**Universität Bielefeld**

CITEC / Faculty of Technology
Ambient Intelligence Group
Dr. Thomas Hermann

Sekretariat: S. Strunk
Office: CITEC-3.311
106-6891

**Exercises of the lecture "Introduction to Data Mining"**

**WS 2018/2019 Excercise sheet 2**

# - Hypothesis testing and dependency analysis -

## Problem 2.1, Kolmogorov-Smirnov-Test

In the file `zahlen.csv` in the e-Learning space (Lernraum) of the ekVV you can find multiple distributions. Test each of these distributions against the normal distribution of the same mean and standard deviation values as the data by using the Kolmogorov-Smirnov-Test.

- Are there distributions that significantly differ from the normal distribution?
- If yes, which are the ones?
- To visualize the results plot a histogram of each distribution. To make the histograms compareable, the distance of the bins should always be 0.1.

The following Python functions may prove helpful to you: `scipy.stats.kstest`, `pylab.load`, `pylab.hist`

**Additional task 1:** If all of the 50 distributions actually were samples of a normal distribution, how often would you still expect to be able to reject the null hypothesis according to the KS-test?

**Addtitional task 2 (optional):** Implement the Kolmogorov-Smirnov-Test in Python and compare its results to the results of the SciPy implementation.

## Problem 2.2, Detection of dependencies between variables

Every year a large amount of money is spent in Germany's universities. A list of all DFG funds can be found at `unis.ods` in the e-Learning space (Lernraum) of the ekVV. Wouldn't it be interesting to know whether there is a dependency between the number of students and the assigned amount of money?

**a)** Fill in the following table using the data from the downloaded .ods-file.

| DFG / Students | > 30.000 | 20.000 - 30.000 | < 20.000 | SUM |
|---|---|---|---|---|
| **> 200 Mio.** | 2 | 2 | 0 | 4 |
| **100 - 200 Mio.** | 5 | 8 | 3 | 16 |
| **< 100 Mio.** | 3 | 5 | 2 | 10 |
| **SUM** | 10 | 15 | 5 | 30 |

The null hypothesis is $H_0$: "The variables DFG and students are independent."

If $H_0$ can be rejected, use the "Cramer's V" to calculate the degree of association and in advance the contingence-coefficient.

**b)** Fill in the following table using the data from the downloaded .ods-file.

| DFG / Etat | > 500 Mio. | 300 - 500 Mio. | < 300 Mio. | SUM |
|---|---|---|---|---|
| **> 200 Mio.** | 2 | 1 | 1 | 4 |
| **100 - 200 Mio.** | 3 | 5 | 5 | 13 |
| **< 100 Mio.** | 0 | 4 | 2 | 6 |
| **SUM** | 5 | 11 | 7 | 23 |

The null hypothesis is $H_0$: "The variables DFG and Etat are independent."

If $H_0$ can be rejected, use the "Cramer's V" to calculate the degree of association and in advance the contingence-coefficient.

Thomas Hermann (thermann@techfak)

Ferdinant Schlatt (fschlatt@techfak)

Lecture Wed 12-14, CITEC lecture hall

Tutorial Tue 12:15-14, H11

office building

Universität Bielefeld

Universitätsstraße 25

33615 Bielefeld

public transportation

light rail 4

from Bahnhof and Jahnplatz