



Workbook Certified Professional Program Data Science

DIKW Academy



Contents

Preface	5
1 Introduction	7
1.1 General course information	7
1.2 Data science in Business	8
1.3 Course learning objectives	8
1.4 Study intructions	9
1.5 Study load	9
1.6 Prerequisite knowledge	9
1.7 Overall assessment	9
2 Block 01 : Course Introduction and Introduction to R and R-Studio	11
2.1 LU01 : Introduction to Datascience	11
2.2 LU02 : Introduction to R and R Studio	12
3 Block 02 : Self Study	15
3.1 Reading assignment	15
3.2 Learning R	15
4 Block 03 : Programming in R	17
4.1 LU03 : Vector calculus, sampling, bootstraps and hypothesis tests	17
4.2 LU04 : Measurement levels & scales, correlation	18
5 Block 04 : Self Study	21
5.1 Reading assignment	21
5.2 Learning R	21
5.3 Optional	21
6 Block 05 : Data Wrangling and visualisation	23
6.1 LU05 : Data Wrangling	23
6.2 LU06 : Data visualisation	24
7 Block 06 : Self Study	27
7.1 Reading assignment	27
7.2 Assignments	27
8 Block 07 : Predictive Analytics and Matrix Algebra	29
8.1 LU07 : Bivariate Analysis	29
8.2 LU08 : Decision Trees	30
9 Block 08 : Self Study	33
9.1 Reading assignment	33
9.2 Assignments	33

10 Block 09 : Introduction Machine Learning	35
10.1 LU09 : Introduction Machine Learning	35
10.2 Guest Lecture	37
11 Block 10 : Self Study	39
11.1 Reading assignment	39
11.2 Assignments	39
12 Break	41
13 Block 11 : Machine Learning - Weak Learners	43
13.1 LU11 : Multivariate Analysis and Matrix Algebra	43
13.2 LU12 : The Classification Task	45

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 which was studied in learning unit 4. In this learning unit, we combine the concepts of correlation and permutation. Similar to bootstrapping, a permutation technique is used to get a more robust estimate of the statistic you are exploring. In this case, it is the correlation 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 -1 and 1 or 0 and 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 = $(1 / \text{length}(x) - 1) * \sum((x - \text{mean}(x)) * (y - \text{mean}(y)))$

$(\text{sd}(x) * \text{sd}(y)) = \text{product standard deviation } x \text{ and standard deviation } y$

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 parameter 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 primary 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 X_i 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

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

Permutation test

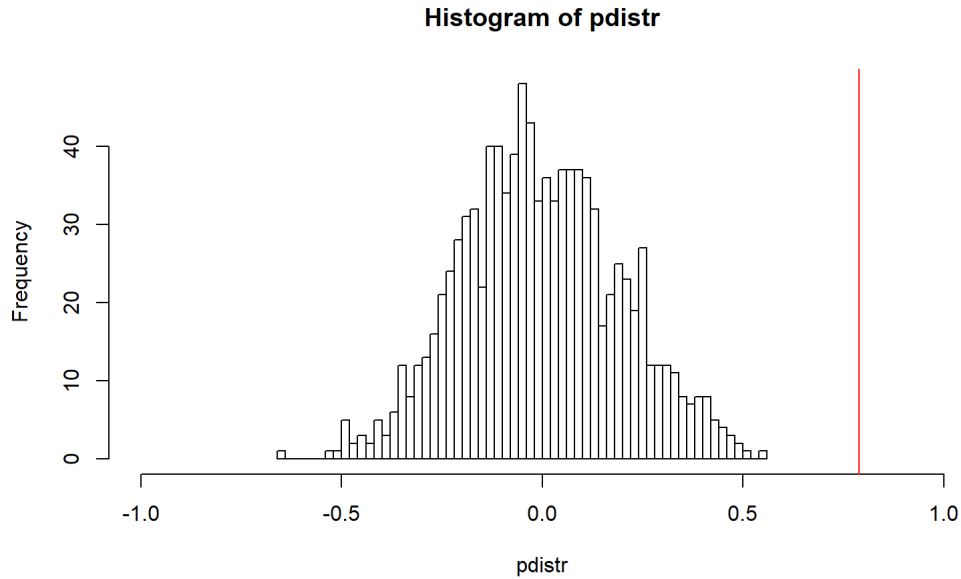
Assumption free test based on permutations of the data vectors.

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

```
pctest <- 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?