Tutorial 8: Statistical Analysis

Informatics 1 Data & Analysis

Week 10, Semester 2, 2013–2014

This worksheet has three parts: tutorial Questions, followed by some Examples and their Solutions.

- Before your tutorial, work through and attempt all of the Questions in the first section.
- The Examples are there for additional preparation, practice, and revision.
- Use the Solutions to check your answers, and read about possible alternatives.

You must bring your answers to the main questions along to your tutorial. You will need to be able to show these to your tutor, and may be exchanging them with other students, so it is best to have them printed out on paper.

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.

Tutorials will not usually cover the Examples, but if you have any questions about those then write them down and ask your tutor, or go along to InfBASE during the week.

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 main tutorial questions, then you may be sent away from the tutorial to do them elsewhere.

Some exercise sheets contain material marked with a star $\star$. These are optional extensions.

Data & Analysis tutorials are not formally 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.

Please send any corrections and suggestions to Ian.Stark@ed.ac.uk

Introduction

In this tutorial you will perform statistical analysis of students' physical exercise, sleep and operating system of choice. This data was collected in the second Inf1-DA lecture this semester using an anonymous questionnaire. That asked students to estimate their average hours of physical exercise per week; hours of sleep the previous night; and to indicate the main operating system they used.

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, calculate estimates for the mean and standard deviation for both daily sleep and weekly exercise hours among all students in the survey.

(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 all the students surveyed. 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

OS X users who exercise at least 7 hours per week	10
OS X users who exercise less than 7 hours per week	14
Non OS X users who exercise at least 7 hours per week	20
Non OS X users who exercise less than 7 hours per week	84
OS X users who exercise at least 10 hours per week	4
OS X users who exercise less than 10 hours per week	20
Non OS X users who exercise at least 10 hours per week	10
Non OS X users who exercise less than 10 hours per week	94

"Non OS X users" here combines the user categories of 'Microsoft Windows', 'Linux', and 'None'.

(a) Compile contingency tables based on these figures to investigate possible correlation between:

- Operating system of choice and exercising at least 7 hours per week;
- Operating system of choice and exercising at least 10 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 OS X users 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 OS X use and reporting 8 hours sleep or less, or more than 8 hours sleep, in the 24 hours before the survey.

Examples

This section contains further exercises on Statistical Analysis. These examples are similar to the main tutorial questions: they involve analysing numerical and categorical data with the use of different statistics, as well as assessing possible correlations through hypothesis testing.

Example 1: Numerical data

A statistical study of former students, 10 years after leaving university, seeks to investigate whether there is any correlation between current salary and exam performance when at university.

(a) What general guidelines should be followed in choosing a sample from the population of former students over which to investigate the correlation? Explain the purpose of these guidelines.

(b) In the event, data is gathered from a sample of 100 former students. The annual salaries are represented as values $x_1, x_2, \ldots, x_{100}$. The corresponding degree marks (as percentages) are represented as values $y_1, y_2, \ldots, y_{100}$. The correlation between salaries and degree marks is to be investigated using Pearson's correlation coefficient, $r_{x,y}$, for which the formula is:

$$\frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{(n-1)s_x s_y}$$

(i) Explain what the symbols n, m_x and s_x stand for in the above formula.

(ii) Give the formulas used to calculate m_x and s_x.

(c) The result of the calculation of $r_{x,y}$, for the data gathered, is 0.270 (to 3 decimal places). The critical values table for Pearson's correlation coefficient (two-tailed test) contains the following entry for $n = 100$.

n	p = 0.1	p=0.05	p = 0.01	p=0.001
100	0.185	0.197	0.256	0.324

Explain in detail what we can conclude about the existence of a correlation in the population between degree performance and salary.

Example 2: Categorical data

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Sold
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

The company plan to use a χ^2 test to investigate whether colour affects sales.

(a) What is the *null hypothesis* for this investigation?

(b) Calculate the table of expected frequencies of sales in each colour.

(c) Give the formula for calculating the χ^2 statistic. Compute χ^2 for the sales data, showing your working.

(d) In this test the data has 3 *degrees of freedom*. Explain what this means.

(e) The critical values for the χ^2 test with three degrees of freedom are as follows.

p	0.1	0.05	0.01	0.001
χ^2	6.25	7.81	11.35	16.27

Based on this information, what can you conclude about selling widgets in coloured boxes?

Example 3: Numerical data

Five CPUs are randomly selected from a batch of 1000 for thermal testing. All are tested at increasingly higher temperatures until they failed, at the following temperatures: $99°, 95°, 92°, 104°$ and $120°$

Compute estimates of the mean and standard deviation of the failure temperatures for the whole batch of CPUs. Show your calculations.

Solutions to Examples

These are not entirely "model" answers; instead, they indicate a possible solution. Remember that not all of these questions will have a single "right" answer. If you have difficulties with a particular example, or have trouble following through the solution, please raise this as a question in your tutorial.

Solution 1

(a) The sample should be small enough that gathering the data is feasible. It should be large enough that analysis of the sample is likely to produce informative results. It should be randomly selected to avoid bias in the sample.

(b) (i) n is the size of the sample, which in this case is 100
m_x is the estimate of the mean of the x values based on the sample
s_x is the estimate of the standard deviation of the y values based on the sample

(ii) The formulas for m_x and s_x respectively are as follows.

$$m_x = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - m_x)^2}{n-1}}$$

(c) Since the value 0.270 is positive, we have detected a positive correlation between salary and exam marks in our data.

Were there no correlation between salary and exam marks in the population (i.e., were the null hypothesis true) the probability of obtaining a value with modulus greater than 0.256 would be 0.01. We thus conclude, with significance $p < 0.01$, that there is likely to be a positive correlation in the population

Since the value 0.270 is less that 0.324, the significance level of $p < 0.001$ is not applicable.

Solution 2

(a) The null hypothesis is that box colour makes no difference to widget sales.

(b) Under the null hypothesis, we expect all frequencies to be equal. The frequency for each colour is the total number sold (1000) divided by the number of colours (4). This gives the following table.

Colour	Sold
Red	250
Yellow	250
Green	250
Blue	250
Total	1000

(c) The χ^2 statistic is computed as follows:

$$\begin{aligned}\chi^2 &= \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \\ &= \frac{15^2}{250} + \frac{25^2}{250} + \frac{15^2}{250} + \frac{25^2}{250} \\ &= 6.8\end{aligned}$$

(d) The only restriction on the four values in the table is that they must add up to the marginal total of 1000. This means that three can be arbitrary, but the fourth is then determined. These are the three degrees of freedom.

(e) The computed χ^2 value of 6.8 lies above the 95% significance level for that statistic. This gives us confidence in rejecting the null hypothesis, and deducing that box colour does affect widget sales.

Solution 3

$$\text{Mean estimator } m = \frac{99 + 95 + 92 + 104 + 120}{5} = 102$$

$$\text{Standard deviation estimator } s = \sqrt{\frac{(99-102)^2 + (95-102)^2 + (92-102)^2 + (104-102)^2 + (120-102)^2}{5-1}}$$

$$= 11.0$$

Notice the denominator of $(5 - 1)$ in the estimate of standard deviation. In this case, the estimate of population deviation, 11.0, is clearly different to the standard deviation of the sample itself, which is 9.86.