

Statistik

CH.1 - Grundlagen

SS 2022 | | Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

- TODO: Prüfungsleistung, weitere Infos/Organisatorisches (Gruppenfindung etc.)

- 1 VL: Intro, Begriffe, Häufigkeit
- 2 VL: Maßzahlen
- 3 VL: Zusammenhang
- 4 VL: Kombinatorik
- 5 VL: Wahrscheinlichkeitsrechnung
- 6 VL: Anwendungen Wahrscheinlichkeitsrechnung
- 7 VL: Wahrscheinlichkeitsverteilungen
- 8 VL: Spezielle Verteilungen
- 9 VL: Schätzung
- 10 VL: Hypothesentests
- 11 VL: Regression
- 12 VL: Multiple Regression
- 13 VL:

Benjamin Buchwitz

- B.Sc. Business Informatics, Berlin School of Economics and Law.
- M.Sc. Management Science, University of Eichstätt-Ingolstadt.
- Ph.D. (Doctorate) at the Chair of Statistics and Quantitative Methods, University of Eichstätt-Ingolstadt.
- Work Experience in Consulting, Information Technology and Renewable Energy Sector, e.g. for IBM, Siemens and own Company.

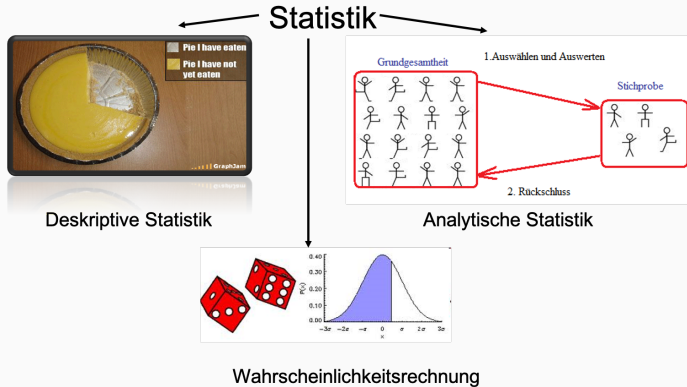
Prof. Dr. Benjamin Buchwitz

- Professur für Data Science und Data Engineering
- Raum 2.3.14, Lindenstr. 53, Meschede
- **Email:** buchwitz.benjamin@fh-swf.de

Was ist Statistik?



- 1 Deskriptive Statistik
- 2 Wahrscheinlichkeitsrechnung
- 3 Induktive/Analytische Statistik



- Häufigkeiten
- Graphische Darstellungen
- Tabellen
- Lagemaße
- Streuungsmaße
- Konzentrationsmaße
- Korrelationskoeffizient

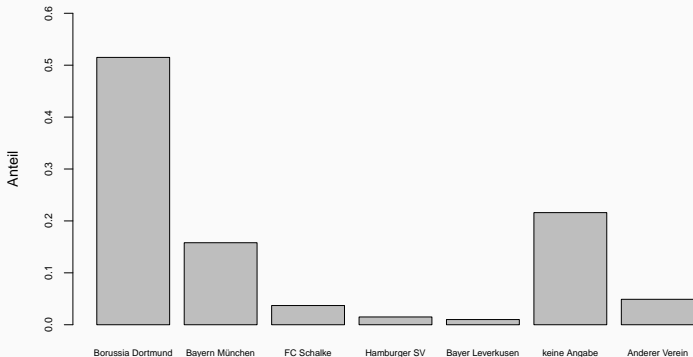
- Angenommen Sie besitzen ein Wertpapier (z.B. Bitcoin), das in den letzten Jahren die folgenden Wertveränderungen durchlaufen hat. Ermitteln Sie die durchschnittliche Wertsteigerung dieses Wertpapiers.

Jahr	Veränderung
2017	12 %
2018	7 %
2019	1 %
2020	4 %
2021	-10 %

Beispiel: Deskriptive Statistik

- Bei einer Befragung von 1195 über 14jährigen Deutschen durch das Institut **promit** im Februar 2011 tippten 51,5 % auf Borussia Dortmund, 15,8 % auf Bayern München, 3,7 % auf FC Schalke 04, 1,5 % auf den Hamburger SV und 1,0 % auf Bayer 04 Leverkusen als Deutschen Fußballmeister 2011. 21,6 % machten keine Angaben, der Rest glaubte an einen anderen Verein als Meister. Stellen Sie die Daten graphisch dar.

Wer wird Deutscher Fußballmeister 2011



- Kombinatorik
- Zufallsexperimente
- Wahrscheinlichkeiten
- Bedingte Wahrscheinlichkeiten
- Zufallsvariablen
- Verteilungsfunktionen
- Spezielle Verteilungen (z.B. Normalverteilung)

- Ein Unternehmen erhält wiederholt Lieferungen von 800 Flaschen zur Verpackung von flüssigem Waschmittel. Mit dem Lieferanten ist vereinbart, dass Lieferungen mit mehr als 2% fehlerhaften Flaschen zurückgewiesen werden dürfen. Aus Zeit- und Kostengründen kann das Unternehmen nicht alle gelieferten Flaschen prüfen. Um zu entscheiden, ob es die Lieferungen zurückweist, verfährt es daher nach folgender Regel:
 - ▶ Der Lieferung werden 50 Flaschen zufällig entnommen und geprüft. Die Lieferung wird zurückgewiesen, wenn mehr als eine Flasche nicht dem vereinbarten Qualitätsstandard entspricht. Wie groß ist die Wahrscheinlichkeit, dass eine gerade noch zulässige Lieferung, d.h. genau 2% fehlerhafte Flaschen, zurückgewiesen wird.

- Stichproben
- Konfidenzintervalle
- Testen von Hypothesen
- Lineare Regression
- Multiple Regression

- Für ein Konzert in Meschede wurden die Karten im Radio verlost. Sie befragen alle 400 Besucher des Konzertes nach ihrem Schokoladenkonsum zu Weihnachten (Dezember 2021). Dabei ermitteln Sie einen durchschnittlichen Schokoladenkonsum von 455g bei einer Standardabweichung von 130g. Mit diesen Daten möchten Sie auf den Schokoladenkonsum der Bevölkerung im Hochsauerlandkreis schließen. Geben Sie an, in welchem Bereich der durchschnittlich Schokoladenkonsum eines HSK-Bürgers mit 90%iger Sicherheit lag.

R Installieren

<https://cloud.r-project.org/>

RStudio installieren

<https://www.rstudio.com/products/rstudio/download/#download>

DataCamp Zugang

Über die Einschreibung bei Moodle erhalten Sie den Zugang zu DataCamp für die begleitenden Online Kurse.

- Ein Programm, das Zahlen als Input nimmt, und daraus Tabellen (und Grafiken) erzeugt?
- Ein(e) (Sammlung von) Programm(en) für Exploration, Inferenz und Modellierung?
- Ein Werkzeug zur Verwaltung, Manipulation und Analyse von Daten?
- Ein Kommunikationsmedium mit CPU (Grafikkarte, Drucker, . . .) des Rechners?

- R ist eine kostenlose Software-Umgebung für statistische Datenanalyse und Graphiken. Es beruht auf einer Implementation der Sprache S. Anfänglich wurde R von Ross Ihaka und Robert Gentleman (Univ. Auckland) entwickelt und wird seit Mitte der 90er Jahre von einem Entwickler-Kollektiv (R-Core) betreut.
- Sehen Sie Statistiksoftware als Hilfsmittel, um dem Computer mitzuteilen, wie Sie Ihre Daten analysieren wollen.
- Für einfache Analysen reichen einfache Programme, aber gerade in Excel haben Sie irgendwann das letzte Untermenü oder die letzte Formel erreicht und können nicht mehr weiter arbeiten.
- R-Befehle sind Ihre Anweisungen an den Computer.

- Ziel 1
- Ziel 2
- Ziel 3

- **Statistische Einheit:** Träger der Information (Merkmalsträger, Untersuchungsobjekt)
- **Grundgesamtheit:** statistische Masse, Population
- **Untersuchungsmerkmal X_i :** Eigenschaft der statistischen Einheit
- **Merkmalsausprägung:** Erscheinungsformen eines Merkmals
- **Beobachtungswert x_i :** Die für das i -te Untersuchungsobjekt beobachtete Ausprägung des Untersuchungsmerkmals

Beispiel: Grundlegende Begriffe

Im Rahmen einer Vollerhebung sollen verschiedene Eigenschaften von allen Mescheder Studierenden dieser Vorlesung in diesem Raum untersucht werden.

- **Grundgesamtheit:** Alle Mescheder Studierende (räumliche Identifikation) dieser Vorlesung (sachliche Identifikation), die heute (zeitliche Identifikation) anwesend sind
- **Merkmalsträger:** Jede/r Studierende
- **Merkmale:** Alter, Geschlecht, Wohnort, Note in Mathe

Merkmalsausprägungen:

	Alter	Geschlecht	Wohnort	Note
Student:in a	19	m	Meschede	2,3
Student:in b	20	f	Meschede	1,3
Student:in c	22	f	Brilon	4,0
Student:in d	25	m	Winterberg	3,0

- **Qualitative Merkmale:** variieren artmäßig
 - ▶ z.B. Geschlecht, Rechtsform von Unternehmen, Haarfarbe etc.
- **Quantitative Merkmale:** variieren der Größe nach
 - ▶ z.B. Alter, Einkommen, Kinderzahl etc.

- **Nominalskala:** für qualitative Merkmale, für die keine sinnvolle Reihenfolge der Ausprägungen gegeben ist.
 - ▶ z.B. Studiengang, Religionszugehörigkeit, Geschlecht
- **Ordinalskala:** für Merkmale mit einer natürlichen Reihenfolge. Die Abstände zwischen den Ausprägungen sind nicht quantifizierbar.
 - ▶ z.B. Schulnoten, Bildungsabschlüsse
- **Metrische Skala:** für Merkmale, bei denen sowohl die Reihenfolge als auch die Abstände zwischen den Ausprägungen sinnvoll definiert sind.
 - ▶ z.B. Umsatz, Alter, Temperatur in °C

- **Diskrete Merkmale:** können nur bestimmte Werte (z.B. nur ganzzahlige) annehmen
 - ▶ z.B. Anzahl der Studierenden an einer Hochschule, Anzahl der Einwohner eines Landes
- **Stetige Merkmale:** können in einem bestimmten Intervall jeden beliebigen Wert annehmen
 - ▶ z.B. Längen, Breiten, Gewichte

- Kann unsere Hochschule eine Untersuchungseinheit sein?
- Nach welchen Kriterien werden Grundgesamtheiten abgegrenzt?
- Was ist der Unterschied zwischen einem Merkmalsträger und einer Merkmalsausprägung?

- Kann unsere Hochschule eine Untersuchungseinheit sein?
 - ▶ Ja, wenn die Grundgesamtheit z.B. aus allen Hochschulen eines Landes besteht.
- Nach welchen Kriterien werden Grundgesamtheiten abgegrenzt?
 - ▶ Räumlich, sachlich, zeitlich
- Was ist der Unterschied zwischen einem Merkmalsträger und einer Merkmalsausprägung?
 - ▶ Merkmalsträger ist die Person oder das Objekt, das untersucht wird, Merkmalsausprägung ist eine der Erscheinungsformen einer Eigenschaft des Merkmalsträgers.