Informatics 1 Data & Analysis
Tutorial 8

Week 11, Semester 2, 2012

- Please attempt question 1 of this worksheet in advance of the tutorial, and bring with you all work. Tutorials cannot function properly unless you do the work in advance. Question 2 will be solved as an exercise in the tutorial itself.
- Data & Analysis tutorial exercises are not assessed, but they 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: if you are ill or otherwise unable to attend one week then email your tutor, and if possible attend another tutorial group in the same week.

Introduction

In this tutorial we will perform statistical analysis over data on students' nationality, physical exercise and alcohol consumption. The data for this tutorial was collected in the first DA lecture on 17th January using an anonymous questionnaire asking students for their: hours of physical exercise in the previous week; units of alcohol consumed in the previous week; whether of UK/Ireland, EU (and not UK/Ireland) or Other nationality.

Reading

For this tutorial, you will need to carry out two specific statistical tests. Look at the following sets of lecture handouts for information about how to do these.

- Pearson's correlation coefficient: Unstructured Data, note 3
- χ^2 (chi-squared) tests: Unstructured Data, note 4

These are available from the course web page. The lecture on χ^2 tests will given on Tuesday 27th March, just before the tutorials. So the question involving this test is being set as an exercise to carry out in tutorials themselves.

You will also need suitable tables of critical values and significance levels. For the χ^2 test, see the lecture slides. For Pearson's correlation coefficient, see the link on the course web page.

Question 1: Statistical analysis of numerical data

This question is to be done in advance of your tutorial. Download the file `data.pdf` from the course homepage. This contains the results of the anonymous questionnaire.

(a) Extract a sample of 10–15 students from this data.

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

(c) Draw a scatter plot showing the units of alcohol and exercise hours for each student in your sample. Visually, does there appear to be any correlation between units of alcohol and exercise hours? If so, is it positive or negative?

(d) Based on your sample, estimate the correlation coefficient between daily units of alcohol and weekly exercise hours for Informatics 1 students. Is there a significant correlation? Is it positive or negative?

Question 2: Statistical analysis of categorical data

This question is to be done in the tutorial. The following are some statistics complied from the data in the file data.pdf

UK/Ireland students who drank some alcohol in the last week	32
UK/Ireland students who drank no alcohol in the last week	12
Foreign students who drank some alcohol in the last week	44
Foreign students who drank no alcohol in the last week	24
UK/Ireland students who drank at least 8 units of alcohol in the last week	23
UK/Ireland students who drank less than 8 units of alcohol in the last week	21
Foreign students who drank at least 8 units of alcohol in the last week	20
Foreign students who drank less than 8 units of alcohol in the last week	48

Here "Foreign students" combines the EU(not UK/Ireland) and Other students.

The purpose of this question is to apply the χ^2 test to investigate the correlation between: (i) nationality (either UK/Ire or foreign) and consumption of some alcohol; and (ii) nationality and the consumption of at least 8 units of alcohol. (The figure 8 has been chosen because it corresponds to drinking an average of more than 1 unit of alcohol per day.)

Split the tutorial group into two teams of students. One team's task is to investigate (i), the other's task is to investigate (ii). Each group should carry out the following steps, using the relevant data.

(a) Compile the contingency tables relevant to (a) or (b), as appropriate.

(b) Calculate the corresponding tables of expected frequencies.

(c) Calculate the corresponding χ^2 values.

(d) Is the χ^2 test reliable? If yes, is there a correlation? At what significance level?

As a group as a whole, address the following questions.

(e) Estimate the mean of the units of alcohol consumed by UK/Ireland students in the previous week and the mean of the units of alcohol consumed by foreign students.

(f) Which information do you find more informative: the answer to question (d) or the answer to question (e)?

(g) Discuss whether the data in the file data.pdf appears to show any correlation between nationality and amount of exercise.