Assignment 8 for Statistical Computing and Empirical Mathods

Dr. Henry WJ Reeve

Teaching block 1 2021

Introduction

This document describes your eighth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommended that you first watch video lectures 18 and 19.

Begin by creating an Rmarkdown document with html output. You are not expected to hand in this piece of work, but it is a good idea to get used to using Rmarkdown.

1 Obstacles to valid scientific inference

For each of the following give (A) an explanation of the concept and (B) an example of a situation (real or hypothetical) where they create a barrier to drawing scientific conclusions based on data. You are encouraged to discuss these concepts with your colleagues.

1. Measurement distortions
2. Selection bias
3. Confounding variables

2 An unpaired t test

In this question the goal is to create a function called **t_test_function()** which implements an unpaired Student's t test, in order to test for a difference of population means between two unpaired samples from two distributions. Your function will play a similar role to the following standard R function:

```
t.test(body_mass_g~species, data=peng_AC,var.equal = TRUE)
```

Begin by creating a data frame called "peng_AC" which is a subset of the Palmer penguins data set consisting of all those penguins which belong to either the "Adelie" or the "Chinstrap" species of penguins.

```
library(palmerpenguins)

peng_AC<-penguins%>%
  drop_na(species,body_mass_g)%>%
  filter(species!="Gentoo")
```

Your function should take in the following arguments:

1. "data" - A data frame argument,
2. "val_col"- A string argument containing a column name for a continuous variable,
3. "group_col" - A string argument containing a column name for a binary variable.

The function should carry out an unpaired Student's t test based on the value of the continuous variable with column name "val_col". The function should begin by partitioning the sample into two groups based on the value of the binary variable named "group_col". Your function should then compute the sample mean, sample variance and sample size for each of these two groups, based upon the variable within the column named "var_col".

Your function should compute a test statistic for the Student's unpaired t-test. In addition, the function should compute the corresponding **p-value**. Finally, your function should compute an estimate for the effect size.

Your function should have the following output:

```
t_test_function(data=peng_AC,val_col="body_mass_g",group_col="species")
```

```
##         t_stat dof     p_val
## 1 -0.5080869 217 0.6119085
```

As an additional challenge you can modify your function so that it takes a fourth argument called "var_equal" which takes a Boolean value. If the input of "var_equal" is the Boolean "TRUE" your function should compute the test-statistic and p-value for an unpaired Student's t test. If, on the other hand, the input of "var_equal" is the Boolean "FALSE" your function should compute the test-statistic and p-value for Welch's t-test. Your function should have the following output:

```
t_test_function(data=peng_AC,val_col="body_mass_g",group_col="species",var_equal=FALSE)
```

```
##         t_stat      dof     p_val
## 1 -0.5430902 152.4548 0.5878608
```

You can compare the output of your function with R's inbuilt **t.test()** function.

3 Statistical hypothesis testing

Explain the following concepts:

1. Null hypothesis
2. Alternative hypothesis

3. Test statistic
4. Type 1 error
5. Type 2 error
6. The size of a test
7. The power of a test
8. The significance level
9. The p-value
10. Effect size

Is the p-value the probability that the null hypothesis is true?

If I conduct a statistical hypothesis test, and my p-value exceeds the significance level, do I have good evidence that the null hypothesis is true?

4 Investigating test size for an unpaired Student's t-test

In this question we shall investigate the performance of an unpaired Student's t test from the perspective of test size. Recall a Type 1 error occurs when we reject the null hypothesis, when the null hypothesis is true. The size of a test is the probability of a Type 1 error. A key property of valid statistical hypothesis tests with a given significance level is that the size of the test does not exceed the significance level.

Note that we can apply unpaired Student's t-test with significance level `alpha` on a samples `sample_0`, `sample_1`:

```
t.test(sample_0,sample_1,var.equal = TRUE, conf.level = 1-alpha)
```

We can apply an unpaired Student's t-test and extract the p-value as follows:

```
t.test(sample_0,sample_1,var.equal = TRUE)$p.value
```

Notice that the significance level wasn't supplied as an argument. Is this a problem?

The following code checks the size of an unpaired Student's t-test with a significance level of $\alpha = 0.05$.

```
num_trials<-10000
sample_size<-30
mu_0<-1
mu_1<-1
sigma_0<-3
sigma_1<-3
alpha<-0.05

set.seed(0) # set random seed for reproducibility

single_alpha_test_size_simulation_df<-data.frame(trial=seq(num_trials))%>%
  mutate(sample_0=map(.x=trial,.f=~rnorm(n=sample_size,mean=mu_0,sd=sigma_0)),
         sample_1=map(.x=trial,.f=~rnorm(n=sample_size,mean=mu_1,sd=sigma_1)))%>%
  # generate random Gaussian samples
  mutate(p_value=pmap(.l=list(trial,sample_0,sample_1),
                    .f=~t.test(..2,..3,var.equal = TRUE)$p.value))%>%
  # generate p values
  mutate(type_1_error=p_value<alpha)

single_alpha_test_size_simulation_df%>%
  pull(type_1_error)%>%
  mean() # estimate of coverage probability
```

Check that you understand the above code.

Modify the above code to explore how the size of the test varies as a function of the significance level α.

5 The power of an unpaired t-test

In this question we shall investigate the performance of an unpaired Student's t test from the perspective of statistical power. Recall that the statistical power of a test is the probability of correctly rejecting the null hypothesis when an alternative hypothesis holds.

Consider a setting in which we have two samples i.i.d with Gaussian distribution. The first sample consists of n_0 observations with population mean μ_0 and population variance σ_0^2. The second sample consists of n_1 observations with population mean μ_1 and population variance σ_1^2.

The following code checks the statistical power of an unpaired Student's t-test in a sample sizes $n_0 = n_1 = 30$, $\mu_0 = 3$, $\mu_1 = 4$, $\sigma_0 = \sigma_1 = 1$ and with a significance level of $\alpha = 0.05$.

```
num_trials<-10000
n_0<-30
n_1<-30
mu_0<-3
mu_1<-4
sigma_0<-2
sigma_1<-2
alpha<-0.05

set.seed(0) # set random seed for reproducibility

data.frame(trial=seq(num_trials))%>%
  mutate(sample_0=map(.x=trial,.f=~rnorm(n=n_0,mean=mu_0,sd=sigma_0)),
         sample_1=map(.x=trial,.f=~rnorm(n=n_1,mean=mu_1,sd=sigma_1)))%>%
  # generate random Gaussian samples
  mutate(p_value=pmap(.l=list(trial,sample_0,sample_1),
                      .f=~t.test(..2,..3,var.equal = TRUE)$p.value))%>%
  # generate p values
  mutate(reject_null=p_value<alpha)%>%
  pull(reject_null)%>%
  mean() # estimate of coverage probability
```

Conduct a simulation study to explore how the statistical power varies as a function of the significance level.

Conduct a simulation study to explore how the statistical power varies as a function of the difference in means $\mu_1 - \mu_0$.

Conduct a simulation study to explore how the statistical power varies as a function of the population standard deviation $\sigma = \sigma_0 = \sigma_1$.

Conduct a simulation study to explore how the statistical power varies as a function of the sample size $n = n_0 = n_1$.

6 Comparing the paired and unpaired t-tests (Optional)

The aim of this question is to explore the benefits of using a paired test when a natural pairing is available. Consider a situation in which we have two i.i.d. samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$.

Suppose that $X_1, \ldots, X_n \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and for each $i = 1, \ldots, n$, we have $Y_i = X_i + Z_i$ where $Z_1, \ldots, Z_n \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ are independent and identically distributed random variables. It follows that $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent and identically distributed with $\mu_Y = \mu_X + \mu_Z$ and $\sigma_Y^2 = \sigma_X^2 + \sigma_Z^2$.

In this situation we only observe the two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. We are interested in performing a statistical hypothesis test to see if $\mu_X \neq \mu_Y$. We have two options here. We could either (1) use the pairing and apply a paired test or (2) ignore the pairing and use an unpaired test. In the console run `?t.test()` to see how to carry out an unpaired and a pared test within R.

Conduct a simulation study to explore the statistical power of these two approaches. You may want to consider a setting in which $n = 30$, $\mu_X = 10$, $\sigma_X = 5$, $\mu_Z = 1$ and $\sigma_Z = 1$. Consider a range of different significance levels α.

7 A chi-squared test of population variance (Optional)

It is recommended that you watch lecture 20 before completing this question.

Suppose we have an i.i.d. sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and a conjectured value for the population variance σ_0^2. We wish to test the null hypothesis that $\sigma^2 = \sigma_0^2$.

Implement a function called `chi_square_test_one_sample_var` which takes as input a sample `sample` and a null value for the variance `sigma_square_null`.

Conduct a simulation study to see how the size of the test varies as a function of the significance level.

Load the "Palmer penguins" library and extract a vector called "bill_adelie" consisting of the bill lengths of the Adelie penguins belonging to the Adelie species.

Suppose we model the sequence of bill lengths as a sample of independent and identically distributed Gaussian random variables $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with population mean μ and population standard deviation σ.

Now apply your function `chi_square_test_one_sample_var` to test the null hypothesis that the population standard deviation is 3mm at a significance level of $\alpha = 10\%$.