Fachhochschule Dortmund
University of Applied Sciences and Arts

Data Science

06: Bivariate characteristics

Klaus Kaiser
WiSe 2024 / 2035

How can we formalize such a correlation?

Describing correlation of two variables with a dimension number (correlation coefficient). There are different ways to compute a correlation coefficient depending on the type of variables.

- Nominal characteristics: Measures of association
- Ordinal characteristics: Rank correlation coefficients
- Metric characteristics: Correlation coefficients

Data Science

Recap Contingency table

The two-dimensional frequency distribution of nominal variables is often represented by a **contingency table.** A $l_X - l_Y$**-contingency** table consists of l_X-rows, for each value of the first variable one, and l_Y-columns, for each value of the second variable.

X \ Y	b_1	b_2	$\dots$	b_{l_Y}
a_1	h_{11}	h_{12}	$\dots$	h_{1l_Y}
a_2	h_{21}	h_{22}	$\dots$	h_{2l_Y}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_{l_X}	h_{l_X1}	h_{l_X2}	$\dots$	$h_{l_Xl_Y}$

- $l_X - l_Y$-contingency table with absolute frequencies

X \ Y	b_1	b_2	$\dots$	b_{l_Y}
a_1	f_{11}	f_{12}	$\dots$	f_{1l_Y}
a_2	f_{21}	f_{22}	$\dots$	f_{2l_Y}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_{l_X}	f_{l_X1}	f_{l_X2}	$\dots$	$f_{l_Xl_Y}$

- $l_X - l_Y$-contingency table with relative frequencies

Data Science

Recap Contingency table

- Let $f_{\bullet j} > 0$, then we define the **conditional frequency** of $X = a_i$ under the condition $Y = b_j$ as

$$f_{X=a_i|Y=b_j} = \frac{f_{ij}}{f_{\bullet j}} \text{ for } i \in \{1, \ldots, l_X\}$$

- Let $f_{i\bullet} > 0$, then we define the **conditional frequency** of $Y = b_j$ under the condition $X = a_i$ as

$$f_{Y=b_j|X=a_i} = \frac{f_{ij}}{f_{i\bullet}} \text{ for } j \in \{1, \ldots, l_Y\}$$

- Note the notation: $AA|BB$ denotes AA holds under the condition that BB holds - in short: AA holds given BB.
- $f_{Y=b_1|X=a_i}, \ldots, f_{Y=b_{l_Y}|X=a_i}$ is called **conditional frequency distribution** of Y under the condition $X = a_i$

Data Science

Recap Contingency table

- The conditional frequency distributions of $X|Y = b_{j_1}$ and $X|Y = b_{j_2}$ for $j_1 \in \{1, \ldots, l_Y\}$ and $j_2 \in \{1, \ldots, l_Y\}$ equal, if for the relative conditional frequencies holds:

$$f_{X|Y=b_{j_1}} = f_{X|Y=b_{j_2}}$$

for all $i = 1, \ldots, l_X$.

The variable X is **empirically independent** of variable Y if all conditional frequency distributions of $X|Y = b_j$ for all $j = 1, \ldots, l_Y$ are equal.

- X is empirically independent of $Y \Leftrightarrow$ Y is empirically independent of X
- X is empirically independent of $Y \Leftrightarrow h_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$

Data Science

Recap Contingency table

Measures of association

- The **Pearson contingency coefficient** K_P is given by

$$K_p = \sqrt{\frac{\chi^2}{\chi^2 + n}} \text{ with } K_P \in \left[0, \sqrt{\frac{\min(l_X - 1, l_Y - 1)}{\min(l_X, l_Y)}}\right]$$

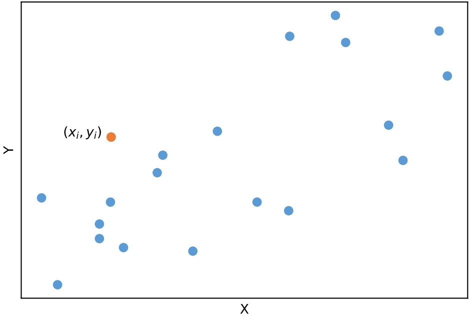
- The **corrected Pearson contingency coefficient** K_P^* is given by

$$K_P^* = \frac{K_P}{\max K_P} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{\frac{\min(l_X, l_Y)}{\min(l_X - 1, l_Y - 1)}} \text{ with } K_P^* \in [0, 1]$$

Data Science

Recap

In a **scatter plot**, every variable of (X, Y) is associated with one axis of Cartesian co-ordinates. Every observation (x_i, y_i) for $i = 1, \ldots, n$ is marked, e.g. with a cross or a point.



Data Science

Recap

The **Pearson correlation** coefficient r_{XY}, also **empirical correlation,** for two variables X and Y is given by

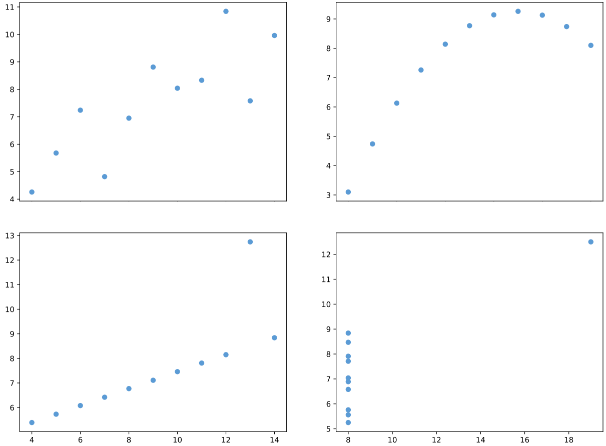
$$r_{XY} = \frac{\tilde{s}_{XY}}{\tilde{s}_X \tilde{s}_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

where $\tilde{s}_X$ and $\tilde{s}_Y$ are the empirical standard deviations of the variable X and Y.

- Correlation coefficient is a dimensionless number
- Correlation coefficient only measures a linear correlation between the two variables

Data Science

Recap



The **Anscombe's quartet** describe four datasets, where all have (nearly) the same descriptive statistics, but are completely different.

Take also the scatter plot into account to verify a possible correlation in the data!

Data Science

Today

Basics of **probability theory**

Data Science

Today:

Up to now: **Describing data** - how does the data look like, are there correlations?

- Data often shows variation, i.e. they scatter around values - this leads to an uncertainty
- Often: random processes are involved!

Probability theory helps us to understand how data was generated and where it stems from! By this, it helps us to understand what will be the most likely result given a setting.

Machine learning Learn from data to predict the most likely result for unknown data!

Probability and Combinatorics

Probability and Combinatorics

Probability

In 1654 Chevalier de Méré set a task to Blaise Pascal

Is it true, that in the following to plays the first one will lead more often to a victory?

- A die is thrown four times: The bet is that at lease one 6 will occur!
- Two dice are thrown 24 times: The bet is that at least one pair of 6s will occur!

First modern, mathematically oriented study of probability

We will try to answer this question in the end of the lecture!

A **random experiment**, is an experiment for which

- conditions are well-defined
- multiple different outcomes are possible
- it is not predictable which outcome will occur

Example A: Simple dice flip

A dice with six sides is thrown. The result can not be predicted in before.

Example A: Color of a passing car

The color of the next passing car is predicted. The result can not be predicted in before.

Data Science

Probability: Event

We define the following:

- An **result** ω is an elementary outcome of a random experiment.
- The **sample space** Ω is the collection of all possible outcomes (result).

$$\Omega = \{\omega | \omega \text{ is possible outcome of the random experiment}\}$$

- An **event** A is a subset of the samples space: $A \subset \Omega$
- An **event** A **occurs,** if the result ω of the random experiment is element of A: $\omega \in A$
- The empty set $\emptyset$ is called **impossible event** and Ω **sure event**

Example A: Simple dice flip

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $A = \{2, 4, 6\}$, $B = \{6\}$

Example A: Color of a passing car

- $\Omega = \{white, red, black, \ldots\}$
- $A = \{black, red, gold\}$, $B = \{darkblue\}$

Data Science

Probability: Event

In the following!

Basic properties of probabilities are defined: Axioms derived by Kolmogoroff in 1933 - up to now: no better way!

This is typically for math: The natural numbers are defined with the help of the Peano-axioms (e.g. every natural number has exactly one successor)

Data Science

Probability: Theory of quantities

Let Ω be a sample space and $A, B \subset \Omega$ two events, $\mathcal{P}(\Omega)$ the power set of Ω.

$\mathcal{P}(\Omega)$ denotes the power set of Ω, i.e. the set of all possible subsets of Ω:

$$\mathcal{P}(\Omega) = \{A | A \subset \Omega\}$$

$A \cap B$	intersection	A and B occur
$A \cup B$	union	A or B or both occur
$A \setminus B$	difference	A occur but not B
$A \Delta B$	symmetric difference	A or B occur, but not both
$\bar{A} = \Omega \setminus A$	complement	the opposite of A occurs

If $A \cap B = \emptyset$, then we call these events disjoint, i.e. If A occurs, then B can not occur and vice versa.

Data Science

Probability: Laplace experiment

A random experiment is called **Laplace experiment**, if all possible results of the experiment have the same chance to occur.

Example A

The dice flip experiment is a Laplace experiment if the dice is fair.

Example A

The passing car experiment is not a Laplace experiment, since every color appears differently often.

Data Science

Probability: Kolmogoroff axioms

A probability measure is a function $P : \mathcal{P}(\Omega) \to \mathbb{R}$ with the following properties:

- $P(A) \geq 0$ for all $A \subset \Omega$
- $P(\Omega) = 1$
- For piece-wise disjoint events $A_1, A_2, \ldots$ there holds
  - $P(A_1 \cup A_2 \cup \cdots \cup A_k) = \sum_{i=1}^{k} P(A_i)$ finite many and
  - $P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$ for countable infinite many.

Some properties of probability measures

- $P(\bar{A}) = 1 - P(A)$ for all $A \in \mathcal{P}(\Omega)$
- $P(\emptyset) = 0$
- If $A \subset B$ then $P(A) \leq P(B)$
- $P(\bar{A}) \leq 1$ for all $A \in \mathcal{P}(\Omega)$

Data Science

Probability: Event

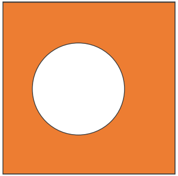
Let $A, B \subset \Omega$ two events of sample space Ω.

If $A \cap B = \emptyset$, then A and B are disjoint:

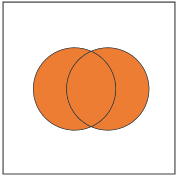
$$P(A \cap B) = 0$$

Data Science

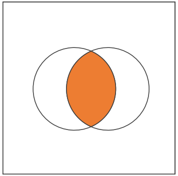
Probability: Event



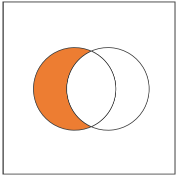
$$P(\bar{A}) = P(\Omega) - P(A)$$
$$= 1 - P(A)$$



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



$$P(A \cap B)$$



$$P(A \setminus B) = P(A) - P(A \cap B)$$

Data Science

Probability: Event

In the case of a Laplace random experiment, every result has the same chance to occur, with $|\Omega| < \infty$ then the **Laplace probability measure** can often be computed by counting:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{\# number of favorable cases}}{\text{\# number of possible cases}}$$

Counting sounds easy, unfortunately this is not always the case - In more complicate cases: Combinatorics!

Example

Trowing a coin (Head (H) and Tail (T)) twice, then there are four possible results (we take the order into account)

$\Omega = \{HH, HT, TH, TT\}$

then

$$\mathcal{P}(\Omega) = \{\emptyset, \{HH\}, \dots, \{TT\}, \{HH, HT\}, \dots, \{HT, THTT\}, \Omega\}.$$

All results have the same probability, i.e.

$$P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}$$

$P(A)$ can be computed for all $A \in \mathcal{P}(\Omega)$, e.g: E=”at least one tail”:

$$P(E) = \frac{|\{HT, TH, TT\}|}{|\Omega|} = \frac{3}{4} = P(HT) + P(TH) + P(TT)$$

Probability and Combinatorics

Combinatorics

Data Science

Combinatorics: Random experiment

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\#\text{ number of favorable cases}}{\#\text{ number of possible cases}}$$

- For simple cases easy: Throwing a dice has 6 possible results and 3 of them are even - probability that the result is even: $\frac{3}{6}$
- But can get more complex: To color three elements, one can choose from 10 colors - what is the probability that at least one element is red?

- The **factorial value** of a natural number $n \in \mathbb{N}$ is given by

$$n! = \prod_{i=1}^{n} i = n \cdot (n-1) \cdot (n-1) \cdot \dots \cdot 2 \cdot 1.$$

  The factorial value of 0 is 1: $0! = 1$

- The **binomial coefficient** "n over k", for $n, k \in \mathbb{N}_0$, is given by

$$\binom{n}{k} = \begin{cases} \frac{n!}{k! \cdot (n-k)!} & n \geq k \\ 0 & n < k \end{cases}$$

Data Science

Combinatorics: Permutation

Consider a tuple of $n \in \mathbb{N}$ objects, i.e. $(x_1, \ldots, x_n)$, with k different values. Then, an order of this set is called **permutation**, e.g. $(x_3, x_1, \ldots, x_4)$.

How many permutations are possible?

1. $k = n$, i.e. no values occurs more than once:

> The number of all possible permutations in the case $n = k$ is
> $$n!.$$

2. $k < n$, i.e. the k-values occur with frequency $n_1, \ldots, n_k$:

> The number of all possible permutations in the case $n < k$ is
> $$\frac{n!}{n_1! \cdot \ldots \cdot n_k!}$$

Example: Lego

There are 3 red and 2 blue Lego stones - how many different colored towers could be build?

$$n = 5, n_1 = 3, n_2 = 2 \Rightarrow \frac{n!}{n_1!n_2!} = \frac{20}{2} = 10$$

Example

There is a box with 20 different Lego stones, how many different choices are possible, when choosing three of them randomly?

$$n = 20, k = 3 \Rightarrow \binom{20}{3} = 1140$$

Number of possible samples with k observations out of n objects is given by the following table

	Repeats allowed	No Repeats ($k \leq n$)
Combinations (order doesn't matter)	$\binom{n+k-1}{k}$	$\binom{n}{k}$
Permutations (order matters)	n^k	$\frac{n!}{(n-k)!}$

Example: Side dishes in a restaurant

How many combinations choosing 2 of 10 side dishes are possible if

- both are of the same size and you are allowed to take one twice?
- both are of the same size, but they should be different?
- the first is larger than the second, but you are allowed to take one twice?
- the first is larger than the second and they should be different?

Example: Side dishes in a restaurant

How many combinations choosing 2 of 10 side dishes are possible if

- both are of the same size and you are allowed to take one twice?
  $\binom{10+2-1}{2} = 55$
- both are of the same size, but they should be different?
  $\binom{10}{2} = 45$
- the first is larger than the second, but you are allowed to take one twice?
  $10^2 = 100$
- the first is larger than the second and they should be different?
  $\frac{10!}{(10-2)!} = 90$

Data Science

Combinatorics: Random experiment

For an operation given as a sequence of r-steps, with:

- the Number of possible chances to finish the first step is given by n_1
- For every possible result of the first step, the number of possible chances to finish the second step is given by n_2
- ...

then the number of all possible sequences to finish the operation is given by

$$n_1 \cdot n_2 \cdot \cdots \cdot n_r$$

Example

How many passwords are possible if

- it has a length of 6
- the first two digits are one of the following letters: A, B, C, D, E, F, G
- the last four digits are numbers, where no number occurs twice

Examples: AG1243, GG5230

Example

How many passwords are possible if

- it has a length of 6
- the first two digits are one of the following letters: A, B, C, D, E, F, G
- the last four digits are numbers, where no number occurs twice

Examples: AG1243, GG5230

- AA: Choosing $k = 2$ out of $n = 7$ with order and repetition: $n^k = 7^2 = 49$
- 1234: Choosing $k = 4$ out of $n = 10$ with order and no repetition: $\frac{n!}{(n-k)!} = \frac{10!}{8!} = 5040$

With the multiplication rule: $49 \cdot 5040 = 246960$ possible passwords.

Conditional probability

Conditional probability

Stochastically independency

Data Science

Stochastically independency

How does the probability of the event A change, if we know, that event B occurred.

Example

- A = "Road is wet", B = "It has rained an hour ago"
- $A \cap B$ = "Road is wet and it has rained an hour ago"
- $A|B$ = "Road is wet given that it has rained an hour ago"

The conditional probability of A under the condition of B (or P of A given B) is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example: Trowing two dices

What is the probability that the first dice shows an odd number given that the sum of both dices is 5?

- $\Omega = \{(1,1),(1,2),\dots,(6,6)\}$, $|\Omega| = 36$
- A : "the first dice shows an odd number" = $\{(1,1),(1,2),\dots,(5,6)\} \Rightarrow P(A) = \frac{18}{36}$
- B : "the sum ob both dices is 5" = $\{(1,4),(2,3),(3,2),(4,1)\} \Rightarrow P(B) = \frac{4}{36}$
- $A \cap B$ = $\{(1,4),(3,2)\} \Rightarrow P(A \cap B) = \frac{2}{36}$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{4} = 0.5$$

Data Science

Stochastically independency

Events A and B are called **(stochastic) independent** if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$

Otherwise A and B are called **(stochastic) dependent.**

The following is equivalent ($\Leftrightarrow$) to A and B are stochastically independent

- $\Leftrightarrow$ $P(B|A) = P(B)$
- $\Leftrightarrow$ $P(A|B) = P(A)$
- $\Leftrightarrow$ $\overline{A}$ and $\overline{B}$ are stochastically independent
- $\Leftrightarrow$ $\overline{A}$ and B are stochastically independent
- $\Leftrightarrow$ A and $\overline{B}$ are stochastically independent

Data Science

Stochastically independency

Example: Thworing two dices

A : ”the first dice shows an odd number” and B : ”the sum ob both dices is 5” is stochastically independent

$$\Leftrightarrow$$

$\overline{A}$: ”the first dice shows an even number” and $\overline{B}$: ”the sum ob both dices is not 5” is stochastically independent

Data Science

Stochastically independency

- Events $A_1, \ldots, A_k$ are called **piece-wise independent** if A_i and A_j are independent for all $i \neq j$.
- Events $A_1, \ldots, A_k$ are **mutual independent** if for every subset $A_{i_1}, \ldots, A_{i_j}$ for $j \leq k$ there holds
$$P(A_{j_1} \cap \cdots \cap A_{i_j}) = P(A_{i_1}) \cdot \cdots \cdot P(A_{i_j}).$$

In 1654 Chevalier de Méré set a task to Blaise Pascal

Is it true, that in the following to plays the first one will lead more often to a victory?

- A die is thrown four times: The bet is that at lease one 6 will occur!
- Two dice are thrown 24 times: The bet is that at least one pair of 6s will occur!

Data Science

Stochastically independency

In 1654 Chevalier de Méré set a task to Blaise Pascal

Is it true, that in the following to plays the first one will lead more often to a victory?

- A die is thrown four times: The bet is that at lease one 6 will occur!
- Two dice are thrown 24 times: The bet is that at least one pair of 6s will occur!

- P(at least one 6 in 4 throws) = 1-P(no 6 in 4 throws)
  = 1-P(no 6 in throw 1)P(no 6 in throw 2)P(no 6 in throw 3)P(no 6 in throw 4)
  = $1 - \left(\frac{5}{6}\right)^4 = 0.518$
- P(at least one pair of 6s in 24 throws) = $1 - \left(\frac{35}{36}\right)^2 4 = 0.491$

Conditional probability

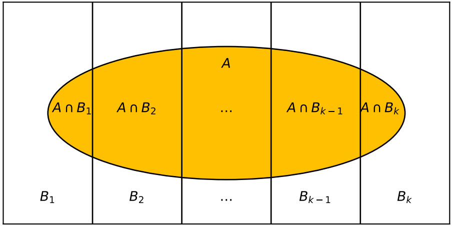
Bayes’ theorem

Data Science

Bayes’ theorem

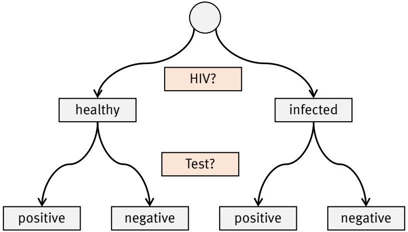
In the following: We assume that there are $B_1, \ldots, B_k$ events which fulfill the following conditions:

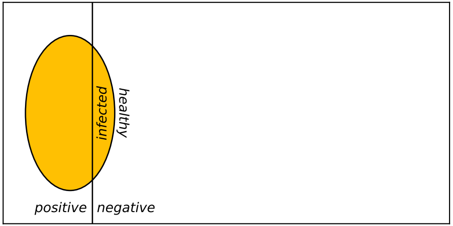
- $B_i \cap B_j = \emptyset$ for $i \neq j$, i.e. the events are piece-wise disjoint
- $\bigcap_{i=1}^{k} B_i = \Omega$, i.e. the events cover the complete sample space



Diagnostic test in medicine

How large is the probability that in the case of a positive HIV test, the patient is actually infected?





Data Science

Bayes' theorem

Law of total probability

Under the given conditions, there holds for every $A \subset \Omega$:

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i)$$

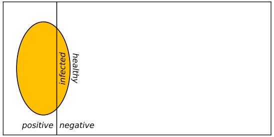
Bayes' Theorem

Under the given conditions, there holds for every $A \subset \Omega$

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

Data Science

Bayes' theorem



- Ω = "group of persons"
- A = "positive result of a test"
- B = "person is infected"

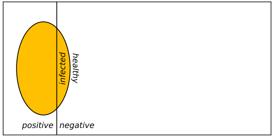
Let assume for a specific test method:

- $P(A|B) = 0.9$: For infected persons, the test is in 90 % of the cases positive
- $P(\overline{A}|\overline{B}) = 0.98$: For healthy persons, the test is in 98 % of the cases negative
- $P(B) = 0.01 \Rightarrow P(\overline{B}) = 0.99$: The proportion of infected persons is 1 %.

Task: What is $P(B|A)$ = "person is infected given a positive test"?

Data Science

Bayes' theorem



- Ω = "group of persons"
- A = "positive result of a test"
- B = "person is infected"

Let assume for a specific test method:

- $P(A|B) = 0.9$: For infected persons, the test is in 90 % of the cases positive
- $P(\bar{A}|\bar{B}) = 0.98$: For healthy persons, the test is in 98 % of the cases negative
- $P(B) = 0.01 \Rightarrow P(\bar{B}) = 0.99$: The proportion of infected persons is 1 %.

Task: What is $P(B|A)$ = "person is infected given a positive test"?

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P((\bar{B}))} = \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.02 \cdot 0.99} = 0.31$$

Summary & Outlook

Data Science

Summary & Outlook: Summary

- You understand the basics of probability theory and are able to apply it
- You know basic ways to compute the number of possibilities and are able to select them depending on the setting
- You know Bayes' theorem and are able to apply it

Random variables

References

Parts of the lecture base on the lecture ”Statistics” (FH Dortmund)

by

Prof. Dr. Sonja Kuhnt and Prof. Dr. Nadja Bauer.