MT 4113: Computing in Statistics, Assignment 3

Setter: Eiren Jacobson

Due: 19 November 2018

Summary

Numerical experiments can be performed using computer simulations. Statisticians commonly use *Monte Carlo simulation*, which involves random sampling from probability distributions, to investigate the properties of statistical methods under different scenarios.

In this assignment, you will choose a dataset from the `dslabs` package, formulate a research question, choose statistical tests that you could use to answer that question, then examine the size and power of those tests when applied to simulated datasets that you have generated.

You will hand it a report that outlines the research question and statistical tests that you chose, describes the methods that you used to conduct a simulation study, and summarizes your findings. We lay out the assignment in more detail below, and then provide example research questions, workflow, and further detail on statistical size and power.

Your assignment

Task 1: Choose a dataset and research question

- Install the package `dslabs` and take a look at the documentation.
- Explore the available datasets and choose one that interests you.
- Formulate a simple research question that could be answered using the dataset. Note that, in this assignment, your task is not to answer this research question. Rather, you will use simulation to investigate the statistical tests you could use to answer the question.

Task 2: Choose statistical tests

- You should choose one parametric and one non-parametric test to investigate.
- The tests should be appropriate for answering your stated research question.

Task 3: Generate and analyze simulated data

- Use the dataset from `dslabs` to inform the properties of your simulated data.
- Develop scenarios for data generation (see workflow for examples).
- Write code to generate simulated data under different scenarios.
- Apply your chosen statistical tests to the simulated data.
- Calculate the size and power of these statistical tests under each of the scenarios you have developed.

Task 4: Summarize your findings

- Report the size and power of your statistical tests.
- Comment on the results of your study.

What to hand in

Please upload a single .zip file to MMS that contains:

- A short report that explains what you did and what you found. This report should be in PDF format and no longer than 5 pages, including references and any tables and figures. You don't need to explain in any depth how the statistical tests work, or what size and power are, but you do need to explain what you did and why, what you found, and what conclusions you draw from your results. Your report should be written as an abbreviated scientific paper and contain an abstract, introduction, methods, results, and conclusions.
- A diagram of your approach, included either as an appendix to the report or as a separate document, that shows you spent time designing the simulation study before you started coding.
- Code needed to conduct the simulations. I should be able to run your scripts and get **exactly** the same results that you did.

Rules and regulations

- You may use base R packages for conducting your simulations (e.g., `t.test` from the `stats` package). Please DO NOT use R packages designed to conduct simulations to complete the assignment (e.g., the `MonteCarlo` or `SimDesign` packages). You may use other packages to assist in writing up your report (e.g., `ggplot2` to produce figures).
- Example datasets, research questions, and statistical tests are provided below. You are welcome to use these examples, but note that points are available for creativity and originality. If you would like to use a publicly-available dataset that is not in the `dslabs` package as the inspiration for your study, you may do so, but please provide a citation.

Marking scheme

This assignment represents 30% of your final grade in the course. A detailed grading rubric is provided in Table 1. This assignment is intentionally more open-ended than previous assignments given in this course, and is an opportunity for you to show us what you have learned. We strongly suggest that you pick a maneageable research question, relatively simple statistical tests, and a small number of scenarios and focus on designing well-documented code and producing a concise report.

Because this assignment counts towards your final grade, it is important that you do not collaborate with others in completing the work. You should be comfortable with the following statement, which you should include at the beginning your report: "I confirm that the following report and associated code is my own work, except where clearly indicated." Marks will be deducted if this statement is not present.

For more information, see the "Academic misconduct' section of the university web site [http://www.st-andrews.ac.uk/students/rules/academicpractice/]. This holds for all assessed project work on this course. Plagiarism cannot be tolerated on this or any other assignment for this module.

Table 1: Detailed grading rubric

Total Possible	Task	Points	Item
2.5	1	2.5	Defined research question
5	2	2.5	Appropriate parametric test
		2.5	Appropriate non-parametric test
30	3	2.5	Diagram of simulation design
		2.5	Explanation of scenarios
		10	Simulation program design and style
		10	Simulation code correctness
		5	Additional points for creativity, efficiency, originality
12.5	4	2.5	Size and power of statistical tests are reported
		5	Explanation of findings
		5	Additional enhancements to report (e.g., figures)
-5	P	-5	Penalty if statement of original work is not included

Examples of possible datasets and research questions

Dataset: `heights`
Research question: On average, are males taller than females?
Statistical tests: t-test and Mann-Whitney U Test

Dataset: `us_contagious_diseases`
Research question: Were rates of Hepatitis A in U.S. states greater in the 20th or 21st century?
Statistical tests: Paired t-test and Wilcoxon Signed-Rank Test

Dataset: `gapminder`
Research question: Is life expectancy the same on all five continents?
Statistical tests: One-way ANOVA and Kruskall-Wallis test

Sketch of example workflow

First, we install and load the `dslabs` package.

```
install.packages("dslabs")
library("dslabs")
help(package = "dslabs")
```

Let's consider the `heights` dataset. Our research question might be, "on average, are males taller than females?" We could use the two-sample *t*-test and Mann-Whitney U Test to answer this question. Looking at the `heights` dataset, we see that the range of values is 50 to 83 cm. The average height of males in the dataset is 69 cm, while the average height of females is 65 cm. We could use these properties to simulate new data using `rnorm()`.

We could consider scenarios where:

- Different numbers of heights are recorded (small vs. large sample size),
- The effect size (true difference between means) is large or small,
- Measurement error is present in the simulated data (e.g., if reported values are always rounded up to the nearest cm, or nearest 5 cm),

and repeatedly (e.g., 1000 times) simulate datasets according to these scenarios.

We would then apply our chosen statistical tests to each simulated dataset, and calculate, for each scenario, the size and power of the tests.

One result might be a table of calculated statistical power with number of samples (10, 100, 1000) in rows and effect size (1 cm, 5 cm, 10 cm) in columns.

Further detail on size and power

Using the same example as above, we have two simulated samples of data, female heights $x_1, x_2, \ldots, x_n$ and male heights $y_1, y_2, \ldots, y_n$, which we assume are a i.i.d. samples from two normal distributions, each with unknown mean and variance.

We wish to test the null hypothesis, H_0, that the samples come from normal distributions that have the same mean, i.e., average height of male and female students is the same $H_0 : \mu_x = \mu_y$. If the null hypothesis is false, then some alternative must be true; if we're working with one-sided tests, we'll assume an alternative hypothesis that the average height of females is less than the average height of males $H_1 : \mu_x < \mu_y$. To test the null hypothesis, we construct the test statistic

$$t = \frac{\bar{x} - \bar{y}}{s_{\bar{x}-\bar{y}}} \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means and s is the standard error of the difference between the means.
If the null hypothesis is true, t follows a Student's t-distribution with mean 0, variance 1 and $n_x + n_y - 2$ degrees of freedom (where n_x and n_y are the sizes of the two samples). The larger the observed value of t, the less likely it is that the null hypothesis is true. Given an observed value of t, we can compute the probability that (or frequency with which) a value this large, or larger, would be observed if H_0 is true – this computed probability is called the *p-value.*

If the observed p-value is equal to or smaller than some pre-defined value, then we judge that it is so unlikely the data could be generated under H_0 that we reject the null hypothesis. The pre-defined value is called the *α-level*, and it's up to us to set the α-level for our test before we see the data. By convention, people usually use $\alpha = 0.05$.

Given this procedure, even if the data really were generated from a two distributions with the same mean (i.e., H_0 was true), we should incorrectly reject H_0 $\alpha \times 100\%$ of the time. Incorrectly rejecting H_0 when it is actually true is referred to as "making a *type 1 error*" (often written type I error).

Sometimes tests don't do what they are supposed to, and incorrectly reject the null too often, or not often enough. This is particularly likely to happen if some of the underlying assumptions are violated (for example, in the case of our t-test, if the data do not come from a normal distribution). It can also happen with small sample sizes for tests that rely on large-sample asymptotics. The *size* of a test is the actual proportion of time that H_0 is incorrectly rejected, given that it is true.

What about if H_0 is actually false and H_1 is true? In this case, we'd be making an error if we failed to reject H_0. Incorrectly failing to reject H_0 when H_0 is actually false is called making a *type 2 error* (often written type II error). The *power* of a test is the proportion of the time we correctly reject H_0 given that H_1 is true.

The size of a test can only be evaluated by generating data from distributions where H_0 is true. By contrast, power can be evaluated over variety of different values of H_1, ranging from values close to H_0 (i.e., μ_x is only just greater than μ_y), where power should be low (close to the size), to values very far from H_1 (i.e., $\mu_x >> mu_y$), where power should be 1. So, you'll need to think about what values of H_1 seem appropriate. $\mu_x - \mu_y$ in this context is sometimes called the *effect size.* Size and (particularly) power are also affected by things like the variance of the two populations, the amount of data in the sample and whether the assumptions of the methods are met or not.