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.
- $\chi^2$ test of significance.

You will also need the following tables: the first shows significance levels for the $\chi^2$ distribution; and the second some critical values for Pearson’s correlation coefficient $\rho$. These show $p$-values (0.10 to 0.001) against degrees of freedom (1 to 4, for $\chi^2$) and sample size (7 to 10, for $\rho$).

<table>
<thead>
<tr>
<th>$\chi^2$</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.71</td>
<td>3.84</td>
<td>6.64</td>
<td>10.83</td>
</tr>
<tr>
<td>2</td>
<td>4.60</td>
<td>5.99</td>
<td>9.21</td>
<td>13.82</td>
</tr>
<tr>
<td>3</td>
<td>6.25</td>
<td>7.82</td>
<td>11.34</td>
<td>16.27</td>
</tr>
<tr>
<td>4</td>
<td>7.78</td>
<td>9.49</td>
<td>13.28</td>
<td>18.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0.669</td>
<td>0.754</td>
<td>0.875</td>
<td>0.951</td>
</tr>
<tr>
<td>8</td>
<td>0.621</td>
<td>0.707</td>
<td>0.834</td>
<td>0.925</td>
</tr>
<tr>
<td>9</td>
<td>0.582</td>
<td>0.666</td>
<td>0.798</td>
<td>0.898</td>
</tr>
<tr>
<td>10</td>
<td>0.549</td>
<td>0.632</td>
<td>0.765</td>
<td>0.872</td>
</tr>
</tbody>
</table>

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”?

(Tutor Notes) The samples should be chosen randomly without replacement. Best of all would be to use some explicit randomisation technique for each of the eight choices. For less effort, one might choose a random entry and then take the next 8 from there. That introduces a small possibility of bias: adjacent submissions might be from students who sit close by in the lecture theatre, for example. Picking just the first or last eight entries can be biased for similar reasons.

Sometimes people actively try to picking a group that contains a good mix of values — say, some Windows, some Linux, some Mac. That’s absolutely not random in this setting, because it’s applying a selection criteria based on the data content. Estimates about the overall population made using this sample will in general not be as reliable as those selected randomly.

In more elaborate statistical tests it can be appropriate to try that kind of intervention so that a sample has a certain characteristic — for example, to match some known gender balance of people in the population — but that’s not the case here.

(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.
The actual statistics for the entire data collection are: exercise mean $\mu$: 4.74; exercise standard deviation $\sigma$: 3.50; sleep mean $\mu$: 7.12; sleep standard deviation $\sigma$: 1.56.

The values given by sampling will depend on the particular choice of sample. The important point is that the standard deviation estimator should be calculated using the formula with denominator $(n-1)$; the questions explicitly specifies to work with a sample rather than the whole dataset. Given the small size of the samples, there may well be quite large discrepancies between the estimates given by the samples and the actual population statistics.

Even the whole population here is itself just a sample, of students who filled out the form. It is worth reflecting on whether it is reasonable to treat this as a sample from some larger population: for example, undergraduate students at Edinburgh. There will be issues of representativeness there — this is only Informatics students, in their first year, who come to lectures at the start of semester, and voluntarily fill in questionnaires.

(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?

(Tutor Notes) This depends very much on the sample chosen, and the visual appearance of the scatter plot. It’s worth writing down any apparent correlation now, though, for comparison with the mathematical analysis in the next question.

(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?

(Tutor Notes) For the whole data set, there is a correlation coefficient of $-0.0111$. This was calculated using the $n$ version of the formula; when estimating this from a smaller sample one should use the $(n-1)$ version. Because samples are small, there will again be discrepancies between the estimated answers and the actual one.

The population size here is 190 and the absolute value of our correlation coefficient is 0.0111. The critical values table for the correlation coefficient identifies values of 0.197 and 0.139 as having significance $p = 0.05$ for $N = 100$ and $N = 200$ respectively. Thus, the significance of our test does not reach the 0.05 significance mark for $N = 190$. This is still true for a lower significance mark of $p = 0.1$ (0.165 and 0.117 for $N = 100$ and $N = 200$, respectively). Hence, there is no evidence of a correlation between exercise and sleep.

An important point is to make sure you refer to the row in the critical value table that corresponds to the sample size, not the row corresponding to the overall population size.

It may be that some samples do suggest a correlation or even a correlation with high significance. In this case it is worth thinking whether this means there “really is” such a correlation.

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


(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.

(Tutor Notes) Here are the contingency tables together with marginals:

<table>
<thead>
<tr>
<th>Sleep ≥ 8 hours</th>
<th>Linux users</th>
<th>Others</th>
<th>Exercise ≥ 11 hours</th>
<th>Sleep ≥ 6</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep ≥ 8 hours</td>
<td>3</td>
<td>66</td>
<td>69</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Sleep &lt; 8 hours</td>
<td>20</td>
<td>101</td>
<td>121</td>
<td>165</td>
<td>15</td>
</tr>
</tbody>
</table>

(b) Calculate the corresponding tables of expected frequencies.

(Tutor Notes) Based on the marginal values, these are the expected frequencies to 2 decimal places:

<table>
<thead>
<tr>
<th>Sleep ≥ 8 hours</th>
<th>Linux users</th>
<th>Others</th>
<th>Sleep ≥ 6</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep ≥ 8 hours</td>
<td>8.35</td>
<td>60.65</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Sleep &lt; 8 hours</td>
<td>14.65</td>
<td>106.35</td>
<td>121</td>
<td></td>
</tr>
</tbody>
</table>

(c) Calculate the corresponding $\chi^2$ values.
(Tutor Notes) $\chi^2$ value for Linux and hours of sleep:

$$\chi^2 = \frac{(-5.35)^2}{8.35} + \frac{5.35^2}{60.65} + \frac{5.35^2}{14.65} + \frac{(-5.35)^2}{106.35}$$

$$= 6.13$$

$\chi^2$ value for hours of exercise and hours of sleep:

$$\chi^2 = \frac{(-2.05)^2}{9.05} + \frac{2.05^2}{0.95} + \frac{2.05^2}{162.95} + \frac{(-2.05)^2}{17.05}$$

$$= 5.19$$

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

(Tutor Notes) The first $\chi^2$ test is reliable, as all expected frequencies in the table of part (b) are above 5. Note that this holds even though one of the observed frequencies (Linux users sleeping 8 or more hours) is below 5. The value obtained suggests a correlation above the 95% significance level between using Linux and sleeping less than 8 hours a night.

These tables have 1 degree of freedom, so the first line of the $\chi^2$ table is the correct one to use. That lists 3.84 as the 95% significance level, and the computed $\chi^2$ statistic of 6.13 is above that.

The value obtained for the second $\chi^2$ test (5.19) is above the 95% significance level, which could suggest a correlation between exercising 11 or more hours per week and sleeping less than 6 hours per night. However, this test is not reliable because in the expected frequencies table of part (b) not all the four expected values are 5 or higher. So we can in fact deduce nothing at all.

(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.

(Tutor Notes) The actual mean values are:

Linux users mean sleep last night 6.76 hours
Non Linux users mean sleep last night 7.17 hours

(f) Which information do you find more informative: the answer to question (d) or the answer to question (e)? For what reasons?
The answer to question (d) is slightly informative: one part is reliable and suggests a correlation; while the other is not reliable due to small numbers. However, even in the first part, suggesting a possible correlation, there is no measure of how much difference this makes: the effect size.

In question (e) the magnitude of the difference in the means between the two data sets is relatively small — less than one third of the standard deviation $\sigma$. This also suggests that Linux and non Linux users sleep for roughly the same amount of hours on average, with Linux users sleeping slightly less. This measure isn’t able to say whether the difference we have found is “real” — just that if it is, then the effect appears fairly small.

The samples of 8 data values chosen at random may turn out to have quite different means: does this indicate a real underlying difference?

There are more sophisticated tests which could compute both statistical significance and effect size to the difference of means in (e).

* (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 $\chi^2$ test to explore whether there is evidence of a correlation here.

(Tutor Notes) Here is the contingency table together with marginals:

<table>
<thead>
<tr>
<th></th>
<th>OS X users</th>
<th>Windows users</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep $\geq$ 7.5 hours</td>
<td>21</td>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td>Sleep $&lt; 7.5$ hours</td>
<td>32</td>
<td>63</td>
<td>22</td>
</tr>
<tr>
<td>$\text{Total}$</td>
<td>53</td>
<td>111</td>
<td>26</td>
</tr>
</tbody>
</table>

Based on the marginal values, these are the expected frequencies to 2 decimal places:

<table>
<thead>
<tr>
<th></th>
<th>OS X users</th>
<th>Windows users</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep $\geq$ 7.5 hours</td>
<td>20.36</td>
<td>42.65</td>
<td>9.99</td>
</tr>
<tr>
<td>Sleep $&lt; 7.5$ hours</td>
<td>32.64</td>
<td>68.35</td>
<td>16.01</td>
</tr>
<tr>
<td>$\text{Total}$</td>
<td>53.00</td>
<td>111.00</td>
<td>26.00</td>
</tr>
</tbody>
</table>

Based on these tables, the $\chi^2$ value is 6.96. These are $(2 \times 3)$-cell tables, with $(2-1) \times (3-1) = 2$ degrees of freedom. Thus the correct line of the $\chi^2$ table to use for evaluating the significance of the values calculated is the second one.

That suggests we have evidence at the 95% level that sleep and operating system choice are correlated. Inspecting the table cells by eye, we might hypothesise that Windows users sleep more than expected, Apple users are in the middle, and others sleep less than most. Or, at least, that’s what they report in ad-hoc anonymous surveys handed round in lectures.

This $\chi^2$ test is reliable, since fewer than 20% of the expected frequencies are below 5 (in fact, none of expected frequencies are below 5, even though one of the observed ones is).
(Tutor Notes) Final Comment on Significance: Some of the tests in this tutorial return statistics that indicate significant correlation. However, please take these with a pinch of salt because of the overarching context that this dataset was searched repeatedly to find anything at all which gave a correlation suitable for writing into a tutorial. There are all sorts of possible tests on the survey data which could have been included in the exercises, but weren’t because they don’t give any hint of a correlation.

This is an instance of a general issue: if you search for twenty different tests, then there’s a good chance one of them will show significance at the 95% level. The problem is known as p-hacking or data dredging and was covered in lectures. In this case, I claim justification as the aim here is to set up a variety of examples for learning the techniques, not to undertake serious investigation into student behaviour. However, when applying statistics for practical ends, it’s important to be aware of the risk of turning up an apparently significant correlation simply by churning through enough possibilities.
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:

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

(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 $n = 100$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p = 0.1$</th>
<th>$p = 0.05$</th>
<th>$p = 0.01$</th>
<th>$p = 0.001$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.185</td>
<td>0.197</td>
<td>0.256</td>
<td>0.324</td>
</tr>
</tbody>
</table>

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: $99^\circ, 95^\circ, 92^\circ, 104^\circ$ and $120^\circ$

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.
<table>
<thead>
<tr>
<th>Colour</th>
<th>Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>235</td>
</tr>
<tr>
<td>Yellow</td>
<td>275</td>
</tr>
<tr>
<td>Green</td>
<td>225</td>
</tr>
<tr>
<td>Blue</td>
<td>265</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
</tr>
</tbody>
</table>

The company plan to use a $\chi^2$ 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 $\chi^2$ statistic. Compute $\chi^2$ 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 $\chi^2$ test with three degrees of freedom are as follows.

<table>
<thead>
<tr>
<th>$p$</th>
<th>0.1</th>
<th>0.05</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>6.25</td>
<td>7.81</td>
<td>11.35</td>
<td>16.27</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Participant</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height/in</td>
<td>67</td>
<td>72</td>
<td>61</td>
<td>72</td>
<td>60</td>
<td>67</td>
</tr>
<tr>
<td>Earnings/kUSD</td>
<td>49</td>
<td>119</td>
<td>24</td>
<td>68</td>
<td>46</td>
<td>53</td>
</tr>
</tbody>
</table>

(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?

(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.

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

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.

<table>
<thead>
<tr>
<th>Colour</th>
<th>Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>250</td>
</tr>
<tr>
<td>Yellow</td>
<td>250</td>
</tr>
<tr>
<td>Green</td>
<td>250</td>
</tr>
<tr>
<td>Blue</td>
<td>250</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
</tr>
</tbody>
</table>

(c) The $\chi^2$ statistic is computed as follows:

$$
\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}
$$

$$
= \frac{15^2}{250} + \frac{25^2}{250} + \frac{15^2}{250} + \frac{25^2}{250}
$$

$$
= 6.8
$$

(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 $\chi^2$ 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.