UNIVERSITÄT
LEIPZIG

10-207-0003: Introduction to Stochastics

Introduction

04.04.2025, Leipzig

Dr. Ing. Andreas Niekler

Computational
Humanities
UNIVERSITÄT LEIPZIG

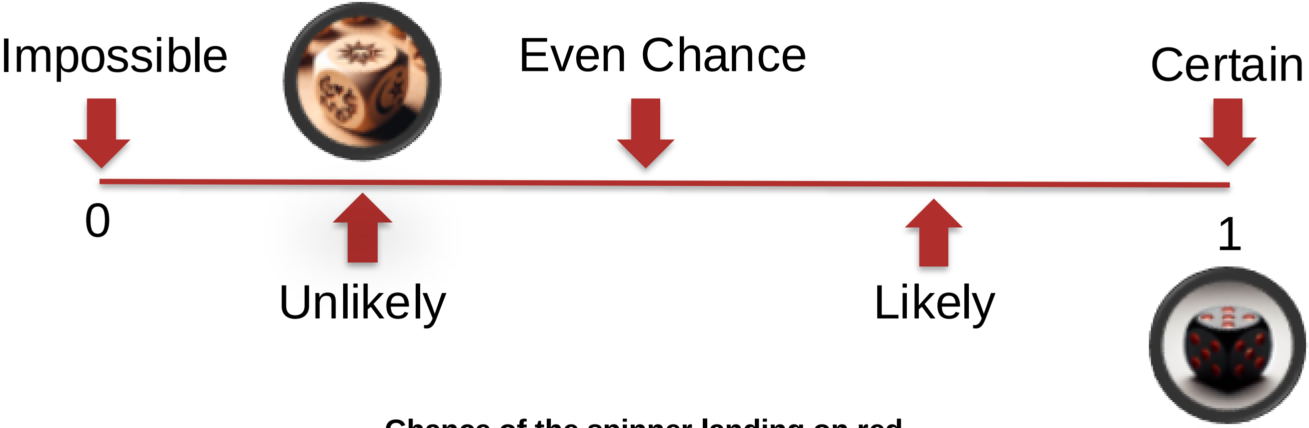
UNIVERSITÄT
LEIPZIG

RANDOMNESS

- Randomness, the **absence of certainty** or predictability in an event or outcome, is a **fundamental concept that permeates** various aspects of our lives.

- By **understanding the concept of randomness** and its pervasiveness in various aspects of life, we gain a **deeper appreciation for the complexity and unpredictability** of our world.

PROBABILITY: A POWERFUL TOOL FOR UNDERSTANDING RANDOMNESS



Impossible

Even Chance

Certain

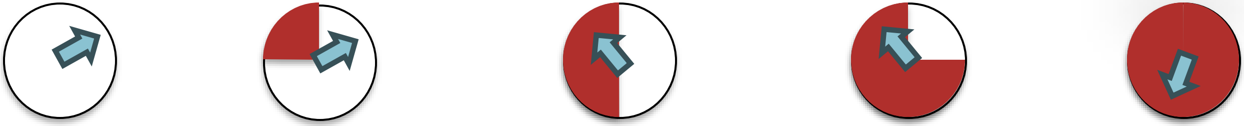
0

1

Unlikely

Likely

Chance of the spinner landing on red



SYLLABUS

1. Empirical research and scale levels
2. Statistical parameters
3. Graphical representation of characteristics / Explorative data analysis
4. The random experiment
5. Combinatorics, permutation
6. Probability theory
7. Probability distributions
8. Central Limit Theorem
9. Confidences
10. Statistical testing
11. Linear Regression
12. Correlation and covariance
13. Logistic regression
14. Bayes theorem

Additional: Entropy, Mutual Information, Maximum Likelihood Estimator, Mathy Stuff

LECTURE

- **Room**: HS 15
- **When**: Wednesday 13:15 – 14:45
- **Who:** Andreas Niekler -- andreas.niekler@uni-leipzig.de

- Form of the lecture will be a **classroom-based presentation**

- 2 SWS = 30h classroom, 50h = individual study time

EXERCISE

- **Room**: SG 3-12
- **When**: Thuesday 9:15 – 10:45
- **Who**: Nicolas Ruth -- nicolas.ruth@informatik.uni-leipzig.de

- The exercise will be a **mixture of exercises and programming** tutorials using Python

- We'll start on the 22nd of April

- There will be **no weekly tests**

- 2 SWS = 30h classroom, 40h = individual study time

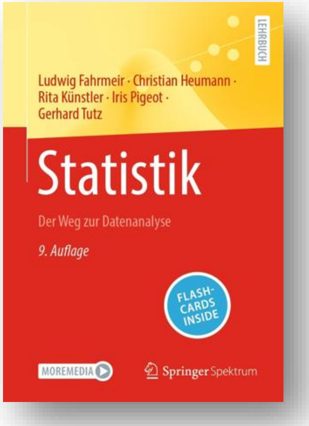
MOODLE COURSE

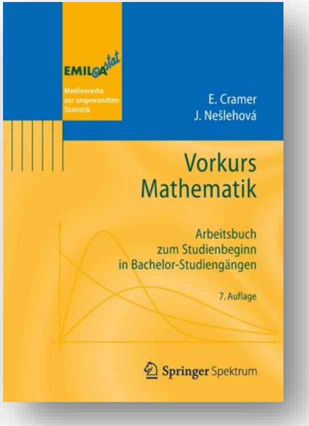
https://moodle2.uni-leipzig.de/course/view.php?id=54754

EXAM

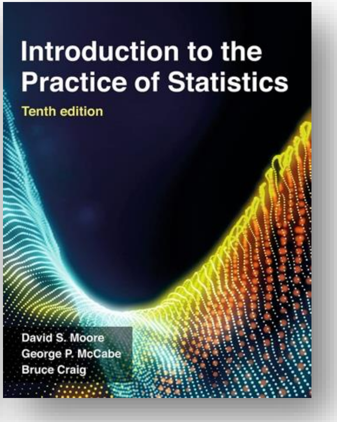
- Written exam -- 60 min
- Last week of the lecture period
- Late deregistration can lead to a failed attempt at the exam

LITERATURE I

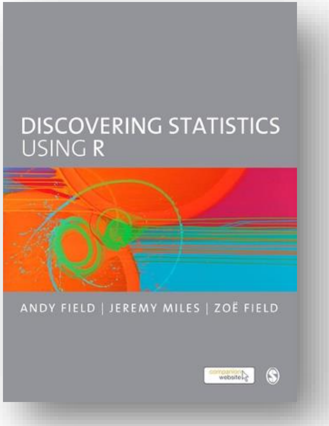


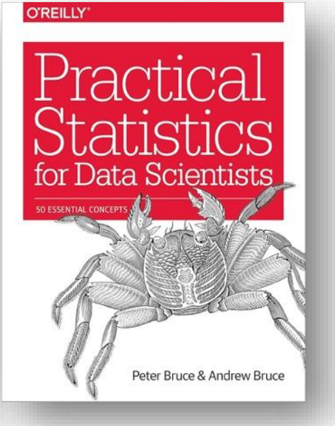

EMILeA-stat
Medienreihe zur angewandten Statistik
E. Cramer
J. Nešlehová
Vorkurs Mathematik
Arbeitsbuch zum Studienbeginn in Bachelor-Studiengängen
7. Auflage
Springer Spektrum

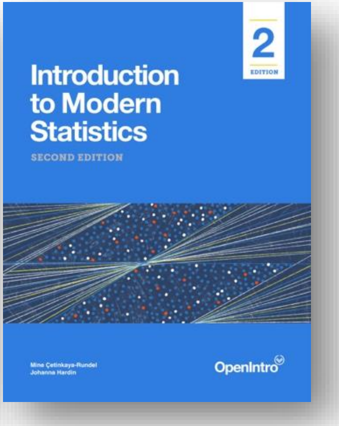
www.lehrbuch-psychologie.de
Bortz
Schuster
Statistik
für Human- und Sozialwissenschaftler
7. Auflage
Springer

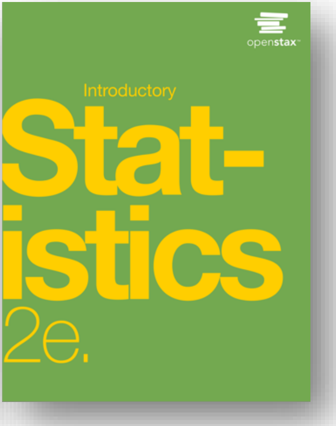


LITERATURE II


DISCOVERING STATISTICS USING R
ANDY FIELD | JEREMY MILES | ZOË FIELD
companion website


O'REILLY
Practical Statistics for Data Scientists
50 ESSENTIAL CONCEPTS
Peter Bruce & Andrew Bruce





WHY STUDY STATISTICS?

- Understanding of **quantitative-empirical methods** and the correct *interpretation of scientific findings*.
- As aspiring scientists, the ability to **generate and analyze data**, often with computational support, is crucial.
- The unique logic of statistical reasoning enables **decision-making under uncertainty** (absence of information).
  - The ongoing development of statistics as a scientific discipline is one of the **key intellectual achievements** of the last century.
  - Current advancements in **AI, information technology, and causality theory would not be possible** without statistical principles and are further advancing the field.

DATA AND STATISTICS

- Statistics is an important **component of empirical scientific work**, which deals with the **creation, interpretation, summarization and presentation of data**.
    - Scientific work aims to consolidate **individual pieces of information** and observations into **generally valid statements**.

👉 **Data:** Facts created by observation or experiments

👉 **Descriptive statistics:** Leads to a clear and descriptive presentation of information

👉 **Inferential statistics:** Makes it possible to test hypotheses against observed reality.

MEASUREMENT

- In a measurement, objects or events (also referred to as the **empirical relative**) are assigned specific numbers according to certain rules (the **numerical relative**).
- **Example (Empirical relative)**:
  - The length of books determined bye eye

MEASUREMENT

👉 In a measurement, objects or events (also referred to as the **empirical relative**) are assigned specific numbers according to certain rules (the **numerical relative**).

- **Example (numerical relative):**
  - The length of books determined bye **PAGES**

896 ▶ 216

OPERATIONALIZATION

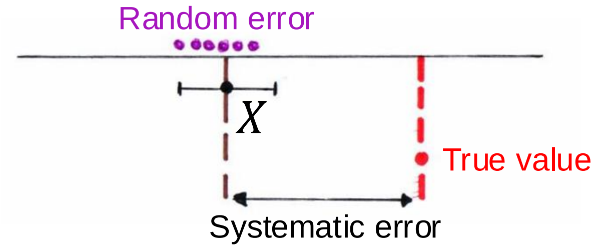
– Operationalization refers to the process of **“making a theoretical construct measurable”**.

👉 **Validity:** Does my measurement instrument really measure what it is supposed to measure, and does my data provide a valid representation of the theoretical construct of interest?

👉 **Reliability:** Is my measuring instrument consistent and does it produce the same results with repeated measurements? Is my data free from random or situational errors?

MEASUREMENT

- $X = T + e_R + e_S$ with
  - X, recorded value
  - T, the true value
  - e_R, random error
  - e_S, systematic error



- Systematic errors always lead to wrong measurements whereas random errors just introduce insecurity

- Valid measurement: $X = T + e_R$

BASIC CONCEPTS I

👉 The objects to which a statistical study refers are called statistical units;

  - depending on the context, the synonymous terms **carriers, study objects, units, cases or or data points** are also used

- Example **Library Checkout**:

  - Type: Senior
  - Checkouts: 5
  - renewals:0
  - Month: Nov
  - Year: 2022

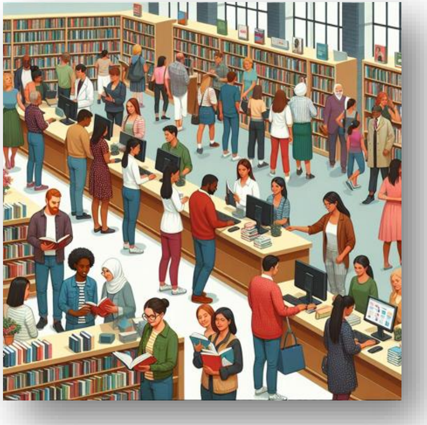
- We call the properties of a data point **feature**, a single manifestation **feature manifestation** and all possible manifestations of a feature **feature space**

BASIC CONCEPTS II

👉 The set of **all statistical units of interest** for a research question forms the **population** of the data.

- Examples of a **finite and concrete** population: All library checkouts in San Francisco, the USA or Europe in the last 50 years

- Examples of an **infinite or hypothetically infinite** population: Periodically repeated measurements such as dayly checkouts or daily temperature measurements or all people who will ever live.



BASIC CONCEPTS III

- A subset of a population, a so-called **subpopulation**, is often considered. Formally, a subpopulation is itself a population, which usually results as a **systematic selection** from a population **regarding a characteristic**.
  - Example: All library checkouts **by students**
- A survey of all variables of interest from all objects in the population is called a **complete survey**.
- Often it does not make sense to carry out a complete survey (not possible, too expensive). Instead, **a sample is drawn from the population**, which is as faithful a reflection of the population as possible, i.e. is **representative of this population**.

BASIC CONCEPTS IV

- By **representative**, we mean that no relevant characteristic is **over- or underrepresented**.

  - The subpopulation of all library checkouts by childs is not representive for the population of all library checkouts (systemic age bias) if we are interested in genre interests or total checkouts

- One way to select a representative sample is through **simple random selection**, i.e., the units are randomly determined so that **each unit has the same chance** of being **selected** in the random sample.

BASIC CONCEPTS V

- A suitable statistical analysis of the observed feature manifestations on the statistical units depends on the **properties of the possible feature manifestations**.

- Features can be distinguished by:
  - Number of manifestations: **discrete or continuous**
  - Type of manifestations: **quantitative or qualitative (categorical)**
  - Interpretation of manifestations: **nominal scale, ordinal scale, interval scale, rational scale**

BASIC CONCEPTS VI

- Features can be distinguished based on whether the possible feature manifestations are either **countable or uncountable**.

👉**Discrete Feature:** A feature is discrete if the possible manifestations can be coded using natural numbers, i.e., they can be counted.

  - Number of checkouts

👉**Continuous Feature:** A feature is continuous if the possible manifestations are no longer countable but uncountable.

  - Age

BASIC CONCEPTS VII

- Every measurement of a continuous feature is always discrete due to **finite measurement accuracy**. In this context, we speak of **quasi-continuous features** if they can still be considered continuous.

- Example of age: A woman with an actual age of 36 years, 3 days and 43,200 seconds is to be measured with accuracy to one year
  - The actual result of the measurement is an interval: [36 + 0 days, 36 + 365 days]
  - Recorded data point manifestation: 36 = 36.0000000 …

- Rule of thumb: If the measurement **error is much smaller than the relevant variation** of the observed manifestations (e.g., between 26 years and 40 years), we can treat quasi-continuous variables as continuous.

  - So, if you measure 50-year intervals, you are recording a discrete variable when measuring age.

BASIC CONCEPTS VIII

- **Quantitative Feature**: The manifestations of a feature are quantitative if they can be meaningfully measured by numbers that describe the **extent or intensity** of a property.

- **Qualitative Feature**: The manifestations of a feature are qualitative or categorical if they are specified in **distinct categories that do not measure the extent** of a property.

  - A qualitative feature describes the differences of a property, i.e., whether a manifestation falls into a certain category (or not), without making any statement about the extent of this property.

  - A feature that has only **two manifestations** (categories) is called **dichotomous**.

  - A discrete feature that has **more than two** manifestations is called **polytomous**.

  - Note: The term qualitative is sometimes a bit misleading, as the measurement of a qualitative feature in a population can indeed have a quantitative character (category count).

SCALES

- **Level of measurement (scales) of a feature:** When we assign certain numbers to the possible feature manifestations during measurement, we speak of a so-called coding. The corresponding **scale** describes what meaning we can attribute to these numbers and what arithmetic operations can be sensibly performed with them.

  - The scale reflects the empirical relations between the feature manifestations.

NOMINAL SCALE

👉 A feature or variable is called nominally scaled if the manifestations are **categories that have no natural order**. The manifestations are **logically mutually exclusive** and can be classified so that the **same manifestations mean the same** and different manifestations mean different.

- For data analysis, numbers are assigned to the manifestations. These numbers have **no intrinsic meaning** and can be assigned arbitrarily **if they are unique**.
-
  - Example of gender: The categories *{female, male, diverse}* can be coded, for example, with *{0, 1, 2}.*
  - At the level of measurement of the nominal scale, all one-to-one recodings are permissible.
  - Example of gender: The categories *{female, male, diverse}* could also be coded with *{1, 0, 2}* or *{21, 2, 31}*, *{−12, 33, 12}*, etc.

ORDINAL SCALE (RANK)

👉 An **ordinally scaled** feature has the same properties as a nominally scaled feature. Additionally, there is a **specific order** or sequence between the manifestations in the ordinal scale. The order allows us to make judgments about the **rank of a manifestation** relative to all other manifestations, such as ‚is older than' or ‚is younger than.

–

  - Age range, school marks (*1* < *2* < *3* < *4* < *5* < *6*)

- It is more informative regarding the order

- You can assign arbitrary numbers to the manifestations if they **reflect the empirical order**.
  - However, the **distances** between different measured values **cannot be meaningfully interpreted**.

INTERVAL SCALE

☞ An interval-scaled feature has the same properties as an ordinally scaled feature. Additionally, the **distances can be meaningfully interpreted** on the interval scale.

- The difference between the years 1500 AD and 1600 AD is the same as the difference between 1600 AD and 1700 AD. However, there is no true zero point in the AD calendar system.
- The year 0 does not mean the absence of time!

RATIONAL SCALE

- A ratio-scaled feature has the same properties as an interval-scaled feature. Additionally, there is a **meaningfully interpretable zero point**, i.e., the measured value 0 can be interpreted.

  – A natural zero point is 0, not born
  – If you are 25 years old and your friend is 23 years old you are 2 years older. This is relatable to the age difference between 40 and 42.
  – You can say that 20 is twice the age of 10.

SCALES OVERVIEW

Type	Comparison	Valid recoding
Nominal	$a = b$	One-to-one
Ordinal	$a < b, a \leq b$	monotonic
Interval	$a - b = c - d$	linear-affine
Rational	$\frac{a}{b} = \frac{c}{d}$	proportional

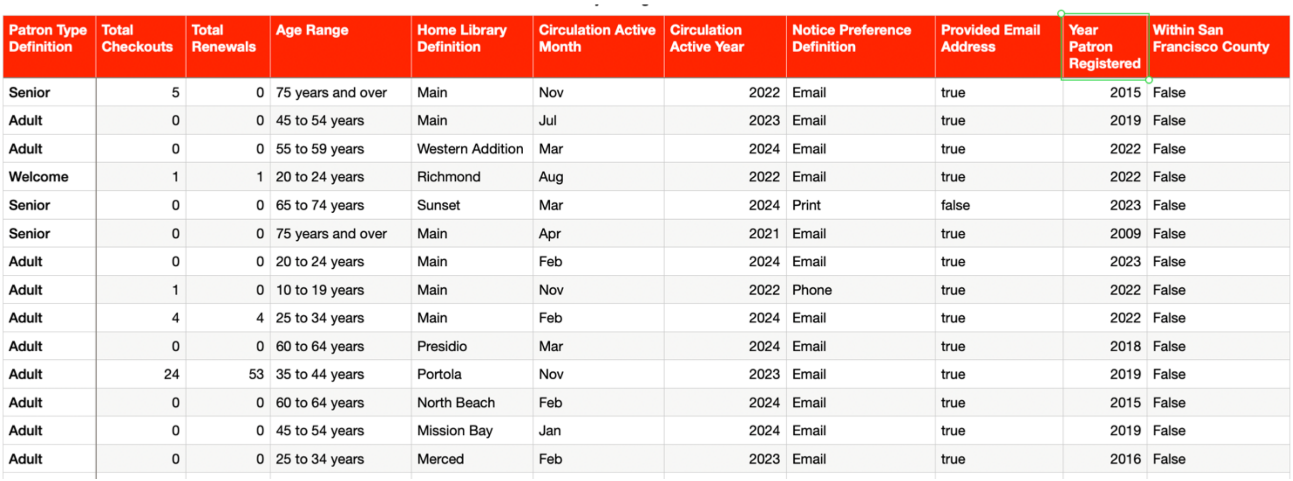
- Scales have a hierarchy:
  - NOIR (noir, french for *black*) = ***N****ominal* < ***O****rdinal* < ***I****nterval* < ***R****ational*
  - The higher the scale level, the more informative it is
  - Always assign the maximum permissible level of measurement

SCALES OVERVIEW

Type	Count	Sort	Differences	Ratios
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Rational	Yes	Yes	Yes	Yes

- **Note:** We often use interval and rational as one scale and call it metric!

TABULAR DATA



Patron Type Definition	Total Checkouts	Total Renewals	Age Range	Home Library Definition	Circulation Active Month	Circulation Active Year	Notice Preference Definition	Provided Email Address	Year Patron Registered	Within San Francisco County
Senior	5	0	75 years and over	Main	Nov	2022	Email	true	2015	False
Adult	0	0	45 to 54 years	Main	Jul	2023	Email	true	2019	False
Adult	0	0	55 to 59 years	Western Addition	Mar	2024	Email	true	2022	False
Welcome	1	1	20 to 24 years	Richmond	Aug	2022	Email	true	2022	False
Senior	0	0	65 to 74 years	Sunset	Mar	2024	Print	false	2023	False
Senior	0	0	75 years and over	Main	Apr	2021	Email	true	2009	False
Adult	0	0	20 to 24 years	Main	Feb	2024	Email	true	2023	False
Adult	1	0	10 to 19 years	Main	Nov	2022	Phone	true	2022	False
Adult	4	4	25 to 34 years	Main	Feb	2024	Email	true	2022	False
Adult	0	0	60 to 64 years	Presidio	Mar	2024	Email	true	2018	False
Adult	24	53	35 to 44 years	Portola	Nov	2023	Email	true	2019	False
Adult	0	0	60 to 64 years	North Beach	Feb	2024	Email	true	2015	False
Adult	0	0	45 to 54 years	Mission Bay	Jan	2024	Email	true	2019	False
Adult	0	0	25 to 34 years	Merced	Feb	2023	Email	true	2016	False

UNIVERSITÄT
LEIPZIG

SEE YA‘LL NEXT WEEK!

Dr. Ing. Andreas Niekler

Computational Humanities

Paulinum, Augustusplatz 10, Raum P 616, 04109 Leipzig

T +49 341 97-32239

andreas.niekler@uni-leipzig.de

https://www.uni-leipzig.de/personenprofil/mitarbeiter/dr-andreas-niekler