Principles of Statistical Data Analysis

Homework 2: simulations and permutations

Academic Year 2020-2021

Practicalities

You must prepare this homework in groups of 2 students and your report should be submitted as a pdf file via Ufora (only a pdf file, no Word file). Please use the following format for the file name

HW2_LastNameStudent1_LastNameStudent2.pdf

In addition to the report you should also upload one R-file with your (cleaned-up and readable) R-code with the file name

HW2_LastNameStudent1_LastNameStudent2.R

Your report should be clear enough so that we do not have to go through your R code. In case your report is not clear, we will look at the R code but account for this in the final grading.

We anticipate that you can finish the assignment in about 8 hours (each). The report should not be longer than 3 pages. You can use the forum on Ufora to find a partner for the home work.

Introduction

In an initial screening of a new type of cognitive behavioral therapy for the treatment of moderate depression, data of a limited number of patients will be collected at the start and the end of the therapy. The outcome of interest is the severity of depression quantified using an inventory score where higher scores indicate more severe depressive symptoms (for simplicity we assume that this outcome is continuous). The psychologists initially want to know if the mean

score differs between the start and the end of the treatment[1]. A team of two statisticians will be in charge of the analysis of the data and when thinking about their strategy they don't agree on the following:

- Statistician A wants to analyze the data with a paired t-test to account for the potential correlation between the two measurements of the same patient.
- Statistician B believes that this correlation will be very weak because the sample is quite homogeneous (all patients exhibit symptoms of moderate depression) and there is a substantial amount of time between both measurements (the therapy takes between six months and a year) and therefore prefers to use a two-sample test as he/she believes this might be beneficial.

The goal of this home work is to set up several simulation studies to examine the properties of the permutation paired and permutation two-sample t-test as a function of the correlation between the pre and post outcomes. In addition to testing, you are also asked to study the properties of asymptotic confidence intervals as a function of this correlation. In the end you are asked to write a recommendation for the two statisticians based on your insights.

Assignment

1. Write two functions `twosample.t(X1, X2, n.perm)` and `paired.t(X1, X2, n.perm)` that return the permutation p-value of a two-sided two-sample t-test and a two-sided paired t-test based on `n.perm` permutations. You should implement this test yourself, so don't use a test from e.g. the coin package. As your answer to this question, copy-paste the (readable) R code of these two functions so that we can check your implementation.
2. Write a function `simulate.data(n, delta, correlation)` that simulates `n` paired observations (so the total number of observations is `2n`), say X_1 and X_2, from a *skewed distribution* where $\mathrm{E}[X_1] - \mathrm{E}[X_2] =$ `delta`. The parameter `correlation` can take on a number form 1 to 5 and is related to the strength of the correlation. More specifically, `correlation = 1` should correspond to $\mathrm{cor}(X_1, X_2) = 0$, `correlation = 2` to $\mathrm{cor}(X_1, X_2) = 0.1$, `correlation = 3` to $\mathrm{cor}(X_1, X_2) = 0.25$, `correlation = 4` to $\mathrm{cor}(X_1, X_2) = 0.5$ and `correlation = 5` to $\mathrm{cor}(X_1, X_2) = 0.7$. Note that many parameters are left free in this function (e.g. the marginal means and variances - you can choose the values yourself for this). As your answer to this question, copy-paste the (readable) R code of this function.

[1] In a later stage more appropriate designs (e.g. randomized controlled trails) will be used to evaluate the therapy, but this is outside the scope of this home work.

Tip: you might want to examine the code below to get inspiration on how you can simulate correlated skewed bivariate data. Note that `cor` below will not exactly equal $\text{cor}(X_1, X_2)$, so you will have to fine tune this. You can do this be setting n to a large number (e.g. 10000) and see what the sample correlation is (which will be very close to $\text{cor}(X_1, X_2)$ because n is large) for a choice of `cor`. It is not mandatory to use this code.

```
library(mvtnorm)
n <- 100
cor <- .1
Y <- rmvnorm(n, mean = c(0,0),
       sigma = matrix(nrow = 2, ncol = 2, c(1, cor, cor, 1)))
Z <- pnorm(Y)
X1 <- qexp(Z[,1])
X2 <- qexp(Z[,2])
```

3. Create a graph where you plot the empirical type I and type II error for each test as a function of the correlation where the correlation takes on the values 0, 0.1, 0.25, 0.5 and 0.7 and for $\alpha = 5\%$. Here n should be equal to 5 (mimicking a very small sample) and to 100 (mimicking a larger sample) and you can choose the number of permutations and Monte-Carlo simulations yourself (the higher the better, but higher also means computationally more intensive, so there is a trade-off). In your answer to this question you should only include two plots: one for $n = 5$ and one for $n = 100$ each with 4 curves (for both errors and for both tests) - this plot should be readable in black-and-white and please provide an informative caption. Make sure that when you simulate data under H_A the type II error is not to close to zero or one (because then it is hard to compare the methods - the `delta` is allowed to be different when $n = 5$ and when $n = 100$).

4. Write down in a few lines what you learn from this graphs and explain what you see.

5. We now turn to confidence intervals and we are first interested in some analytical results.

   (a) Work out the expectation and variance of the sample mean difference $\bar{X}_1 - \bar{X}_2$ assuming a) independence and b) dependence between X_1 and X_2. Tip: for at least one of the two scenarios the solution can be found in the course notes on hypothesis testing.

   (b) Use the above expressions to write down the formula of an $(1-\alpha) \times 100\%$ confidence interval for $\text{E}[X_1] - \text{E}[X_2]$ assuming a) independence and b) dependence between X_1 and X_2. When deriving the confidence intervals, you can assume that the variables are normally distributed or you can rely on large-sample approximations.

6. Simulate data under H_0 or H_A (you can choose and you can reuse the code from the previous questions) for $n = 100$ and compute the empirical coverage and average width of a 95% confidence interval for $\mathrm{E}[X_1] - \mathrm{E}[X_2]$ (using the two confidence intervals you have derived in the previous question - so you implement these confidence intervals in R yourself). Make two graphs: one with the empirical coverage as a function of the correlation and one with the average width as a function of the correlation where the correlation takes on the values 0, 0.1, 0.25, 0.5 and 0.7.

7. Write down in a few lines what you learn from these graphs and explain what you see.

8. Write down in a few lines some general recommendations to help the two statisticians.