Introduction to Statistical Analysis

Mark Dunning and Sarah Vowler

Last modified: 08 Feb 2018

Introduction

In this practical, we will use several 'real-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.

Parametric Tests

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 a certain 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://bioinformatics.cruk.cam.ac.uk/stats/OneSampleTest/

To import the file `diseaseX.csv`; you will need to select the `Choose File` option from the `Data Input` tab and navigate to where the course data are located on your laptop. The right-hand panel of the `Data Input` tab should update to show the Heights of various individuals in the study.

Also, on the `Data Input` tab you will need to change the value of **Hypothesized mean** to the correct value.

Question: What are your null and alternative hypotheses?

A histogram and boxplot of the `Height` variable will be automatically generated for you. To view it, click on the ***Data Distribution***. You can toggle whether to overlay a density plot on top of the boxplot, or choose different bin sizes for the histogram.

Question: Do the data look normally distributed? Based on the plots, is the parametric one-sample t –test appropriate?

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** by clicking the ***Statistical Analysis*** tab.

Question: 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. 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://bioinformatics.cruk.cam.ac.uk/stats/TwoSampleTest/

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

Question: What are your null and alternative hypotheses?

Histograms and boxplots to compare the two groups will be created for you automatically. You can also see a basic numerical summary of the data distribution.

Question: Do the data look normally distributed for each cell-type? Is the independent t-test appropriate? What statistics are appropriate to report the location (mean or median) and spread (sd or IQR) of the data?

In order to apply the correct statistical test, we need to test to see if the variances of the two groups are comparable. This is tested for us automatically in the Shiny app. Click the ***Statistical Analysis*** tab to see the result of the "F-test". However, it is often easier to eye-ball the data to assess the variances.

Question: What do you conclude from the p-value of this test. Does this agree with your impression of the variances from the boxplot and histograms? How does it influence what test to use?

Now use the appropriate two-sample t-test to compare the durations of the two groups.

Question: What is your value of the test statistic? What is the p-value? How do you interpret the p-value?

---

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.

These data can be analysed online at http://bioinformatics.cruk.cam.ac.uk/stats/TwoSampleTest/

The data are contained in the file `bloodplasmacancer2.csv`. Import the data, making sure that ***`1st column is a factor`*** is *not* ticked. Now choose whether you will be performing a paired test or not by ticking the **Paired Samples?** box under ***Are your samples paired?***.

Question: What are your null and alternative hypotheses?

View the histogram and boxplot of the paired differences on the ***Differences*** tab.

Question: Do the differences look normally distributed? Is the paired t –test appropriate?

We are interested in seeing whether there is a decrease in the bone marrow micro vessel density after treatment with a stem cell transplant.

Question: Is this a one-tailed or two-tailed test?

Now select the correct options in the ***Statistical Analysis*** tab in order to perform the analysis. Ensure you select the one- or two-tailed test as appropriate.

Question: What is the mean difference? What is your value of t? What is the p-value? How do you interpret the p-value? Why does ticking / unticking the equal variances have no effect?

---

4. Gene Expression in Breast Cancer patients

A gene expression study was performed on patients categorised into positive and negative Estrogen Receptor (ER) groups. It is well-known that ER positive patients have more treatment options available and thus have more better prognosis.

The gene NIBP was measured as part of this study and the results are available in the file `NIBP.expression.csv`. We are interested to see if the expression level of the gene is different between ER positive and negative patients.

Question: What are your null and alternative hypotheses?

Now conduct an independent two-sample t-test to see if there is a difference in expression between the two groups.

Question: What is the p-value from the test? Do we achieve statistical significance at the 0.05 level?

Look closely at data distribution, calculated means for each group and the estimated confidence interval

---

Question: Is the finding likely to hold Biological significance? Would you be willing to put further resources into validating the finding?
