Open Universiteit
www.ou.nl

Workbook Certified Professional Program Data Science

DIKW Academy

Contents

Chapter 8

Block 07 : Predictive Analytics and Matrix Algebra

8.1 LU07 : Bivariate Analysis

Correlation is not cause, it is just a 'music of chance'. - Siri Hustvedt

8.1.1 Introduction

In this Learning Unit, you will learn to perform bivariate data analysis. Key concept in this unit is the concept of correlation wich was studied in learning unit 4. In this learning unit, we combine the concepts of correlation and permutation. Similar to bootstrapping, a permutationm technique is used to get a more robust estimate of the statistic you are exploring. In this case, it is the correlaton measure.

8.1.2 Learning Objectives

- What is a permutation (test)?
- What is correlation and how to apply it?
- How to apply hierarchical clustering

8.1.3 Study Core

8.1.3.1 Recap Correlation

In Learning Unit 4, correlation is defined as a single number to express the relationship between 2 characteristics. This is a coefficient that varies between c (-1, 1) or c (0, 1). If we elaborate a little bit more, correlation can be defined as an *index* that measures the strength and direction of *linear relationship* between variables. This linear relationship is described using the Covariance measure.

Thus covariance = (1 / length(x) - 1) * sum((x - mean(x)) * (y - mean(y)))

(sd(x) * sd(y)) = product standard deviation x and standard deviation y

Pearson correlation is covariance / product both standard deviations

Another thing you should notice: correlation is simply used to describe data. If a correlation is found between two variables, it doesn't tell you anything about causation between the two variables. The golden rule of correlation is simple and even sort of rhymes: **correlation is not causation**.

8.1.3.2 The concept of Permutation

In Learning unit 3, bootstrapping has been covered as a parameter free method to get robust estimates of all kinds of statistics. Similar to bootstrapping is permutation testing also a paramter free method which can be used to get more robust estimates. The method is based on considering all different permutations (reorderings)

8.1.3.3 Combined - A randomization test using permutation

When we apply randomization tests to bivariate data, our primarily goal is to test a null hypothesis, usually that correlation = 0. We do this by holding one variable (e.g. X) constant, and permuting the other variable (Y) against it. Because each Xi is randomly paired with a value of Y, the expected value of the correlation is 0. By repeating this process a large number of times, we can build up the sampling distribution of correlation for that situation in which the true value of correlation is 0.0. We can either create confidence limits on correlation (a strange undertaking in this case), or we can increment a counter to record the number of times a correlation coefficient from a bivariate population where correlation = 0.0 exceed the obtained sample correlation (for either a one- or a two-tailed test.)

The sample correlation is 0.259, but is this value statistically significantly different from zero? To answer that, we will perform a permutation test by creating 5000 permuted samples of the two variables and then calculate the correlation between them for each sample. These 5000 values represent the distribution of correlations when the null hypothesis is true (i.e. the two variables are not correlated). We can then compare the original sample correlation with this distribution to determine if the value is too extreme to be explained by null hypothesis.

8.1.4 Reading Assignment:

- Study the slidedeck Slidedeck

8.2 LU08 : Decision Trees

The leader of a company needs a decision tree in his head - if this happens, we go this way, but if it winds up like that, then we go this other way. - Sean Parker

8.2.1 Introduction

Decision trees have been around for a in machine learning since the early eighties. .jm

8.2.2 Learning Objectives

- Undestand the concept of decision trees
- Know different split criteria
- Understand the concept of overfitting
- Being able to “prune” the tree

Permutation test

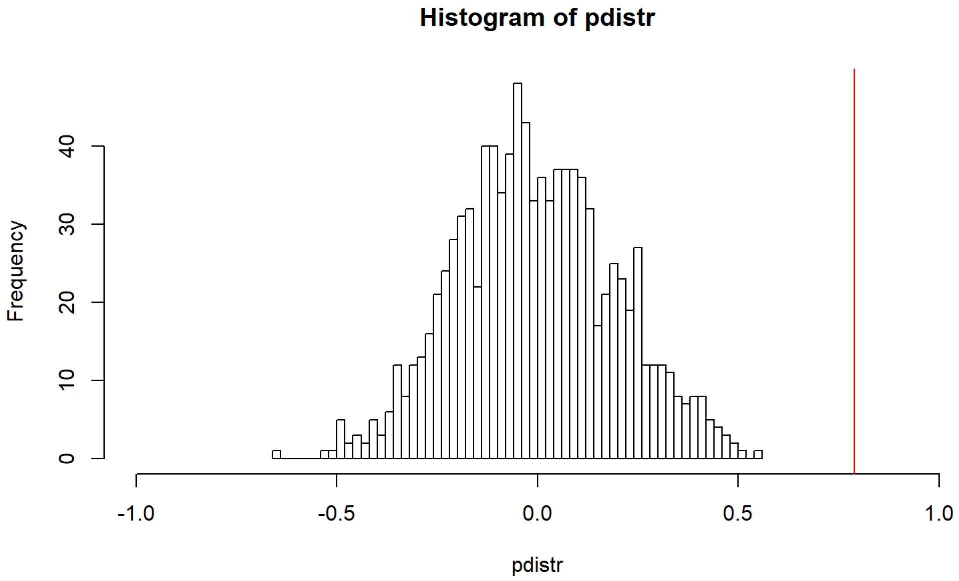
Assumption free test based on permutations of the data vectors.

Or, is it a coincidence that the combination pairs of the two data vectors occur together?

```r
ptest <- function(x, y, n_rep){
  # in sample observedd correlation
  c_obs <- gen_c(x, y)
  # correlation between permutation of x and y
  pdistr <- replicate(n_rep, gen_c(sample(x), y))
  # histogram of correlation between permutations of x and y
  hist(pdistr, xlim = c(-1, 1), breaks=50)
  # observed correlation as red line
  abline(v = c_obs, col = "red")
}
```

Permutation test result

```
ptest(attitude[ ,1] , attitude[ ,2], 999)
```



Exercise

Perform the permutation test on other combinations of attitude data vectors.

Which correlations are statistically significantly different from zero?