Introduction to Statistical Analysis

Mark Dunning and Sarah Vowler

Last modified: 03 Nov 2015

Contents

1 Introduction

In this practical, we will use several 'read-life' datasets to demonstrate some of the concepts you have seen in the lectures. We will guide you through how to analyse these datasets in Shiny and the kinds of questions you should be asking yourself when faced with similar data.

To answer the questions in this practical we will be using apps that we have developed using the ***Shiny*** add-on for the *R* statistical package. **R** is a freely-available open-source software that is popular within academic and commercial communities. The functionality within the software compares favourably with other statistical packages (SAS, SPSS and Stata). The downside is that **R** has a steep learning-curve and requires a basic familiarity with command-line software. To ease the transition we have chosen to present this course using a series of online tools that will allow you to perform statistical analysis without having to worry about learning R. At the same time, the R code required for the analysis will be recorded in the background. You will therefore be able to repeat the analysis at a later date, or pass-on to others. As you gain familiarity with R through other courses, you will see how the code generated by Shiny can be adapted to your own needs.

2 T-tests practical

2.1 The effect of disease on height

A scientist knows that the mean height of females in England is 165cm and wants to know whether her patients with disease X have heights that differ significantly from the population mean - we will use a one-sample t-test to test this. The data are contained in the file `diseaseX.csv` and can be analysed online at:-

`http://bioinf-rstud001:3838/OneSampleTest/`

a) What are your null and alternative hypotheses?

To import the file `diseaseX.csv` into ***Shiny*** you will need to select the `Choose File` option and navigate to where the course data are located on your laptop. You can use the ***The data*** tab to check that the data has be imported correctly.

b) A histogram of the `Height` variable will be automatically generate for you. To view it, click on the ***Data Distribution***

Do the data look normally distributed? Based on the histogram, is the one-sample t –test appropriate?

c) We are interested in knowing whether the mean height in our sample of patients with disease X is different from that of the general population. Perform a **one-sample t-test**

Remember to change the value of ***True mean***.

What is the mean height in your sample? What is your value of t? What is the p-value? How do you interpret the p-value?

2.2 Biological processes duration

In the file `bp_times.csv`, we have the durations of a biological process for two samples of wild-type and knock-out cells (times in seconds). We are interested in seeing whether there is a difference in the durations for the two types of cells – we shall use an **independent t-test** to compare the two cell-types.

These data can be analysed online at http://bioinf-rstud001:3838/TwoSampleTest/

a) What are your null and alternative hypotheses?

Import the data using ***Choose File*** as before. Make sure that the ***1st column is a factor?*** checkbox is ticked.

b) Histograms to compare the two groups will be created for you automatically. Do the data look normally distributed for each cell-type? Is the independent t-test appropriate.
c) Use the Numerical statistics analysis to compute descriptive statistics for each group. \

Given the distribution of your data, which statistics might you report to summarise your data? Look at and compare the 95% confidence intervals of the mean durations of the two cell-types.

Do they overlap?

d) In order to apply the correct statistical test, we need to test to see if the variances of the two groups are comparable.

What do you conclude from the p-value of this test. How does it influence what test to use?

e) Use the appropriate test to compare the durations of the two groups.

Is a Welch's correction needed? What is your value of t? What is the p-value? How do you interpret the p-value? Is this in agreement with the 95% confidence intervals?

2.3 Blood vessel formation

In blood plasma cancer, there is an increase in blood vessel formation in the bone marrow. A stem cell transplant can be used as a treatment for blood plasma cancer. The bone marrow micro vessel density was measured before and after treatment for 7 patients with blood plasma cancer.

We are interested in seeing whether there is a decrease in the bone marrow micro vessel density after treatment with a stem cell transplant. We will use a paired two-sample t-test to compare the before and after bone marrow micro vessel densities.

The data are contained in the file `bloodplasmacancer2.csv`. These data can be analysed online at http://bioinf-rstud001:3838/TwoSampleTest/

a) What are your null and alternative hypotheses?

Import the data and create a column of differences (after-before).

b) Plot a histogram of the differences. Do the data look normally distributed? Is the paired t –test appropriate?
c) We are interested in seeing whether there is a decrease in the bone marrow micro vessel density after treatment with a stem cell transplant. Is this a one-tailed or two-tailed test?
d) Compare the durations before and after values. Ensure you select the one- or two-tailed test as appropriate. What is the mean difference? What is your value of t? What is the p-value? How do you interpret the p-value?

3 Tests for categorical variables

3.1 Nucleotide frequency

In **Table 1**, we have the frequencies of the four nucleotides in two sequences. We are interested in comparing the nucleotide proportions of the two sequences.

a) What are your null and alternative hypotheses?

We can analyse these data online

Note that you do not need to enter the totals.

What is your value of your Chi-squared statistic and its corresponding p-value? How do you interpret the result?

3.2 Disease association

Table 2 gives the frequencies of wild-type and knock-out mice developing a disease thought to be associated to the absence of the knock-out gene.

a) What are your null and alternative hypotheses?
b) What are your expected frequencies?

Enter the data as before;\

c) Select the **Fisher's exact test** option to compare the proportion of mice in each group that developed the disease. What is your p-value? How do you interpret the result?