# Inf1B Data and Analysis
# Tutorial 9 (week 11)
# Collocations and Statistical Analysis

Hutchins-Korte and Simpson

12 March 2008

- 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:* lecture notes 8, 10 and 13.

Consider the following text:

> *Toe nail a 2.5 inch screw through front and back edge of leg into bottom of arm rest. Screw a 3 inch screw through the triangle block into bottom of arm rest. Do the same with the left arm rest and left front leg. Wipe up any glue seepage with wet cloth. Peel tape off that is attached to inside surface of both legs.*

From "Assembly Instruction for Adirondack Chair"

## Question 1

**(a)** Tabulate the bigram frequencies for this text, for all bigram frequencies strictly greater than 1, in an ordered list.

**(b)** Let bigram $A$ be 'arm rest' and bigram $B$ be 'of arm'. Based on your knowledge of English, which of these, if either, would you say is a collocation?

## Question 2

In this question, we use the $\chi^2$-test, based on the quoted text above, to test the significance of bigrams $A$ and $B$ as collocations.

**(a)** Compile contingency tables for each of the bigrams $A$ and $B$.

**(b)** Describe what each of the four entries in the contingency table for bigram $A$ represents.

**(c)** Compile expected frequency tables for bigrams $A$ and $B$.

**(d)** Calculate $\chi^2$-values for bigrams $A$ and $B$.

**(e)** You are told that the critical value, with 1 degree of freedom, for the $\chi^2$ test is 3.84, for $p = 0.05$, and 6.64, for $p = 0.01$. How would you interpret your results from (d)?

**(f)** What criticisms are applicable to the method applied above as a test for whether bigrams $A$ and $B$ are collocations? Would these criticisms still be valid if the test had been carried out on the entire British National Corpus rather than on the quoted text?