Introduction to Statistical Analysis

Mark Dunning, Dominique-Laurent Couturier and Sarah Vowler

Last modified: 23 Oct 2017

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

Statistical distributions and central limit theorem

Question (i):

The tab **Estimated coverage of Student's CI** in the shiny app **central-limit-theorem** displays the confidence intervals of 100 simulated datasets.

1. Assuming that the simulated data are normally distributed, what is the probability of the **true** mean belonging to a confidence interval?
2. Let X denote a random variable that equals 1 if the **true mean belongs to the confidence interval** and 0 otherwise. What is the distribution of X?
3. What is the probability that 0 confidence intervals out of 50 contain the **true mean** if data are normally distributed?

Question (ii):

Using the shiny app **central limit theorem**, answer the following questions: http://bioinformatics.cruk.cam.ac.uk/apps/stats/central-limit-theorem/

1. Simulate **1000 samples** of **size n=10** of **Poisson** random variates, first assuming a **mean of 0.25**, and then assuming a **mean of 100**. Compare the coverage level of Student's confidence intervals for the mean of these 2 simulation sets: How do you explain that the latter is better than the first one?
2. Now consider **zero-inflated Poisson** variates with a **mean of 30** and a **10% probability of belonging to the clump-at-zero**. Can you think of a random variable having such a distribution? How large should the sample size be for the Student's confidence intervals to have good properties?

3. A student lost a few points in the statistic exams as the use of Student's confidence intervals for the probability of success of a **Bernoulli** variable with **pi = 40%** and a **n=100** was not considered as suitable. Should he/she contact the University to dispute his/her mark?

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?

Non-Parametric Tests

5. Vitamin D levels

The file `vitd.csv` contains data on vitamin D levels for subjects with (“Y”), and without (“N”) fibrosis. To import these data, you will need to select the ***1st column in a factor*** option.

Question: State the null and alternative hypothesises

Question: Examine the distribution of the data. Why doesn't a parametric analysis seem appropriate?

By un-ticking the ***Use Parametric Test?*** option in the ***Statistical Analysis*** tab you will see the results of a Mann-Whitney U (/ Wilxcoxon rank-sum test) test.

Question: How do you interpret the value of the test?

6. 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. One twin suffered Sudden Infant Death Syndrom (SIDS), and the other twin did not. The hypothesis to be tested is that the SIDS child of each pair had a lower birth-weight.

Question: State the null and alternative hypothesises

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

Question: Would be appropriate to treat these as paired or independant samples?

Recall from the lectures that for paired data we have to consider whether the differences are symmetrical about zero. Carry out both the sign-test and Wilcoxon signed rank tests on the data.

Question: Do both tests draw the same conclusion about the data? Which test is the most appropriate?

Tests for categorical variables

7. Nucleotide frequency

In the following table 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
Sequence 1	273	233	236	258
Sequence 2	281	246	244	229

Question: What are your null and alternative hypotheses?

We can analyse these data online in the Shiny app by entering the counts into the table on the left hand side. Note that the table needs to be resized to increase the number of columns.

http://bioinformatics.cruk.cam.ac.uk/stats/contingency-table

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

8. Disease association

The following table 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.

	WT	KO	Total
Disease	1	7	8
No disease	9	3	12
Total	10	10	20

Question: What are your null and alternative hypotheses?

Question: What are your expected frequencies?

Enter the data into the Shiny app as before. Select the **Fisher's exact test** option to compare the proportion of mice in each group that developed the disease.

Question: What is your p-value? How do you interpret the result?

Small-Group Exercise: Choosing a test

In this section, we invite you to form small groups to select a dataset and discuss what methods/tests you would use to analyse those data.

You can use this interactive document to record your observations:

https://public.etherpad-mozilla.org/p/2017-10-23-intro-to-stats

Dataset 1: Plant Growth `data1.csv`

Darwin (1876) studied the growth of *pairs* of zea may (aka corn) seedlings, one produced by cross-fertilization and the other produced by self-fertilization, but otherwise grown under identical conditions. His goal was to demonstrate the greater vigour of the cross-fertilized plants. The data recorded are the final height (inches, to the nearest 1/8th) of the plants in each pair.

Is there evidence to support the hypothesis of greater growth in Cross-fertilized plants?

Dataset 2: Florence Nightingale `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?

Dataset 4: Effect of Autism drug `data4.csv`

Consider a clinical investigation to assess the effectiveness of a new drug designed to reduce repetitive behaviors in children affected with autism. If the drug is effective, children will exhibit fewer repetitive behaviors on treatment as compared to when they are untreated. A total of 8 children with autism enroll in the study. Each child is observed by the study psychologist for a period of 3 hours both before treatment and then again after taking the new drug for 1 week. The time that each child is engaged in repetitive behavior during each 3 hour observation period is measured. Repetitive behavior is scored on a scale of 0 to 100 and scores represent the percent of the observation time in which the child is engaged in repetitive behavior. For example, a score of 0 indicates that during the entire observation period the child did not engage in repetitive behavior while a score of 100 indicates that the child was constantly engaged in repetitive behavior.

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

Dataset 5: 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?

Dataset 6: 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?

Dataset 7: 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?

Dataset 8: Salaries for Professors `data8.csv`

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members. (salary given as nine-month salary, in dollars.)

Is there evidence that Female professors are paid differently to their Male counterparts?