from the population of all economists) offers the following values for the growth rate of an economy in the next year:

$$
\begin{array}{cccccccc}
1.3 & 3.8 & 4.1 & 2.6 & 2.4 & 2.2 & 3.4 & 5.1 \\
1.8 & 2.7 & 3.1 & 2.3 & 3.7 & 2.5 & 4.1 & 4.7 \\
2.2 & 1.9 & 3.6 & 2.8 & 4.3 & 3.1 & 4.2 & 4.6 \\
3.4 & 3.9 & 2.9 & 1.9 & 3.3 & 6.2 & 5.4 & 3.3 \\
4.5 & 5.2 & 3.1 & 2.5 & 3.3 & 3.4 & 4.4 & 5.2
\end{array}
$$

    i. Construct and label a histogram for this dataset. What is the median? Work out the upper and lower quartiles.

    ii. Look at the shape of your histogram (but do not do any further calculations). Explain briefly which of the following statements you would expect to be true:

- the population mean forecast would be greater than the population median forecast

- the population mean forecast would be equal to the population median forecast

- the population mean forecast would be less than the population median forecast.

(12 marks)

(b) Random samples are taken from two populations with distributions $N(\mu_x, \sigma^2)$ and $N(\mu_y, \sigma^2)$, i.e. their variances are the same. The summary statistics for the two samples are shown in the following table:

	Sample size, n	Sample mean, m	Sample variance s^2
x-data	19	7.0	1.69
y-data	25	5.1	2.56

    i. Compute a 95% confidence interval for the difference $\mu_x - \mu_y$ between the two population means. Does the result support the view that there is no true difference between the population means? (Explain your reasoning!)

    ii. You are asked to carry out a formal hypothesis test. Set out the null and alternative hypotheses and give the p-value for the difference given. Does this agree with your earlier findings? Why is this?

(13 marks)

4. (a) A market research company has conducted a survey of adults in two large towns, either side of an international border, in order to judge attitudes towards a controversial internationally-broadcast celebrity television programme.

The following table shows some of the information obtained by the survey:

	Town A	Town Z
Sample size	40	40
Sample number approving of the programme	24	22

   i. Conduct a formal hypothesis test, at the 5% significance level, of the claim that the population proportions approving the programme in the two towns are equal.
   ii. Would your conclusion be the same if, in both towns, the sample sizes had been 100 (with the same sample proportions of approvals)?
   iii. And if they had been 400, again with the same sample proportions?

(10 marks)

(b) The following table, published in *USA Today*, lists divorce rates and mobility rates for different regions of the USA. Divorce rate is measured as the annual number of divorces per 1000 population and the mobility rate is the percentage of people living in a different household from five years before.

Region	Mobility rate, x	Divorce rate, y
New England	41	4.0
Middle Atlantic	37	3.4
East North Central	44	5.1
West North Central	46	4.6
South Atlantic	47	5.6
East South Central	44	6.0
West South Central	50	6.5
Mountain	57	7.6
Pacific	56	5.9
$\sum$ of data	422	48.7
$\sum$ of squared data	20132	276.91

   i. Using the graph paper supplied, and using mobility rate as the x-variable and divorce rate as the y-variable, draw a scatter diagram of these data and comment.
   ii. Calculate the regression equation and give the expected divorce rate if the mobility rate was 40. Note: $\sum xy = 2341.6$.
   iii. Why might it be reasonable to use divorce rate as the y-variable?
   iv. Calculate the correlation coefficient and comment.
   v. A colleague disagrees and wants to use the mobility rate as the dependent variable. What would this imply?

(15 marks)

END OF PAPER