Informatics 1B
Data and Analysis Lab 3: Unstructured Data

Manuel Marques Pita
Edited by Gaya Nadarajan

for Week 10 beginning March 12, 2007

1 MATLAB

The goal of this lab session is to give you an introduction to MATLAB. This application has a large number of different functionalities. In this course it will be used for doing numerical computations on matrices and vectors, compute basic statistics and represent information graphically.

Please download the tutorial available at

`http://www.inf.ed.ac.uk/teaching/courses/inf1/da/labs/primer40.pdf`

and work through it, focusing on sections 1 to 12. If you decide to print this out, please print it from the command line using the command

```
lpr -Z2up -Pat8 primer40.pdf
```

More on-line help could be found in the 'Matrices and Arrays' section of the MathWorks support page:

`http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.html`

To start MATLAB from your terminal, type the following from your command prompt:

```
matlab &
```

Once you have worked through the MATLAB tutorial, proceed to work on the following exercises

2 Basic Chi-Squared in MatLab

For a 2x2 matrix containing the rows `((A,B);(C,D))`, χ^2 is calculated as

$$\chi^2 = \frac{(AD-BC)^2 N}{(A+B)(C+D)(A+C)(B+D)}$$

Where N = A + B + C + D.

Question 1: Provide the code to calculate this statistical value for a given 2x2 matrix in MATLAB.
[Hint: Create and instantiate variables A, B, C, D with some given values before providing the formula for χ^2]

3 Proportions

When gathering data for the study of the significance of some variable, we will sometimes be interested in *ratios* and *relative proportions.* Consider the following example,

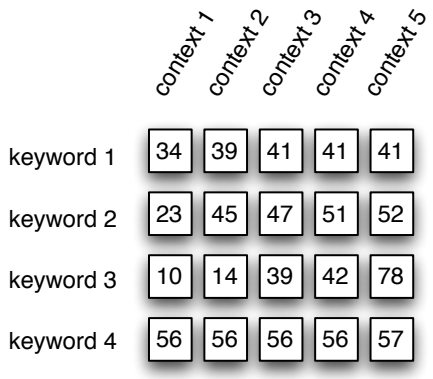

Figure 1: Collected data about occurrences of four keywords in five context

Figure 1 shows the results of collecting some data concerning the number of occurrences of certain keywords in a number of different contexts. There are four different keywords and five contexts. The contexts are not independent of each other: context 1 is a subset of context 2, context 2 is a subset of context 3 and so on. Context 5 is the superset that contains all the other contexts.

Question 2: Taking this into account, what is the total number of *observations* for each keyword? How would you code this in MATLAB?

Question 3: What are the relative proportions of each keyword within each context? (calculate the proportions in MATLAB).

[Hint: Use loops and matrices/arrays where necessary]

Question 4: Looking at the proportions you have calculated, what can you say about the relationship between different keywords and contexts?