UNIVERSITY OF DUBLIN

TRINITY COLLEGE

Faculty of Engineering and Systems Sciences

Department of Computer Science

B. A. (Mod.) Computer Science **Trinity Term 2000**

3BA1 Statistics and Numerical Analysis

Thursday 1st June **Luce Hall** **9.30 – 12.30**

Dr. K. Mosurski and Professor J.G. Byrne

Answer five questions, at least one of which is from Section B. Cambridge Elementary Statistical Tables are available from the Invigilator. Calculators may be used. Use separate answer books for each section. A table of formulae applicable to section A is attached.

SECTION A

1. (a) A PC attempts to connect to a network via a telephone line. On each trial there is probability of 0.4 that a connection will be established.

(i) If n attempts are made, what probability distribution describes the number of successful connections?

(ii) The automatic dialler is set up to make 5 attempts before it gives up. What is the probability that it makes a connection?

(iii) A user, desperate to make the connection, keeps invoking the dialler until she makes the connection. What is the probability distribution of the number of times she has to invoke the dialler?

(iv) What is the probability that the dialler will have to be invoked more than twice before a connection is made?

(b) Before being used in the manufacturing process, all ICs are tested to establish that they work correctly. The test, however, is not perfect. While all ICs that work correctly pass the test, 5% of faulty ICs also pass the test. From past experience it is known that 20% of all ICs are faulty.

(i) Calculate the proportion of ICs that pass the test.

(ii) If an IC actually passes the test, what is the probability that it is working correctly?

(c) Let X be Normally distributed with a mean of 3 and a variance of 4. What is the probability that X is greater than 1.5?

2. A researcher has developed a new automatic English to Japanese translator. In a test, 100 random sentences were submitted to the translator. The Japanese text was then read by a bilingual, native speaker and she awarded a score 0 = nonsense to 10 = perfect for each attempt. Later it was established that a translation that was awarded a score of 4 or more could actually be understood by a native speaker.

These data were input into MINITAB and descriptive statistics were calculated as:

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Score	100	6.330	8.000	6.478	3.638	0.364

Variable	Minimum	Maximum	Q1	Q3
Score	0.000	10.000	2.250	9.000

(a) Explain how the Mean and Median are calculated. Comment on the observed difference between these summaries.

(b) What is the variance of the data?

(c) Assuming that the score may be regarded as Normally distributed with mean and standard deviation as calculated from this experiment, obtain an estimate of the probability that a random sentence, when translated, can be understood by a Japanese native speaker (i.e. score ≥ 4).

Contd.

(d) The command TALLY was used to obtain a frequency count of the score variable:

Score	Count
0	13
1	8
2	4
3	1
4	3
5	3
6	5
7	8
8	10
9	29
10	16

N= 100

Sketch a rough histogram or bar chart of these data. Comment on the Normality assumption in (c) above. Obtain a direct estimate of P(score≥4)

(e) Compute an approximate 95% confidence interval for the mean score. Explain why this approximation is still valid despite the conclusions in (d) above.

3. Data have been collected on the effects of alcohol on the subjects' ability to perform a task. Subjects were asked to click the mouse inside a disc that appeared on a computer screen. The time to perform the task was recorded. Subjects in one group (group B) were asked to drink 4 units of alcohol before performing the task.

(a) The data were entered into MINITAB and descriptive statistics were calculated for both groups.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
alcohol	12	147.50	148.00	147.60	2.81	0.81
No alcoh	8	145.12	145.00	145.12	2.47	0.87

Based on this information, carry out an appropriate statistical test to ascertain whether alcohol influences response time.

(b) One of the experimenters in (a) above reported the story to a friend who is a statistician and expressed his disappointment at the fact that the result was inconclusive. The friend responded that he should have designed a paired comparison experiment and if he did this she would perform the test for him

as long as she was allowed to keep the left over alcohol. Only five subjects were tested this time. The data were:

Subject	Before Alcohol	After Alcohol
Andy	144	147
Mary	148	151
John	152	156
Kevin	135	139
Lisa	137	140

Explain why this is a better way to test for effects of alcohol than the previous design. Carry out an appropriate test.

4. A small department of a large organisation is responsible for the maintenance of PCs. To enable them to manage their operations they would like to know how many PCs are likely to need attention at any given time. It has been suggested that a possible model for this is the number of "problems" during a week, which should be Poisson distributed. From past records, data have been extracted on the number of "problems" for every week in the past two years.

The table below gives the frequencies of the numbers of weeks in which a given number of breakdowns occured over the past two years (100 weeks). The last column shows the Poisson probabilities based on the mean estimated from the data.

Mean problems per week = 1.87

problems	Count	Poisson Probability
0	23	0.15
1	27	0.29
2	17	0.27
3	17	0.17
4	8	0.08
5	5	0.03
6	3	0.01
N=	100	

Carry out a Chi-squared goodness of fit test on these data. What are you testing? What is your conclusion?

(b) There were 187 breakdowns during this period, 87 in the first year and 100 in the second. The researchers distinguish three categories of break down:

Software problem, Peripheral configuration and Hardware failure. They would like to know if the pattern of the types of failure has changed in the last two years. The table below reports the frequencies of failure by type by year:

	Software	Peripheral	Hardware	Total
Year 1	55	20	12	87
Year 2	60	30	10	100
Total	115	50	22	

The table below give the Chi-square contributions under the assumption of independence.

Chi-square contributions		
0.042	0.457	0.304
0.036	0.397	0.264

Complete the test. What has testing for statistical independence to do with whether patterns change or not?

5. A new algorithm solves LPs (linear programs) in linear time as a function of the number of constraints. Twenty standard problems were solved using software implementing the new algorithm and the time to complete the task was recorded.

A linear regression of time against number of constraints was carried out in Data Desk.

Dependent variable is:Time
No Selector

R squared = 77.6% R squared (adjusted) = 76.4%
s = 3.107 with 20 - 2 = 18 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	602.426	1	602.426	62.4
Residual	173.792	18	9.65508	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	4.84672	1.909	2.54	0.0206
Constraints	2.14486	0.2715	7.9	< 0.0001

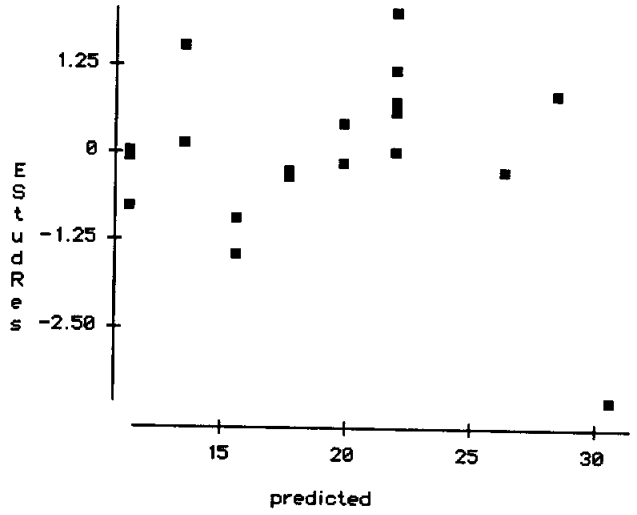
(a) Write down the estimated equation. What interpretation would you put on the coefficients?

(b) Obtain an estimate for the average time the software takes to solve 10 constraint problems, and 100 constraint problems. You are told that the test problems ranged from 3 to 12 constraints, comment.

(c) Obtain a confidence interval for the time the software takes to solve a particular 10 constraint problem.

Note that: $\sum^{20}(x_i - \bar{x})^2 = 131,\ \bar{x} = 6.6$.

(d) Studentized residuals were plotted against predicted values. Discuss.



6. (a) X and Y are two random variables with the following probabilities:

		X		
		0	1	2
Y	0	0.4	0.2	0.1
	1	0.1	0.1	0.1

Compute P(Y=y), P(Y=y|X=1), covariance of X and Y.

(b) A regression was carried out to obtain a relationship between the access time of a hard disk and three predicting factors, fragmentation index in the range 0 to 1, % full (0 -100%) and Operating system used (1 for PC DOS and 0 for Windows 95). Quadratic terms were included for fragmentation (Frag2) and full (Full2).

Dependent variable is:Access Time
No Selector

R squared = 94.1% R squared (adjusted) = 92.9%
s = 3.012 with 30 - 6 = 24 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3499.93	5	699.986	77.1
Residual	217.784	24	9.07432	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	7.16584	2.899	2.47	0.0209
PCDos	19.99	1.218	16.4	< 0.0001
Fragment	-7.08155	9.867	-0.718	0.4799
Frag2	16.8584	9.383	1.8	0.0850
Full%	-0.013674	0.07692	-0.178	0.8604
Full2	0.000331976	0.0007661	0.433	0.6686

(i) Explain the criterion you might use to decide which variables should be dropped from the equation and which should be retained. Say what should be the next step in developing an equation for access time.

(ii) What does "R squared" measure? Comment on the value obtained above.

(iii) A range of different models was considered for these data (see Model 1 and Model 2 overleaf) and two candidates emerged as possibilities:

(1) Write down the equations for both models and obtain estimated access times for machines running Windows 95 that have fragmentation indices of 0.5 and 0.05. Comment on the results.

(2) Explain how it can happen that both models can fit the data (almost) equally well.

If you were forced to choose which model to use to make predictions, which would you select and why?

(Note: No uniquely correct answer here: discuss the pros and cons of the model you choose.)

Model 1.

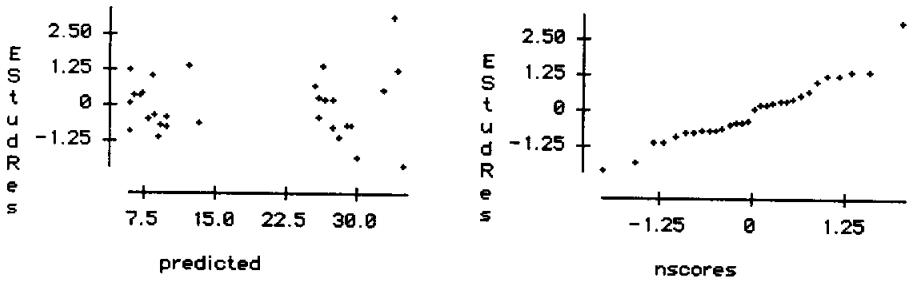
Dependent variable is: Access Time
No Selector

R squared = 93.0% R squared (adjusted) = 92.5%
s = 3.094 with 30 - 3 = 27 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3459.21	2	1729.61	181
Residual	258.502	27	9.57414	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	4.02534	1.217	3.31	0.0027
PCDos	20.9868	1.131	18.6	< 0.0001
Fragment	10.3393	2.128	4.86	< 0.0001

Residual plots



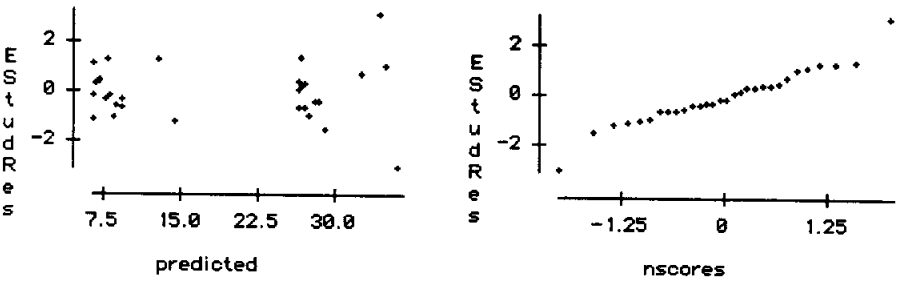
Model 2

Dependent variable is: Access Time
No Selector

R squared = 93.7% R squared (adjusted) = 93.2%
s = 2.946 with 30 - 3 = 27 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3483.38	2	1741.69	201
Residual	234.328	27	8.67883	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	6.06144	0.8844	6.85	< 0.0001
PCDos	20.4309	1.077	19	< 0.0001
Frag2	10.4657	1.949	5.37	< 0.0001



Note: if both Fragment and Frag2 are included in a model, neither of their coefficients is significantly different from 0.

SECTION B

7. Derive Newton's method for finding a root of a single non-linear equation. Using a Taylor series expansion show that it has quadratic convergence.

   Discuss, with the aid of diagrams, some difficulties which can arise with the method, especially with regard to the starting guess x_0?

8. Write down the formula for Lagrange interpolation of the data points $(x_i, f(x_i))$, $i = 0$ to n.

   What is its complexity? Discuss briefly some of the problems which can arise in Lagrange interpolation.

   Derive the Barycentric form of Lagrange interpolation. In what ways is it better than Lagrange interpolation?

Some of these formulae may be useful for questions in Section A.

Statistical Distributions:

Name	$p(x), f_X(x)$	E(X)	V(X)
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	np	np(1 - p)
Geometric	$p(1-p)^{x-1}$	$\frac{1}{p}$	$\frac{1-p}{p}$
Poisson	$e^{-\lambda}\frac{\lambda^x}{x!}$	λ	λ
Exponential	$\lambda e^{\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$

Rules of probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad , \quad P(A) = \sum_i P(A|B_i)P(B_i) \quad , \quad P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)}$$

If A and B independent $P(A|B)=P(A)$.

Statistical Estimation and Testing:

$$\bar{x} = \frac{1}{n}\sum_i x_i \quad , \quad s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \quad , \quad z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \quad , \quad t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

$$sd(\bar{x}) = \frac{\sigma}{\sqrt{n}} \quad , \quad sd(\bar{x}-\bar{y}) = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}, \text{ pooled estimate of } \sigma \quad s = \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2}}$$

Contribution to Chi - square $x_i = \frac{(x_i - e_i)^2}{e_i}$

ctd.

Linear Regression

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} \quad , \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad , \quad s^2 = \frac{SSE}{n\text{-}2}$$

$$sd(\hat{\beta}) = \frac{s}{\sum (x_i - \bar{x})^2} \quad , \qquad sd(\text{predicted mean at } x) = s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$sd(\text{individual prediction at } x) = s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$