UNIVERSITY OF DUBLIN

TRINITY COLLEGE

Faculty of Engineering and Systems Sciences

Department of Computer Science

B. A. (Mod.) Computer Science
J. S. Examination **Trinity Term 1998**

3BA1 Statistics and Numerical Analysis

Thursday 21st May **Exam Hall** **14.00 - 17.00**

Dr. S.P. Wilson 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.

SECTION A

1. (a) A software company is testing an anti-virus utility to see if it can detect a new virus. If the virus is present, the utility will detect it with probability 0.95. However, the utility sometimes yields a 'false positive'; this means that if the virus is not present, there is a probability 0.01 that the utility will claim to have detected it.

It is thought that 20% of all computers are infected with the virus.

(i) If the utility is run on a computer, what is the probability that the utility will claim to have detected the virus?

(ii) Suppose the utility claims to have detected the virus. Using Bayes Law, calculate the probability that the virus is actually present.

(b) A communications tower is located on top of a hill and is subject to lightning strikes. The computers in the control centre are protected from lightning by circuit breakers. There is a probability 0.01 that a lightning strike will be sufficiently powerful to trip the circuit breakers and shut down the tower. In one year, the tower was hit 200 times.

(i) What is the probability distribution for the number of times the circuit breakers are tripped?

(ii) What is the probability that the circuit breakers are tripped once in the year?

(iii) What is the expected number of trips per year? What is the variance?

(c) The time between consecutive lightning strikes on the communications tower is exponentially distributed with a mean of 2 days.

(i) Write down the probability density function for the time between consecutive strikes.

(ii) What is the probability that the time between lightning strikes is more than 1 day?

(iii) What is the probability that the time between lightning strikes is between 2 and 4 days?

2. A company is developing a new generation of read-write CD drives. There is concern that the drives are not reliable enough. In a test, a batch of 20 drives was used for one week and the number of times that each failed to read or write successfully was recorded as follows: 8, 2, 0, 0, 3, 1, 1, 3, 3, 6, 7, 5, 2, 0, 1, 4, 2, 2, 1, 5.

(a) These data were analysed with Data Desk, which produced the following output:

Summary statistics for No. of failures	
Mean	2.8000000
Median	2
StdDev	2.3530495
Lower Quartile	1
Upper Quartile	4.7500000

(i) Explain each of the terms Mean, Median, StdDev, Lower Quartile and Upper Quartile.

(ii) What is the Interquartile Range of the data?

(b) Form a 99% confidence interval for the mean number of failures per drive per week. How do you interpret this interval?

(c) From a consumer survey, the company believes that an unacceptable mean number of failures per drive is more than 2 per week. The product development manager conducts a test of hypothesis to test for this. She uses Data Desk, which produces the following output:

t-Test of Individual μ's

Individual Alpha Level 0.05
Ho: $\mu = 2$ Ha: $\mu > 2$

Test Ho: μ(No. of failures) = 2 vs Ha: μ(No. of failures) > 2
Sample Mean = 2.8000000 t-Statistic = 1.52 w/19 df
Fail to reject Ho at Alpha = 0.05
p = 0.0724

(i) Say how the test statistic value of 1.52 was calculated (you do not have to do the actual calculations).

(ii) The level of significance of this test is 5%. Define what the level of significance of the test is.

(iii) The final statement in the Data Desk output is "p = 0.0724". Explain what this value is.

(iv) Is the mean number of failures per week of the new CD drive more than 2? Explain your reasoning.

3. (a) A medical researcher is trying to compare the effectiveness of two different drugs, A and B, for stomach ulcers. He takes two groups of patients who are suffering from stomach ulcers: the first group is of 12 patients and the second is of 15.

The first group is given drug A for 1 month, and the difference between the number of ulcers before and after the month of treatment is recorded. The mean of the 12 differences is –3.2 and the standard deviation is 0.6.

The second group is given drug B for 1 month, and the difference between the number of ulcers before and after the month of treatment is recorded. The mean of the 15 differences is –0.4 and the standard deviation is 1.1.

(i) What is the pooled estimate of the standard deviation in change in number of ulcers from the two groups?

(ii) A test of hypothesis is conducted to see if the two treatments result in different reductions in the number of ulcers. Write down the null and alternate hypotheses for this test.

(iii) The test of hypothesis is conducted, and the test statistic value found to be 7.91. Say how this value was calculated (you do not have to carry out the calculations).

(iv) What is the conclusion from the test? Explain your reasoning.

(b) In another experiment, a different drug for treating stomach ulcers was tested on 10 patients. The number of ulcers counted in the stomach both before and after treatment with the drug was observed as follows:

Patient	1	2	3	4	5	6	7	8	9	10
Ulcers before	12	21	18	26	10	3	20	19	33	23
Ulcers after	5	13	10	11	3	3	9	2	21	15

(i) The researcher conducting this experiment wants to know if there has been a decrease in number of ulcers after the treatment. She decides to use a paired t-test to test this

hypothesis. Briefly explain why a paired t-test is appropriate in this case.

(ii) Write down the null and alternate hypotheses for the test.

(iii) The researcher calculates a test statistic value of 6.18. Say how this value was calculated (you do not have to carry out the calculations).

(iv) Write down the conclusion of the test.

4. (a) You are estimating the speed of new database software to process a request for information. In an initial set of 4 queries, the times to process the requests are 8, 6, 3 and 9 seconds.

(i) Calculate the sample standard deviation of the 4 observed times.

(ii) The systems manager wants to be 95% confident that the mean time to process a request is estimated to within 0.5 seconds. How many more requests should you time in order to obtain this level of accuracy in your estimation?

(b) A company makes computer monitors. It has two manufacturers, A and B, for the cathode ray tube. As part of its quality control procedures it tests all its monitors to see if their contrast meets specifications. It takes a sample of 100 monitors and classifies them according to the manufacturer and according to whether they pass or fail the contrast test. This data is put into the following contingency table:

	Manufacturer	
	A	B
Pass	15	32
Fail	25	28

(i) The company wants to conduct a test of hypothesis to see if there is an association between the manufacturer of the cathode ray tube and its contrast quality. Write down the appropriate null hypothesis for this test.

(ii) If the null hypothesis that you have written in part (i) is true, how many of the 100 monitors would you *expect* to classify

as coming from manufacturer A and passing the contrast test?

(iii) A χ^2 test is used to test the null hypothesis, and the value of the test statistic is found to be 2.415. Explain how this value was calculated (you do not have to do the calculations).

(iv) Is there evidence for an association between manufacturer and contrast quality? Explain your reasoning.

5. (a) X and Y have the following joint probability distribution:

		X		
		1	2	3
Y	1	0.2	0.1	0.1
	2	0.1	0.1	0.4

(i) Calculate the marginal distribution of X.

(ii) What is $P(Y = 2 \mid X = 1)$?

(iii) The correlation between X and Y is 0.421. What is the formula for correlation? What does its value say about the relationship between X and Y?

(b) A pollution monitoring station at the side of a motorway measures levels of nitrous oxide. At the same location, a traffic monitor counts the number of vehicles using the motorway. The following data were collected on amount of traffic and average nitrous oxide levels over 10 one hour periods during the day:

Nitrous Oxide (parts per million)	7.1	12.3	2.4	4.9	8.0	11.9	8.5	7.6	4.1	6.8
No. of vehicles ('000)	3.1	7.2	1.5	2.5	4.1	5.9	4.3	4.2	2.0	3.2

The County Council is interested in predicting nitrous oxide levels from the traffic intensity. It uses Data Desk to analyse the data, with the following output:

Dependent variable is Nitrous oxide level
10 total cases

R squared = 95.3% R squared (adjusted) = 94.7%
s = 0.7207 with 10 - 2 = 8 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	84.2888	1	84.2888	162
Residual	4.15524	8	0.519405	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	0.712070	0.5695	1.25	0.2465
Traffic	1.74946	0.1373	12.7	≤ 0.0001

(i) What is the least squares line of nitrous oxide level on traffic? Briefly explain how this line has been calculated.

(ii) Using the information provided by Data Desk, conduct a test of hypothesis to see if the slope of the regression line differs significantly from 0.

(iii) Comment on the strength of the linear relationship between nitrous oxide level and traffic.

(iv) In 5 years time, peak traffic intensity is predicted to be 8 thousand vehicles per hour. What is your prediction of the nitrous oxide level at this level of traffic? Briefly explain why you should be cautious about using this prediction.

6. (a) Describe the linear congruence method of generating pseud-random numbers.

(b) Describe how, given a stream of pseudo-random numbers, you would simulate values from a binomial distribution with parameters n and p.

(c) Devise a scheme for generating values from an exponential distribution with distribution function $F(x) = 1 - e^{-x}$.

(d) Devise a scheme, using the rejection method, to simulate values from a distribution with density function $f(x) = 2x, \quad 0 \le x \le 1$.

SECTION B

7.

(i) Using the exchange scheme, exchange x_2 and y_3 in the following equations:

$$y_1 = 2x_1 + 4x_2 + x_3$$
$$y_2 = 6x_1 - 3x_2 + 2x_3$$
$$y_3 = 4x_1 - x_2 + 2x_3$$

(ii) Solve the following linear programme either graphically or using the

Simplex algorithm:

$$\text{Maximise } 10x_1 + 12x_2$$

Subject to the constraints

$$5x_1 + 17x_2 \leq 170$$
$$6x_1 + 11x_2 \leq 132$$
$$35x_1 + 18x_2 \leq 630$$
$$x_1 \geq 0$$
$$x_2 \geq 0$$

8. Define "interpolation" and write down the Lagrangean interpolation

formula for a set of data $(x_i, f(x_i))$ $i = 0...n$

What is the complexity of this formula?

Derive the Barycentric form of this interpolation formula. What is its complexity?