UNIVERSITY OF LONDON

GOLDSMITHS' COLLEGE

B. Sc. Examination 2003

STATISTICS

ST51005A (ST154) Statistics

Duration: 3 hours

Date and time:

Answer ELEVEN questions, which are worth nine marks each. Full marks will be given for complete answers to ELEVEN questions.

Electronic calculators may be used. The make and model should be specified on the script. The calculator must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

NOTE: Full details of all calculations are to be shown; pre-programmed statistical tests and procedures on a calculator, apart from mean and standard deviation, must not be used.

WHITE, YEATS & SKIPWORTH: Tables for Statisticians to be provided.

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Question 1 Hydrogen chloride (HCl) levels in the stratosphere are related to ozone depletion. The data below are levels of HCl above 12km altitude both before and after the eruption of the volcano El Chicon in Mexico (in 10^{15} molecules per square centimetre):

Preeruption	Posteruption
0.78	1.46
1.26	1.40
0.75	1.29
0.69	1.22
0.88	1.63
1.56	1.90
1.18	1.24
	1.45
	1.79

(a) *Without losing any information*, construct an appropriate plot to compare HCl levels before and after eruption. [6]

(b) Comment on the inapplicability of the two-sample t-test or the hypothesis that there is no difference in mean HCl levels. [3]

Question 2 Using the data of question 1,

(a) compute the upper and lower quartiles of each data set. [4]

(b) Hence compute the semiinterquartile ranges. [2]

(c) Using the quartiles only, comment on the evidence for differences in HCl levels. [3]

Question 3 (a) Define the *standard deviation* s of a sample of data $x_1, \ldots, x_n$. [2]

(b) The *mean absolute deviation* t of such a sample is defined by

$$t = \frac{1}{n} \sum_{i=1}^{n} \mid x_i - \bar{x} \mid,$$

where $\bar{x}$ is the sample mean. Why is s rather than t used in tests of hypotheses? [1]

(c) Show that for $n = 2$

$$t = \frac{1}{2} \mid x_1 - x_2 \mid, \qquad \textit{whereas} \quad s = \frac{1}{\sqrt{2}} \mid x_1 - x_2 \mid . \qquad [6]$$

Question 4 Let E and F be any two events. Write down using set notation the events

(a) both E and F

(b) at least one of E and F

(c) neither E nor F

(d) at most one of E and F. [4]

Hence find their probabilities if $P[E] = 0.4$, $P[F] = 0.7$ and $P[E \cap \bar{F}] = 0.3$, (where bar denotes the complement). [5]

Question 5 A test for errors in an audit of accounts is said to be 90% effective in the sense that if there is an error (event E), the probability is 0.9 that the test will report 'error detected' (event D). However it is admitted that that there is a small chance (0.01) that the test will report an error even when one does not exist. Suppose that you judge there is a 40% chance of an error before the test. By how much does your probability increase if the test reports 'error detected'? [5]

By how much does it decrease if the test does not so report? [4]

Question 6 Random digits are such that every digit (0 to 9) has an equal chance of appearing, and different digits are independent. Let X be the random digit to appear next in a sequence.

(a) Name the distribution of X. [1]

(b) Find the mean of X. [3]

(c) Find the variance of X. [5]

Question 7 Ecologists use the number of reported sightings of rare species to estimate the population size. Suppose that the number X of reported sightings for blue whales is recorded and assume that it follows a Poisson distribution with a mean of 2.5 sightings per week. Find the probability that

(a) fewer than two sightings are reported in a week [2]

(b) fewer than four sightings are reported in a two week period [3]

(c) at most 7 sightings are reported in a week [2]

(d) at most 7 sightings are reported in a two week period [2]

Question 8 The probability density function (pdf) of a continuous random variable X is given by

$$f(x) = k(1-x), \qquad 0 < x < 1, \quad \text{zero otherwise.}$$

(a) Show that k=2 for $f(x)$ to be a proper pdf. [3]

(b) Find the mean $E[X]$. [3]

(c) Find $P[X \leq 1/2]$. [3]

Question 9 The weight X of breakfast cereal in a packet is normally distributed with mean 500g and standard deviation 10g. What is the probability that the packet contains

(a) between 490g and 510g? [2]

(b) less than 475g? [3]

Ten such packets are randomly sampled from a large batch of production. What is the probability that they have a total weight less than 4.95kg? [4]

Question 10 A random sample of size n is drawn from a population with mean μ and variance σ^2. Answer the following questions about the sampling distribution of $\bar{X}$, the sample mean:

(a) What is the expected value of $\bar{X}$? [1]

(b) What is the standard deviation of $\bar{X}$? [2]

(c) What further assumption (if any) do we need for the statistic

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

to have a standard normal distribution for all n? [2]

(d) If you replaced σ in Z above by S, the sample standard deviation, would the resulting statistic have a standard normal distribution with your further assumption (yes or no)? [1]

(e) Answer the previous question with a short statement if it is known further that n is very large. [3]

Question 11 A random sample of 10 economists were asked to predict the perentage growth in the consumer price index over the following year and the results were:

3.6 3.1 3.9 3.7 3.5 3.7 3.4 3.0 3.6 3.4

The sample mean is 3.49 and the sample variance is 0.077.

(a) Assuming that these forecasts are from a normal population with mean μ and variance 1, is there any evidence against the hypothesis that μ is equal to 4? [4]

(b) Test the above assumption about the variance and comment briefly. [5]

Question 12 The following table shows a summary of data (percentage scores) collected from a reading test at the end of a learning period where the children had been taught according to either a new or standard (std) method

	Sample	
	New	Std
Size	10	12
Mean	76	72
S.D.	5.2068	5.2915

(a) Is there any evidence that the population means are different? Answer this question on the assumption that we have two independent random samples from normal populations with equal (but unknown) variances. [7]

(b) Which (if any) of 90% and 95% confidence intervals for the difference in means contain the value zero? [2]

Question 13 In a household survey, 200 families with three children were considered and the number of girls in each family was recorded:

number of girls	0	1	2	3
observed frequency	15	71	98	16

(a) Assuming that births of either sex are equally likely and that successive births are independent, calculate the corresponding expected frequencies. [4]

(b) Do the discrepancies provide evidence to cast doubt on the above assumptions? Show all reasoning. [5]

Question 14 The marginal distributions of two jointly discrete random variables X and Y are given by

$$P[X=-1]=0.3, \qquad P[X=0]=0.3, \qquad P[X=1]=0.4,$$

$$P[Y=-1]=0.1, \qquad P[Y=0]=0.4, \qquad P[Y=1]=0.5.$$

(a) Find a, b, c, d in the following table of joint probabilities: [6]

		Y		
		-1	0	1
	-1	0.1	0.1	0.1
X	0	b	0.1	a
	1	c	d	0.2

(b) Hence find $E[XY]$. [3]

Question 15 In a study to determine the influence of training on the time required to do an assembly job, 15 new employees were given amounts of training ranging from 3 to 12 hours (assigned at random), After training, their time to complete the job were recorded. Let x denote the duration in training (in hours) and let y denote the time taken to do the job (in minutes). The following summary statistics were obtained:

$CS(x,y)=-57.2$, $CS(x,x)=33.6$, $CS(y,y)=160.2$, $\bar{x}=7.2$, $\bar{y}=45.6$

(a) Why are these data NOT suitable for a correlation analysis? [2]

(b) Determine the equation of the best fitting straight line to these data. [5]

(c) A new employee is given 5 hours training. Estimate how long he will take to complete the task. [2]