Introduction to Statistical Analysis

Mark Dunning, Rosemary Tate and Sarah Vowler

Last modified: 13 Oct 2016

Contents

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.

The datasets you will need for this practical should be ***downloaded and unzipped now***:- https://rawgit.com/bioinformatics-core-shared-training/IntroductionToStats/master/CourseData.zip

T-tests practical: Parametric Tests

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://bioinformatics.cruk.cam.ac.uk/stats/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 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***.

b) 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.

Do the data look normally distributed? Based on the plots, is the parametric 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** by clicking the ***Statistical Analysis*** tab.

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?

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/

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 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.

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?

c) 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.

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

d) 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?

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

a) What are your null and alternative hypotheses?

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?***.

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

T-tests practical: Non-Parametric Tests

Birth-weight of twins

Dr D. R. Peterson of the Department of Epidemiology, University of Washington, collected the data found in file `twins.csv`. It consists of the birth-weights of each of 20 dizygous twins. The hypothesis to be tested is that the SIDS child of each pair has a lower birth-weight.

a. Construct the null and alternative hypothesises

b. Decide on the level of significance to be used and whether the test should be one-sided or two-sided.

c. Carry out both the sign and Wilcoxon signed rank tests on the data. Do both tests draw the same conclusion about the data? Which test is the most appropriate?

Vitamin D levels

The file `vitd.csv` contains data on vitamin D levels for subjects with fibrosis.

a. Use the Mann-Whitney U and t-tests to compare vitamin D levels between those with and without fibrosis. Interpret the results from both tests. Do both tests reach the same conclusion?

b. Which test is the more appropriate?

Tests for categorical variables

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	C	G	T	Total
Sequence 1	273.00	233.00	236.00	258.00	1000.00
Sequence 2	281.00	246.00	244.00	229.00	1000.00
Total	554.00	479.00	480.00	487.00	2000.00

Table 1: Nucleotide frequencies for two sequences

a) What are your null and alternative hypotheses?

We can analyse these data online in the Shiny app, by modifying the contents of the ***Enter your data as a table*** box. Columns need to be separated by a '-' character, and rows by a '|'. You can check that you have entered the data correctly by looking at ***The data*** tab. You do not need to calculate row or column totals.

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?

	WT	KO	Total
Disease	1.00	7.00	8.00
No Disease	9.00	3.00	12.00
Total	10.00	10.00	20.00

Table 2: Frequencies of wild-type and knock-out mice developing disease

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 into the Shiny app 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?

Choosing a test

In this section, we introduce several datasets and will invite you in groups to select a dataset and discuss what methods / tests you would use to analyse those data.

Dataset 1: Barley yields `data1.csv`

In this data set named, the barley yield in years 1931 and 1932 of the same field are recorded.

Is there evidence for a difference in yield between 1931 and 1932?

Dataset 2 `data2.csv`

In the history of data visualization, Florence Nightingale is best remembered for her role as a social activist and her view that statistical data, presented in charts and diagrams, could be used as powerful arguments for medical reform.

After witnessing deplorable sanitary conditions in the Crimea, she wrote several influential texts (Nightingale, 1858, 1859), including polar-area graphs (sometimes called "Coxcombs" or rose diagrams), showing the number of deaths in the Crimean from battle compared to disease or preventable causes that could be reduced by better battlefield nursing care.

Her Diagram of the Causes of Mortality in the Army in the East showed that most of the British soldiers who died during the Crimean War died of sickness rather than of wounds or other causes. It also showed that the death rate was higher in the first year of the war, before a Sanitary Commissioners arrived in March 1855 to improve hygiene in the camps and hospitals.

Do the data support the claim that deaths due to avoidable causes decreased after a change in regime?

Dataset 3: Effect of bran on diet: `data3.csv`

The addition of bran to the diet has been reported to benefit patients with diverticulosis. Several different bran preparations are available, and a clinician wants to test the efficacy of two of them on patients, since favourable claims have been made for each. Among the consequences of administering bran that requires testing is the transit time through the alimentary canal. By random allocation the clinician selects two groups of patients aged 40-64 with diverticulosis of comparable severity. Sample 1 contains 15 patients who are given treatment A, and sample 2 contains 12 patients who are given treatment B.

Does transit time differ in the two groups of patients taking these two preparations?

Dataset4: Effect of Autism drug `data4.csv`

A new chemotherapy treatment is proposed for patients with breast cancer. Investigators are concerned with patient's ability to tolerate the treatment and assess their quality of life both before and after receiving the new chemotherapy treatment. Quality of life (QOL) is measured on an ordinal scale and for analysis purposes, numbers are assigned to each response category as follows: 1=Poor, 2= Fair, 3=Good, 4= Very Good, 5 = Excellent.

Is there statistically significant improvement in repetitive behavior after 1 week of treatment?

Dataset5: CD4 `data5.csv`

CD4 cells are carried in the blood as part of the human immune system. One of the effects of the HIV virus is that these cells die. The count of CD4 cells is used in determining the onset of full-blown AIDS in a patient. In this study of the effectiveness of a new anti-viral drug on HIV, 20 HIV-positive patients had their CD4 counts recorded and then were put on a course of treatment with this drug. After using the drug for one year, their CD4 counts were again recorded.

Do patients taking the drug have increased CD4 counts?

Dataset6: Drink Driving `data6.csv`

Drunk driving is one of the main causes of car accidents. Interviews with drunk drivers who were involved in accidents and survived revealed that one of the main problems is that drivers do not realize that they are impaired, thinking "I only had 1-2 drinks ... I am OK to drive."

A sample of 100 drivers was chosen, and their reaction times in an obstacle course were measured *before* and *after* drinking two beers. The purpose of this study was to check whether drivers are impaired after drinking two beers

Does drinking beer alter the reaction time of the driver?

Dataset7: Pollution in Trees `data7.csv`

Laureysens et al. (2004) measured metal content in the wood of 13 poplar clones growing in a polluted area, once in August and once in November. Concentrations of aluminum (in micrograms of Al per gram of wood) are shown below.

Is there any evidence for an increase in pollution between November and August?

Dataset8: Colon cancer `data8.csv`

The tumor and the normal counter-part samples were prospectively collected from 9 patients who underwent surgical resection at the INT-MI. Neoplastic samples were obtained from the central area of the neoplasia, avoiding to select necrotic material or transition zones with healthy mucosa. Samples of colonic healthy mucosa were resected at least 20 centimeters far from the neoplasia and distant from the surgical resection margins. Tissue samples were stored in liquid nitrogen until RNA extraction. Total RNA was extracted from 10–20 mg of tumor samples and from 30–40 mg of normal samples.

The gene expression measurements for a paticular gene were extracted from the data.

Is this gene differentially-expressed between tumours and normals?