Informatics 1 Data & Analysis
Tutorial 8

Week 10, Semester 2, 2013

- You must prepare for the tutorial by attempting the questions on this worksheet in advance. Bring with you a copy of your work, including printouts of code and other results.

  If you cannot do some questions, write down what it is that you find challenging and use this to ask your tutor in the meeting.

  It's important both for your learning and other students in the group that you come to tutorials properly prepared. If you have not attempted the exercise sheet, then you may be sent away from the tutorial to do it elsewhere.

- Some exercise sheets contain material marked with a star ⋆. These are optional extensions.

- 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 sleep. The data for this tutorial was collected in the first Inf1-DA lecture this semester using an anonymous questionnaire. This asked students to estimate their average hours of physical exercise per week; estimated hours of sleep the previous night; and to indicate their nationality among various categories.

Background

For this tutorial, you will need to carry out specific statistical tests.

- Estimation of population mean and variance from a sample.
- Pearson's correlation coefficient.
- χ^2 test of significance.

You can find lecture slides presenting these on the course web page.

You will also need the following tables: significance levels for the χ^2 distribution and critical values for Pearson's correlation coefficient ρ. These show p-values (0.10 to 0.001) against degrees of freedom (1 to 4, for χ^2) and sample size (7 to 10, for ρ).

χ^2	0.10	0.05	0.01	0.001
1	2.71	3.84	6.64	10.83
2	4.60	5.99	9.21	13.82
3	6.25	7.82	11.34	16.27
4	7.78	9.49	13.28	18.47

ρ	0.10	0.05	0.01	0.001
7	0.669	0.754	0.875	0.951
8	0.621	0.707	0.834	0.925
9	0.582	0.666	0.798	0.898
10	0.549	0.632	0.765	0.872

Question 1: Statistical analysis of numerical data

Download the file data.pdf from the course homepage. This contains the results of the anonymous questionnaire.

(a) Extract a random sample of 8 students from this data.

(b) Based on your sample, estimate the population mean and standard deviation for both daily sleep and weekly exercise hours.

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

(d) Based on your sample, estimate the correlation coefficient between daily sleep 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

The following are some statistics from the the file data.pdf

EU students who exercise at least 2.5 hours per week	109
EU students who exercise less than 2.5 hours per week	21
Not EU students who exercise at least 2.5 hours per week	8
Not EU students who exercise less than 2.5 hours per week	6
EU students who exercise at least 5 hours per week	69
EU students who exercise less than 5 hours per week	62
Not EU students who exercise at least 5 hours per week	7
Not EU students who exercise less than 5 hours per week	7

Here “EU students” here combines the ’Scottish’, ’UK, not Scottish’ and ’EU, not UK’ categories; while ’not EU’ contains all other nationalities.

(a) Compile the relevant contingency tables to investigate correlation between:

- Nationality and exercising at least 2.5 hours per week;
- Nationality and exercising at least 5 hours per week.

(b) Calculate the corresponding tables of expected frequencies.

(c) Calculate the corresponding χ^2 values.

(d) Are the two χ^2 tests reliable? If yes, are there correlations? At what significance levels?

(e) Using two samples of 8 students each, estimate the mean weekly exercise of EU students and the mean weekly exercise of other students.

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

⋆ **(g)** Revisit the data file and look for a correlation between EU/not EU and reporting 7 hours sleep or less, or more than 7 hours sleep, in the 24 hours before the survey.