UNIVERSITY OF ABERDEEN **SESSION 2004–2005**

DEGREE EXAMINATION

ST1505 Understanding Data

Tuesday 31 May 2005 (9 am—11 am)

Only calculators approved by the Department of Mathematical Sciences may be used in this examination. Calculator memories must be clear at the start of the examination.

Marks may be deducted for answers that do not show clearly how the solution is reached.

Answer ALL questions.

1. Below is a back to back stem and leaf plot comparing reaction times (in hundredths of a second) for two groups of volunteers (denoted by A and B). There are 50 individuals in each group.

A	Group	B
1	4	111222457899
988443	5	00123345566666777899
9999886665422200	6	03445578
9976665432211100	7	000238
953322100	8	03
63	9	6
	10	6

Leaf Unit = 1·0

(a) Calculate the median reaction time for both groups.

(b) Without performing any further calculations, state which of the following statements is correct for group B:

(i) the median is less than the mean;

(ii) the median is about equal to the mean;

(iii) the median is greater than the mean.

(c) Without performing any further calculations, state which of the following numbers is approximately equal to the standard deviation of group A: 2·1, 3·4, 5·3, 6·8 or 10·7.

(d) Calculate the quartiles for both groups and suggest a suitable graphical method to compare the quartiles of groups A and B. **[9]**

2. In an introductory class to statistics, the number of males and females is as shown in the frequency table below

Gender	Frequency
Male	17
Female	23
Total	40

(a) A student is selected at random. What is the probability the selected student is a female?

(b) A student is selected at random and removed from the class. A second student is then selected. Find the probability that the first student selected is female and the second is male.

Draw a tree diagram with the probabilities associated with all possible outcomes. **[7]**

3. According to the *Daily Racing Form* the probability is about 0.60 that the favourite in a horse race will finish in the money (first, second or third place).

(a) Consider the random variable X, the number of times the favourite finishes in the money in the next four races. What is the distribution of X?

(b) Tabulate the probability distribution of X.

(c) What is the probability that the favourite finishes in the money:

(i) exactly twice?

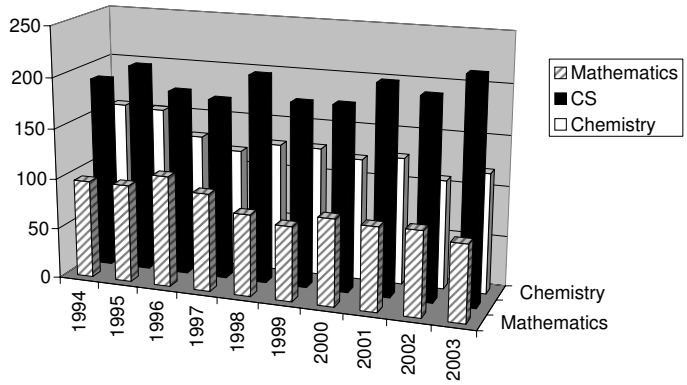
(ii) exactly four times?

(iii) at least one time?

(iv) between two and four times inclusive?

(d) Calculate the mean and variance of X. **[11]**

4. In a meeting, a Scottish Executive minister shows the following barchart illustrating the numbers of students applying to study maths, computer science and chemistry over a 10 year period.



(a) List at least two problems with this graph.

(b) Show how the data can be displayed in a more appropriate way graphically. **[4]**

5. A company that produces snack foods uses a machine to package 454 gms bags of pretzels. We assume that the net weights are normally distributed and that the population standard deviation of all such weights is 7.8 gms.

(a) Calculate the probability that a randomly selected bag will weight:

(i) more that 460 gms,

(ii) between 430 and 450 gms,

(iii) exactly 455 gms.

(b) Consider a simple random sample of 25 bags of pretzels with mean weight equal to 450 gms. Does this data provide sufficient evidence to conclude that the packing machine is not working properly? Formulate the hypothesis and calculate the p-value of the test. State conclusions. **[11]**

6. Neurosurgeons want to determine whether a dynamic system reduces the operative time relative to a static system. The data collected for this is presented below:

Dynamic							Static		
370	360	510	445	295	315	490	430	445	455
345	450	505	335	280	325	500	455	490	535

(a) Do the data provide sufficient evidence to conclude that the mean operative system have different mean times? State conclusions for level of significance $\alpha = 5\%$.

(b) State the assumptions you have made in your test in (a). **[7]**

continued...

7. The table below summarises data on health improvement of 60 patients assigned to a new treatment and 40 patients assigned to a standard treatment.

	improved	not improved	**Total**
new treatment	40		
standard treatment		15	
Total			

(a) Complete the table above and explain how the frequencies of each cell are calculated under the assumption of independence between health improvement and type of treatment.

A chi-squared test of independence between health improvement and type of treatment was performed. Consider the results below:

```
Chi-Square Test

Expected counts:
        improved    not improved
    1       39            21
    2       26            14


Chi-Sq = 0.0064 + <a> + 0.00962 + 0.0179 = <b>
DF = <c>
```

(b) Which test statistic was calculated here? Explain.

(c) Showing necessary working, find the missing values `<a>, <b> and <c>`. Note that `<a>` refers to cell 'not improved' and 'standard treatment'.

(d) Is there any evidence of an association between health improvement and type of treatment? Report the results of this test briefly in a non-technical way. **[10]**

8. A random sample of nine custom homes currently for sale provided the following information on size and price:

size (x)	26	27	33	29	29	34	30	40	22
price (y)	290	305	325	327	356	411	488	554	246

$\sum_{i=1}^{9} x_i = 270$, $\sum_{i=1}^{9} y_i = 3302$, $\sum_{i=1}^{9} x_i y_i = 102493$, $\sum_{i=1}^{9} {x_i}^2 = 8316$, $\sum_{i=1}^{9} {y_i}^2 = 1290912$.

(a) Determine the regression equation for the data.

(b) Calculate the coefficient of determination R^2 and describe the apparent relationship between size and price of the custom homes.

(c) Do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0?

(d) Calculate the 95% confidence interval for the population slope β. State conclusions. **[11]**