Inf1B Data and Analysis
Tutorial 8 (week 10)

13 March 2009

- Please answer all questions on this worksheet in advance of the tutorial, and bring with you all work. Tutorials cannot function properly unless you do the work in advance.
- Data & Analysis tutorial exercises are not assessed, but are a compulsory and important part of the course. If you do not do the exercises then you are unlikely to pass the exam.
- Attendance at tutorials is obligatory; please let your tutor know if you cannot attend.
- *Background Reading:* Slides for Parts IV and V.

Question 1: Information Retrieval

You are looking for a document on `Economic Recession in Scotland` in a huge collection of documents. Incidentally, you decide to search using the terms: **economy**, **recession**, **Scotland**, **banks** and **business** using an *information retrieval system* and you find three possible documents. You are given the frequency of each of the terms in each document, as shown below:

Terms	economy	Scotland	recession	banks	business
Document 1	10	8	0	2	1
Document 2	0	0	9	9	8
Document 3	2	2	4	4	6
Query	1	1	1	1	1

(a) Which measure would you use with this information to determine which of the 3 documents is the best match for the query?

(b) Compute the measure provided in (a) for all three documents.

(c) Based on your results of (b), which document is the best match? Why?

(d) Do you agree with the results of this analysis? What are the strengths and weaknesses of the measure you used?

Question 2: Statistical Analysis of Data

Refer to the on-line data file `data.pdf` on the Data & Analysis course homepage.

(a) Extract a sample of of 5–10 students from this data.

(b) Based on your sample, estimate the mean, and standard deviations for both weekly alcohol consumption and weekly exercise hours.

(c) Based on your sample, estimate the Pearson correlation coefficient to compare weekly alcohol consumption and weekly exercise hours. Is there a significant correlation? Is it positive or negative?