Tutorial 9: Statistical Analysis

Informatics 1 Data & Analysis

Week 11, Semester 2, 2017/18

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. If you get stuck or need help then ask a question on *Piazza*.
- 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 post a question on *Piazza*.

It's important both for your learning and other students in the group that you come to tutorials properly prepared. Students who have not attempted the main tutorial questions will be sent away from the tutorial to do them elsewhere and return later.

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

Data & Analysis tutorials are not formally assessed, but the content is examinable and they are an 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 from Inf1-DA students in 2017 using an anonymous questionnaire. That asked them to estimate their average hours of physical exercise in a week; hours of sleep the previous night; and to indicate the main operating system used.

You will need to carry out specific statistical tests on this data.

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

You will also need the following tables: the first shows significance levels for the χ^2 distribution; and the second some 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 survey.pdf from the course web pages. This contains the results of the anonymous questionnaire.

The file has information from a population of 190 students. In this question you are going to simulate the situation where information about the whole dataset is unknown and you need to estimate it from a small sample. It's quite likely that each person in your tutorial group will come up with different estimates as you each take a different sample.

(a) Extract a random sample of 8 students from this data. How did you choose a "random sample"?

(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) Use your sample to 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 survey.pdf file.

Linux users who slept 8 hours or more the previous night	3
Linux users who slept less than 8 hours the previous night	20
Non-Linux users who slept 8 hours or more the previous night	66
Non-Linux users who slept less than 8 hours the previous night	101
Students who exercise ≥ 11 hours and who slept ≥ 6 hours the previous night	7
Students who exercise < 11 hours and who slept ≥ 6 hours the previous night	165
Students who exercise ≥ 11 hours and who slept < 6 hours the previous night	3
Students who exercise < 11 hours and who slept < 6 hours the previous night	15

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

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

- Using Linux and sleeping at least 8 hours the previous night;
- Exercising at least 11 hours per week and sleeping at least 6 hours the previous night.

(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 hours of sleep of Linux users and the mean hours of sleep of other students.

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

$\star$ **(g)** Revisit the data file and look for any evidence of correlation between choice of operating system — OS X, Windows, or something else — and reporting 7.5 hours or more of sleep the night before. For this you need to go through the survey.pdf file and build up a single contingency table with two rows and three columns. How many degrees of freedom does this table have? Carry out a χ^2 test to explore whether there is evidence of a correlation here.

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: 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.

Example 3: 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

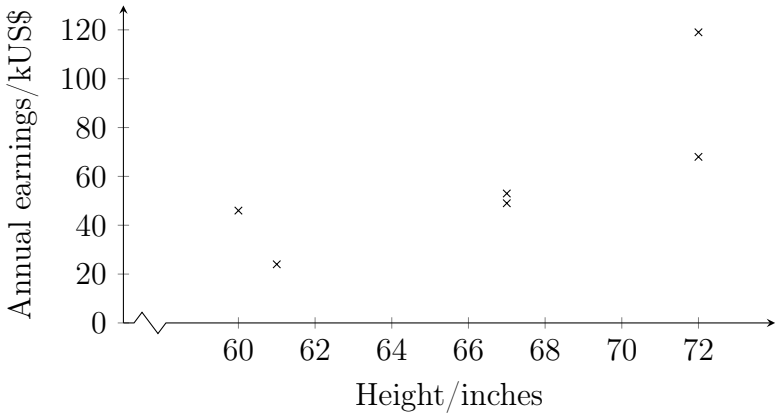
Based on this information, what can you conclude about selling widgets in coloured boxes?

Example 4: Correlation and causation

Suppose that the following data has been collected in a small survey to explore a possible relationship between people's physical height and their annual earnings.

Participant	A	B	C	D	E	F
Height/in	67	72	61	72	60	67
Earnings/kUSD	49	119	24	68	46	53

(Data based on Judge & Cable, 2004; earnings adjusted to 2002 US dollars)



(a) Calculate the mean and standard deviation of height and earnings for these six people.

(b) The six survey participants have been selected at random from a much larger population. Calculate estimates for the mean and standard deviation of the whole population.

(c) The following equation gives an estimate for the Pearson's correlation coefficient of a whole population based on sample values.

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

Use this to estimate the correlation coefficient between height and earnings in the population from which the sample above was taken.

(d) You are asked to test the hypothesis that there is a positive correlation between height and earnings. Use your answer from (c) and the table below to answer the following questions, in each case explaining how you arrive at your answer:

(i) Does this sample show positive correlation between height and earnings?

(ii) Is it statistically significant?

Critical values for Pearson's correlation

two-tail	$p = 0.20$	$p = 0.10$	$p = 0.01$	$p = 0.001$
one-tail	$p = 0.10$	$p = 0.05$	$p = 0.005$	$p = 0.0005$
$N = 4$	0.800	0.900	0.990	0.999
$N = 5$	0.687	0.805	0.959	0.991
$N = 6$	0.608	0.729	0.917	0.974
$N = 7$	0.551	0.669	0.875	0.951

(e) In an actual sample of over 4000 people from the US National Longitudinal Survey, reported in a 2004 meta-analysis by Judge & Cable, there was an observed correlation between height and earnings, with statistical significance at the 99% level.

This might suggest a causal relationship between height and earnings. Give three distinct kinds of causal dependency which would lead to an observed correlation between these two factors.

Judge, Timothy A. and Cable, Daniel M. *The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model.* Journal of Applied Psychology, Vol. 89(3), June 2004, 428–441. https://doi.org/10.1037/0021-9010.89.3.428

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 x 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

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

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.

Solution 3

(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 90% significance level for that statistic. This gives us some confidence in rejecting the null hypothesis, and deducing that box colour does affect widget sales.

Solution 4

(a) For this sample of six people, their heights have mean 66.5 inches and standard deviation 4.72 inches; their earnings have mean $\$59,800$ and standard deviation $\$29,500$.

(b) The sample means are appropriate estimators of the population mean height 66.5 inches and mean earnings $\$59,800$.

Appropriate estimates for the population standard deviation are 5.18 inches height and $\$32,300$ earnings. This is using an $(n-1)$ denominator to account for the fact that we are using a small sample to estimate the value for a larger population.

(c) The estimate of correlation coefficient in the population as a whole is +0.78.

(d) **(i)** Yes, the sample does show a positive correlation, as the coefficient is greater than zero.

(ii) Yes, this is statistically significant, at the 95% level, as it exceeds the critical one-tail value for $p = 0.05$ over $N = 6$ samples.

(e) Each of the following causal dependencies could lead to an observed correlation:

- If earnings influence height (for example, over the long term wealth might be linked to better diet and growth).

- If height influences earnings (for example, if employers are more inclined to hire and promote taller people).
- If some other factor influences both height and earnings (for example, the wealth of the state someone lives in).

One factor that might be an influence on both height and earnings is an individual's sex. In fact, the researchers in this paper had already controlled for this in their analysis: the correlation given was calculated after removing any impact of sex.