▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Statistik

Christian Reinboth
M.Sc., Dipl.-Wi.Inf.(FH)

Sommersemester 2022
Berufsbegleitender Bachelorstudiengang Betriebswirtschaftslehre

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

H.G. Wells

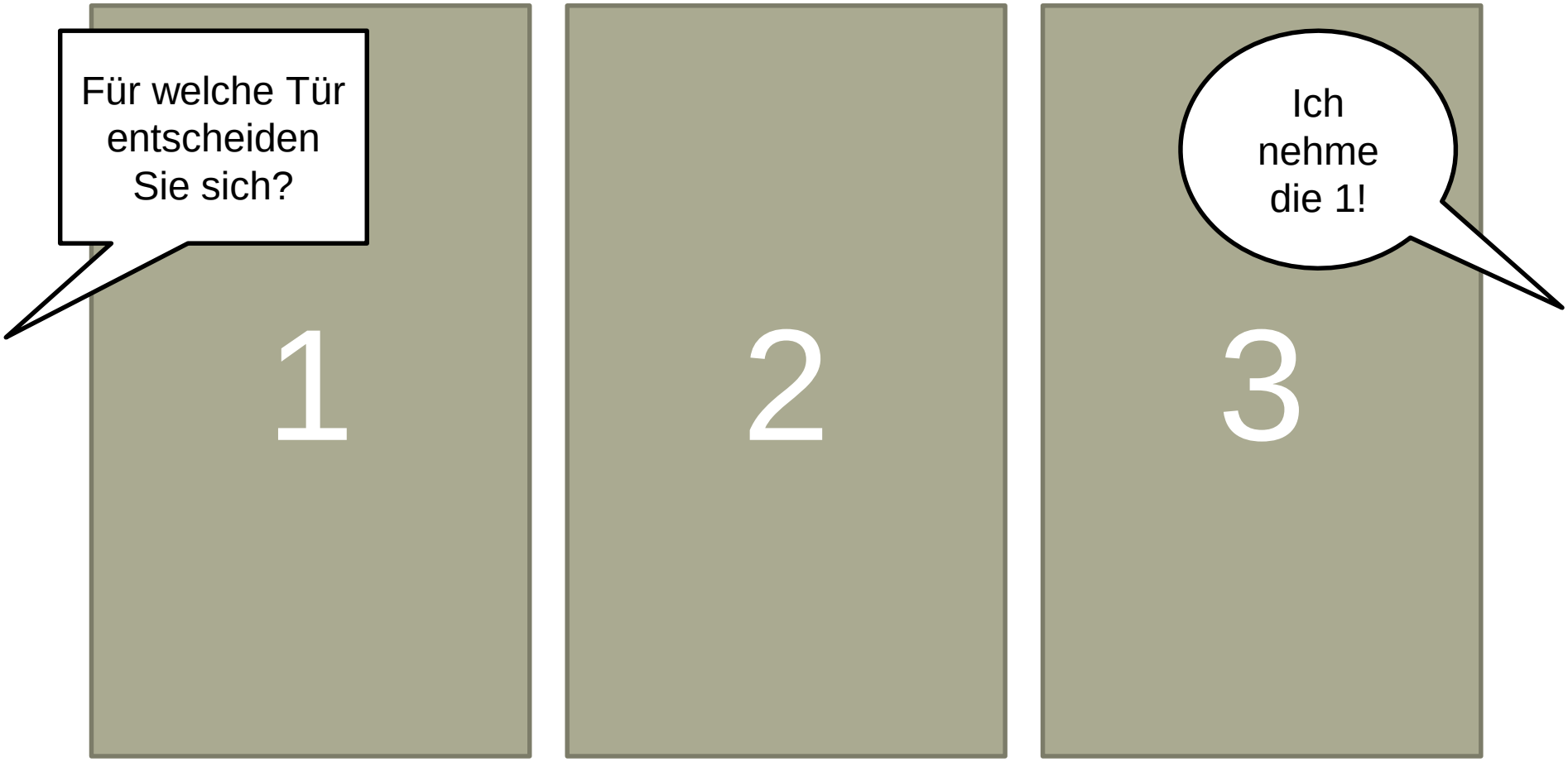
Eisbrecheraufgabe

Auftakt mit Spaß: Das Ziegenproblem

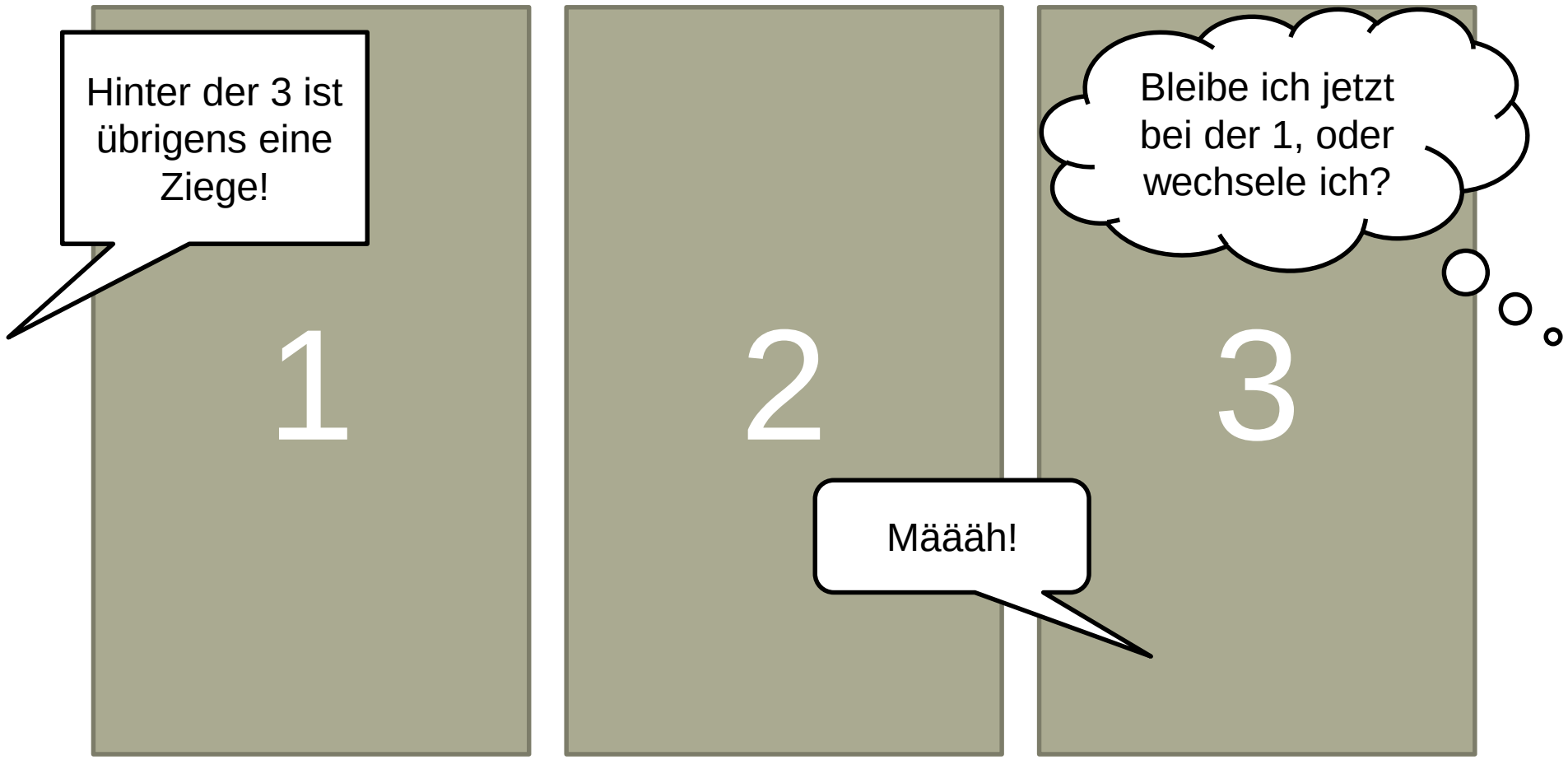
▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Für welche Tür sollte man sich entscheiden?



Für welche Tür sollte man sich entscheiden?



Einige interessante Fragestellungen

- Das Ziegenproblem lässt sich nahezu beliebig weiterdiskutieren...
  - Würde ein neuer Kandidat auf der Bühne erscheinen, nachdem sich der erste Kandidat bereits endgültig für eine Tür entschieden hat – könnte dieser sich mit einer 50/50-Siegwahrscheinlichkeit zwischen den verbliebenen Türen entscheiden?
  - Wenn von Anfang an zwei Kandidaten/innen mitspielen, von denen eine/r Tür 1 und eine/r Tür 2 wählt – können sich dann wirklich die Gewinnchancen beider erhöhen, wenn sie auf die jeweils andere Tür wechseln, nachdem Tür 3 geöffnet wurde?
  - ...

Lesetipp

Noch viel mehr Varianten in: „Das Ziegenproblem – Denken in Wahrscheinlichkeiten“ von Gero von Randow (rororo-Verlag, 2004)

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Statistik I

Christian Reinboth
M.Sc., Dipl.-Wi.Inf.(FH)

Sommersemester 2022
Berufsbegleitender Bachelorstudiengang Betriebswirtschaftslehre

Statistik
Wesentliche Kursinhalte (1)

- Kurzvorstellung
- Organisatorisches
- Bücher und Software

- Grundlagen

Statistik I

  - Einordnung
  - Grundbegriffe
  - Skalenniveaus
  - Variablentypen

- Qualitative und quantitative Forschung
  - Unterschiede
  - Vor- und Nachteile
  - Methoden der Datenerhebung
  - Methoden der Datenauswertung

- Erhebungsplanung und -durchführung
  - Erhebungsarten
    - Zufällige Auswahl
    - Klumpenstichprobe
    - Willkürliche Auswahl
    - Auswahl typischer Fälle
    - Konzentrationsverfahren
    - Mindeststichprobengröße
  - Gütekriterien
    - Bedeutung
    - Validität
    - Reliabilität
    - Objektivität
    - Repräsentativität
    - Sonstige Gütekriterien

- Gutes Fragebogendesign
  - Zieldefinition
  - Anschreiben
  - Incentivierung
  - Frageformulierung
  - Gängige Fragetypen

- Deskriptive Statistik
  - Häufigkeiten
    - Häufigkeiten
    - Häufigkeitstabellen
    - Bildung von Klassen
    - Verteilungsfunktion
    - Summenfunktion

Statistik
Wesentliche Kursinhalte (2)

- Statistische Lagemaße
  - Statistische Lagemaße
  - Arithmetisches Mittel
  - Median
  - Quartile
  - Modus
- Dispersionsparameter
  - Dispersionsparameter
  - Spannweite
  - Interquartilsabstand
  - Fünf-Werte-Zusammenfassung
  - Varianz
  - Standardabweichung
  - Variationskoeffizient
- Verteilungsmaße
  - Verteilungsmaße
  - Momentenkoeffizient
  - Quartilskoeffizient
  - Kurtosis / Exzeß
- Korrelationskoeffizienten
  - Korrelationskoeffizienten
  - Korrelation und Kausalität
  - Bravais-Pearson-Koeffizient
  - Rangkorrelationskoeffizienten
  - Spearman-Koeffizient
  - Kendall-Koeffizient
- Explorative Statistik
  - Grafische Darstellungen
    - Box-Whisker-Plot
    - Stem-and-Leaf-Plot
    - Objektivität von Grafiken
  - Ausreißer und fehlende Werte

Statistik II

Statistik
Wesentliche Kursinhalte (2)

Statistik II

- Induktive Statistik
  - Lineare Regression
    - Zielstellung
    - Voraussetzungen
    - Interdependenzproblem
    - Methode der kl. Quadrate
    - Ergebnisinterpretation
    - Bestimmtheitsmaß
  - Statistische Testverfahren
    - Statistische Tests
    - Chi-Quadrat-Test
    - Alpha-Fehlerinflation

- Mengenlehre
  - Mengenlehre
  - Logische Operatoren
  - Kommutativgesetz
  - Assoziativgesetz
  - Distributivgesetz
  - De Morgansche Regel
  - Venn-Diagramme
- Wahrscheinlichkeitslehre
  - Laplace-Wahrscheinlichkeit
  - Axiome von Kolmogoroff
  - Additionssatz
  - Multiplikationssatz
  - Pfaddiagramme
  - Kombinatorik
  - Satz von Bayes

- Konfidenzintervalle
- Statistische Software
  - Kostenlose Software
  - Einführung in R
- Klausurvorbereitung
  - Übungsaufgaben
  - Probeklausur
  - Fragestunde

Kurzvorstellung
Arbeit, Forschung und Lehre

Arbeit bei der HarzOptics GmbH

- An-Institut der HS Harz (seit 2007)
- Gegründet 2006, 4 Mitarbeiter/innen
- Entwicklung optischer Messverfahren zur Qualitätssicherung in der Luftfahrt
- Projektierung des Breitbandausbaus im Auftrag von Kreisen und Kommunen
- Fernlehrgang „Technische Optik“

Arbeit an der Hochschule Harz

- Seit 2010 Forschung im Bereich AAL und Telepflege, seit 2013 Fundraising

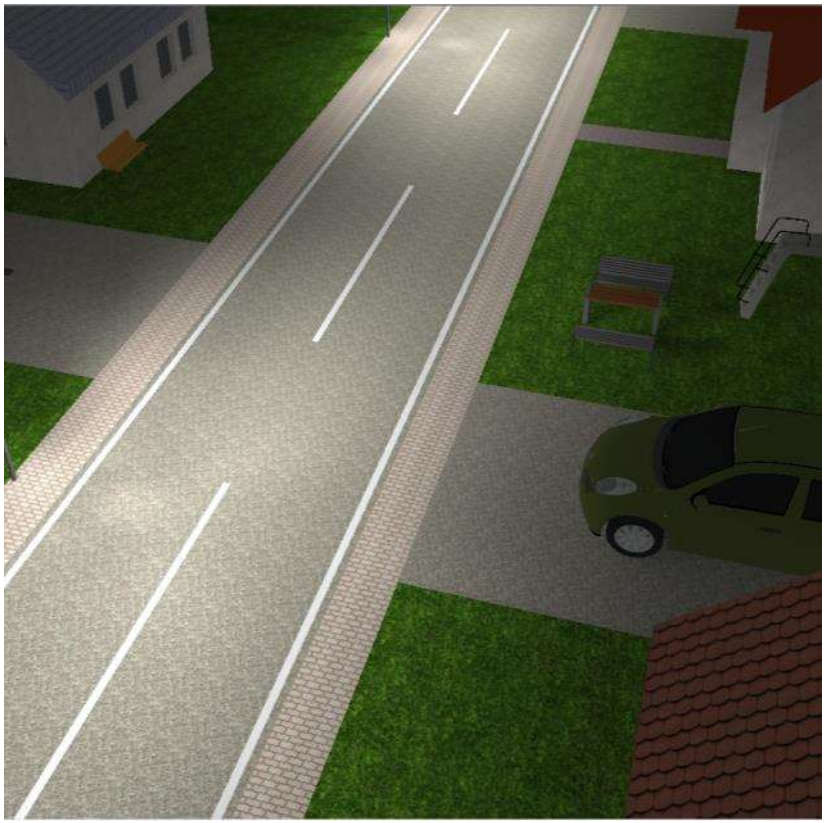
- IHK-Forschungspreis 2006
- 3. Platz Hugo-Junkers-Preis 2008
- 3. Platz Hugo-Junkers-Preis 2012
- NoAE Innovation Award 2011/2012

Bisherige Lehrerfahrung

- Lehrbeauftragter an der HS Harz von 2006 bis 2010 und seit 2015 (Statistik, Marktforschung, SPSS, HTML, BIS und strategisches Informationsmanagement)
- Dozent für die Harzer Hochschulgruppe (2007 - 2008) und die Sternwarte Sankt Andreasberg / VHS Goslar (2011 - 2013)

Mein zentrales Forschungsthema
Umweltfreundliche Beleuchtungsplanung

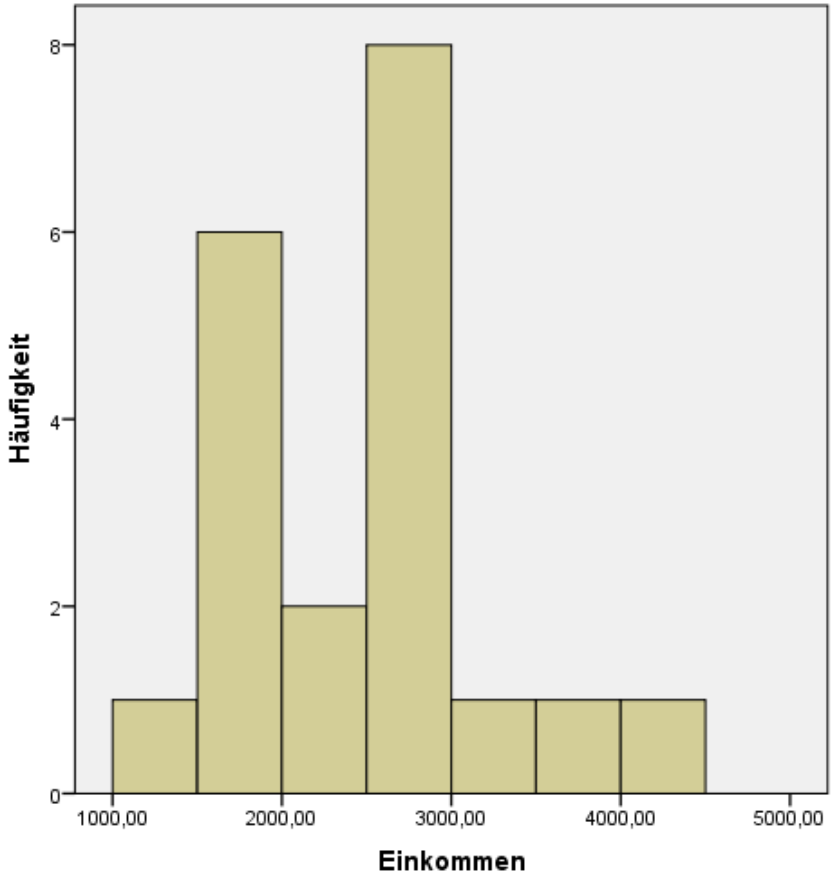
Innenraumsimulation mit DIALux (Sternwarte Sankt Andreasberg)



Außenraumsimulation mit DIALux (Ortsteil Freiheit in Osterode)

Organisatorisches

Wie wird dieser Kurs ablaufen?



- Beherrschung der Grundbegriffe von Statistik und Wahrscheinlichkeitslehre
- Sichere deskriptive Analyse von Daten
- Grundkenntnisse über statistische Testverfahren und die univariate lineare Regressionsanalyse

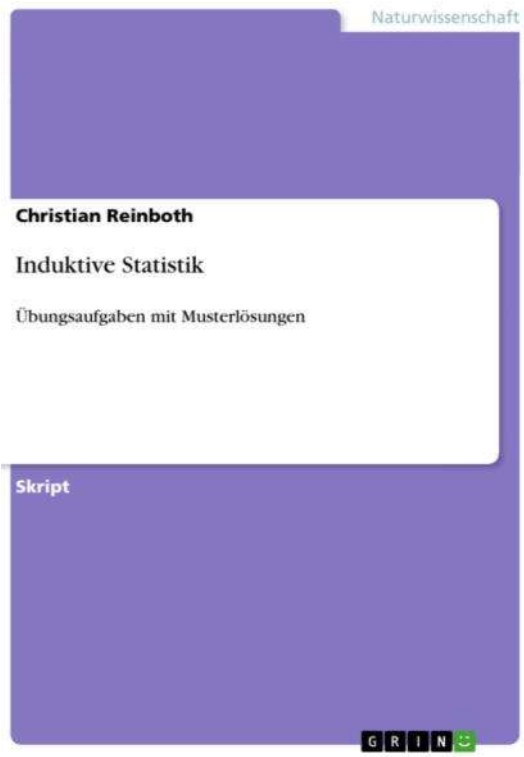
- Vorlesung mit eingestreuten Übungen
- Übungsaufgaben zur eigenständigen Vorbereitung der Abschlussprüfung

- Klausuren über 60 und 120 Minuten

Empfohlene Literatur
(Weitere Hinweise in der Modulbeschreibung)

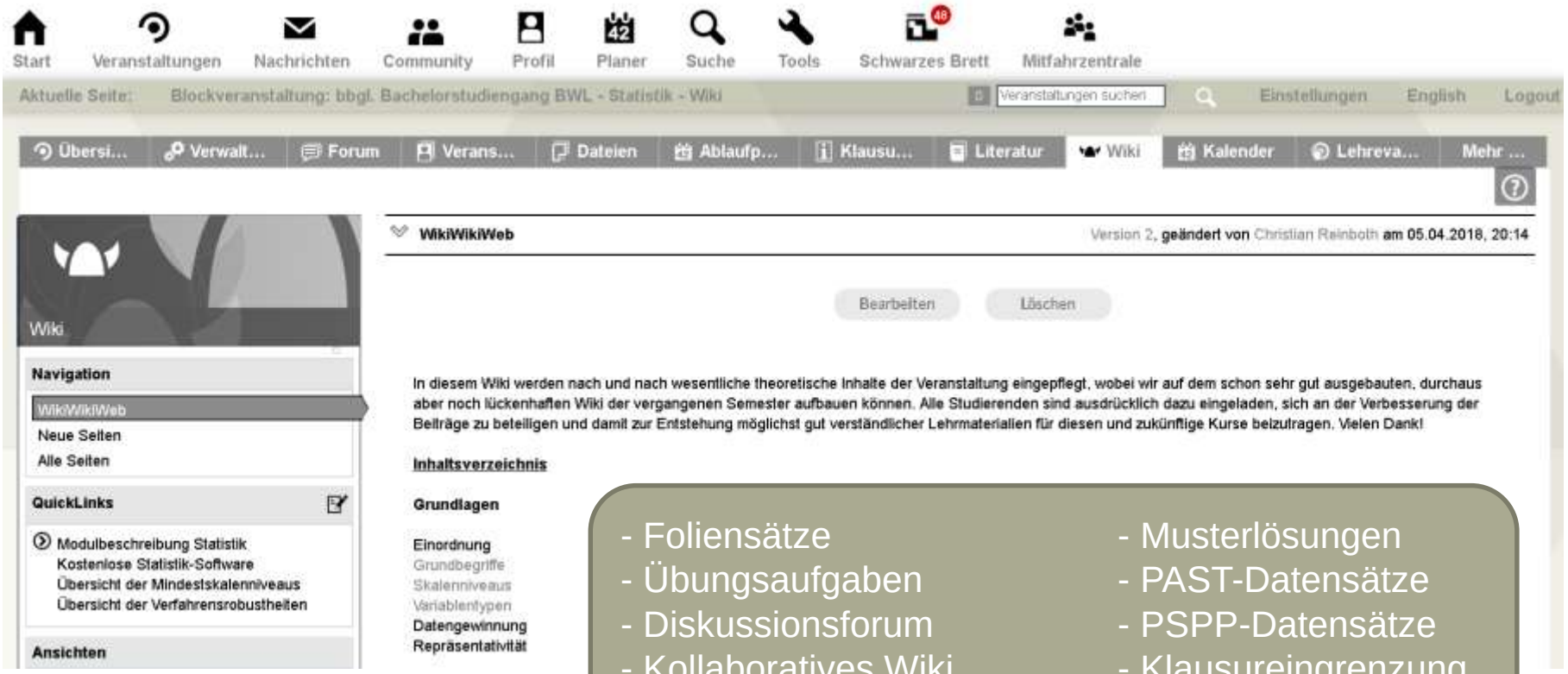
I. Rößler & A. Ungerer: Statistik für Wirtschaftswissenschaftler. Eine anwendungsorientierte Darstellung, Springer-Verlag, 4. Auflage, Luxemburg, 2014, ISBN: 978-3-642-41259-2

G. Bourier: Beschreibende Statistik. Praxisorientierte Einführung mit Aufgaben und Lösungen, Gabler-Verlag, 9. Auflage, Wiesbaden, 2011, ISBN: 978-3-8349-2763-7



C. Reinboth: Induktive Statistik – Übungsaufgaben mit Musterlösungen, eBook, GRIN-Verlag für wissenschaftliche Texte, 75 Seiten, München, 2013, ISBN: 978-3-656-53867-7

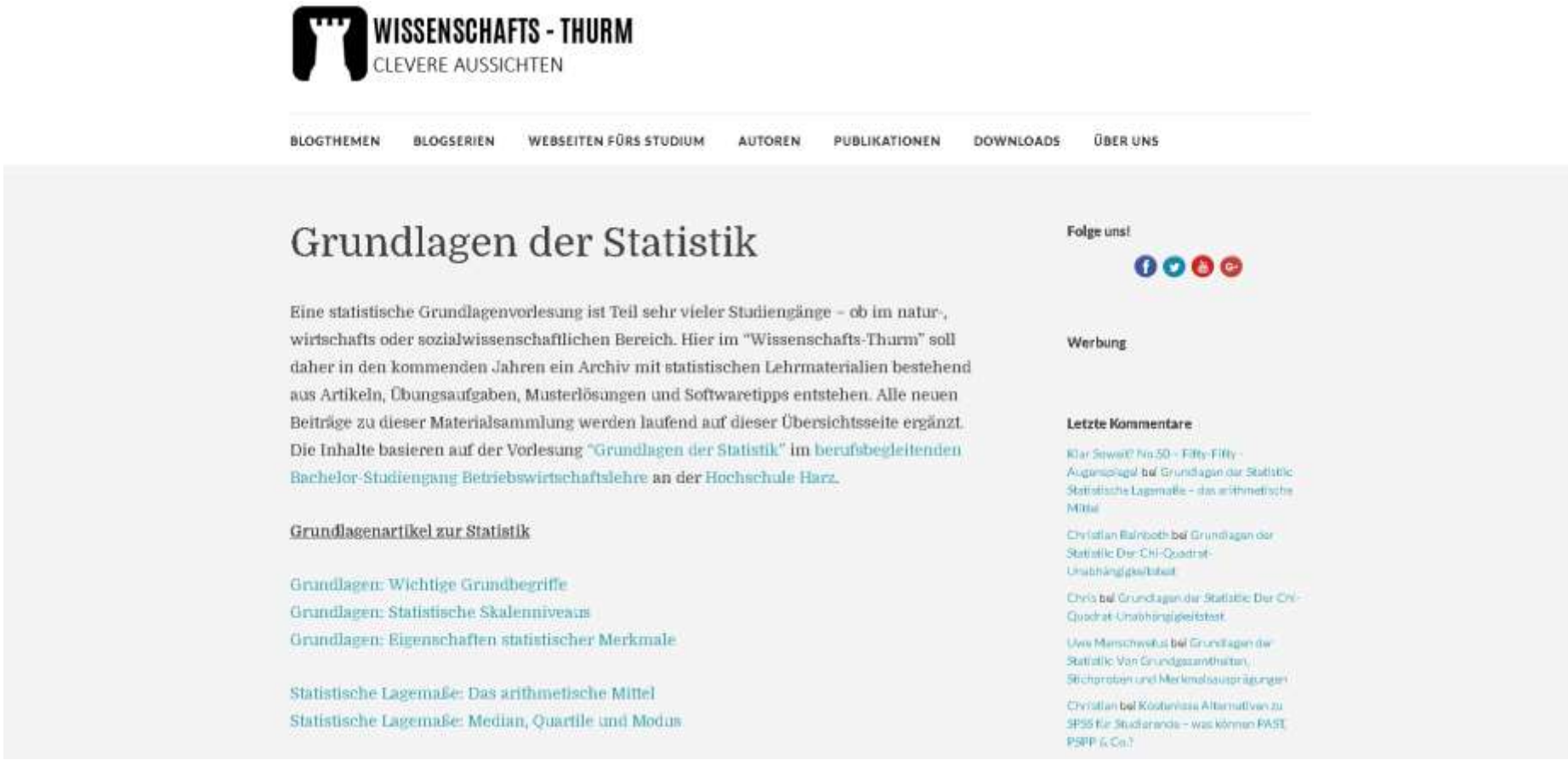
Nutzung von Stud.IP
https://studip.hs-harz.de



- Foliensätze
- Übungsaufgaben
- Diskussionsforum
- Kollaboratives Wiki
- Formelsammlungen
- Musterlösungen
- PAST-Datensätze
- PSPP-Datensätze
- Klausureingrenzung
- ...und vieles mehr...

Begleitender Vorlesungsblog im „Thurm“

https://wissenschafts-thurm.de/grundlagen-der-statistik/



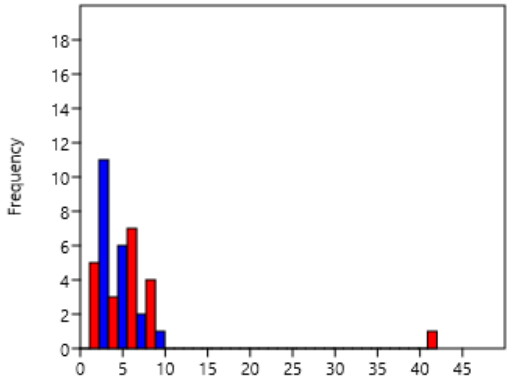
Was ist SPSS?
Statistical Package for Social Sciences

- **SPSS** ist eines der **marktführenden Softwareprodukte** für Datenanalysen in der Sozial- und Gesundheitswissenschaft sowie in der Markt- und Meinungsforschung

- Es wurde 1983 von SPSS Inc. Entwickelt (Ausgründung der Stanford University)
- Der Name wechselte mehrfach von „Statistical Package for Social Sciences“ über „Superior Performing Software System“ und „Predictive Analysis Software“ (PASW) bis zu IBM SPSS STATISTICS seit der Übernahme von SPSS Inc. durch IBM in 2009

www.ibm.com/software/de/analytics/spss/

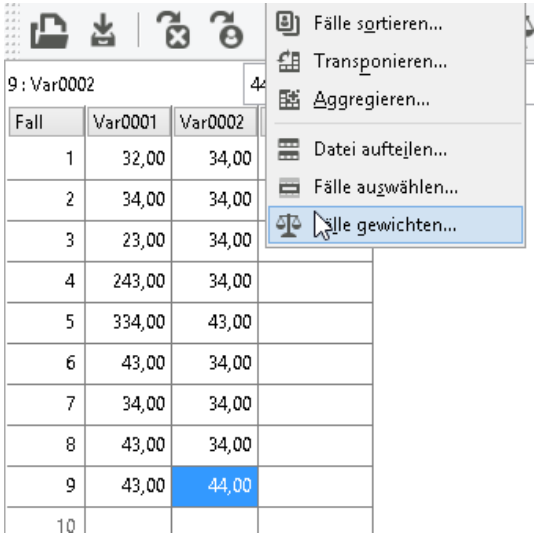
Empfehlenswerte freie Statistik-Software
(Kategorie: Allgemeine Datenanalyse)



PAST (Windows, Mac)

- Paleontological Statistics Software Package for Education and Data Analysis (Universities of Copenhagen and Oslo)

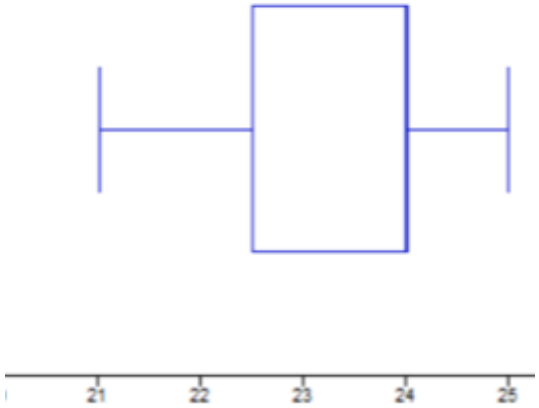
http://folk.uio.no/ohammer/past/



PSPP (Windows, Mac, Linux)

- Open Source-“Nachbau“ von SPSS
- Identische Funktionen und Bedienung, „Look & Feel“ ist sehr gut vergleichbar

https://www.gnu.org/software/pspp/

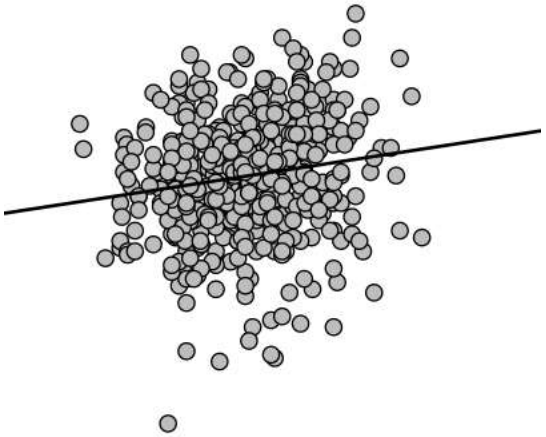


SSP (Windows, Mac)

- Smith's Statistical Package
- “Ein-Mann-Entwicklung” von Prof. Gary Smith vom Pomona College

http://economics-files.pomona.edu/GarySmith/StatSite/ssp.html

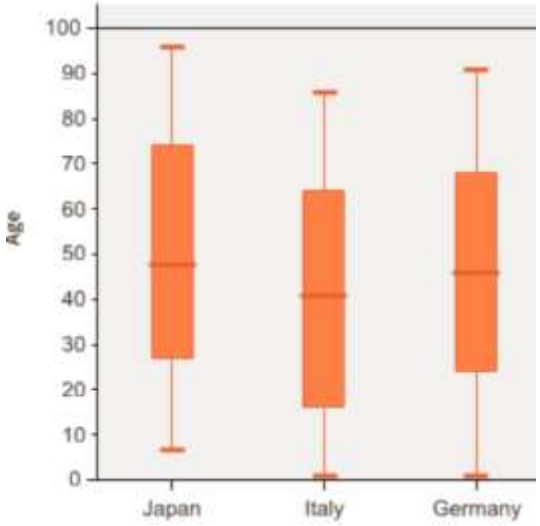
Empfehlenswerte freie Statistik-Software
(Kategorie: Spezielle Anforderungen)



JASP (Windows, Mac, Linux)

- Just Another Stats Program
- Bietet liquiden Output, der sich mit jedem Klick ändert (ideal für Lerner)

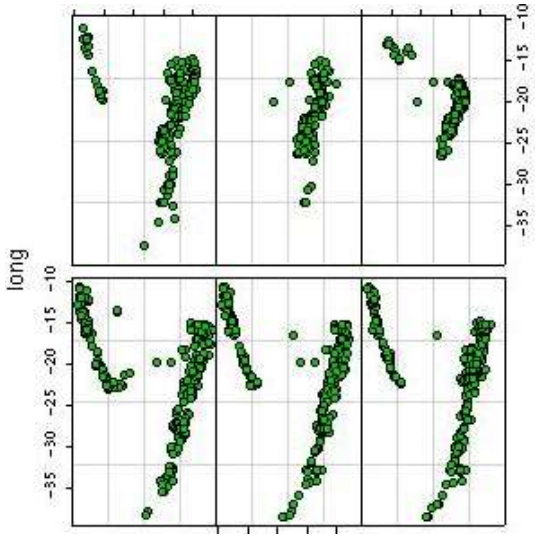
https://jasp-stats.org



SOFA (Windows, Mac, Linux)

- Statistics Open For All
- Bietet vielfältige Möglichkeiten der grafischen Aufbereitung von Daten

http://www.sofastatistics.com



MacANOVA (Windows, Mac, Linux)

- Entwickelt an der Uni Minnesota
- Der Schwerpunkt der Software liegt auf der Varianzanalyse (ANOVA)

http://www.stat.umn.edu/macanova/

Softwarealternativen zu SPSS

Für Übungen am heimischen Rechner…

Software	URL	System(e)
PSPP	https://www.gnu.org/software/pspp/	Alle
PAST	http://folk.uio.no/ohammer/past/	Win, Mac
SSP	http://economics-files.pomona.edu/GarySmith/StatSite/ssp.html	Win, Mac
SOFA	http://www.sofastatistics.com	Alle
SciLab	http://www.scilab.org	Alle
FreeMat	http://freemat.sourceforge.net	Alle
Gnumeric	http://www.gnumeric.org	Linux

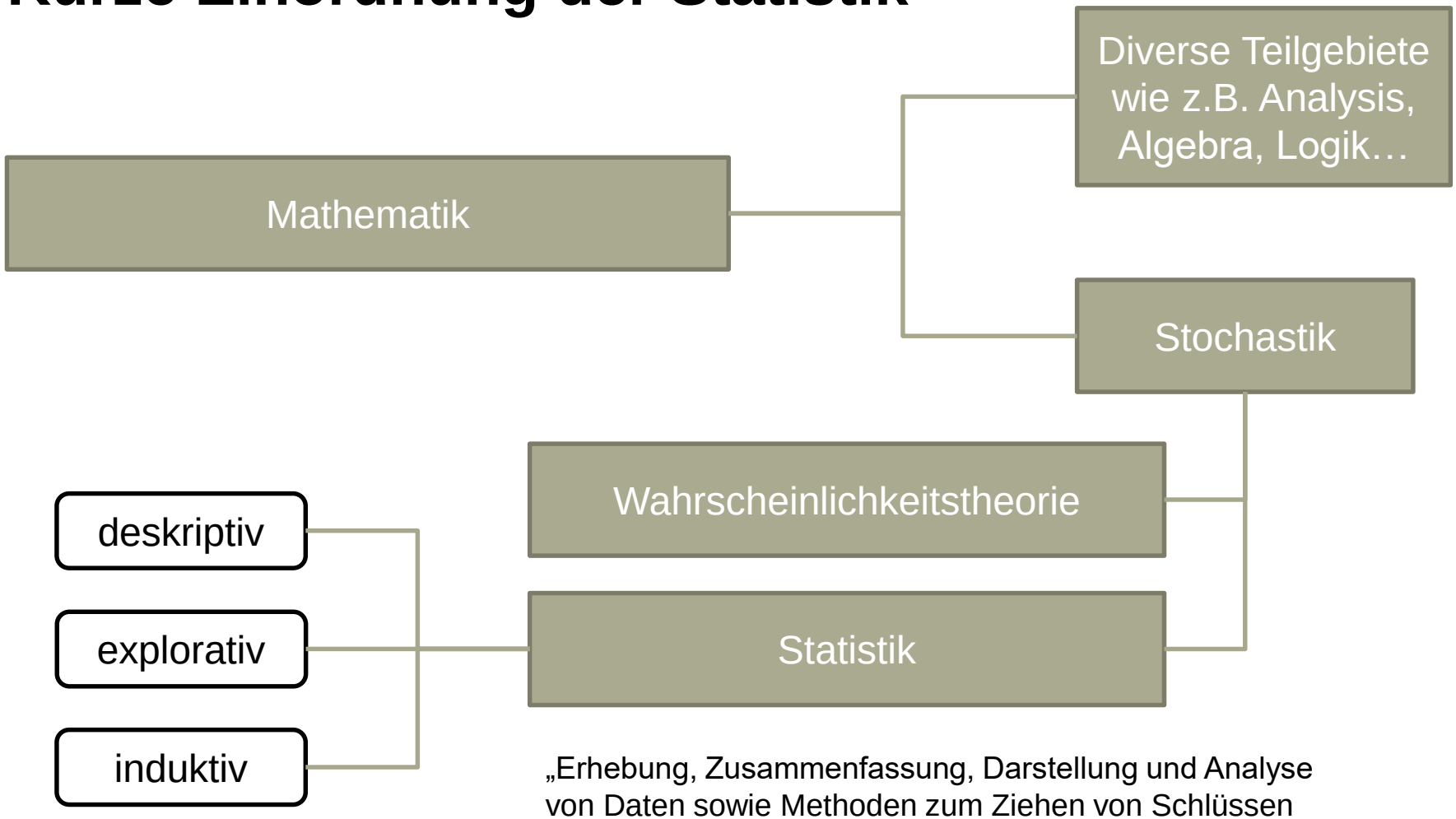
Teil I

Grundlagen

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

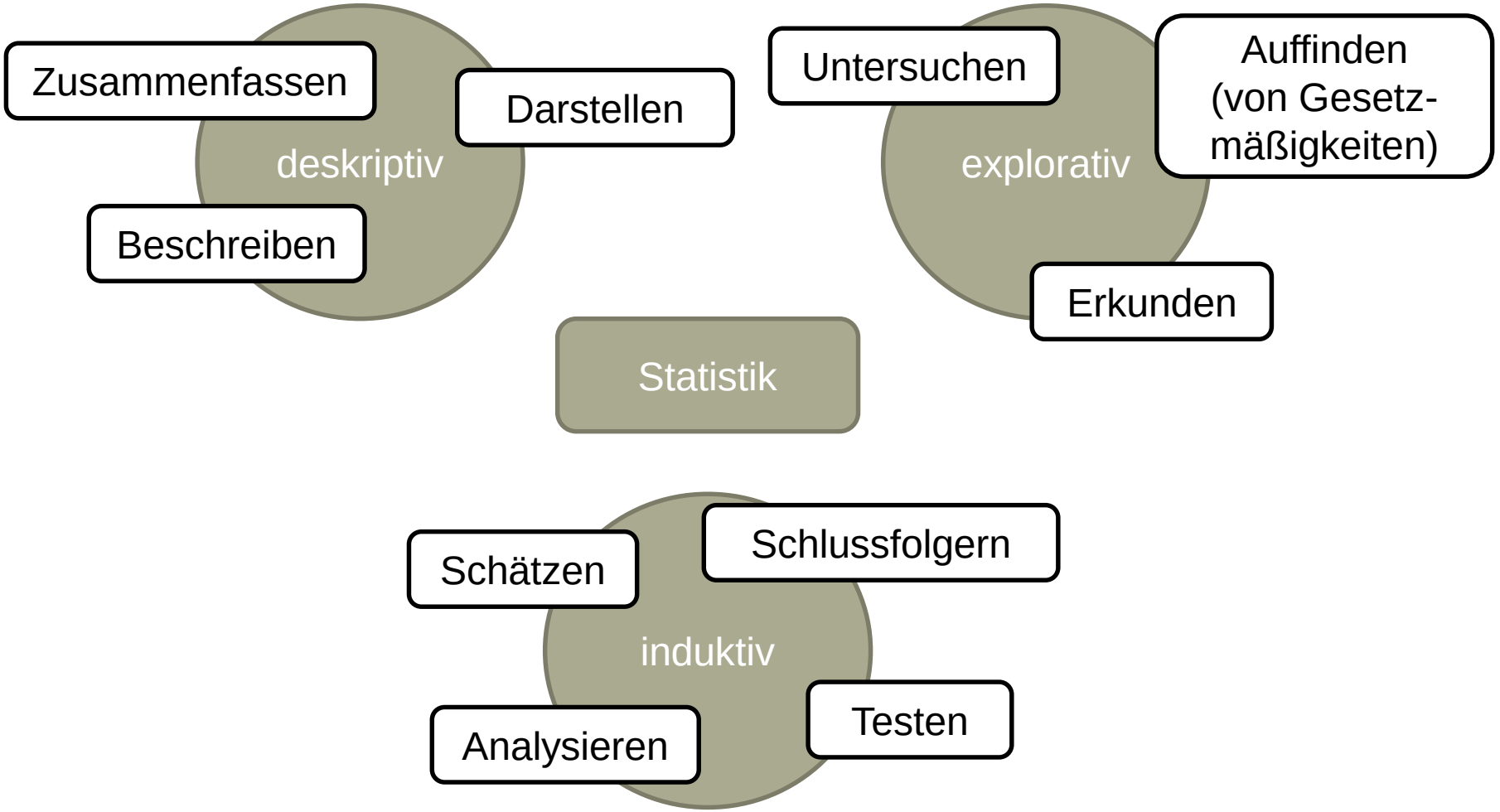
Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Kurze Einordnung der Statistik

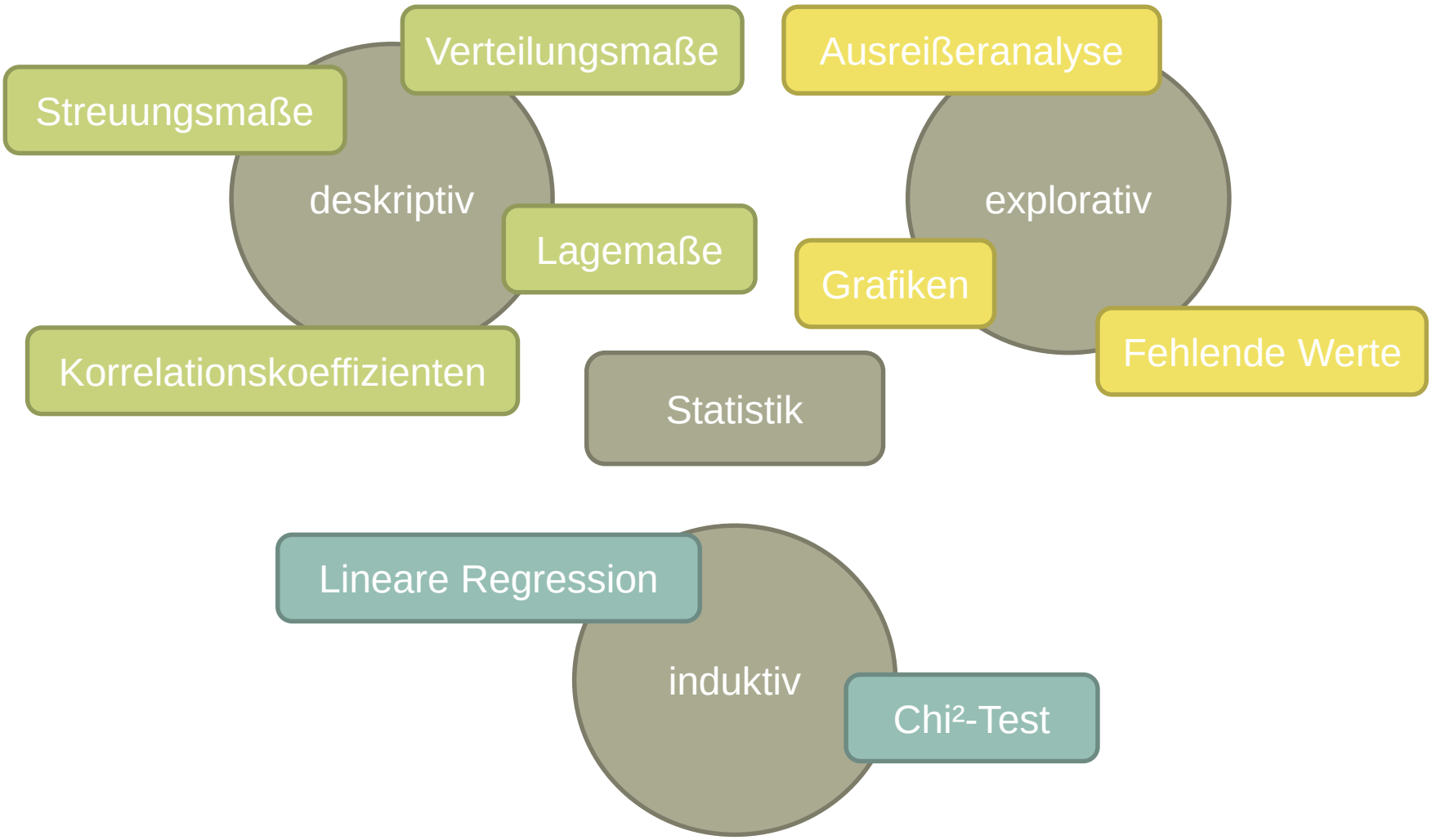


„Erhebung, Zusammenfassung, Darstellung und Analyse von Daten sowie Methoden zum Ziehen von Schlüssen auf Grundlage von Daten“ - Tobias Hell, LFU Innsbruck

Kurze Einordnung der Statistik

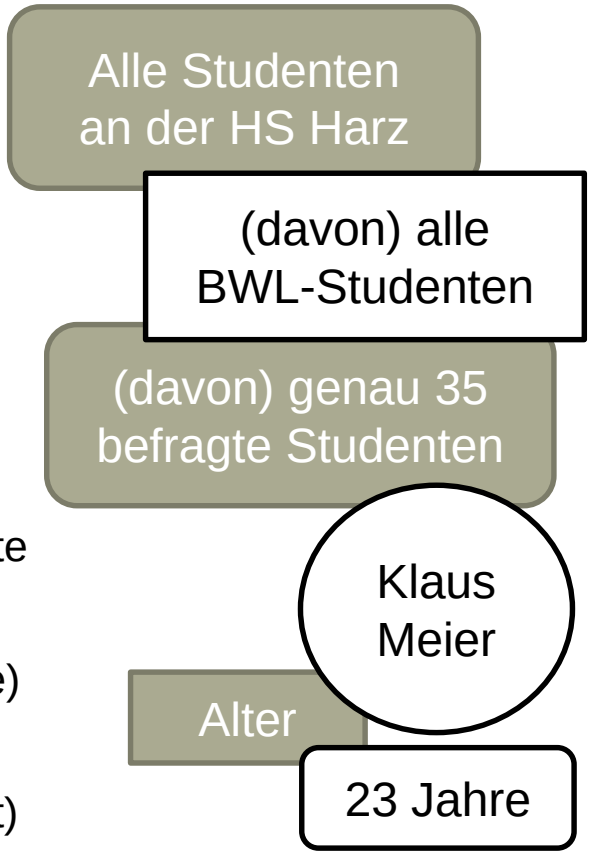


Kurze Einordnung der Statistik



Grundbegriffe der Statistik
Wer erinnert sich noch?

- **Grundgesamtheit / Population**
  = Menge aller relevanten statistischen Einheiten
- **Teilgesamtheit / Teilpopulation**
  = Betrachtete Teilmenge einer Grundgesamtheit
- **Stichprobe**
  = Real untersuchte Teilmenge einer Grundgesamtheit
- **Statistische Einheiten**
  = Einzelne im Rahmen einer Erhebung untersuchte Objekte
- **Merkmal**
  = Interessierende Größe der statistischen Einheit (Variable)
- **Ausprägung**
  = konkreter Merkmalswert einer statistischen Einheit (Wert)



Übung: Grundbegriffe der Statistik

- Eine Wohnungsbaugesellschaft will aus der Menge all ihrer Mieterinnen und Mieter diejenigen mit einem Alter oberhalb von 65 Jahren zum Thema „seniorenfreundliches Wohnen“ befragen. Hierzu werden per Zufall 150 ältere Mieterinnen und Mieter aus der Kundenkartei herausgesucht und angeschrieben. Gefragt wird unter anderem nach der persönlichen Einschätzung von barrierefreien Korridoren, wobei lediglich einer der Befragten angab, dass diese für ihn „überhaupt nicht von Bedeutung“ sei.

  - Grundgesamtheit:
  - Teilgesamtheit:
  - Stichprobe:
  - Statistische Einheit(en):
  - Merkmal:
  - Ausprägung:

Übung: Grundbegriffe der Statistik

- Eine Wohnungsbaugesellschaft will aus der Menge all ihrer Mieterinnen und Mieter diejenigen mit einem Alter oberhalb von 65 Jahren zum Thema „seniorenfreundliches Wohnen“ befragen. Hierzu werden per Zufall 150 ältere Mieterinnen und Mieter aus der Kundenkartei herausgesucht und angeschrieben. Gefragt wird unter anderem nach der persönlichen Einschätzung von barrierefreien Korridoren, wobei lediglich einer der Befragten angab, dass diese für ihn „überhaupt nicht von Bedeutung“ sei.

  - Grundgesamtheit: Alle Mieterinnen und Mieter der Wohnungsbaugesellschaft
  - Teilgesamtheit: Nur ältere Mieterinnen und Mieter oberhalb von 65 Jahren
  - Stichprobe: 150 per Zufall selektierte ältere Mieterinnen und Mieter
  - Statistische Einheit(en): Einzelne befragte Mieterinnen und Mieter
  - Merkmal: Persönliche Einschätzung von barrierefreien Korridoren
  - Ausprägung: Ist für Befragten „überhaupt nicht von Bedeutung“

Statistische Skalenniveaus

Welches Informationsniveau haben Daten?

- **Nominalskala**
  - Daten sind nur Bezeichnungen ohne Rangordnung
  - Feststellbar ist lediglich Gleichheit oder Ungleichheit

  Geschlecht, Telefonnummern, Kontonummern, Geschmack…

- **Ordinalskala**
  - Daten weisen eine natürliche (!) Rangordnung auf
  - Abstände zwischen Daten sind nicht interpretierbar

  Schulnoten, Präferenzrangfolgen, Dienstränge, Zufriedenheiten…

- **Intervallskala**
  - Daten können in eine Rangordnung gebracht werden
  - Abstände zwischen Daten sind ebenfalls interpretierbar

  Temperaturen in Celsius oder Fahrenheit, Jahreszahlen…

- **Verhältnisskala**
  - Genau wie Intervallskala – nur mit natürlichem Nullpunkt

  Temperaturen in Kelvin, Zeit, Streckenlängen, Wassertiefen…

Diskrete und stetige Variablen
Wie viele Ausprägungen gibt es?

- **Diskrete Variablen („zählen“)**
  - Endlich oder abzählbar unendlich viele Ausprägungen
  - Variablen mit nur zwei Ausprägungen sind dichotom

Augen beim Würfeln, Kinderzahl, Haarfarbe, Geschlecht, Berufe...

Was bedeutet „abzählbar unendlich“?

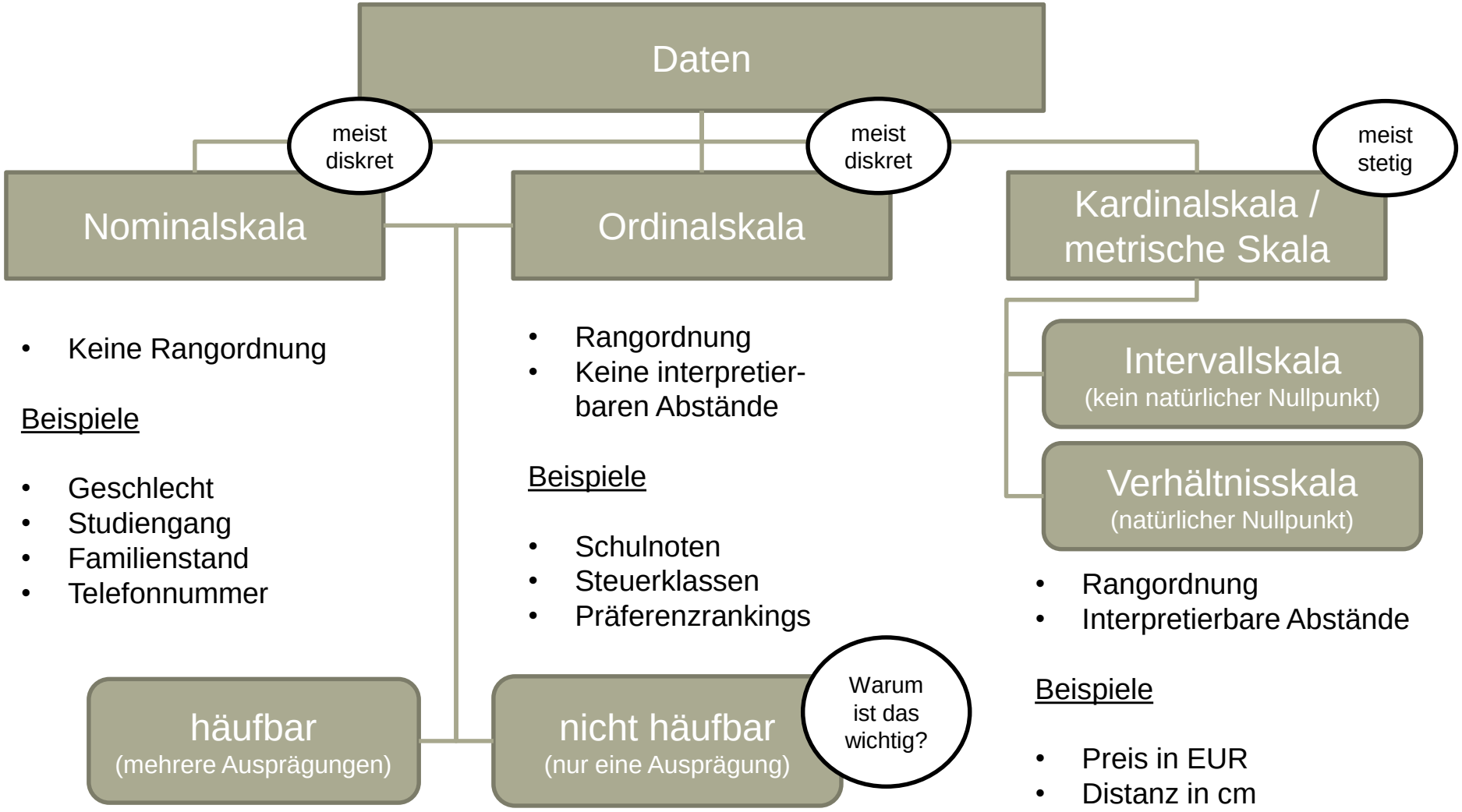
- **Stetige Variablen („messen“)**
  - Alle Werte eines Intervalls sind mögliche Ausprägungen
  - Die Zahl möglicher Ausprägungen ist somit unendlich

Wassertiefe, Luftfeuchtigkeit, Wassertemperatur, Zeitintervall...

- **Quasi-stetige Variablen („ungenau messen“)**
  - Diskrete Variablen mit sehr vielen Ausprägungen werden in der Praxis oft wie stetige Variablen behandelt (und damit „quasi-verstetigt“)
  - Quasi-stetig sind auch stetige Variablen, die nur diskret genau gemessen werden können

Nettoeinkommen, Produktpreise...

Skalenniveaus und Variablentypen



Übung: Skalenniveaus und Variablentypen

- Wassertiefe eines Schwimmbeckens
- Telefonnummern von Versandkunden
- Geschmacksrichtungen von Speiseeis
- Schulnoten auf einer Skala von 1 bis 6
- Abstand zwischen zwei Gebäuden in cm
- Preis eines Neuwagens in Euro und Cent
- Haarfarbe von Kundinnen im Friseursalon
- Temperatur eines glimmenden Holzscheits
- Produktwertung auf einer Skala von 1 bis 5
- Klausurnoten auf einer Skala von 1,0 bis 5,0

Übung: Skalenniveaus und Variablentypen

– Wassertiefe eines Schwimmbeckens	metrisch, stetig
– Telefonnummern von Versandkunden	nominal, diskret
– Geschmacksrichtungen von Speiseeis	nominal, diskret
– Schulnoten auf einer Skala von 1 bis 6	ordinal, diskret
– Abstand zwischen zwei Gebäuden in cm	metrisch, stetig
– Preis eines Neuwagens in Euro und Cent	metrisch, diskret
– Haarfarbe von Kundinnen im Friseursalon	nominal, diskret
– Temperatur eines glimmenden Holzscheits	metrisch, stetig
– Produktwertung auf einer Skala von 1 bis 5	ordinal, diskret
– Klausurnoten auf einer Skala von 1,0 bis 5,0	ordinal, diskret

Teil II

Qualitative und quantitative Forschung

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Was unterscheidet beide Ansätze?

Kombination via „mixed methods“

Quantitative Forschung

- Hypothesen werden vorab festgelegt und überprüft
- Erkenntnisse aus der Stichprobe sollen für Grundgesamtheit gelten
- Im Vordergrund steht die (hoffentlich) objektive Perspektive der Forschenden

Qualitative Forschung

- Hypothesen werden neu aus erhobenen Daten entwickelt
- Erkenntnisse aus Erhebungen werden nicht verallgemeinert
- Im Vordergrund steht die (gewollt) subjektive Perspektive der Betroffenen

Beispielhafte Erhebungsverfahren

Quantitative Forschung

- Versuche
- Experimente
- Befragungen
- Beobachtungen
- Automatische Erfassung

Qualitative Forschung

- Interviews
- Shadowing
- Delphi-Verfahren
- Einzelfallanalysen
- Gruppendiskussionen

Forscher entscheiden, was wichtig ist

Betroffene entscheiden, was wichtig ist

Beispielhafte Auswertungsverfahren

Quantitative Forschung

- Clusteranalyse
- Varianzanalyse
- Faktorenanalyse
- Statistische Tests
- Regressionsanalyse
- Answer-Tree-Verfahren

Kernkompetenz: Mathematik / Statistik

Qualitative Forschung

- Laddering
- Diskursanalyse
- Kategorisierung
- Narrative Analyse
- Konversationsanalyse
- Hermeneutische Analyse

Kernkompetenz: Text- / Inhaltsanalyse

Qualitative Methodik: Das Delphi-Verfahren

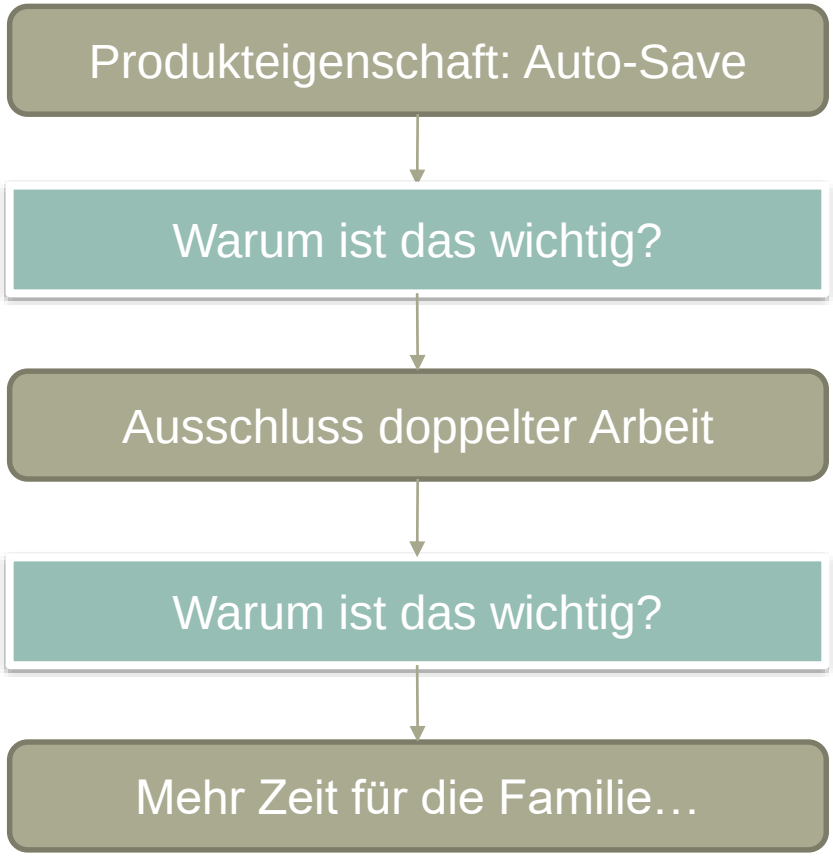
Ruinen von Delphi (Foto © luvmyslr!, lizensiert unter CC BY-ND 2.0)

- Entwickelt durch die RAND Corporation in den 1960ern
- Ziel: Konsensbildung in einer anonymen (von Dominanzen freien) Expertengruppe
- (Grobes) Vorgehen: Thesen kursieren über mehrere Runden in einer Expertengruppe, bis sich iterativ ein Grundkonsens formiert

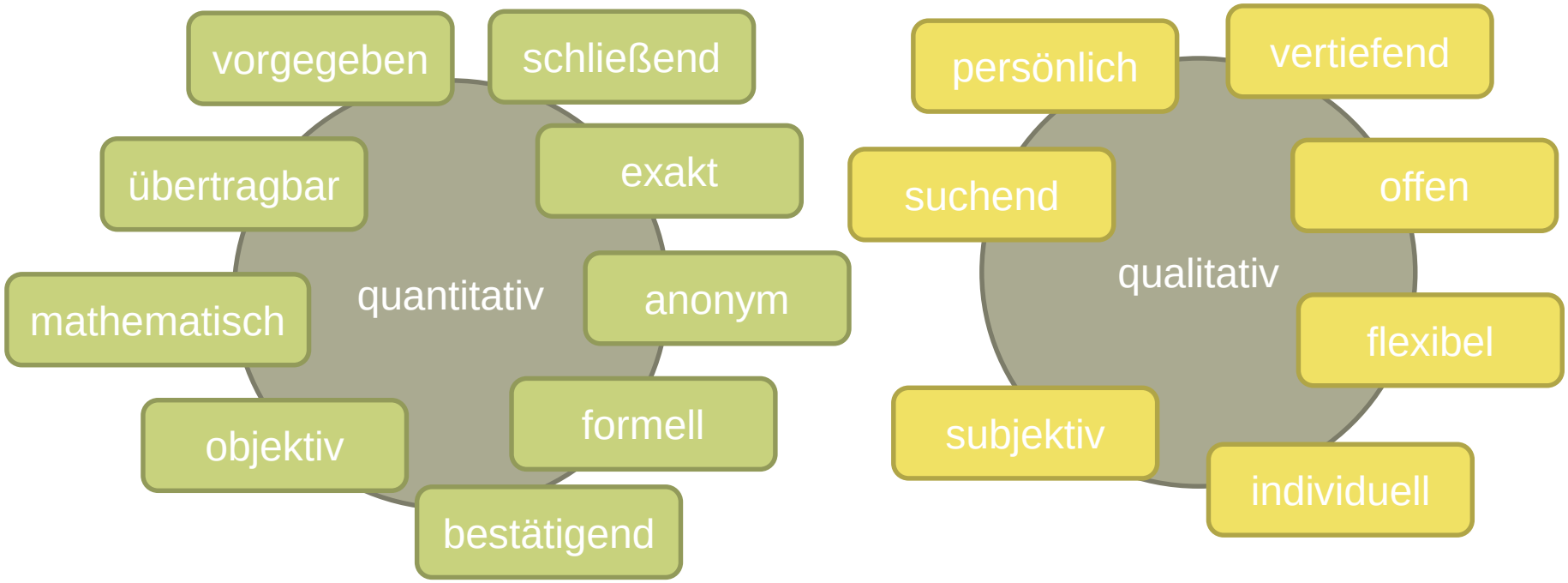
Qualitative Methodik: Das Laddering

– Interview-Methode der 1980er, die dem Ziel der Aufdeckung des subjektiven Kundennutzens von Produkteigenschaften dient

– Grundgedanke: Es wird immer weiter nach dem Nutzen gefragt, bis keine tiefere Antwortebene mehr erreichbar ist → unbewusste Motive und verdeckte Einstellungen der Befragten werden offengelegt



Quantitative vs. qualitative Forschung



Liefert: Zahlen, Daten und Fakten

Liefert: Verständnis und Ideen

Teil III

Planung und Durchführung quantitativer Erhebungen

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Planung und Durchführung quantitativer Erhebungen

Formen der Stichprobenziehung

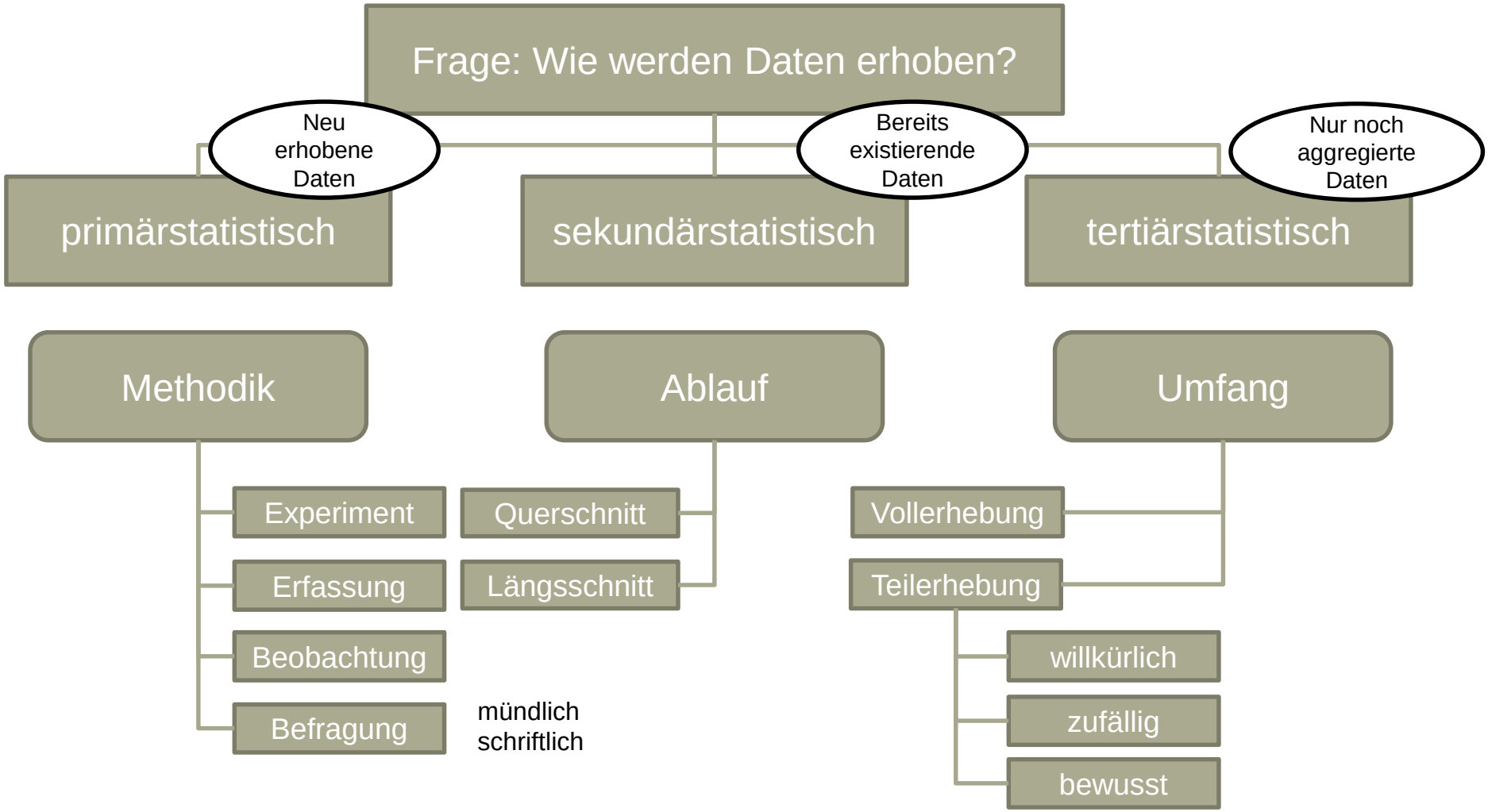
▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Die Phasen der Markt- und Meinungsforschung

Definition	Formulierung der Fragestellung und Erstellung des Forschungsdesigns
Design	Festlegung der Datenquellen und der zu verwendenden Methoden
Datengewinnung	Durchführung von Beobachtungen, Befragungen oder Experimenten
Datenanalyse	Datenbereinigung, Kodierung, Auswertung und Interpretation
Dokumentation	Erstellung des Abschlussberichts und Präsentation der Ergebnisse

Methoden der Datengewinnung



Methoden der Stichprobenziehung (1)

- Willkürliche Auswahl
  - z.B. willkürliche Ansprache von Passantinnen und Passanten in der Fußgängerzone oder von Teilnehmerinnen und Teilnehmern einer Demo; empirisch wertlos (es sei denn für qualitative Vorstudien)
- Zufallsauswahl
  - Einfache Zufallsstichprobe: Jedes Element der Grundgesamtheit hat die exakt gleiche Chance, in die Stichprobe aufgenommen zu werden (z.B. Zufallsauswahl aus einem Register aller Kundinnen und Kunden)

Exkurs: Willkür ≠ Zufall

Methoden der Stichprobenziehung (2)

- Geschichtete Zufallsstichprobe: Durchführung mehrerer einfacher Zufallsstichproben in disjunkten Schichten der Grundgesamtheit (z.B. aus kinderlosen Familien und aus Familien mit Kindern)

- Klumpenstichprobe: Unterteilung einer Grundgesamtheit in natürliche Klumpen auf Basis eines einzelnen Merkmals und anschließende Vollerhebung innerhalb dieser Klumpen (z.B. Untersuchung von Planquadraten auf einer Landkarte)

  [Das Risiko bei diesem Verfahren besteht insbesondere in der irrtümlichen Auswahl nichtrepräsentativer Klumpen]

Methoden der Stichprobenziehung (3)

- Bewusste Auswahl

  - Quotenstichprobe: Konstruktion einer Stichprobe, die bestimmte Merkmale perfekt abbildet, auf Basis dieser Merkmale (z.B. Befragung von Akademikern und Nichtakademikern nach Bevölkerungsanteilen)

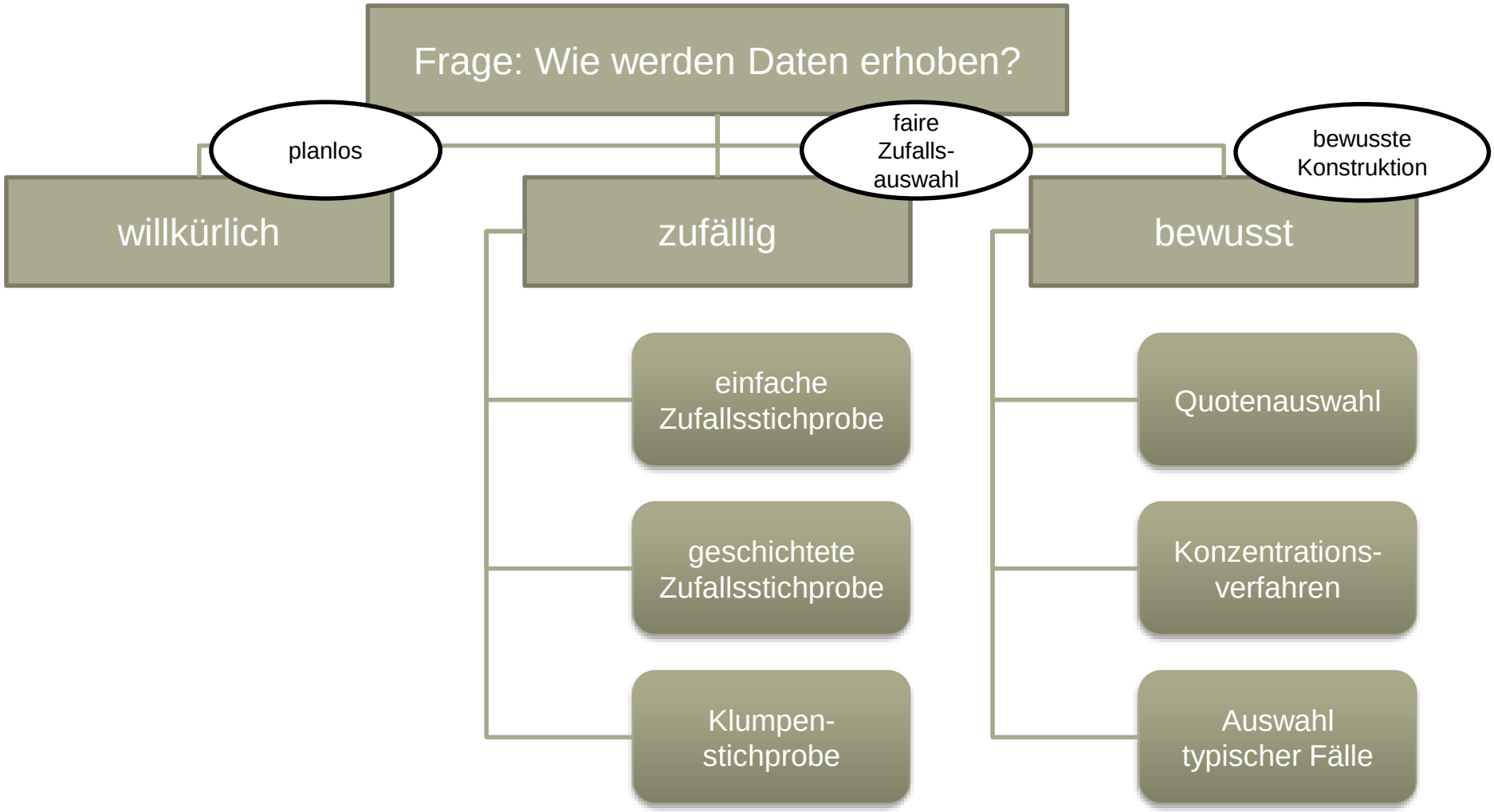
    [Das Problem bei diesem Verfahren besteht insbesondere im stetig schwindenden Spielraum bei der Auswahl der „letzten Fälle“, die oft eine Vielzahl von Merkmalsbedingungen zu erfüllen haben, darunter ggf. auch seltene oder unmögliche Merkmalskombinationen]

Methoden der Stichprobenziehung (4)

- Konzentrationsverfahren: Konzentration auf besonders relevante Teilgesamtheiten (z.B. vorrangige Befragung von Großkunden in einer Kundenbefragung, um deren Bedeutung widerzuspiegeln)

- Auswahl typischer Fälle: (Möglichst objektive) Auswahl „typischer“ Fälle (etwa typischer Kundinnen und Kunden, typischer Studierender oder typischer Mitarbeiterinnen und Mitarbeiter) und deren möglichst vollumfängliche Untersuchung

Methoden der Stichprobenziehung (5)



Wie groß sollte meine Stichprobe sein?

- Stichproben sind nur (streng) repräsentativ, wenn sie drei Bedingungen erfüllen:
  - Echte **Zufallsauswahl** aus einer vollständig erfassten Grundgesamtheit
  - Generierung einer Stichprobe mit ausreichendem **Stichprobenumfang**
  - Hohe **Rücklaufquote** idealerweise von 90% und mehr der Probanden
- Wie man sich leicht vorstellen kann, ist eine Auswahl von 3 Personen aus 1.000 nicht repräsentativ – auch dann nicht, wenn es sich um eine echte Zufallsauswahl handelt und alle 3 Probanden/innen an der Erhebung teilnehmen (100% Rücklauf)
- Da Zufallsauswahl und Rücklaufquote bereits in Statistik I besprochen wurden, bleibt für Statistik II nun nur noch eine offene Frage: **Welchen Umfang sollte eine Zufallsstichprobe mindestens haben?**

Eine Möglichkeit (von vielen): Cochran-Formel

– William G. Cochran entwickelte 1963 die nach ihm benannte Formel basierend auf dem bereits bekannten Prinzip der Konfidenzintervalle

- n = Stichprobenumfang (Zielgröße)
- N = Größe der Grundgesamtheit (z.B. 10.000)
- e = Breite des Konfidenzintervalls (z.B. +/- 5%)
- p = Stichprobenanteil (z.B. 20%)
- q = (1-p) (ergibt sich)
- Z = Z-Wert aus der Standard-normalverteilung für die gewollte Sicherheit des Konfidenzintervalls (z.B. 1,96 bei 95%)

$$n = \frac{\frac{Z^2 * p * q}{e^2}}{1 + \frac{\frac{Z^2 * p * q}{e^2} - 1}{N}}$$

– Ist der Stichprobenanteil (der Anteil an Probanden/innen, welche die untersuchte Merkmalsausprägung aufweisen) unbekannt – was häufig der Fall ist – setzt man mit p=0,5 den konservativsten Schätzwert (maximale Stichprobengröße) ein

Beispielrechnung nach Cochran

– Gegeben sei eine Grundgesamtheit von 50.000 Personen (N), ein unbekannter Stichprobenanteil (p=0,5; q=0,5), sowie eine gewünschte Intervallbreite von +/- 5% um den Stichprobenanteilswert (e=0,05) bei 95%iger Sicherheit (Z=1,96)

$$n = \frac{\frac{Z^2 * p * q}{e^2}}{1 + \frac{\frac{Z^2 * p * q}{e^2} - 1}{N}} = \frac{\frac{1{,}96^2 * 0{,}5 * 0{,}5}{0{,}05^2}}{1 + \frac{\frac{1{,}96^2 * 0{,}5 * 0{,}5}{0{,}05^2} - 1}{50000}} = 381{,}23$$

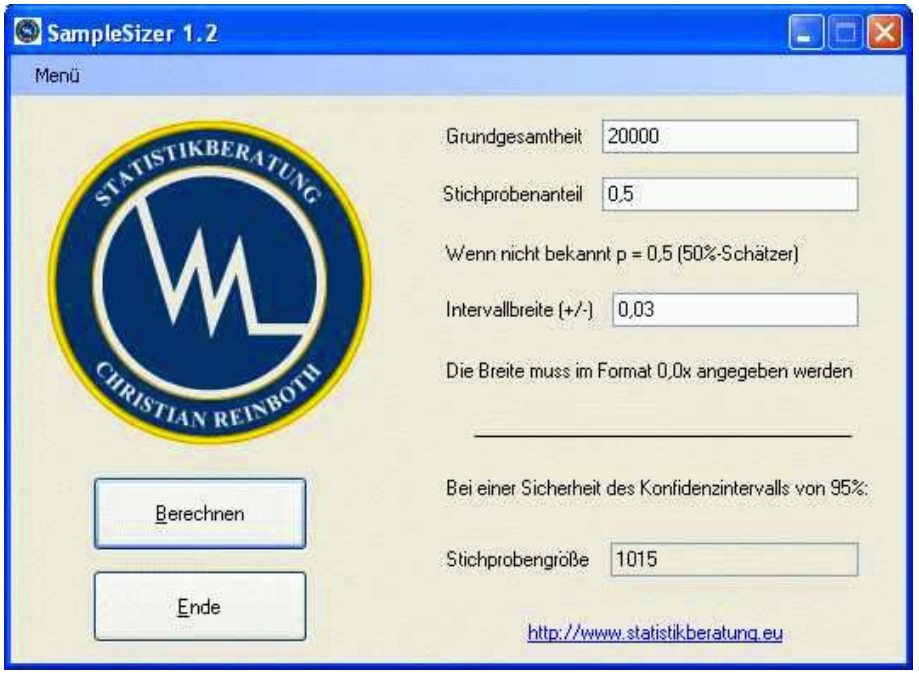
Aufrunden!

– Interpretation: Bei einer Grundgesamtheit von 50.000 Personen wären mindestens 382 Personen zu befragen, wenn man sich zu 95% sicher sein möchte, dass der reale Anteilswert um maximal +/- 5% vom Stichprobenwert abweicht

Der Sample Sizer als Stichproben-Tool

- Was passiert eigentlich bei....
  - größerer Grundgesamtheit?
  - kleinerer Grundgesamtheit?
  - bekannten Anteilswerten?
  - kleinerer Intervallbreite?
  - größerer Intervallbreite?
  - kleinerer Sicherheit?
  - größerer Sicherheit?

- Nicht zulässig ist natürlich das nachträgliche „Anpassen“ der Parameter an das gewollte Ergebnis / die realisierbare Größe



Kostenloser Download unter:
https://www.hs-harz.de/creinboth/lehre/

Planung und Durchführung quantitativer Erhebungen

Wesentliche Gütekriterien

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Die Gütekriterien quantitativer Forschung (1)

Objektivität

Messe ich „fair“ und unabhängig?

– Erhebungen sind objektiv, wenn sie frei von subjektiven Einflüssen sind, die Ergebnisse also nicht von den durchführenden Personen abhängen

– Objektiver Idealfall: Bei der Verwendung des gleichen Messinstruments gelangen unterschiedliche Personen zu den exakt gleichen Resultaten

– Es wird (je nach Stadium der Erhebung) in Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität unterschieden

Die Gütekriterien quantitativer Forschung (2)

Reliabilität

Messe ich sicher und zuverlässig?

– Die Reliabilität bezeichnet den Grad der Zuverlässigkeit, mit der ein Merkmal erfasst wird – grundsätzlich sollte das Ergebnis möglichst unabhängig von einem konkreten Mess-/Erhebungsvorgang sein

– Reliabler Idealfall: Solange sich die Ausprägung eines Merkmals nicht ändert, führen Messungen mit einem reliablen Instrument immer wieder zu identischen Ergebnissen

– Fehlende Werte reduzieren die Reliabilität einer Erhebung

Die Gütekriterien quantitativer Forschung (3)

Validität

Messe ich, was ich messen will?

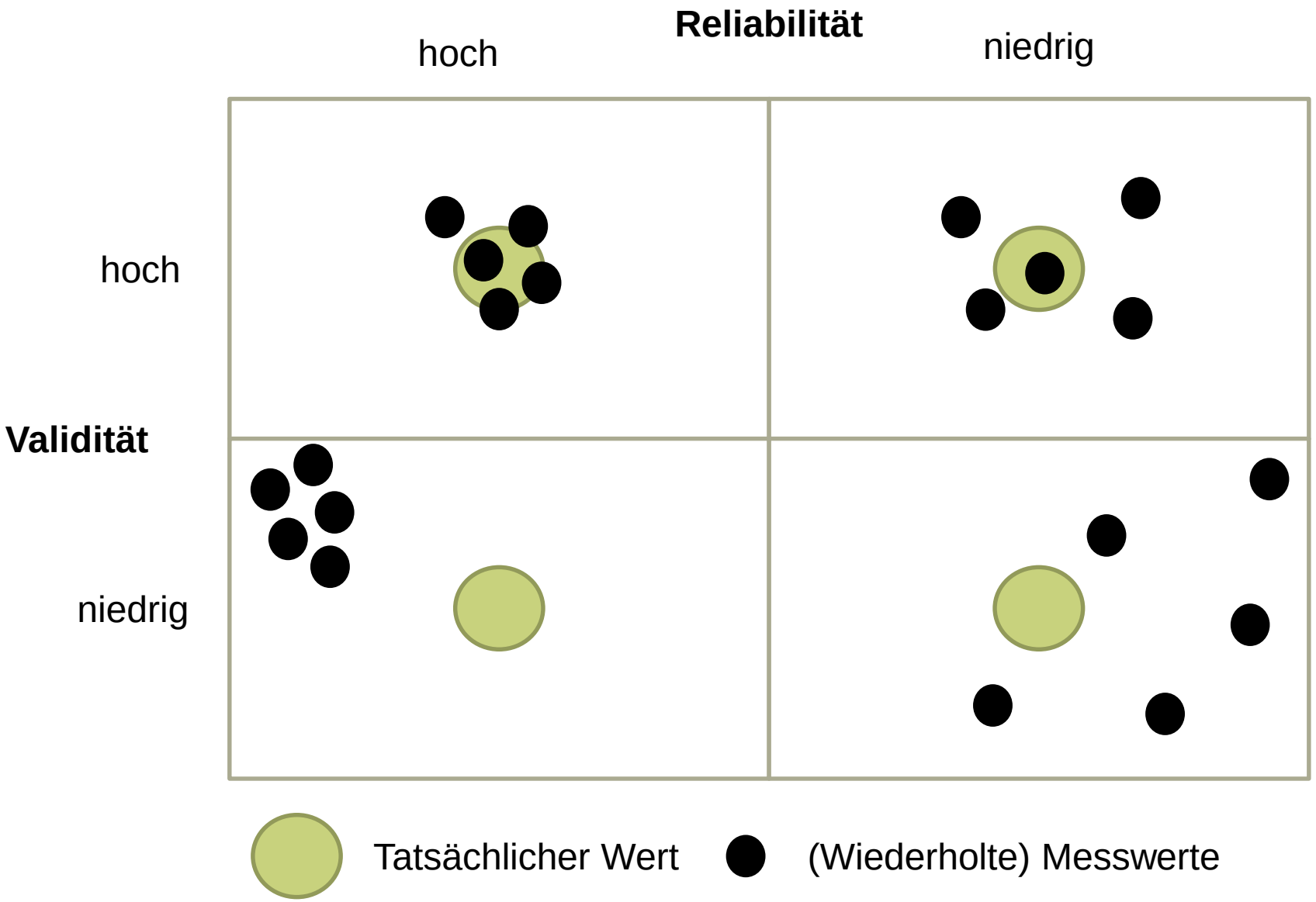
– Eine Messung ist dann valide, wenn sie das Merkmal misst, welches gemessen werden soll

Für alle Merkmale gleich schwierig?

– Es ist zwischen interner und externer Validität zu unterscheiden

Wider-spruch?

  – Interne Validität: Alle Störvariablen sind ausgeschaltet, so dass nur die zu untersuchenden Merkmale erfasst werden (möglichst kontrollierte Umgebung)
  – Externe Validität: Die Ergebnisse sind möglichst gut generalisierbar, lassen sich also auf andere Situationen übertragen (möglichst natürliche Umgebung)


Reliabilität
hoch
niedrig
hoch
Validität
niedrig
Tatsächlicher Wert
(Wiederholte) Messwerte

Weitere Kriterien für die Güte erhobener Daten

- Relevanz für den Untersuchungsgegenstand
- Vollständigkeit und Korrektheit der Datenerfassung
- Aktualität („Nichts ist so alt wie die Zeitung von gestern…“)
- Weiterverwendbarkeit der Daten (Open Access, Datenschutz…)

Tauchen in den Medien nicht immer zwei andere Kriterien auf…?

- Repräsentativität der erhobenen Daten → mehr dazu gleich
- Signifikanz der durchgeführten Tests → mehr dazu später

"Die drei R der guten quantitativen Forschung sind Repräsentativität, Reproduzierbarkeit und R-gebnisoffenheit."

Lars Fischer

Wann sind Daten repräsentativ?
Nicht immer stimmt die Behauptung...

– Eine Stichprobe ist **repräsentativ**, wenn sie alle für die Grundgesamtheit charakteristischen Merkmale und Merkmalskombinationen getreu der realen relativen Häufigkeiten in der Grundgesamtheit aufweist, d.h. ein **exaktes Merkmalsabbild der Grundgesamtheit** darstellt

– Der Begriff hat eine hohe **Suggestivwirkung** und wird in der Praxis der Markt- und Meinungsforschung leider sehr häufig zu Unrecht verwendet

– Faustregel: Der Begriff sollte nur verwendet werden, wenn eine faire statistische Zufallsauswahl ausreichenden Umfangs mit sehr hoher (idealerweise maximaler) Rücklaufquote aus einer klar definierten Grundgesamtheit vorliegt

Exkurs: Das Literary Digest Disaster von 1936

Franklin D. Roosevelt (1882 - 1945) [Foto: U.S. National Archives, Public Domain]

Alfred Landon (1887 - 1987) [Foto: Library of Congress, Public Domain´]

George Gallup (1901 - 1984) [Foto: Wikimedia, gemeinfrei]

Karte: Matté, Public Domain

Die große Reproduktionskrise der Psychologie

„Ob einem jemand sympathisch erscheint, entscheidet sich in den ersten 30 Sekunden!“

„Jüngere Geschwister sind oft durchsetzungsstärker!“

„Ein höherer Blutzucker steigert die Fähigkeit zur Konzentration!“

- 2015: Wiederholung von 100 psychologischen Experimenten
  → in nur 39% aller Fälle ließ sich das Ergebnis reproduzieren
- Mögliche Ursachen

p <= 0.05? Hurra!

  - Publication Bias: Nur signifikante Ergebnisse werden veröffentlicht
  - Häufig viel zu kleine Stichproben
  - Bevölkerung entwickelt sich weiter
    → Effekte sind daher nicht statisch

Planung und Durchführung quantitativer Erhebungen

Was macht gutes Frage(bogen)design aus?

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Warum überhaupt eine schriftliche Befragung?

- Der Hauptunterschied zwischen einer schriftlichen und anderen Befragungstypen ist der fehlende Interviewer

- Der größte Vorteil eines persönlichen Interviews ist die Möglichkeit der individuellen Anpassung an die Situation

  - Überraschende Antworten lassen sich hinterfragen
  - Emotionale Widerstände lassen sich ausräumen
  - Eine Beeinflussung durch Dritte ist ausgeschlossen

- Der Verzicht auf einen Interviewer bedeutet Nachteile

Warum überhaupt eine schriftliche Befragung?

- Ein Interviewer kann aber auch ein Problem sein
  - Er kann dem Probanden unsympathisch sein
  - Er kann bei der Befragung Fehler machen
  - Er kann den Probanden beeinflussen
- Das größte Problem ist das der sozialen Erwünschtheit
  - Probanden versuchen unbewusst, sich nach den subjektiv wahrgenommenen Ansichten des Interviewers zu richten (Grundlage ist das „geschätzt sein wollen“ der meisten)

Warum überhaupt eine schriftliche Befragung?

- Zwei wesentliche Ursachen für erfolglose Gespräche sind die sogenannten Interviewbarrieren und Interviewblocker
  - Interviewbarrieren: Schlechter erster Eindruck des Partners, „gewollte Wahrnehmung“ durch den Interviewer, schlechte Erfahrungen mit Befragungen etc. → unbewusste Einflüsse
  - Interviewblocker: Rückmeldungen des Interviewers während des Interviews („Sind Sie sicher?“), Diagnose des Probanden („Sie sind sehr emotional“) etc. → bewusstes Fehlverhalten

Warum überhaupt eine schriftliche Befragung?

- Diese Probleme sind beim Fragebogen ausgeräumt
  - Es gibt keine (unmittelbare) Beeinflussung durch den Interviewer
  - Für alle Probanden verläuft die Befragung (halbwegs) identisch
  - Die Abwesenheit des Interviewers erhöht außerdem die Chancen dafür, dass Probanden auf sensible Fragen wie zum Beispiel zu Einkommen, Aussehen oder Sexualverhalten antworten – und dass die Antworten kaum durch den Effekt der sozialen Erwünschtheit beeinflusst werden
- **Gelingensvoraussetzung ist gutes Fragen- und Fragebogendesign**

„The quality of the questions asked will have an impact on the quality of the answers received.“

Lee Smith

Bevor man mit dem Schreiben anfängt…

- Welche Fragestellungen sollen beantwortet werden?
- Wie viel Zeit haben die Probanden für die Befragung?
- Wie sollen die Daten anschließend ausgewertet werden?

- Beim Fragebogendesign sind viele Aspekte zu berücksichtigen:
  - Anschreiben, Datenschutzerklärung, Eisbrecherfrage, Design…
- Aus Zeitgründen konzentrieren wir uns nachfolgend auf typische Fehler bei der Formulierung von Fragen und deren Vermeidung

Typische Fehler: Die Doppelfrage

Welches ist der höchste Bildungsabschluss, über den Sie verfügen oder den Sie derzeit anstreben?

– Diese Frage kombiniert zwei Fragestellungen miteinander → bei der Auswertung bleibt unklar, auf welche der beiden Fragen der jeweilige Proband bzw. die Probandin wirklich geantwortet hat

– Lösung: Werden mehrere Antworten benötigt (Länge des Fragebogens beachten!), sind immer auch mehrere Fragen zu stellen

Typische Fehler: Fehlende Antwortoptionen

Welches ist ihr derzeit höchster akademischer Abschluss?

a) Bachelor b) Master c) Magister d) Promotion

– Werden wirklich nur Akademiker*innen befragt? Falls nicht: Was sollen Befragte ohne akademischen Abschluss ankreuzen? Wie unterscheidet man sie von Auskunftsverweigerern?

– Lösung: Gründliche Prüfung aller Antwortmöglichkeit und Einbau einer Non-Option zur Unterscheidung zwischen Personen, die die Frage nicht beantworten wollten und solchen, die es nicht konnten

Typische Fehler: Kategorieüberschneidungen

Welcher Altersgruppe gehören Sie an?

a) 15 – 20 Jahre *b) 20 – 25 Jahre* *c) 25 – 30 Jahre*

– Wie schon bei Doppelfragen müssen manche Probandinnen und Probanden auch bei sich überschneidenden Antwortkategorien praktisch willkürlich entscheiden, wo sie ihr Kreuz setzen

– Lösung: Vorgegebene Antwortkategorien dürfen sich – insbesondere dann, wenn nur eine Antwort zugelassen ist – niemals überschneiden

Typische Fehler: Viel zu viele Fragen

– Angenommen, Sie würden für die Teilnahme an einer Befragung ein Incentive (z.B. einen Einkaufsgutschein) im Wert von 15 Euro erhalten. Wie viel Minuten würden Sie maximal aufwenden wollen?

– [Zusatzfrage: Beeinflusst das verfügbare Einkommen die Bewertung der Angemessenheit von Incentives? Wozu könnte das führen?]

– Lösung: Der Fragebogen sollte niemals länger sein, als unbedingt erforderlich (Lean Design) → darüber hinaus haben Zeitangaben zu Beginn der Befragung unbedingt realistisch zu sein (Frustabbrüche)

Typische Fehler: Falsches Skalenniveau

Bewerten Sie Ihre Zufriedenheit mit dem Produkt auf einer Schulnoten-Skala von 1 (sehr gut) bis 6 (ungenügend).

– An dieser Frage ist ja eigentlich gar nichts falsch…

– Aber: Wenn die Daten in eine Varianzanalyse einfließen sollen, ist die Skala falsch gewählt, da sie ordinale Daten produziert

– Lösung: Bei der Fragenformulierung ist stets zur berücksichtigen, wie die erhobenen Daten im Nachgang ausgewertet werden sollen

Typische Fehler: Abschreckende Fragen

Geben Sie bitte Ihr Jahresbruttoeinkommen (möglichst genaue Angabe) aus nichtselbständiger Tätigkeit für das Jahr 2015 an.

– Bei bestimmten Fragen muss man damit rechnen, dass ein Großteil der Probandinnen und Probanden diese nicht beantworten wird – oder die Befragung schlimmstenfalls sogar verärgert abbricht

– Lösung: Wenn solche Fragen gestellt werden müssen, sind sie am Ende der Befragung zu stellen → dadurch minimiert sich das Risiko eines Totalabbruchs der Erhebung

Typische Fehler: Unverständliche Begriffe

Rechnen Sie sich selbst der Altersgruppe der ‘Best Ager’ zu?

– Befragte, die z.B. einen Fachbegriff nicht kennen, werden eine Frage möglicherweise falsch beantworten (da sie sie falsch interpretieren) oder überspringen, obwohl sie sie eigentlich beantworten könnten

– Lösung: Fachbegriffe sollten nur dann ohne Erläuterungen verwendet werden, wenn man Expertinnen und Experten zum Thema befragt → für alle anderen gilt: Umschreiben oder erklären

Typische Fehler: Zu allgemeine Fragen

Was halten Sie von Umweltschutz?

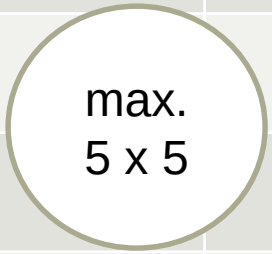
a) Finde ich eher wichtig. *b) Finde ich weniger wichtig.*

– Der Bezugsrahmen dieser Frage bleibt unklar: Will man die Meinung der Probanden zum Umweltschutz im eigenen Lebensumfeld, in der Industrie oder zur Umweltschutzpolitik der Regierung erfragen?

– Lösung: Fragen sollten stets möglichst eindeutig formuliert werden
→ eine der größten Herausforderungen bei der Fragenerstellung

Typische Fehler: Zu große Matrizen

	Note 1	Note 2	Note 3	Note 4	Note 5	Note 6
Autokorrektur						
Displaygröße						
Schnellwahltaste						
App-Verwaltung						
Arbeitsspeicher						
Tastengröße						
Lautstärkeregler						



Typische Fehler: Leading Questions

Die Vorratsdatenspeicherung (VDS) gilt als effizientes Instrument gegen internationalen Terrorismus und Menschenhandel. Finden Sie, dass das Bundesverfassungsgericht das Verbot der VDS angesichts der jüngsten Gewalttaten wieder aufheben sollte?

– Warum ist “Wie würden Sie die Beziehung zu Ihrem Ehepartner beschreiben?” eine bessere Frageformulierung als “Welche Probleme haben Sie mit Ihrem Ehepartner?”

– Lösung: Fragen sollten stets möglichst neutral formuliert werden
  → eine der größten Herausforderungen bei der Fragenerstellung

Teil IV

Deskriptive Statistik

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Deskriptive Statistik

Häufigkeiten

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Absolute und relative Häufigkeiten

- **Absolute Häufigkeit:** Die Anzahl an statistischen Einheiten, die hinsichtlich eines Merkmals die gleiche Ausprägung besitzen (Ergebnis einer einfachen Zählung)

- **Relative Häufigkeit:** Die Anzahl an statistischen Einheiten, die hinsichtlich eines Merkmals die gleiche Ausprägung besitzen, im Verhältnis zur Gesamtzahl der statistischen Einheiten (d.h. der prozentuale Anteil der absoluten Häufigkeit)

- Die Gesamtzahl aller absoluten bzw. relativen Häufigkeiten (in einer Tabelle oder einer Grafik) wird als **absolute bzw. relative Häufigkeitsverteilung** bezeichnet

- Beispiel: 25 Studierende werden nach ihrem Alter befragt. Von diesen 25 geben 13 an, derzeit 24 Jahre alt zu sein. Die absolute Häufigkeit der Altersausprägung „24“ liegt daher bei 13, die relative Häufigkeit dagegen bei 0,52 bzw. 52% (13/25)

Beispiel für eine Häufigkeitstabelle

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

Sind Häufigkeitstabellen eher bei mehr oder eher bei weniger Ausprägungen aussagekräftig?

Bildung von Klassen

– Liegen in einem Datensatz sehr viele Ausprägungen vor, lohnt sich unter Umständen eine Klassenbildung, d.h. die Unterteilung der Daten in Klassen (idealerweise gleicher Breite – dazu in einigen Wochen mehr)

– Bei der Klassenbildung ist zu berücksichtigen, dass eindeutig definiert werden muss, zu welcher Klasse die Elemente der jeweiligen Klassengrenzen gehören

$$K_1 = [g_0, g_1); K_2 = [g_1, g_2); \ldots K_j = [g_{j-1}, g_j)$$

Warum liegt die Grenze der zweiten Klasse bei 28 statt 27 Jahren?

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
[20 – 24) Jahre	9	0,36	36,00%
[24 – 28) Jahre	16	0,64	64,00%
Σ	25	1,00	100,00%

Empirische Verteilungsfunktion

– Mit Hilfe der empirischen Verteilungsfunktion lässt sich die Frage beantworten, welcher Anteil der Daten eine Grenze (nicht) überschreitet bzw. unterschreitet:
F(x) = Welcher Anteil der Daten ist kleiner oder gleich x? („höchstens x“)
(z.B.: Welcher Anteil der befragten Studierenden ist höchstens 23 Jahre alt?)

$$F(x) = \begin{cases} 0 & \text{für } x < a_1 \\ f(a_1) + \ldots + f(a_j) = \sum_{i=1}^{j} f_i & \text{für } a_j \leq x \text{ und } a_{j+1} > x \\ 1 & \text{für } x \geq a_k \end{cases}$$

– Für alle Werte kleiner als die kleinste Ausprägung ist F(x) = 0
– Für alle Werte größer als die größte Ausprägung ist F(x) = 1

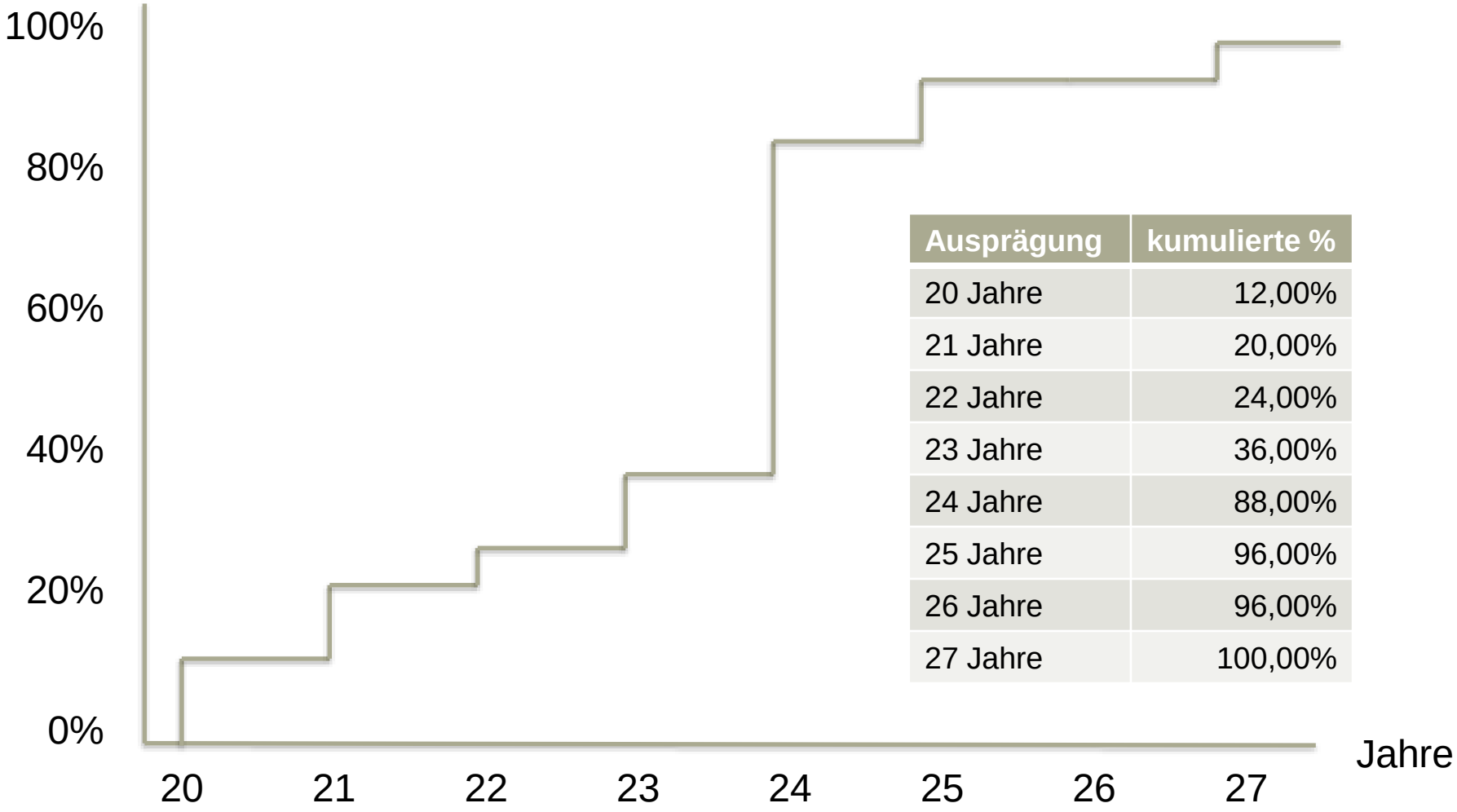
– Die empirische Verteilungsfunktion lässt sich grafisch (Treppendiagramm) oder tabellarisch (Tabelle mit kumulierten absoluten / relativen Häufigkeiten) darstellen

Beispiel für eine Kumulationstabelle

Ausprägung	kumulierte abs. Häufigkeit	kumulierte rel. Häufigkeit	kumulierte %
20 Jahre	3	0,12	12,00%
21 Jahre	5	0,20	20,00%
22 Jahre	6	0,24	24,00%
23 Jahre	9	0,36	36,00%
24 Jahre	22	0,88	88,00%
25 Jahre	24	0,96	96,00%
26 Jahre	24	0,96	96,00%
27 Jahre	25	1,00	100,00%
Σ	25	1,00	100,00%

Welcher Anteil der befragten Studierenden ist höchstens 23 Jahre alt?

Beispiel für ein Treppendiagramm



Ausprägung	kumulierte %
20 Jahre	12,00%
21 Jahre	20,00%
22 Jahre	24,00%
23 Jahre	36,00%
24 Jahre	88,00%
25 Jahre	96,00%
26 Jahre	96,00%
27 Jahre	100,00%

Übung: Rechnen mit der Verteilungsfunktion

- Frage: Welcher Anteil der befragten Studierenden ist höchstens 23 Jahre alt?

  - Lösungsmöglichkeit 1: Ablesen aus der Kumulationstabelle (36%)
  - Lösungsmöglichkeit 2: Berechnung mit der Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x < a_1 \\ f(a_1) + \ldots + f(a_j) = \sum_{i=1}^{j} f_i & \text{für } a_j \leq x \text{ und } a_{j+1} > x \\ 1 & \text{für } x \geq a_k \end{cases}$$

Übung: Rechnen mit der Verteilungsfunktion

- Frage: Welcher Anteil der befragten Studierenden ist höchstens 23 Jahre alt?

  - Lösungsmöglichkeit 1: Ablesen aus der Kumulationstabelle (36%)
  - Lösungsmöglichkeit 2: Berechnung mit der Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x < a_1 \\ f(a_1) + \ldots + f(a_j) = \sum_{i=1}^{j} f_i & \text{für } a_j \leq x \text{ und } a_{j+1} > x \\ 1 & \text{für } x \geq a_k \end{cases}$$

$$F(23) = f(20) + f(21) + f(22) + f(23)$$
$$= 0{,}12 + 0{,}08 + 0{,}04 + 0{,}12 = 0{,}36 = 36\%$$

Summenfunktion

– Bei klassierten Daten wird die empirische Verteilungsfunktion als stetige empirische Verteilungsfunktion oder als Summenfunktion bezeichnet

$$F(x) = \begin{cases} 0 & \text{für} \quad x \leq g_0 \\ F(g_{i-1}) + \dfrac{x - g_{i-1}}{d_i} * f_i & \text{für } g_{i-1} < x \leq g_i \\ 1 & \text{für} \quad x \geq g_k \end{cases}$$

Annahme: Die Werte innerhalb jeder Klasse sind gleichmäßig verteilt

(1) Zunächst wird der Wert der empirischen Verteilungsfunktion bis zum Ende der Klasse berechnet, die vor der Klasse liegt, welche den gesuchten Wert enthält
(2) Anschließend wird die Differenz zwischen gesuchtem Wert und unterer Klassengrenze in der nächsten Klasse berechnet, durch die Klassenbreite geteilt und abschließend mit der relativen Häufigkeit dieser Klasse multipliziert
(3) Zum Schluss werden beide Summen miteinander addiert

Übung: Rechnen mit der Summenfunktion

– Frage: Welcher Anteil der befragten Studierenden ist höchstens 25 Jahre alt?

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
[20 – 24) Jahre	9	0,36	36,00%
[24 – 28) Jahre	16	0,64	64,00%
Σ	25	1,00	100,00%

$$F(x) = \begin{cases} 0 & \text{für} \quad x \leq g_0 \\ F(g_{i-1}) + \dfrac{x - g_{i-1}}{d_i} * f_i & \text{für } g_{i-1} < x \leq g_i \\ 1 & \text{für} \quad x \geq g_k \end{cases}$$

Übung: Rechnen mit der Summenfunktion

– Frage: Welcher Anteil der befragten Studierenden ist höchstens 25 Jahre alt?

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
[20 – 24) Jahre	9	0,36	36,00%
[24 – 28) Jahre	16	0,64	64,00%
Σ	25	1,00	100,00%

$$F(g_{i-1}) = F(24) = 0{,}36$$

$$\frac{x - g_{i-1}}{di} * fi = \frac{25-24}{4} * 0{,}64 = 0{,}16$$

$$0{,}36 + 0{,}16 = 0{,}52 = 52\%$$

Wie kommt es zu der Abweichung im Vergleich zur Berechnung mit den nichtklassierten Daten?

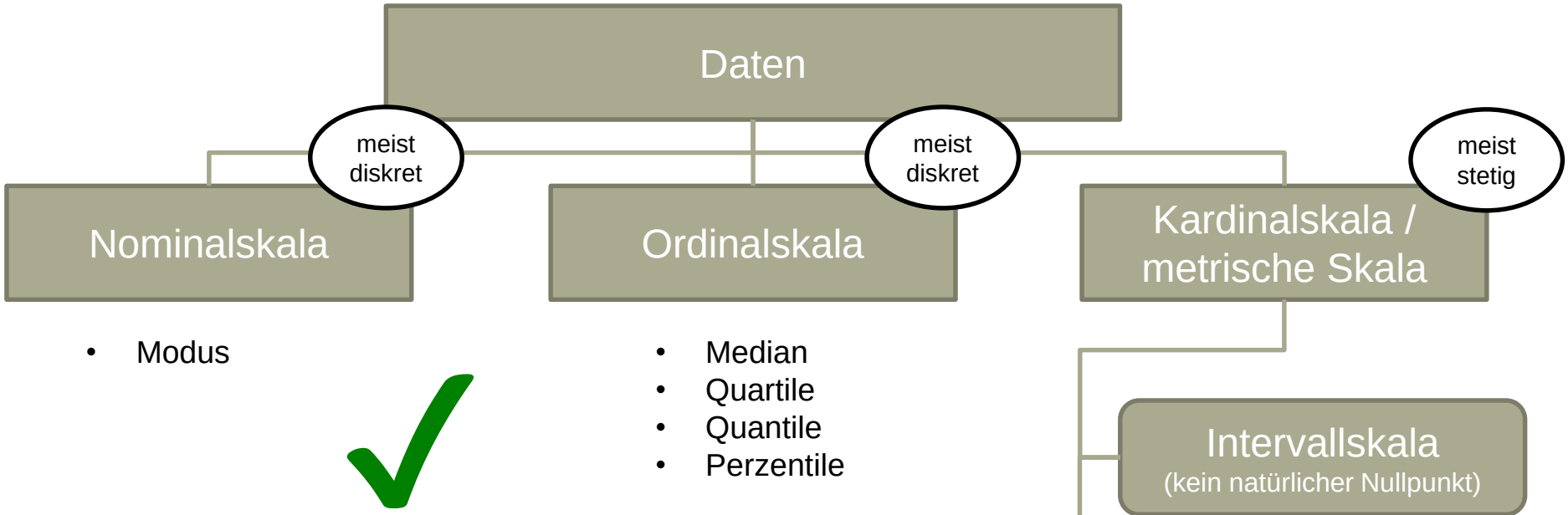
Deskriptive Statistik

Lagemaße / Maße der zentralen Tendenz

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Lagemaße / Maße der zentralen Tendenz



„Abwärts-, aber nicht aufwärtskompatibel“

Lagemaße, welche ein niedriges Skalenniveau voraussetzen, können problemlos für Datensätze auf einem höheren Skalenniveau berechnet werden

Das arithmetische Mittel

- Das arithmetische Mittel ist das **bekannteste statistische Lagemaß** (Standardmittel)
- Es kann **nur** für metrisch skalierte Daten berechnet werden (Intervall-/Verhältnisskala)
  - Vorsicht: SPSS „berechnet“ das arithmetische Mittel auch für nichtmetrische Daten
  - Anwender/innen benötigen daher Methodenkenntnisse (typischer Fehler: Schulnoten)
- Liegen von einem metrischen Merkmal x insgesamt n Werte vor, berechnet sich das arithmetische Mittel auf Basis dieser Formel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Das arithmetische Mittel ist **nicht robust**, d.h. empfindlich gegenüber Ausreißern:

(1, 2, 3, 4) → (1+2+3+4) / 4 = 2,5
(1, 2, 3, 50) → (1+2+3+50) / 4 = 14

Ursache: Jeder Wert in der Verteilung beeinflusst das Mittel gleichermaßen

Exkurs: Lebenserwartung im Mittelalter

Ausschnitt aus dem Dresdner Totentanz von 1534 (Wikimedia, gemeinfrei)

Übung: Arithmetisches Mittel

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

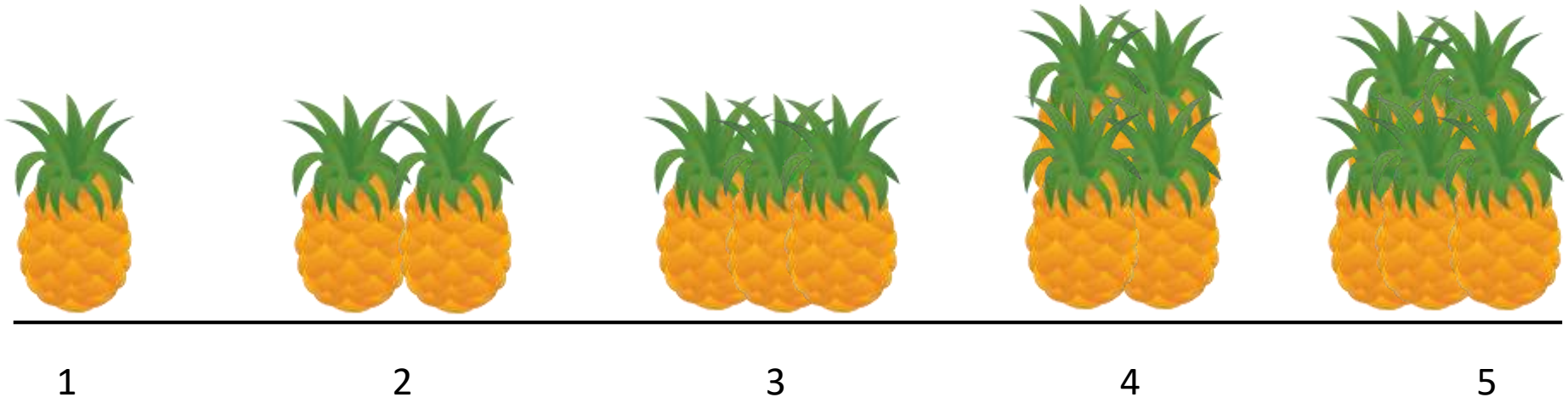
Übung: Arithmetisches Mittel

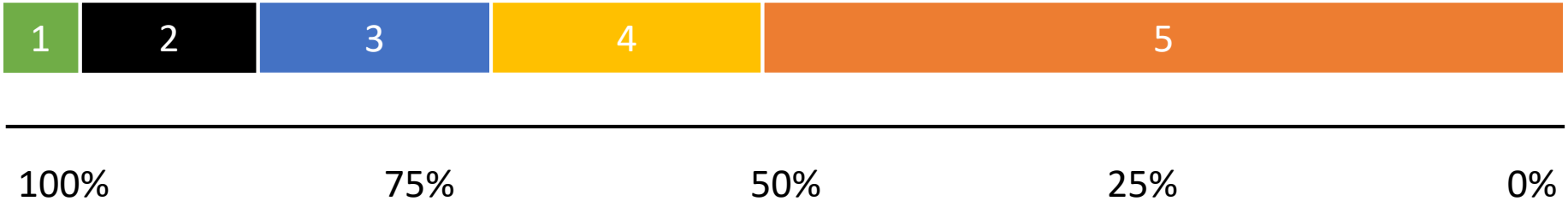
Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad x = \frac{(20+20+20+\ldots+25+25+27)}{25} = \frac{582}{25} = 23{,}28$$

Schulnoten und das arithmetische Mittel

Ein (leider) nie endendes Missverständnis





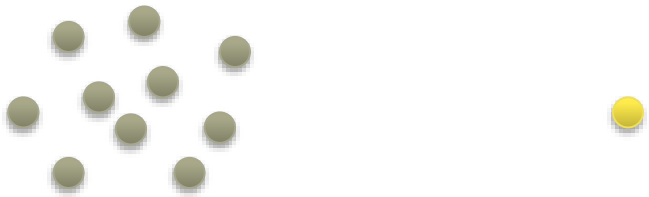
Getrimmtes arithmetisches Mittel

– Treten in einem Datensatz einzelne besonders große oder kleine Werte auf (sogenannte **Ausreißer**), verzerren diese das arithmetische Mittel erheblich

– Möglich ist in diesen Fällen entweder ein Ausweichen auf ein anderes Maß der zentralen Tendenz oder die Berechnung des **getrimmten arithmetischen Mittels**

– Hierfür werden beispielsweise die 2% oder 5% der **jeweils größten und kleinsten Werte** aus dem Datensatz entfernt, bevor das arithmetische Mittel berechnet wird

– Nachteil: Da nicht nur die Ausreißer entfernt werden, sondern die Trimmung symmetrisch erfolgt, kann es zur Entfernung nicht-extremer Werte kommen



Der Median

- Der Median ist derjenige Wert, der **in der Mitte der geordneten Verteilung** liegt
- Die Berechnung des Medians setzt daher mindestens ordinalskalierte Daten voraus
- Bei einer ungeraden Anzahl an Werten wird der mittlere Wert der geordneten Verteilung gewählt

$$x_{med} = x_{(\frac{n+1}{2})}$$

- Bei einer geraden Anzahl an Werten wird das arithmetische Mittel der mittleren Werte gewählt

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

- Der Median ist **äußerst robust**, d.h. er wird von Ausreißern kaum beeinflusst:

(1, 2, 3, 4) → Median: 2,5
(1, 2, 3, 50) → Median: 2,5

Ursache: Nur zwei Werte (bzw. ein Wert) gehen in die Berechnung ein

Übung: Median

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

ungerade Anzahl an Werten (25): $x_{med} = x_{(\frac{n+1}{2})}$

Übung: Median

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

$$x_{med} = x_{(\frac{25+1}{2})} = x_{13} = 24$$

Lässt sich dieses Ergebnis auch direkt aus der Tabelle ablesen?

20; 20; 20; 21; 21; 22; 23; 23; 23; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 25; 25; 27

Die Perzentilwerte

- Perzentilwerte sind Werte, **unterhalb derer ein definierter Anteil aller Werte liegt**

- Für die Perzentilberechnung müssen **mindestens ordinalskalierte Daten** vorliegen

- Der bekannteste Perzentilwert ist das 50%-Perzentil – der bereits bekannte **Median**
- Häufig erfolgt eine „Vierteilung“ des Wertebereichs mit den sogenannten **Quartilen**:
  - 25%-Perzentil (25% aller Werte liegen unterhalb dieses Wertes, 75% liegen oberhalb)
  - 50%-Perzentil – Median (50% aller Werte liegen unter- bzw. oberhalb dieses Wertes)
  - 75%-Perzentil (75% aller Werte liegen unterhalb dieses Wertes, 25% liegen oberhalb)

- Die Quartile spielen u.a. für die **Bildung von Box-Plots** (Grenzen der Box) sowie für die Unterscheidung in **Ausreißer und Extremwerte** (IQR) eine Rolle

- Wie der Median sind auch die restlichen Perzentile **robust gegenüber Ausreißern**

Die Perzentilwerte

– Die Berechnung von Perzentilwerten erfolgt gemäß folgender Formel(n):

– Ergibt (n * p) keinen ganzzahligen Wert, ist k die auf (n * p) folgende ganze Zahl $x_p = x_{(k)}$

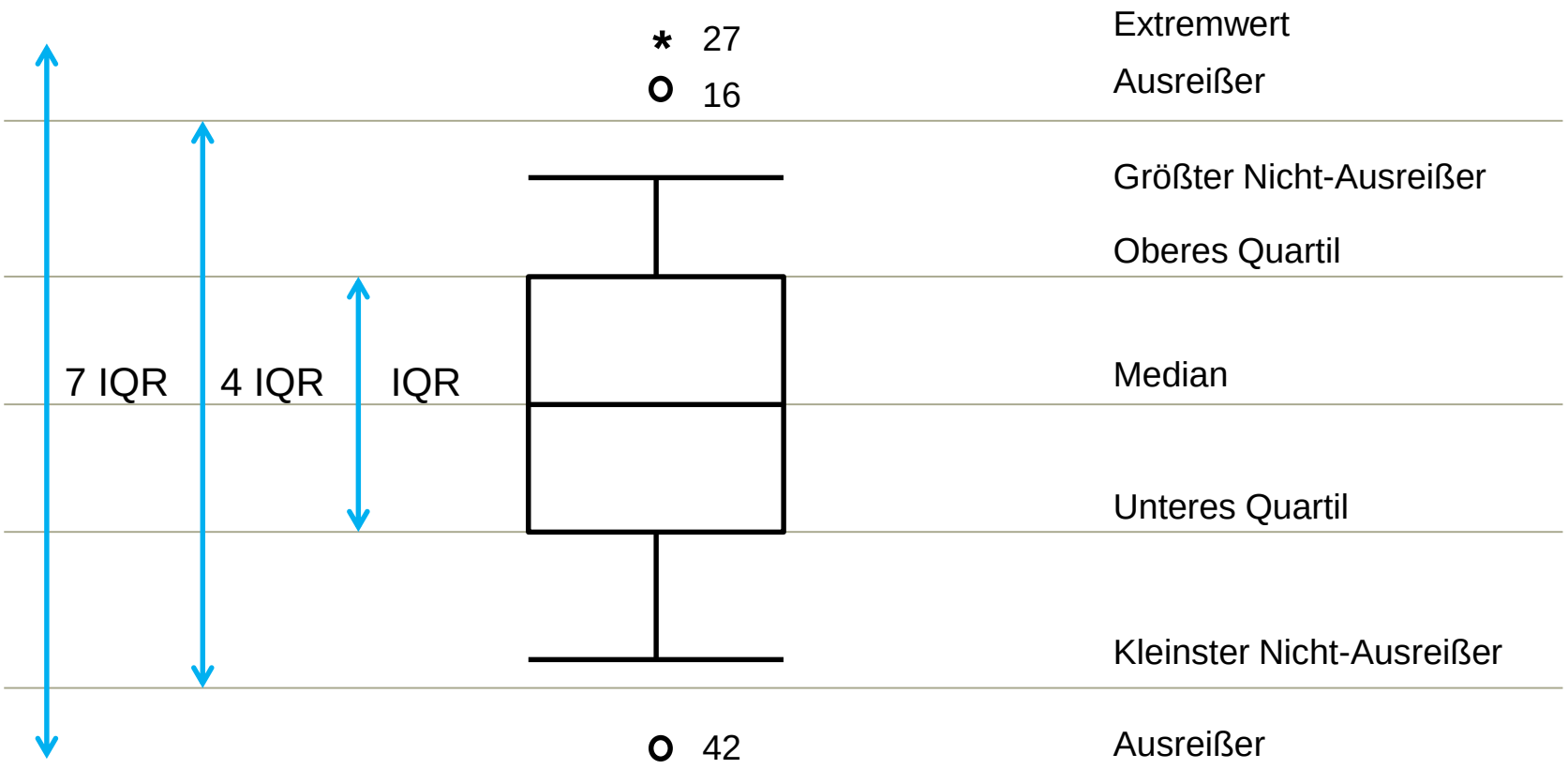
– Ergibt (n * p) einen ganzzahligen Wert, entspricht k dem Ergebnis von (n * p) $x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$

(1) Der gewünschte Perzentilwert (z.B. 0,25 für das 25%-Perzentil) wird mit der Anzahl der Werte im Datensatz (n) multipliziert. In vielen Fällen kommt dabei ein ungerader Wert heraus, der auf den nächsthöheren Wert (k) aufzurunden ist. Der gesuchte Perzentilwert entspricht in diesen Fällen dem k-ten Wert im Datensatz.

(2) Für den Fall, dass sich bei der Multiplikation von n und p doch einmal eine gerade Zahl (k) ergeben sollte, wird das arithmetische Mittel des k-ten Wertes im Datensatz und des auf den k-ten Wert folgenden Wertes im Datensatz berechnet.

Perzentilwerte und Box-Plots

- Box-Plots bieten einen Verteilungsüberblick und gestatten Verteilungsvergleiche
- Wesentliche Konstruktionsgröße ist der Interquartilsabstand (IQR = $x_{0,75} - x_{0,25}$)



Übung: Quartile

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

Bei der Multiplikation von n und p ergeben sich ausschließlich nicht ganzzahlige Werte, daher gilt:

$$x_p = x_{(k)}$$

$x_{0,25} =$

$x_{0,50} =$

$x_{0,75} =$

Übung: Quartile

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

$x_{0,25}$ = 20; 20; 20; 21; 21; 22; 23; 23; 23; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 25; 25; 27 (7)

$x_{0,50}$ = 20; 20; 20; 21; 21; 22; 23; 23; 23; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 25; 25; 27 (13)

$x_{0,75}$ = 20; 20; 20; 21; 21; 22; 23; 23; 23; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 24; 25; 25; 27 (19)

Der Modus

- Der Modus (oder Modalwert) ist der in den Daten **am häufigsten auftretende Wert**
- Bei klassierten Daten entspricht der Modus die **Klassenmitte** der Klasse mit den meisten Fällen (dies gilt allerdings nur beim Vorliegen gleichbreiter Klassen)

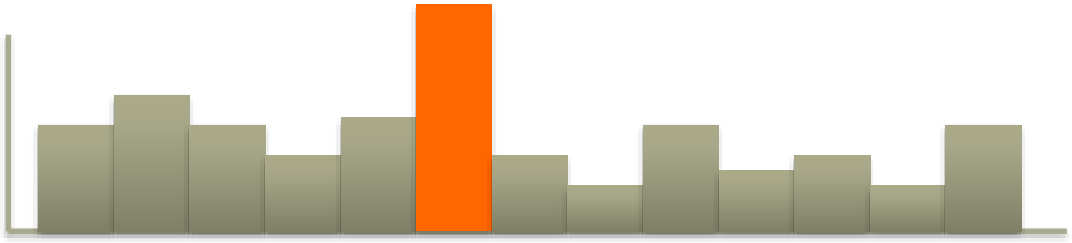
- Der Modus eignet sich vor allem für diskrete Daten (Punktwahrscheinlichkeit)

  - Er wird v.a. für nominalskalierte Daten gebildet, für die sich kein anderes Lagemaß eignet
  - Bei metrisch skalierten Daten kann der Modus über gleichbreite Klassen gebildet werden (in dem Fall entspricht der Modus der Klassenmitte der Klasse mit den meisten Werten)

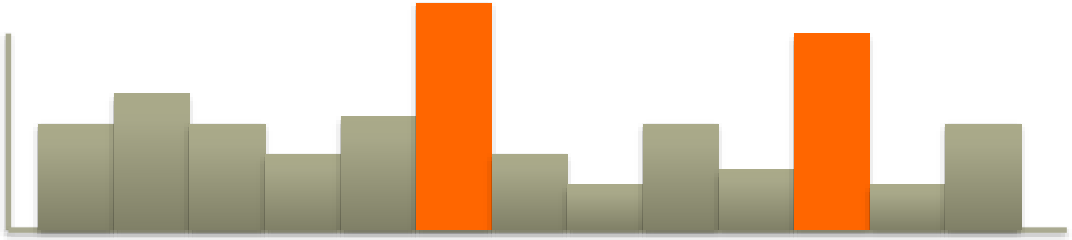
- Vorteil: Der Modus ist ohne Rechnung erkennbar und lässt sich leicht bestimmen
- Nachteil: Der Modus ist nur interpretierbar, wenn **ein klares Maximum** existiert

- Achtung**:** Sind in einem diskreten Datensatz mehrere Werte mit gleicher Häufigkeit vertreten, gibt SPSS nur den in der Häufigkeitstabelle zuoberst stehenden Wert aus

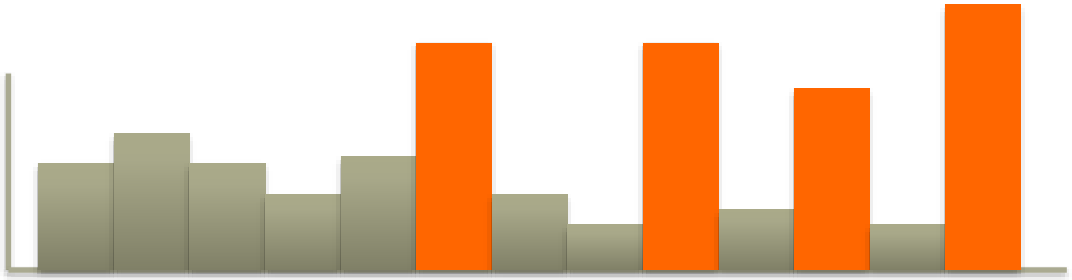
Modus und Verteilungsform



Unimodale Verteilung



Bimodale Verteilung



Multimodale Verteilung

Übung: Modus

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

$$x_{\text{mod}} = 24$$

Warum?

Zusammenfassung der Lagemaße

- Lagemaße beschreiben das **Zentrum einer Verteilung**

- **Arithmetisches Mittel**

  - Sogenanntes „Standardmittel“
  - Nicht robust gegenüber Ausreißern
  - Daten müssen stets metrisch skaliert sein

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Getrimmtes arithmetisches Mittel**

  - Arithmetisches Mittel nach Entfernung einiger Randdaten
  - Trimmung der Daten erfolgt stets beidseitig symmetrisch
  - Ziel ist die Verringerung des Einflusses von Ausreißern

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Zusammenfassung der Lagemaße

- **Median**
  - Mittlerer Wert der geordneten Verteilung
  - Von Ausreißern praktisch nicht beeinflussbar
  - Daten müssen mindestens ordinalskaliert sein
  - Für gerade und ungerade n existieren zwei Formeln

$$x_{med} = x_{(\frac{n+1}{2})}$$

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

- **Perzentile**
  - „Verallgemeinerung“ des Medians
  - Anstelle von 50% werden andere Prozentzahlen gewählt
  - In der Praxis spielen vor allem Quantile und Quartile eine Rolle
  - Für ganzzahlige und nicht ganzzahlige (n*p) existieren zwei Formeln

$$x_p = x_{(k)}$$

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

Zusammenfassung der Lagemaße

- **Modus**
  - Am häufigsten auftretender Wert in den Daten
  - Kann schon für nominalskalierte Werte bestimmt werden
  - Nur sinnvoll, wenn ein einzelnes, klares Maximum vorliegt

$$x_{\text{mod}} = a_{x\max}$$

- **Geometrisches Mittel**
  - Lagemaß für relative Veränderungen (Wachstum)
  - In solchen Fällen das einzig zugelassene Lagemaß
  - Faktoren können unterschiedlich gewichtet werden

$$\overline{x}_{geom} = \sqrt[n]{x_1 \ldots x_n}$$

- **Harmonisches Mittel**

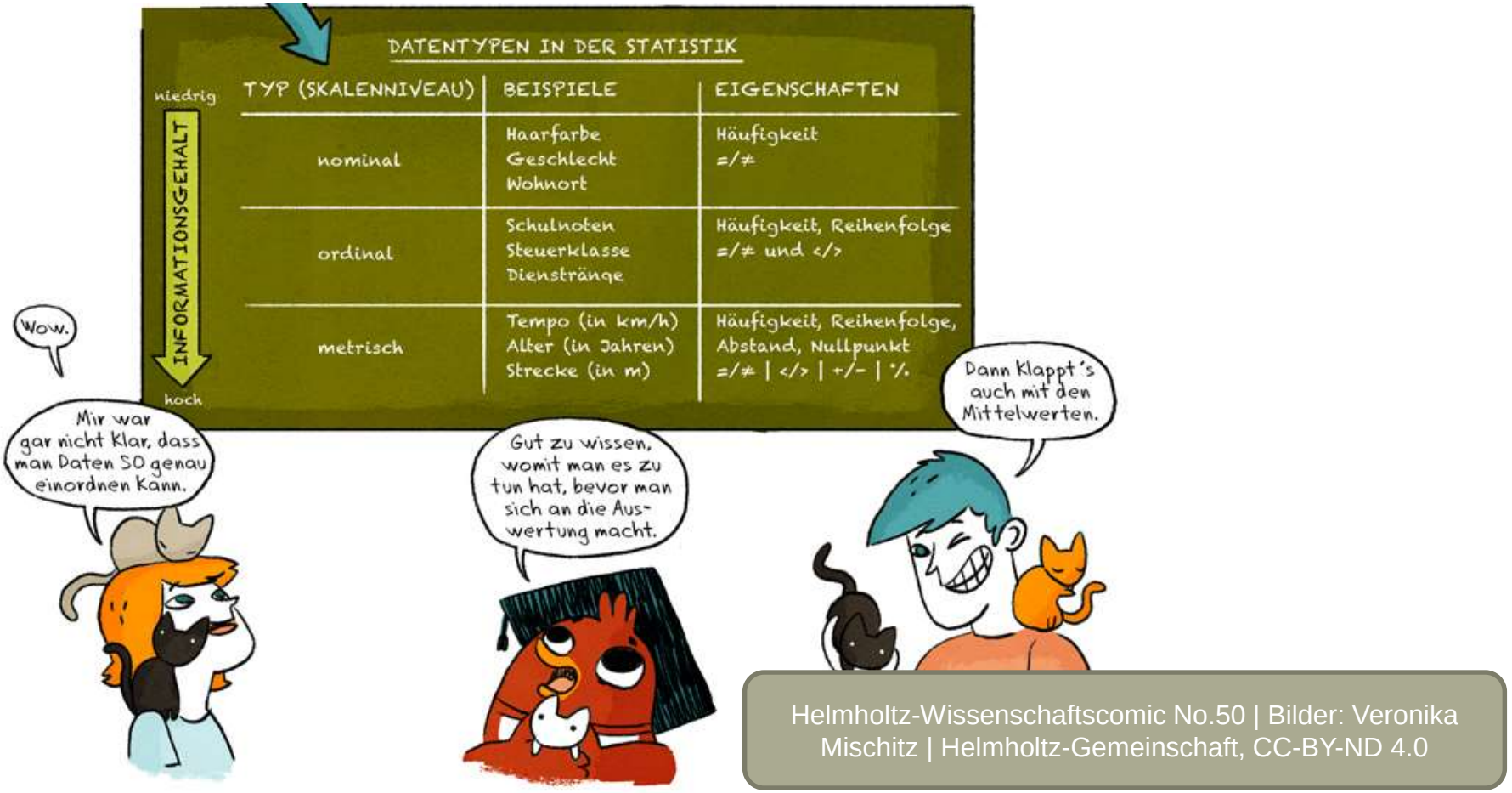
Warum funktioniert das a.M. hier nicht?

$$\overline{x}_{har} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

  - Kommt bei Quotienten zum Einsatz (z.B. Geschwindigkeiten)
  - Kann analog zum geometrischen Mittel gewichtet werden

Helmholz-Wissenschaftscomic: Mittelwerte

https://blogs.helmholtz.de/augenspiegel/2018/03/klar-soweit-no-50/



Helmholtz-Wissenschaftscomic No.50 | Bilder: Veronika Mischitz | Helmholtz-Gemeinschaft, CC-BY-ND 4.0

Das „SPSS-Analyseproblem“

- **SPSS führt JEDE Analyse unabhängig von den Voraussetzungen durch!**

- ...also auch die Berechnung des arithmetischen Mittels

  - ... aus Schulnoten
  - ... aus Geschlechtern
  - ... aus Kontonummern
  - ... aus Telefonnummern
  - ... aus Präferenzrängen

- Bei komplexen Verfahren sind noch weit schlimmere „Vergehen“ denkbar
- Die fachlichen Kenntnisse der Anwender/innen sind daher entscheidend

- **Darum: KEINE Analyse ohne vorherige Prüfung der Voraussetzungen!**

Übung: Maße der zentralen Tendenz

– Berechne: Arith. Mittel, um 5% getrimmtes arith. Mittel, Median und Modus

Schulnote	Anzahl	Schulnote	Anzahl
1	5	4	4
2	8	5	3
3	12	6	1

– Berechne: Arith. Mittel, um 5% getrimmtes arith. Mittel, Median und Modus

Alter	Anzahl	Alter	Anzahl
40	3	34	1
39	4	33	3
38	2	32	4
37	6	31	2
36	2	30	5
35	1	29	1

Übung: Maße der zentralen Tendenz

– Berechne: Arith. Mittel, um 5% getrimmtes arith. Mittel, Median und Modus

Schulnote	Anzahl	Schulnote	Anzahl
1	5	4	4
2	8	5	3
3	12	6	1

$$x_{med} = 3$$

$$x_{\text{mod}} = 3$$

– Berechne: Arith. Mittel, um 5% getrimmtes arith. Mittel, Median und Modus

Alter	Anzahl	Alter	Anzahl
40	3	34	1
39	4	33	3
38	2	32	4
37	6	31	2
36	2	30	5
35	1	29	1

Warum kein Modus?

$$\bar{x} = 34{,}79$$

$$\bar{x}_{get} = 34{,}80$$

$$x_{med} = 35{,}50$$

Deskriptive Statistik

Streuungsmaße / Dispersionsparameter

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Wozu werden Streuungsmaße benötigt?

Mitarbeiter Abt. A	Einkommen	Mitarbeiter Abt. B	Einkommen
MA 1	2.500,00 Euro	MA 1	4.130,00 Euro
MA 2	2.550,00 Euro	MA 2	1.060,00 Euro
MA 3	2.480,00 Euro	MA 3	1.110,00 Euro
MA 4	2.630,00 Euro	MA 4	5.020,00 Euro
MA 5	3.000,00 Euro	MA 5	4.000,00 Euro
MA 6	2.210,00 Euro	MA 6	1.250,00 Euro
Summe	15.370,00 Euro	Summe	16.570,00 Euro
Arithmetisches Mittel	2.561,67 Euro	Arithmetisches Mittel	2.761,67 Euro

Sollte man die Mittelwerte direkt miteinander vergleichen?

Die Spannweite

– Die Spannweite ist als der **absolute Abstand** zwischen dem jeweils kleinsten (Minimum) und größten (Maximum) Wert im untersuchten Datensatz definiert

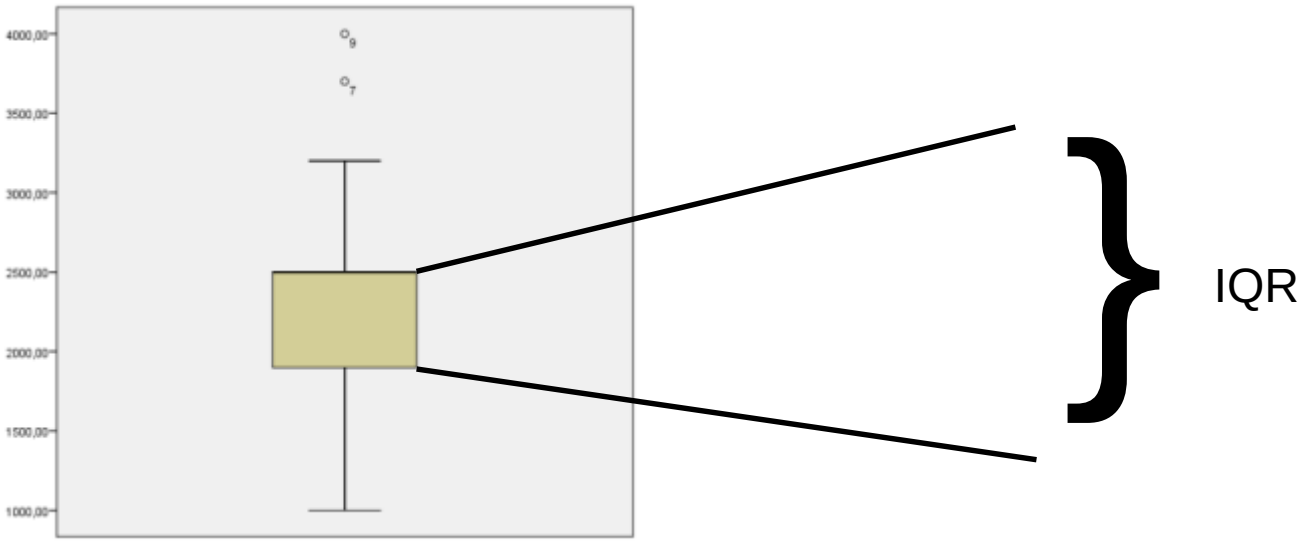
– Die Spannweite ist als Streuungsmaß in den meisten Fällen ungenügend, da sie – soweit vorhanden – **extrem stark von Ausreißern beeinflusst wird**
– Existieren an beiden Verteilungsrändern Ausreißer, wird der Wert der Spannweite tatsächlich sogar ausschließlich (!) durch diese bestimmt

(1, 2, 3, 4, 5) → Spannweite: 4
(1, 2, 3, 4, 50) → Spannweite: 49

Der Interquartilsabstand

– Der **Interquartilsabstand** (IQR = Inter-Quartile Range) ist definiert als der Abstand zwischen dem oberen (75%) und dem unteren Quartil (25%)

– Da die Quartile bekanntlich nicht von Ausreißern beeinflusst werden können, ist der IQR als Streuungsmaß deutlich robuster als die Spannweite
– Quartile, Minimum und Maximum bilden die Fünf-Werte-Zusammenfassung



Varianz und Standardabweichung

Durch-schnitt-liche Ab-weichung

- Die Varianz (bzw. empirische Varianz) ist das **meistgenutzte Streuungsmaß**
- Sie berechnet sich als Summe der quadrierten Abweichungen der Einzelwerte (Ausgleich negativer und positiver Abweichungen) vom arithmetischen Mittel, geteilt durch die Gesamtzahl aller Werte

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  - Bei der Berechnung der Stichproben-Varianz (SPSS) stehen die Freiheitsgrade im Nenner
  - Die Varianz wird immer kleiner, je näher die Einzelwerte am arithmetischen Mittel liegen
  - Sind alle Werte mit dem Mittel identisch (keine Streuung), ergibt sich eine Nullvarianz

- Bei der Interpretation ist zu beachten, dass mit **quadrierten Werten** gerechnet wird

  - Auch die Varianz ist also in der quadrierten Einheit dimensioniert (z.B. in €² statt in €)
  - Die **Standardabweichung** als Quadratwurzel der Varianz erleichtert die Interpretation

Übung: Varianz und Standardabweichung

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

$$s^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n}(x_1 - \bar{x})^2$$

$$\bar{x} = 23{,}28$$

Übung: Varianz und Standardabweichung

$$(20-23{,}28)^2 = 10{,}7584$$
$$(20-23{,}28)^2 = 10{,}7584$$
$$\ldots$$
$$(25-23{,}28)^2 = 2{,}9584$$
$$(27-23{,}28)^2 = 13{,}8384$$

$$\sum = 71{,}04$$

$$\frac{71{,}04}{25} = 2{,}8416$$

$$20^2 = 400$$
$$20^2 = 400$$
$$\ldots$$
$$25^2 = 625$$
$$27^2 = 729$$

$$\sum = 13620$$

$$\frac{13620}{25} = 544{,}80$$

$$23{,}28^2 = 541{,}9584$$

$$544{,}80 - 541{,}9584 = 2{,}8416$$

$$s^2 = 2{,}8416$$
$$s = 1{,}6857$$

Wie sind die Ergebnisse zu interpretieren?

In welcher Einheit stehen die Ergebnisse?

Streuungsmaße / Dispersionsparameter

- Streuungsmaße geben Auskunft darüber, **wie stark Daten um das Zentrum einer Verteilung** (Mittelwert) **streuen**

- **Empirische Varianz**

  - Mittlere quadrierte Abweichung vom arithmetischen Mittel
  - Kann daher nur für metrisch skalierte Daten berechnet werden
  - Varianz ist nicht robust, d.h. empfindlich gegenüber Ausreißern
  - Die hier dargestellte Formel ist die vereinfachte Rechenvariante

$$s^2 = \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right\} - \bar{x}^2$$

- **Standardabweichung**

  - Durch die Quadratur ist die Varianz schwer interpretierbar, da sie sich in Einheiten wie z.B. €² oder Stunden² ausdrückt
  - Die Standardabweichung ist die positive Wurzel der Varianz

$$s = +\sqrt{s^2}$$

Streuungsmaße / Dispersionsparameter

- **Variationskoeffizient**

  - Streuungen in unterschiedlichen Einheiten sind nicht vergleichbar
  - Beispiel: Währungsschwankungen in verschiedenen Währungen
  - Ist der Mittelwert positiv, können die Daten aber normiert werden
  - Der entstehende Variationskoeffizient gestattet direkte Vergleiche

$$v = \frac{s}{\bar{x}} \qquad \bar{x} > 0$$

- **Spannweite**

  - Differenz zwischen größtem und kleinstem Wert
  - In die Berechnung fließen also nur wenige Daten ein
  - Ausreißer beeinflussen die Spannweite daher erheblich

$$d_s = x_{\max} - x_{\min}$$

Streuungsmaße / Dispersionsparameter

- **Interquartilsabstand (IQR)**
  - Der IQR ist der Abstand zwischen oberem und unterem Quartil
  - Er wird für Box-Plot und Fünf-Werte-Zusammenfassung benötigt

$$IQR = x_{0,75} - x_{0,25}$$

- **Fünf-Werte-Zusammenfassung**
  - Hochkomprimierte Darstellung von Streuung und Lage einer Verteilung, bestehend aus dem Minimum, dem Maximum und den drei Quartilen

$$\left[x_{\min}; x_{0,25}; x_{med}; x_{0,75}; x_{\max}\right]$$

Übung: Streuungsmaße

– Berechne: Spannweite, IQR, Varianz und Standardabweichung

Schulnote	Anzahl	Schulnote	Anzahl
1	5	4	4
2	8	5	3
3	12	6	1

– Berechne: Spannweite, IQR, Varianz und Standardabweichung

Alter	Anzahl	Alter	Anzahl
40	3	34	1
39	4	33	3
38	2	32	4
37	6	31	2
36	2	30	5
35	1	29	1

Übung: Streuungsmaße

– Berechne: Spannweite, IQR, Varianz und Standardabweichung

Schulnote	Anzahl	Schulnote	Anzahl
1	5	4	4
2	8	5	3
3	12	6	1

$$IQR = (3-2) = 1$$

– Berechne: Spannweite, IQR, Varianz und Standardabweichung

Alter	Anzahl	Alter	Anzahl
40	3	34	1
39	4	33	3
38	2	32	4
37	6	31	2
36	2	30	5
35	1	29	1

$$d_s = (40-29) = 11$$

$$IQR = (38-32) = 6$$

$$s^2 = 12{,}45$$

$$s = 3{,}53$$

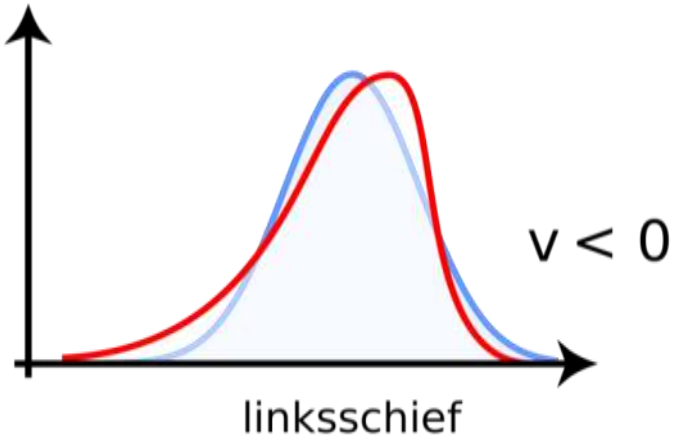
Deskriptive Statistik

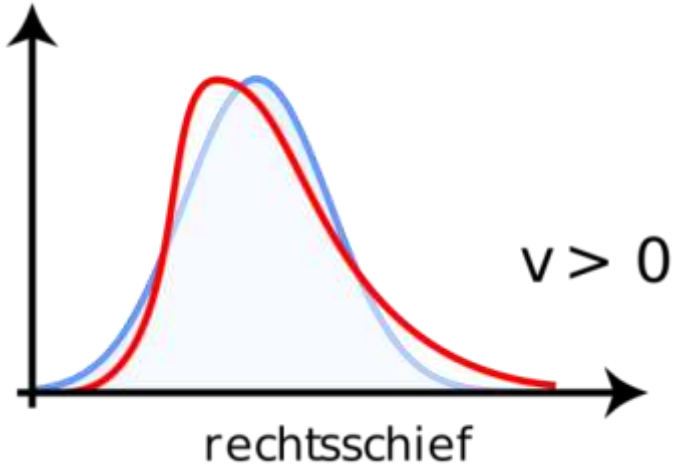
Schiefe und Wölbung

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Schiefe und Wölbung





- Verteilungen können nach **Schiefe** unterschieden werden
  - Symmetrische Verteilungen (spiegelbildlich)
  - Linkssteile und rechtsschiefe Verteilungen
  - Rechtssteile und linksschiefe Verteilungen
- Zudem kann nach der **Wölbung** unterschieden werden
  - Der Wölbungsgrad entspricht der Wölbung einer Normalverteilung
  - Die Wölbung verläuft flacher als die Wölbung einer Normalverteilung
  - Die Wölbung verläuft spitzer als die Wölbung einer Normalverteilung

Quelle: Wikimedia Commons / User: Christian Schirm / Lizenz: gemeinfrei

Schiefe und Wölbung

- **Momentenkoeffizient der Schiefe**
  - Abweichung der Verteilung von der symmetrischen Form
  - Die Daten müssen dabei mindestens intervallskaliert sein
  - Es ergeben sich positive Werte für linkssteile Verteilungen und negative Werte für rechtssteile Verteilungen sowie Werte nahe 0 für symmetrische Verteilungen

$$g_m = \frac{m_3}{s^3}$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

$$s^3 = \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^3$$

- **Quartilskoeffizient der Schiefe**
  - Koeffizient wird mit den Quartilen gebildet
  - Daten müssen daher lediglich ordinalskaliert sein
  - Interpretation ist identisch zum Momentenkoeffizient

$$g_{0,25} = \frac{(x_{0,75} - x_{med}) - (x_{med} - x_{0,25})}{x_{0,75} - x_{0,25}}$$

Was passiert bei IQR=0?

Wichtig: Beide Maßzahlen für die Schiefe sind lediglich für unimodale Verteilungen sinnvoll interpretierbar!

Schiefe und Wölbung

- **Kurtosis / Exzeß**
  - Abweichung der Wölbung von der einer Normalverteilung
  - Es ergeben sich positive Werte für spitze Verteilungen und negative Werte für flache Verteilungen

$$g_k = \frac{m_4}{s^4} - 3$$

$$m_4 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^4$$

$$s^4 = \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^4$$

Auch die Kurtosis ist nur bei einer unimodalen Verteilung sinnvoll interpretierbar

Lagemaße und Box-Plots

– Aus der Lage des Medians im Box-Plot lässt ebenfalls die Verteilungsform ablesen

Symmetrische Verteilung

Linkssteile Verteilung

Rechtssteile Verteilung

Lagemaße und Verteilungsformen

Lagemaß	min. Skalenniveau
Modalwert	Nominalskalenniveau
Median / Perzentile	Ordinalskalenniveau
Arithmetisches Mittel	Metrisches Skalenniveau

Verhältnis der Lagemaße	Form der Verteilung
$\bar{x} \approx x_{med} \approx x_{\text{mod}}$	Symmetrische Verteilung
$\bar{x} < x_{med} < x_{\text{mod}}$	Rechtssteile Verteilung
$\bar{x} > x_{med} > x_{\text{mod}}$	Linkssteile Verteilung

Übung: Quartilskoeffizient und Kurtosis

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

$$g_{0,25} = \frac{(x_{0,75} - x_{med}) - (x_{med} - x_{0,25})}{x_{0,75} - x_{0,25}}$$

$$g_k = \frac{m_4}{s^4} - 3$$

$$m_4 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^4$$

$$s^4 = \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^4$$

Übung: Quartilskoeffizient und Kurtosis

$$g_{0,25} = \frac{(x_{0,75} - x_{med}) - (x_{med} - x_{0,25})}{x_{0,75} - x_{0,25}}$$

$$x_{0,25} = 23$$

$$x_{0,50} = 24$$

$$x_{0,75} = 24$$

$$g_{0,25} = \frac{(24-24)-(24-23)}{24-23}$$

$$g_{0,25} = \frac{-1}{1} = -1$$

Wie sind die Ergebnisse zu interpretieren?

$$g_k = \frac{m_4}{s^4} - 3$$

$$m_4 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^4$$

$$s^4 = \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^4$$

$$m_4 = \frac{1}{25} * 616{,}47 = 24{,}66$$

$$s^4 = \sqrt{2{,}84}^4 = 8{,}07$$

$$g_k = \frac{24{,}66}{8{,}07} - 3 = 0{,}55$$

Deskriptive Statistik

Korrelationskoeffizienten

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Wie lassen sich Zusammenhänge aufspüren?

- Für zwei Variablen X und Y kann dann ein Zusammenhang unterstellt werden (dieser **muss aber real nicht existieren**), wenn sie sich gleichmäßig verändern
  - Gleichsinnig = wird X größer wird Y größer; wird X kleiner wird Y kleiner
  - Gegensinnig = wird X größer wird Y kleiner; wird X kleiner wird Y größer
- Die Berechnung von **Korrelationskennzahlen** orientiert sich am Skalenniveau
  - Nominalskalenniveau: Chi²-Koeffizient
  - Ordinalskalenniveau: Spearman, Kendall
  - Metrisches Skalenniveau: Bravais-Pearson
- Grundsätzlich immer möglich ist auch eine **grafische Analyse der Daten**
  - Diskrete Daten: Gruppierte Balkendiagramme, Bedingte Balkendiagramme
  - Stetige Daten: Zwei- und dreidimensionale Streudiagramme, Scatterplot-Matrix

Analyse bivariater Zusammenhänge

Frage: Liegt in einem bivariaten Datensatz ein Zusammenhang vor?

grafisch	nominalskaliert	ordinalskaliert	metrisch
stetig: Streudiagramm, Scatterplot-Matrix	Chi^2-Koeffizient	Konkordanz-koeffizient nach Kendall	Bravais-Pearson-Korrelations-koeffizient
diskret: Balkendiagramme (gruppiert, bedingt)		Rangkorrelations-koeffizient nach Spearman	

Der Bravais-Pearson-Korrelationskoeffizient

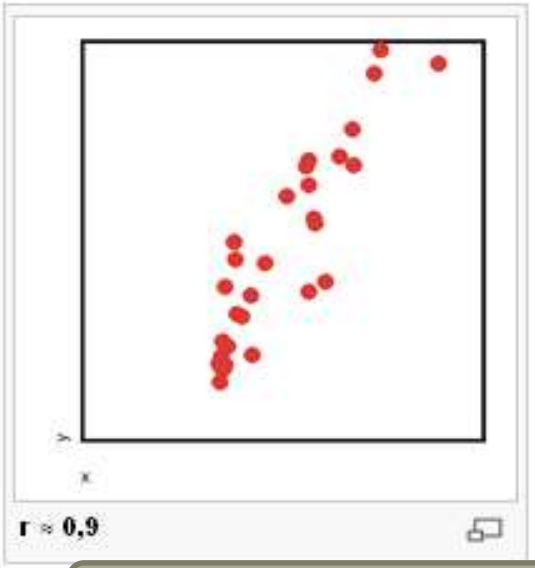
- Für metrisch skalierte Merkmale wird in den meisten Fällen der Bravais-Pearson-Korrelationskoeffizient berechnet (obwohl auch andere Koeffizienten möglich sind)

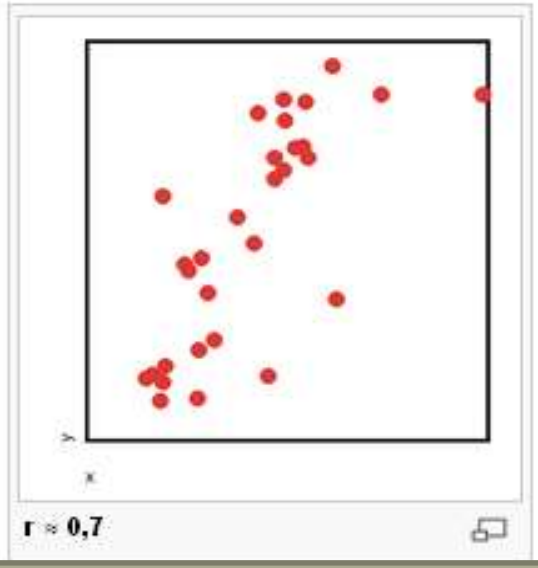
- Bei der Interpretation zu beachten: Der Bravais-Pearson-Korrelationskoeffizient misst **ausschließlich den linearen Zusammenhang** zwischen zwei Variablen
- Nicht-lineare (z.B. quadratische oder logarithmische) Zusammenhänge werden somit nicht aufgedeckt, auch wenn sie stark oder sogar vollkommen sein sollten

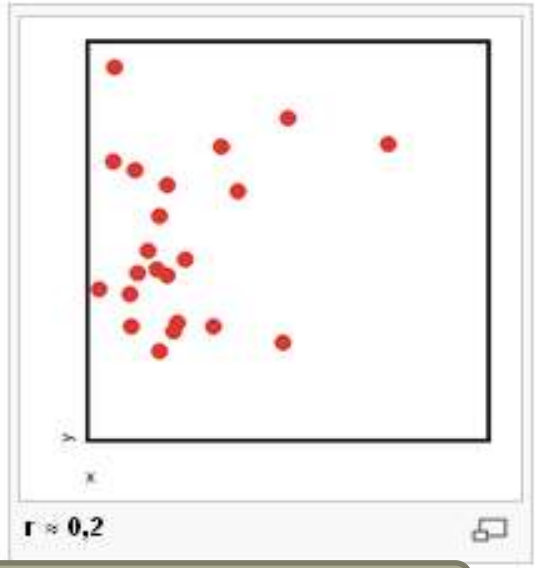
$$r = \frac{\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^{-2}} * \sqrt{\sum_{i=1}^{n}(y_i^2) - n * \bar{y}^{-2}}}$$

Der Bravais-Pearson-Korrelationskoeffizient

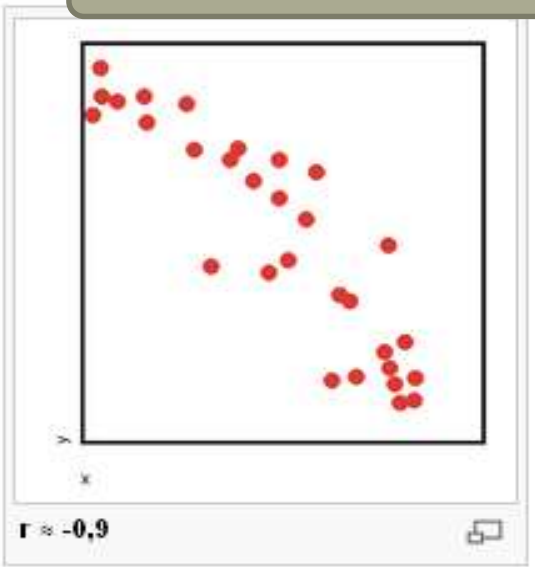
- Der Koeffizient r kann Werte zwischen -1 und +1 annehmen
  - Bei positiven Werten liegt ein positiver Zusammenhang vor, d.h. die Wertepaare liegen auf einer steigenden Gerade
  - Bei negativen Werten liegt ein negativer Zusammenhang vor, d.h. die Wertepaare liegen auf einer fallenden Gerade
  - Werte nahe Null deuten darauf hin, dass keinerlei lineare Korrelation zwischen den beiden Variablen vorliegt
- Interpretation des Betrags (!) von r
  - r = 0 = keine Korrelation
  - 0 < r < 0,5 = schwache Korrelation
  - 0,5 <= r < 0,8 = mittlere Korrelation
  - 0,8 <= r < 1 = starke Korrelation
  - r = 1 = perfekte Korrelation

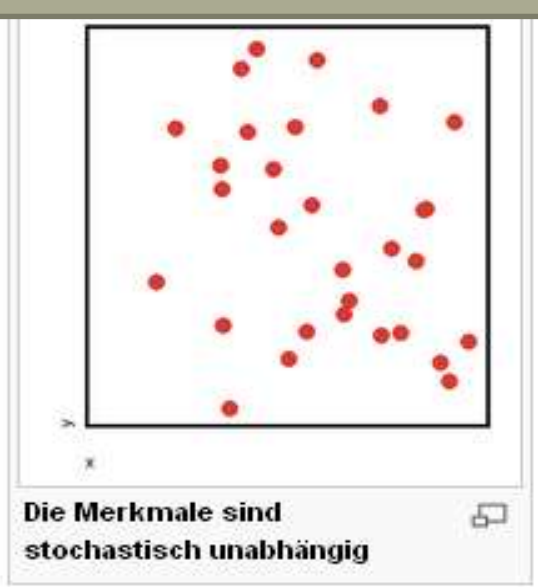


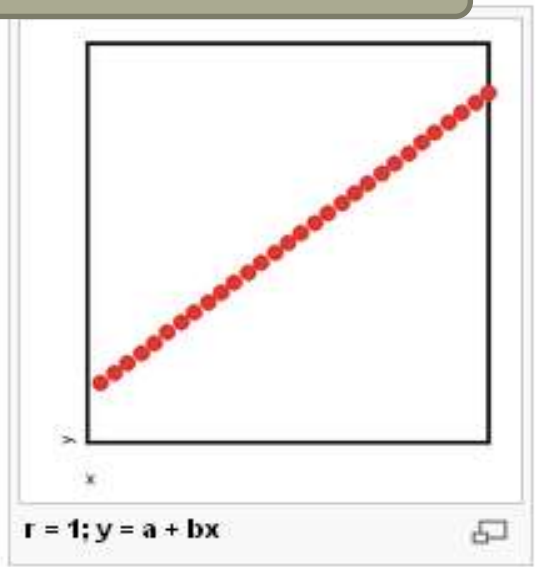




Quelle: WikiBooks / User: Philipendula / Lizenz: GNU-Lizenz für freie Dokumentationen







Empfohlene Hilfstabelle für die Berechnung

Nr.	x	y	x^2	y^2	(x*y)
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
Σ	…	…	…	…	…

$$r = \frac{\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^{-2}} * \sqrt{\sum_{i=1}^{n}(y_i^2) - n * \bar{y}^{-2}}}$$

Übung: B-P-K

Befragte/r	Größe (m)	Gewicht (kg)
1	1,55	64
2	1,68	72
3	1,72	71
4	1,73	75
5	1,82	102
6	1,81	98
7	1,66	71
8	1,78	78
9	1,73	77
10	1,59	69

Übung: B-P-K

Welche Größen müssen wir ermitteln?

$$\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y} = ?$$

$$\sqrt{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^{-2}} = ?$$

$$\sqrt{\sum_{i=1}^{n}(y_i^2) - n * \bar{y}^{-2}} = ?$$

Welche Hilfsgrößen benötigen wir?

$$\bar{x} = 1{,}707$$

$$\bar{y} = 77{,}7$$

$$n = 10$$

$$r = \frac{\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^{-2}} * \sqrt{\sum_{i=1}^{n}(y_i^2) - n * \bar{y}^{-2}}}$$

Übung: B-P-K

Nr.	x	y	x^2	y^2	(x*y)
1	1,55	64	2,4025	4096	99,2
2	1,68	72	2,8224	5184	120,96
3	1,72	71	2,9584	5041	122,12
4	1,73	75	2,9929	5625	129,75
5	1,82	102	3,3124	10404	185,64
6	1,81	98	3,2761	9604	177,38
7	1,66	71	2,7556	5041	117,86
8	1,78	78	3,1684	6084	138,84
9	1,73	77	2,9929	5929	133,21
10	1,59	69	2,5281	4761	109,71
Σ	17,07	777	29,2097	61769	1334,67

Übung: B-P-K

$$\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y} = ?$$

$$(1{,}55 * 64) = 99{,}2$$

$$(1{,}68 * 72) = 120{,}96$$

...

$$(1{,}73 * 77) = 133{,}21$$

$$(1{,}59 * 69) = 109{,}71$$

$$\sum = 1334{,}67$$

$$\bar{x} = 1{,}707$$

$$\bar{y} = 77{,}7$$

$$n = 10$$

$$1334{,}67 - (10 * 1{,}707 * 77{,}7)$$

$$= 8{,}331$$

Übung: B-P-K

$$\sqrt{\sum_{i=1}^{n}(x_i^2) - n * \overline{x}^2} = ?$$

$$\sqrt{\sum_{i=1}^{n}(y_i^2) - n * \overline{y}^2} = ?$$

$$\overline{x} = 1{,}707$$

$$\overline{y} = 77{,}7$$

$$n = 10$$

$$\sum_{i=1}^{n} x_i^2 = 29{,}2097$$

$$\sum_{i=1}^{n} y_i^2 = 61769$$

$$\sqrt{\sum_{i=1}^{n}(x_i^2) - n * \overline{x}^2} = \sqrt{29{,}2097 - 10 * 1{,}707^2} = 0{,}2667$$

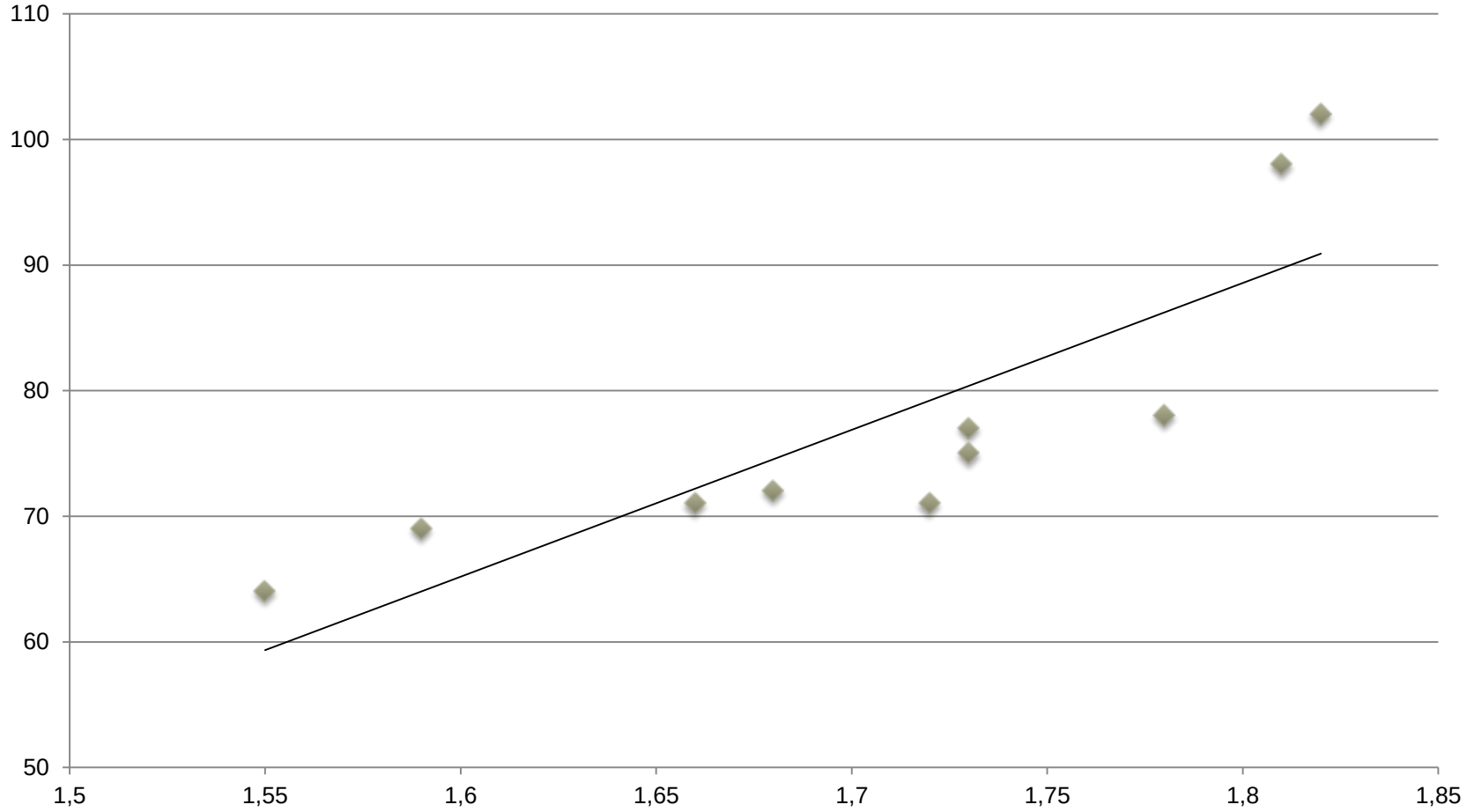
$$\sqrt{\sum_{i=1}^{n}(y_i^2) - n * \overline{y}^2} = \sqrt{61769 - 10 * 77{,}7^2} = 37{,}3644$$

Übung: B-P-K

$$r = \frac{\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^2} * \sqrt{\sum_{i=1}^{n}(y_i^2) - n * \bar{y}^2}}$$

$$r = \frac{8{,}331}{0{,}2667 * 37{,}3644} = \frac{8{,}331}{9{,}9651} = 0{,}8360$$

Übung: B-P-K



Spearman-Rangkorrelationskoeffizient

- Für **ordinalskalierte Merkmale** bieten sich zwei Zusammenhangsmaße an:
  - Der Rangkorrelationskoeffizient nach Spearman
  - Der Konkordanzkoeffizient nach Kendall
- Der Rangkorrelationskoeffizient nach Spearman misst den **monotonen Zusammenhang zweier Variablen**

$$rho = 1 - \frac{6 * \sum d_i^2}{(n^2 - 1) * n}$$

  - Für die n Datenpaare werden dabei innerhalb jeder Variablen zunächst Ränge gebildet
  - Die kleinste Ausprägung von X erhält den Wert 1, die zweitkleinste den Wert 2 etc. pp.
  - Für Y wird identisch vorgegangen, auch hier erhält die kleinste Ausprägung die 1 etc.
  - Anschließend werden die Rangdifferenzen d der jeweiligen Datenpaare gebildet
  - Auf Basis dieser Differenzwerte lässt sich dann der Rangkorrelationskoeffizient (nach obenstehender Formel) berechnen

Spearman-Rangkorrelationskoeffizient

- Die Ergebnisse liegen stets zwischen -1 und +1

  - rho > 0 = **gleichsinniger monotoner Zusammenhang**
    (große X-Werte gehen mit großen Y-Werten einher und umgekehrt)
  - rho ~ 0 = es besteht **kein monotoner Zusammenhang**
    (damit kann auch kein linearer bestehen!)
  - rho < 0 = **gegensinniger monotoner Zusammenhang**
    (große X-Werte gehen mit kleinen Y-Werten einher und umgekehrt)

- Wichtig: Das Verfahren liefert nur dann genaue Resultate, wenn **keine Rangplatzbindungen** (die sogenannten ties) auftreten

- Haben Beobachtungen identische Werte, ordnet man allen identischen Daten einen Durchschnittsrang zu

Übung: Spearman

Befragte/r	Größe (m)	Gewicht (kg)
1	1,55	64
2	1,68	72
3	1,72	71
4	1,73	75
5	1,82	102
6	1,81	98
7	1,66	71
8	1,78	78
9	1,73	77
10	1,59	69

Übung: Spearman

Nr.	x	rg (x)	y	rg (y)	d	d^2
1	1,55	1	64	1	0	0
2	1,68	4	72	5	-1	1
3	1,72	5	71	3,5	1,5	2,25
4	1,73	6,5	75	6	0,5	0,25
5	1,82	10	102	10	0	0
6	1,81	9	98	9	0	0
7	1,66	3	71	3,5	-0,5	0,25
8	1,78	8	78	8	0	0
9	1,73	6,5	77	7	-0,5	0,25
10	1,59	2	69	2	0	0
Σ	//	//	//	//	//	4

Übung: Spearman

$$rho = 1 - \frac{6 * \sum d_i^2}{(n^2 - 1) * n}$$

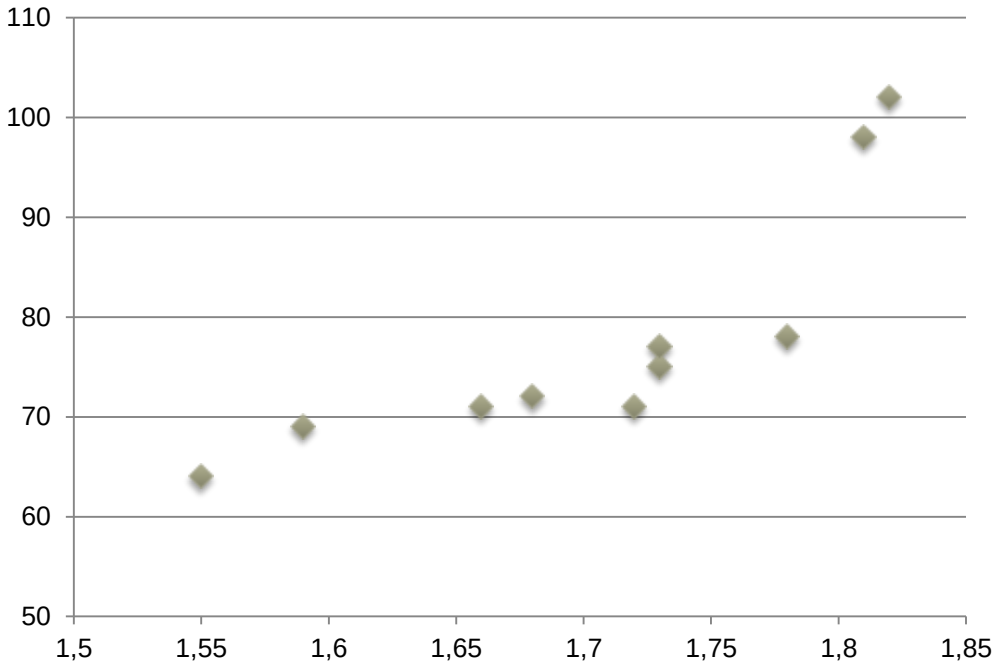
$$rho = 1 - \frac{6 * 4}{(10^2 - 1) * 10}$$

$$rho = 1 - \frac{24}{990}$$

$$rho = 1 - 0{,}024$$

$$rho = 0{,}976$$

Passt das Ergebnis
zum Streudiagramm?



Konkordanzkoeffizient nach Kendall

- Alternativ zu Spearman kann für Ordinaldaten auch Kendalls tau berechnet werden
- Die Berechnung benötigt die Anzahl konkordanter (K) und diskordanter (D) Paare
  - Zur Bestimmung der Paare wird eine der Datenreihen nach der Größe geordnet
  - Anschließend wird untersucht, inwieweit sich die zweite Datenreihe „mitsortiert“ hat
- Für jedes Datenpaar aus den beiden Datenreihen (yi, yj) mit i < j gilt:
  - ist yi < yj, so ist das Paar konkordant (K)
  - ist yi > yj, so ist das Paar diskordant (D)
  - ist yi = yj, so liegt eine Bindung vor (wird nicht mitgezählt)

$$tau = \frac{2*(K-D)}{n*(n-1)}$$

- Sind alle Paare entsprechend untersucht worden, wird tau (Formel) berechnet
  - Auch hier gilt, dass das Ergebnis nur Bestand hat, **wenn keine Bindungen auftreten**
  - Einige wenige Bindungen können ignoriert werden, da sie das Ergebnis kaum verzerren

Konkordanzkoeffizient nach Kendall

- Die Ergebnisse liegen stets zwischen -1 und +1
  - tau > 0 = **gleichsinniger monotoner Zusammenhang**
    (große X-Werte gehen mit großen Y-Werten einher und umgekehrt)
  - tau ~ 0 = es besteht **kein monotoner Zusammenhang**
    (damit kann auch kein linearer bestehen!)
  - tau < 0 = **gegenseitiger monotoner Zusammenhang**
    (große X-Werte gehen mit kleinen Y-Werten einher und umgekehrt)
- Bei der Interpretation von Korrelationskoeffizienten ist zu beachten:
  - Sowohl mit Spearman als auch mit Kendall können nur monotone Zusammenhänge identifiziert werden, mit dem B-P-K nur lineare
  - Ein niedriger Korrelationskoeffizient bedeutet daher nicht, dass keine andere Korrelation (z.B. eine logarithmische) in den Daten zu finden ist

Übung: Kendall

Befragte/r	Größe (m)	Gewicht (kg)
1	1,55	64
2	1,68	72
3	1,72	71
4	1,73	75
5	1,82	102
6	1,81	98
7	1,66	71
8	1,78	78
9	1,73	77
10	1,59	69

Übung: Kendall

Nr.	x	rg (x)	y	rg (y)	K	D
1	1,55	1	64	1	9	0
2	1,59	2	69	2	8	0
3	1,66	3	71	3,5	6	0
4	1,68	4	72	5	5	1
5	1,72	5	71	3,5	5	0
6	1,73	6,5	75	6	4	0
7	1,73	6,5	77	7	3	0
8	1,78	8	78	8	2	0
9	1,81	9	98	9	1	0
10	1,82	10	102	10	-	-
Σ	//	//	//	//	43	1

9 x K
0 x D

Übung: Kendall

$$K = 43$$

$$D = 1$$

$$n = 10$$

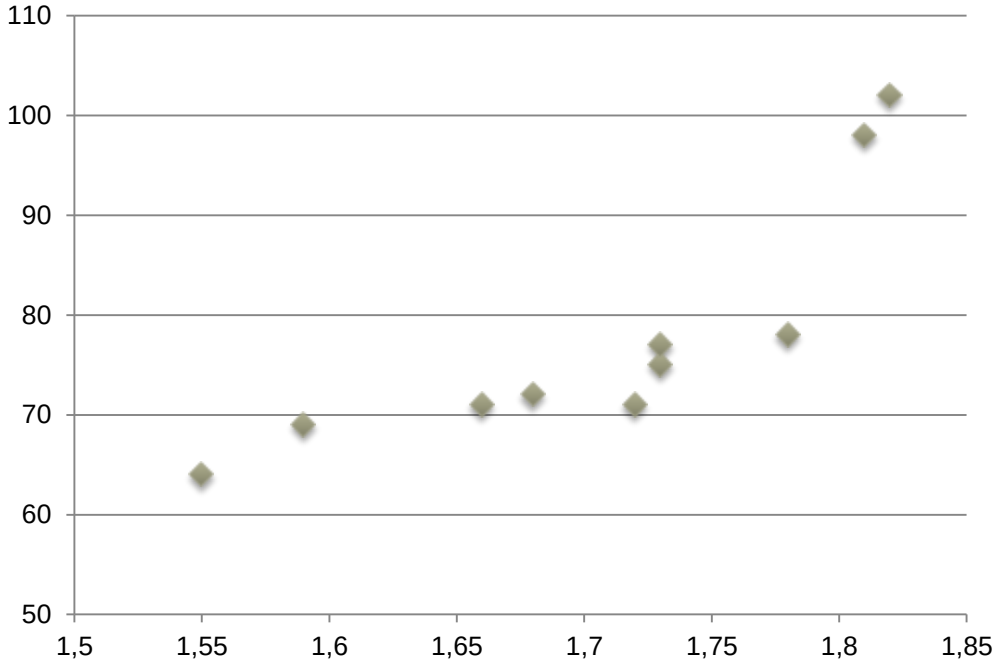
Passt das Ergebnis zum Streudiagramm?

$$tau = \frac{2*(K-D)}{n*(n-1)}$$

$$tau = \frac{2*(43-1)}{10*(10-1)}$$

$$tau = \frac{84}{90}$$

$$tau = 0{,}9\overline{3}$$



Korrelation ist nicht gleich Kausalität

- Eine über einen Korrelationskoeffizienten identifizierte Korrelation sollte näher untersucht, dabei jedoch **niemals inhaltlich interpretiert werden**
- Grund dafür ist, dass eine Korrelation nicht notwendigerweise auf einem Ursache-Wirkungs-Zusammenhang beruht – auch wenn es in vielen Fällen leider äußerst verführerisch ist, diese Annahme zu treffen

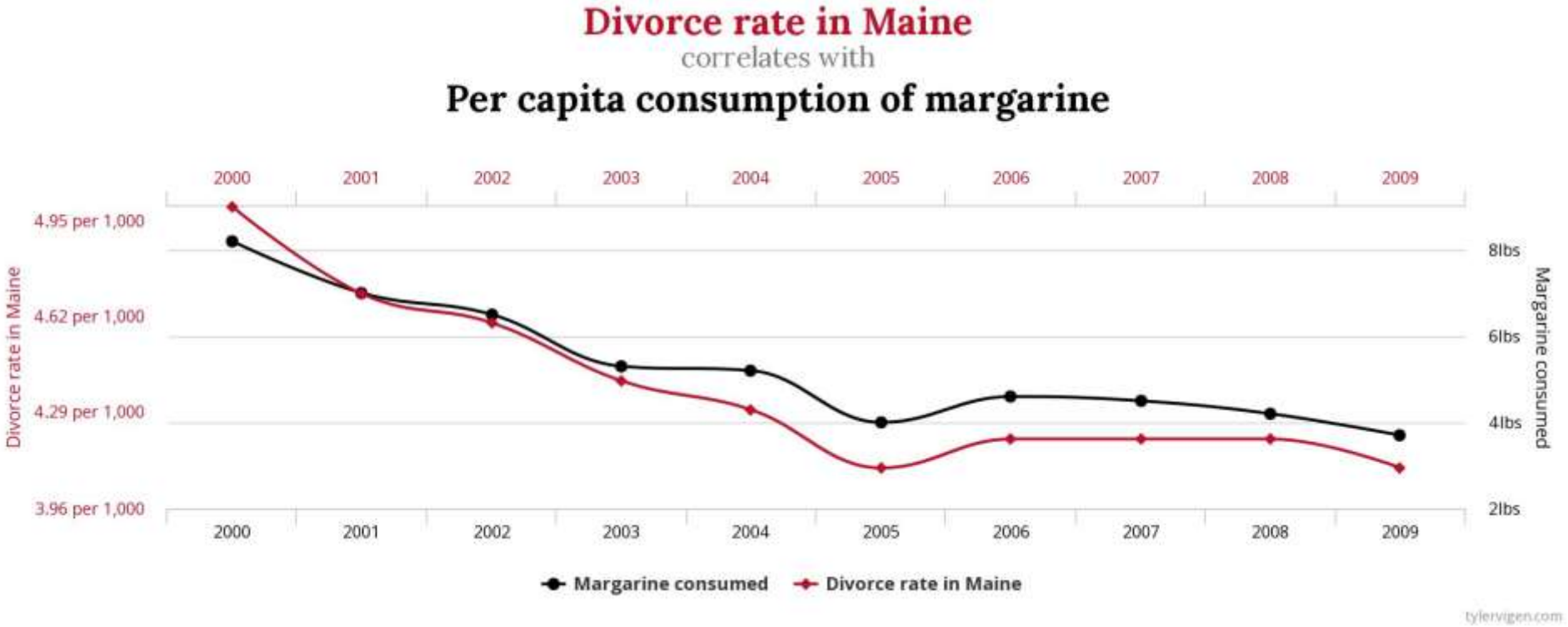
- Tatsächlich kann es verschiedene Erklärungen für Korrelationen geben
  - Einseitiger Zusammenhang: X beeinflusst Y bzw. Y beeinflusst X
  - Beidseitiger Zusammenhang: X und Y beeinflussen sich gegenseitig
  - Es handelt sich um einen reinen Zufallseffekt in den Daten (Scheinkorrelation)
  - Eine dritte Variable (Z) beeinflusst X und Y gleichermaßen (Scheinkorrelation)
- Ein klassisches Beispiel für eine Scheinkorrelation ist die Korrelation zwischen Storchenzahl und Geburtenquote (verbunden über die Variable „Urbanisierung“)

"One of the first things taught in introductory statistics textbooks is that correlation is not causation. It is also one of the first things forgotten."

Thomas Sowell

Korrelation ist nicht gleich Kausalität

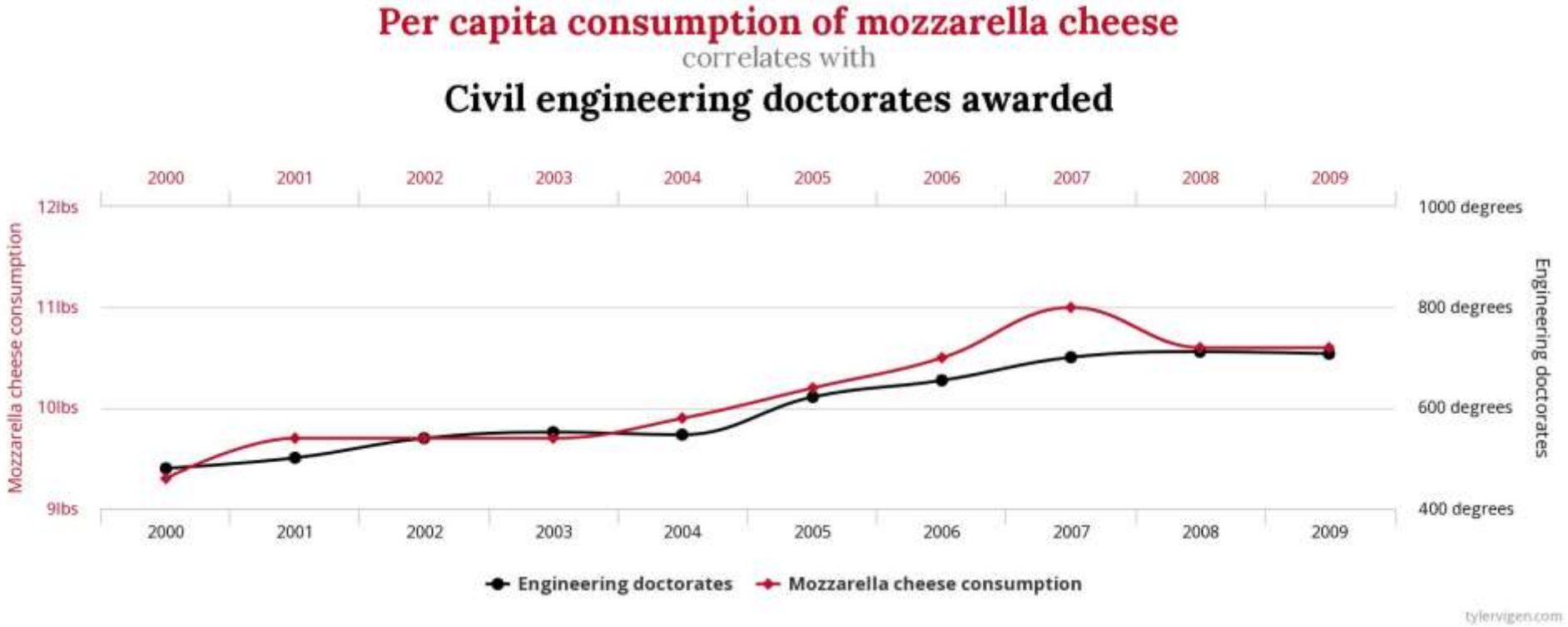
Scheidungsrate vs. Magarinekonsum



http://www.tylervigen.com/spurious-correlations

Korrelation ist nicht gleich Kausalität

Mozzarellakonsum vs. Doktorarbeiten



http://www.tylervigen.com/spurious-correlations

Teil V

Explorative Statistik

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

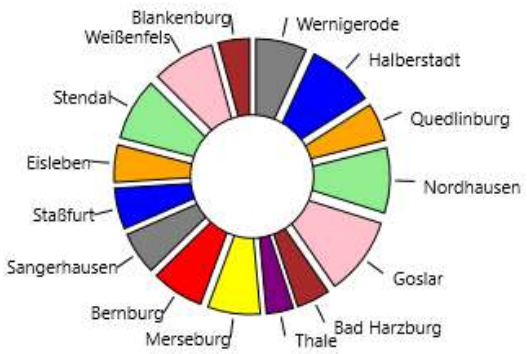
Explorative Statistik

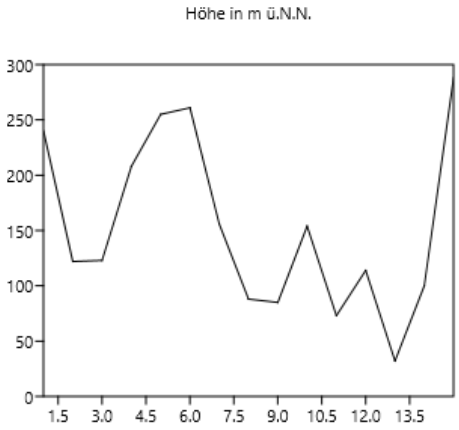
Ausgewählte grafische Darstellungsformen

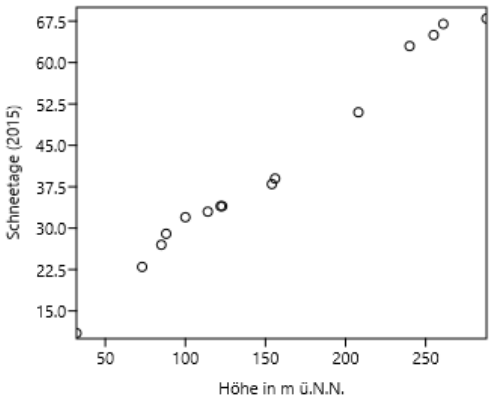
▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

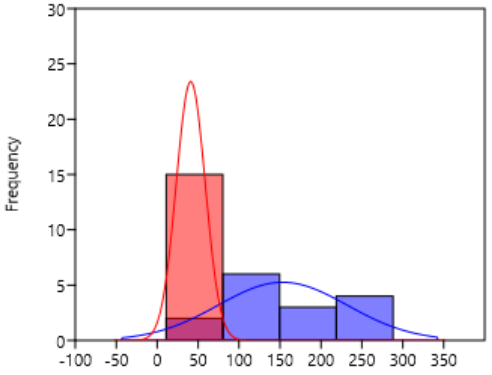
Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

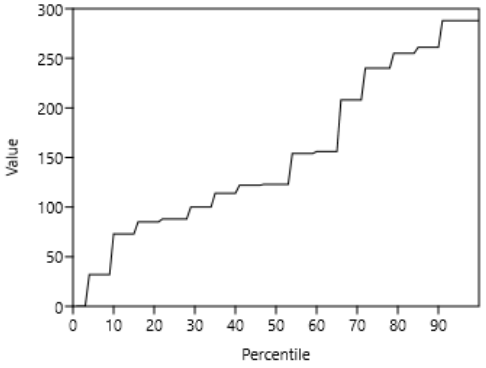
Die große Vielfalt statistischer Grafiken…

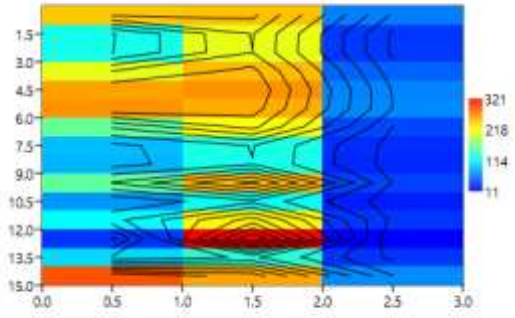




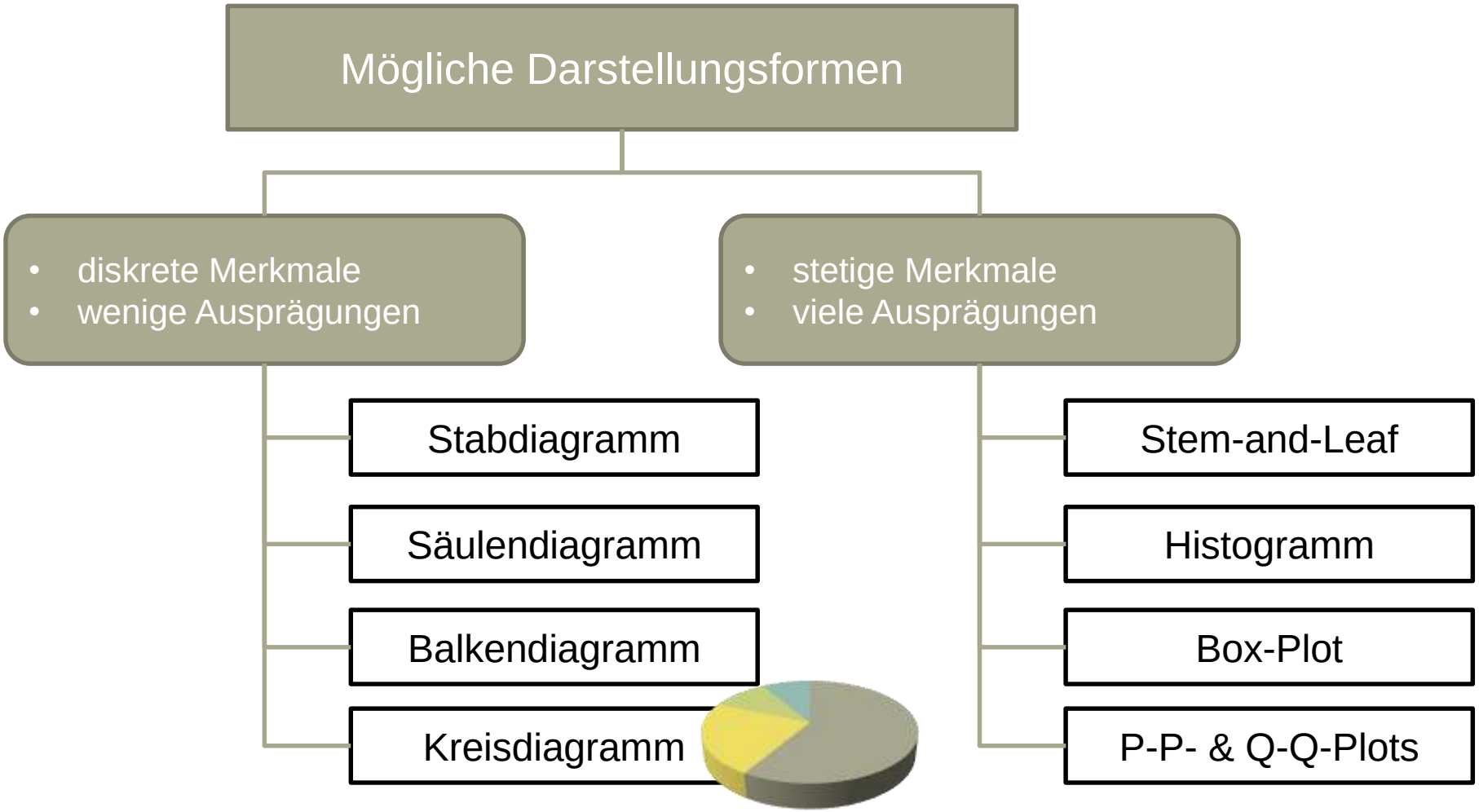




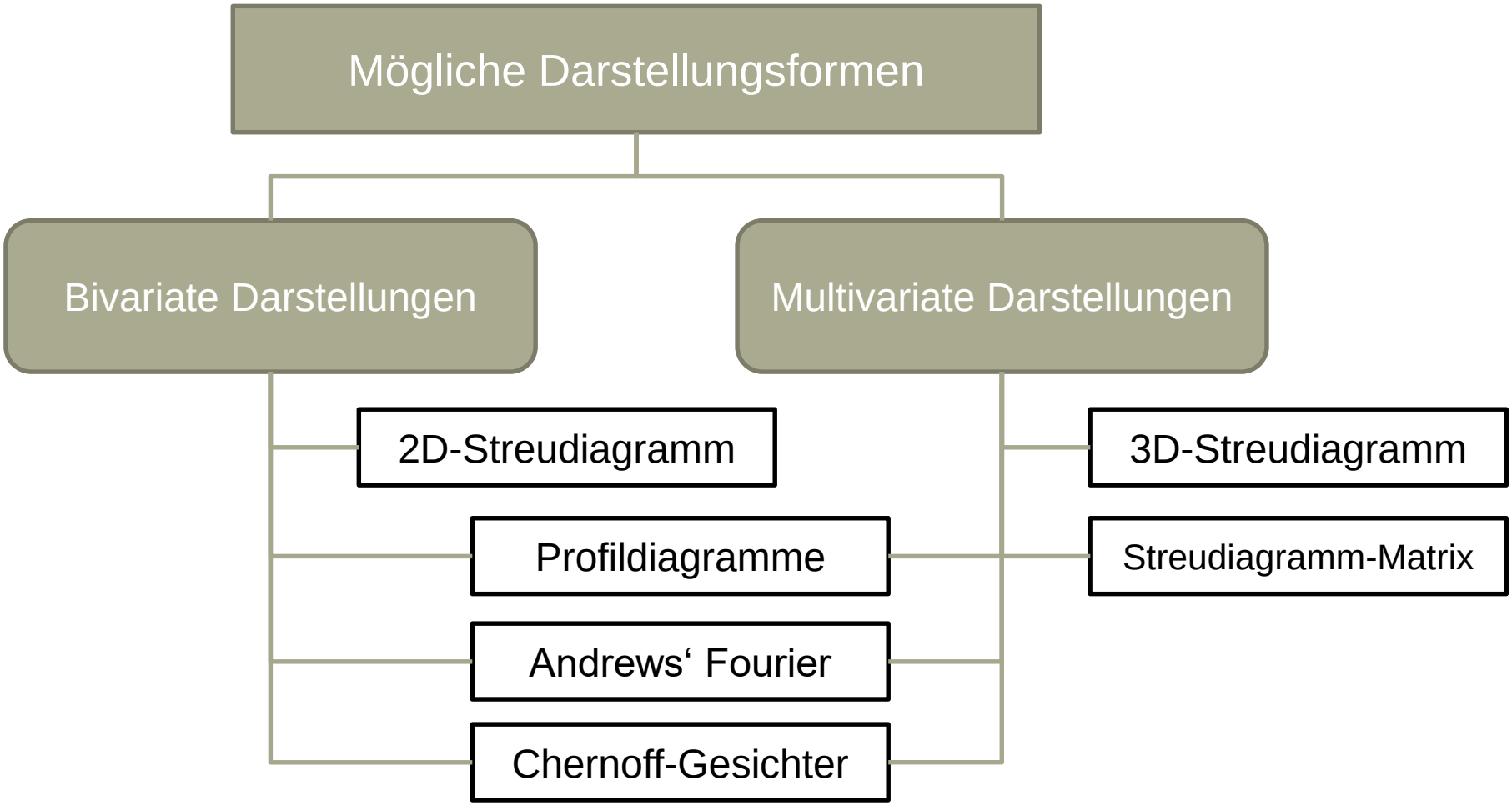




Grafische Darstellung univariater Daten



Grafische Darstellung multivariater Daten



Stem-and-Leaf-Plots

– Die **Stem-and-Leaf-Plots** (Stamm-Blatt-Diagramme) eignen sich im Gegensatz zu Kreis- und Balkendiagrammen vor allem zur Darstellung stetiger Merkmale

– Der große **Vorteil gegenüber jeder anderen grafischen Darstellungsform** ist, dass die Originaldaten (zumindest bis zu einer gewissen Genauigkeit) noch aus dem Diagramm abgelesen werden können

– Das Diagramm ist aufgebaut wie ein gekipptes Histogramm, d.h. flächenproportional
– Der „Stamm“ besteht aus der ersten Ziffer, die „Blätter“ aus der jeweils folgenden
– Sehr große oder sehr kleine Zahlen (Ausreißer) können auf- oder abgerundet sowie als Extremwerte ausgewiesen oder aus der Grafik gestrichen werden

– Stem-and-Leaf-Plots können – neben den Box-Plots – bemerkenswert gut dazu genutzt werden, um zwei **Verteilungen miteinander zu vergleichen**

Stem-and-Leaf-Plots

```
1          |          1 1 1 2 2 3 4 5 7 7
2          |          2 2 4
3          |          3 3 3 4 5 8 8
4          |          1 2 9 9 9 9

2 Extremes

Stem width:          10
Each leaf:           1 case(s)
```

Singulärer Stem-and-Leaf-Plot

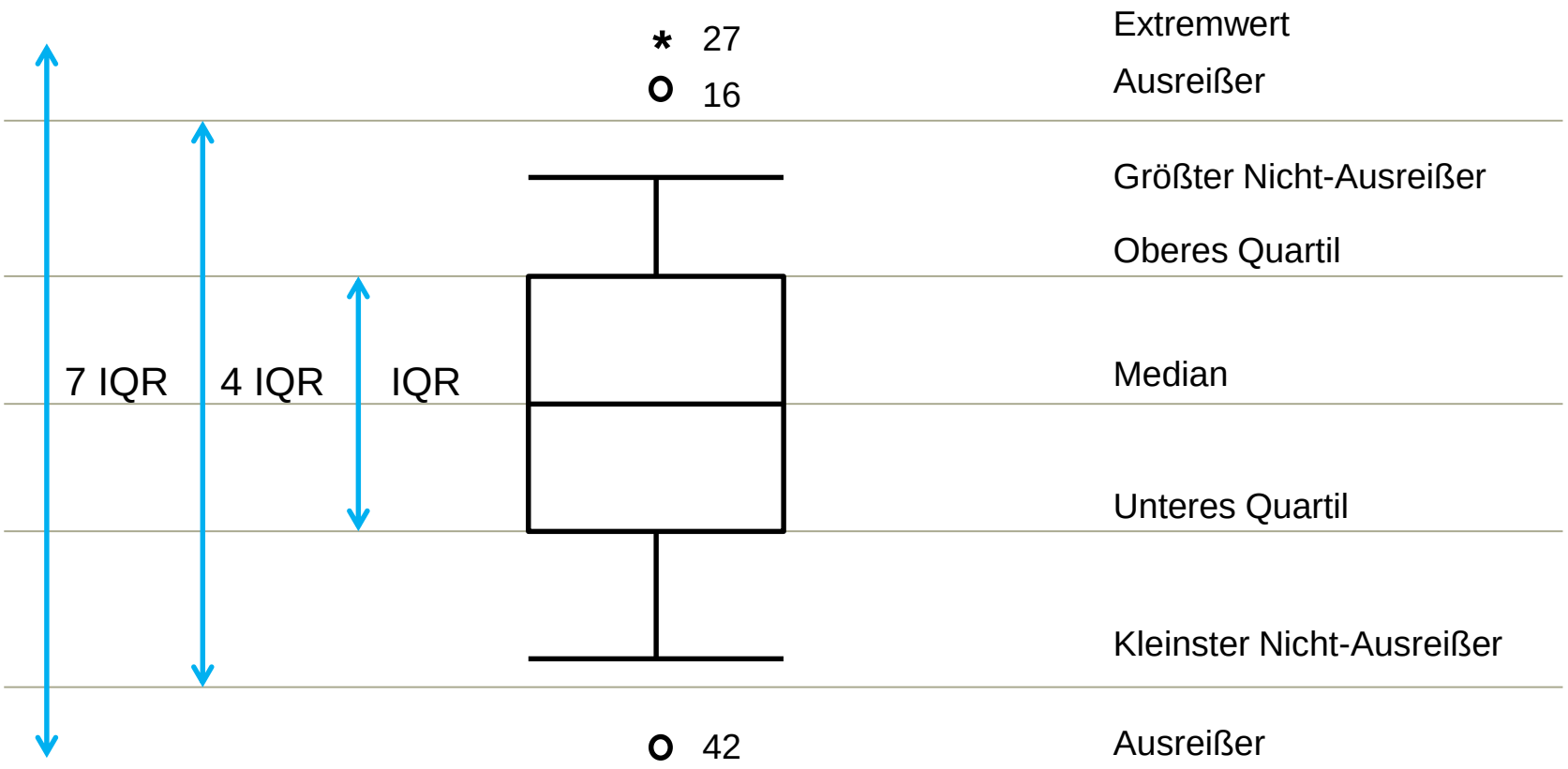
Vergleichender Stem-and-Leaf-Plot

Datensatz A		Datensatz B
8 8 8 3 2	1	1 1 1 2 2 3 4 5 7 7
2 1	2	2 2 4
9 5 4 43 3	3	3 3 3 4 5 8 8
4 3 32 1	4	1 2 9 9 9 9
3 Extremes		2 Extremes

Stem width: 10
Each leaf: 1 case(s)

Box-Plots

- Box-Plots bieten einen Verteilungsüberblick und gestatten Verteilungsvergleiche
- Sie stellen Lage und Streuung dar und dienen zudem der Ausreißeridentifikation



Box-Plots

– Aus der Lage des Medians im Box-Plot lässt sich die Form einer Verteilung ablesen

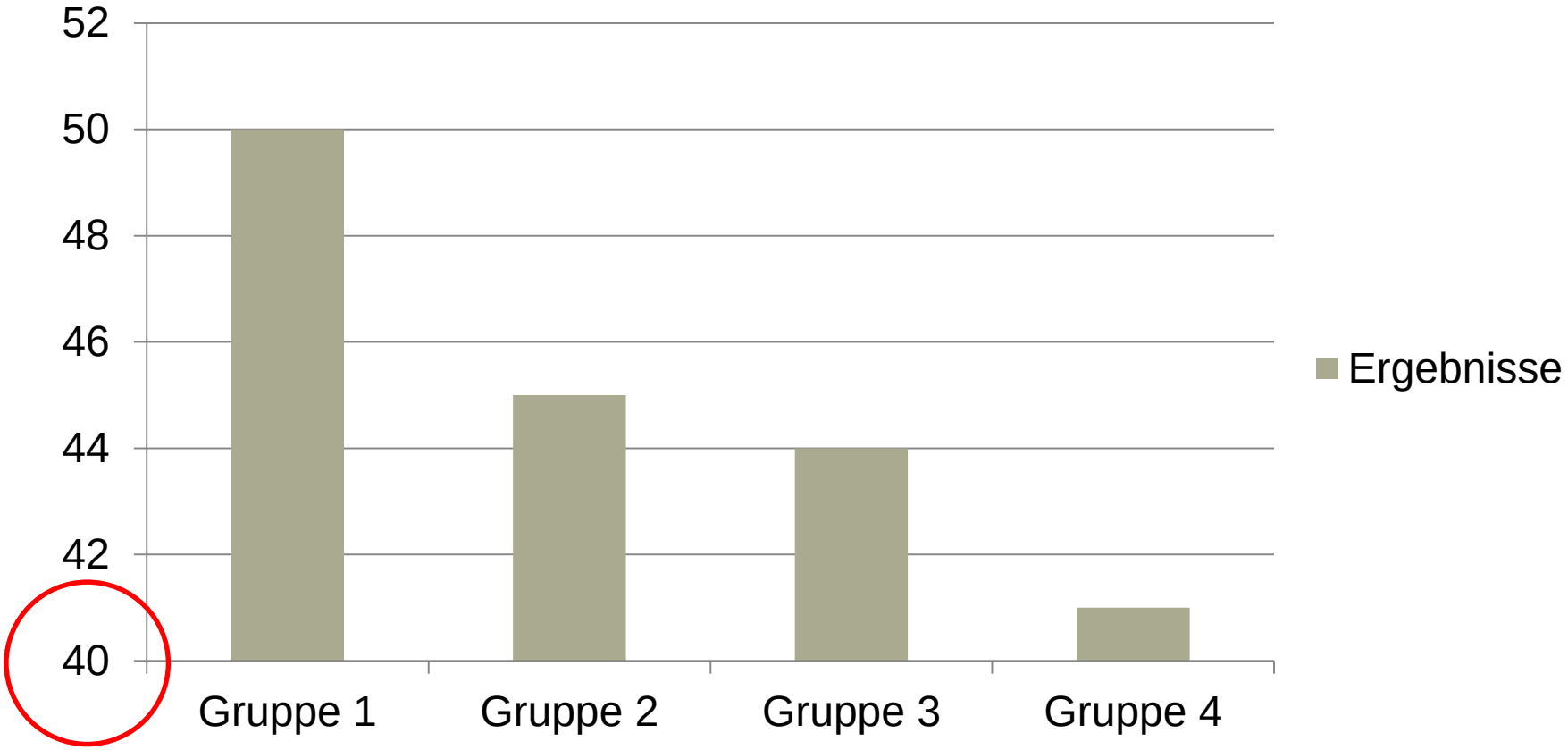
Symmetrische Verteilung

Linkssteile Verteilung

Rechtssteile Verteilung

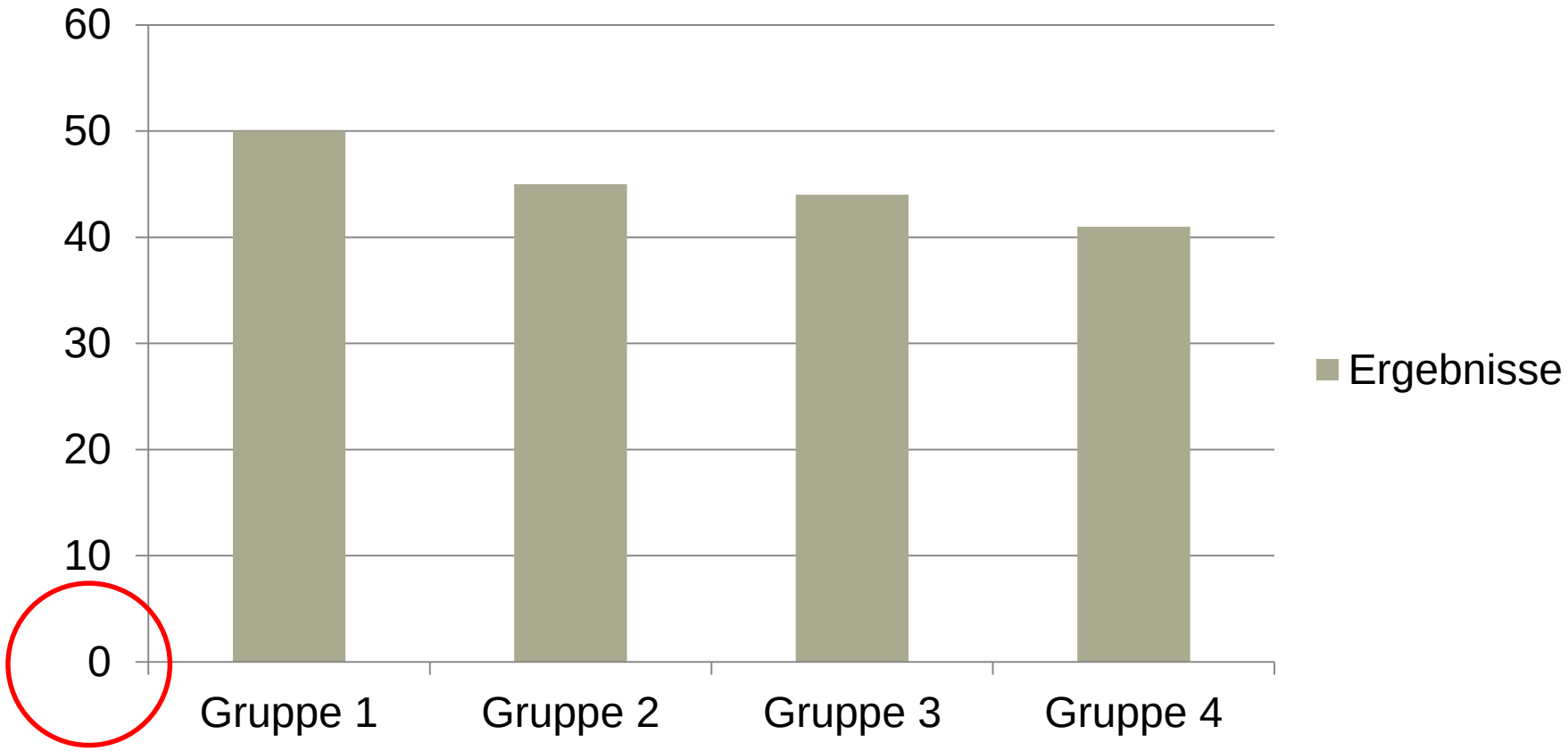
Wie leicht sind statistische Diagramme manipulierbar?

Trick 17: Die leicht übersehene Achsenverkürzung



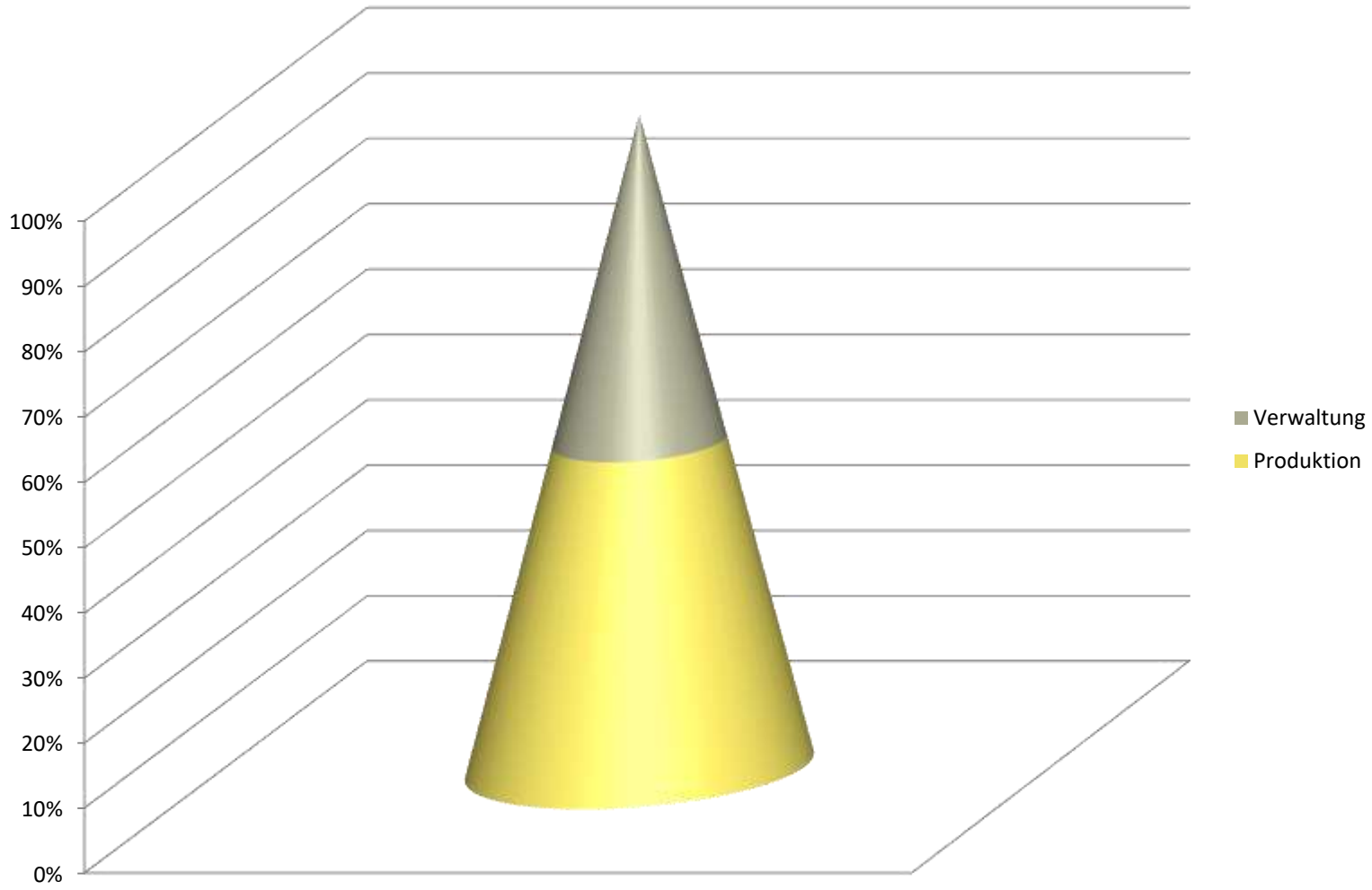
Wie leicht sind statistische Diagramme manipulierbar?

...und schon sind die Unterschiede viel geringer...

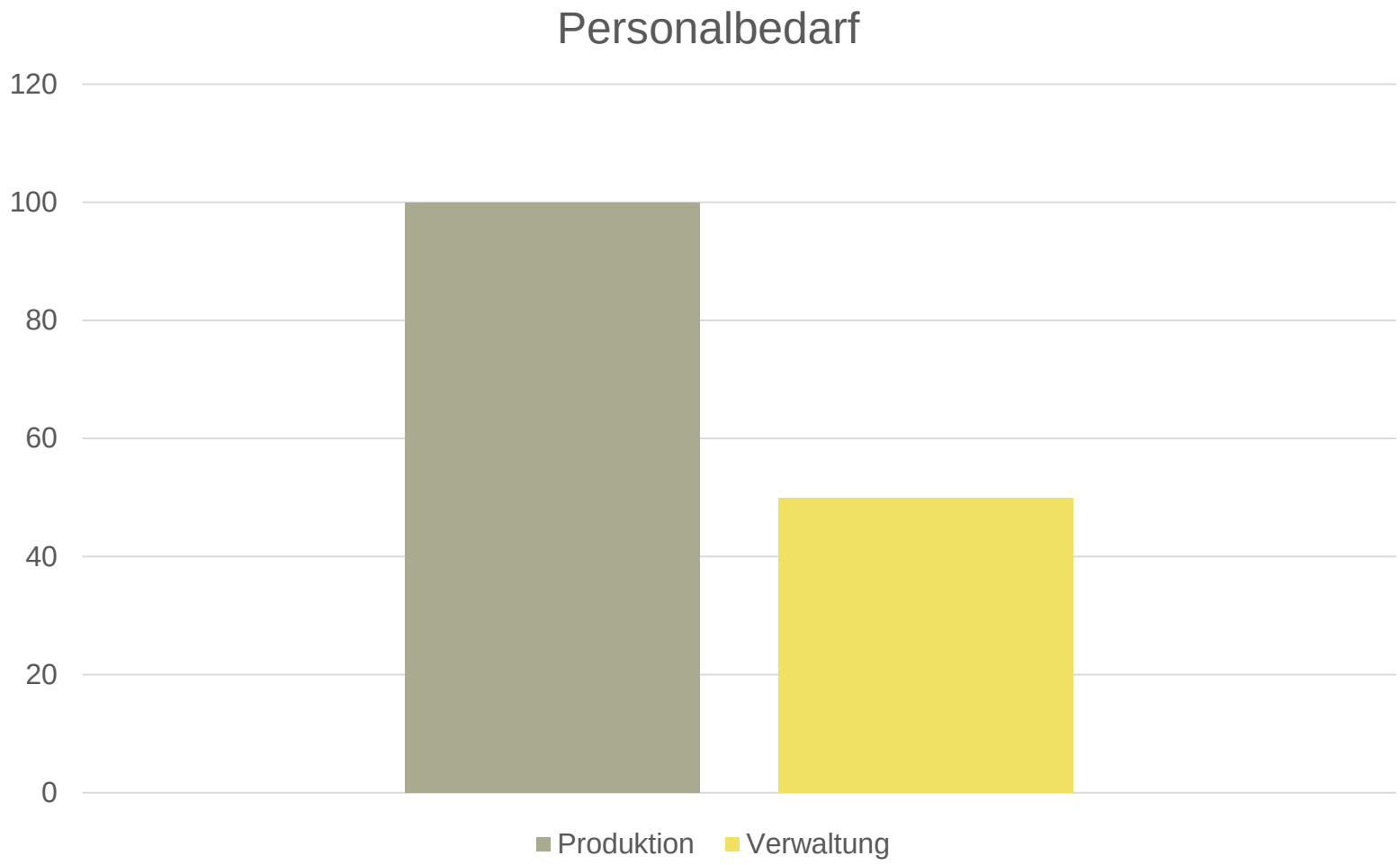


Wie 3D-Kegel die Realität verzerren

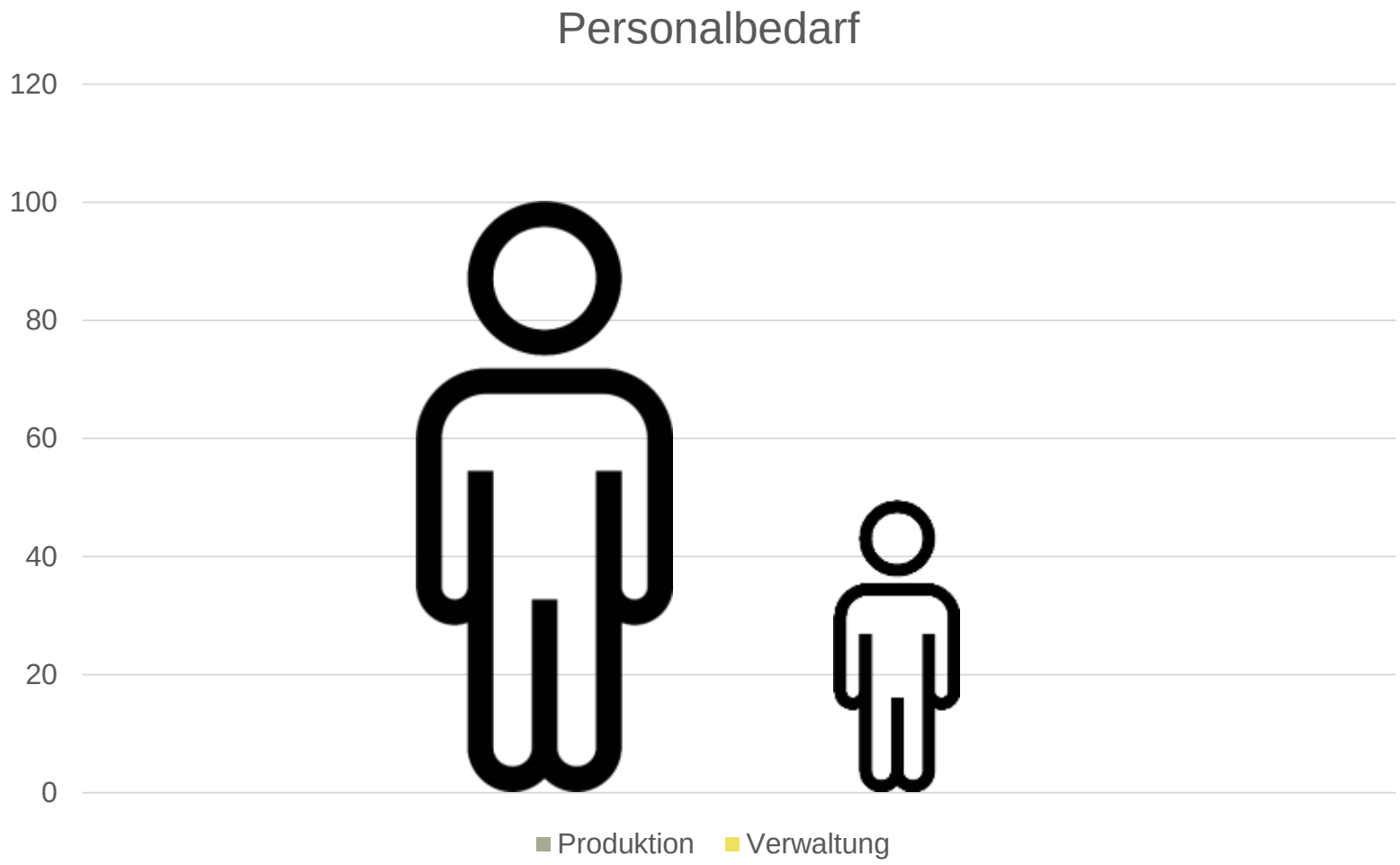
Wie 3D-Kegel die Realität verzerren



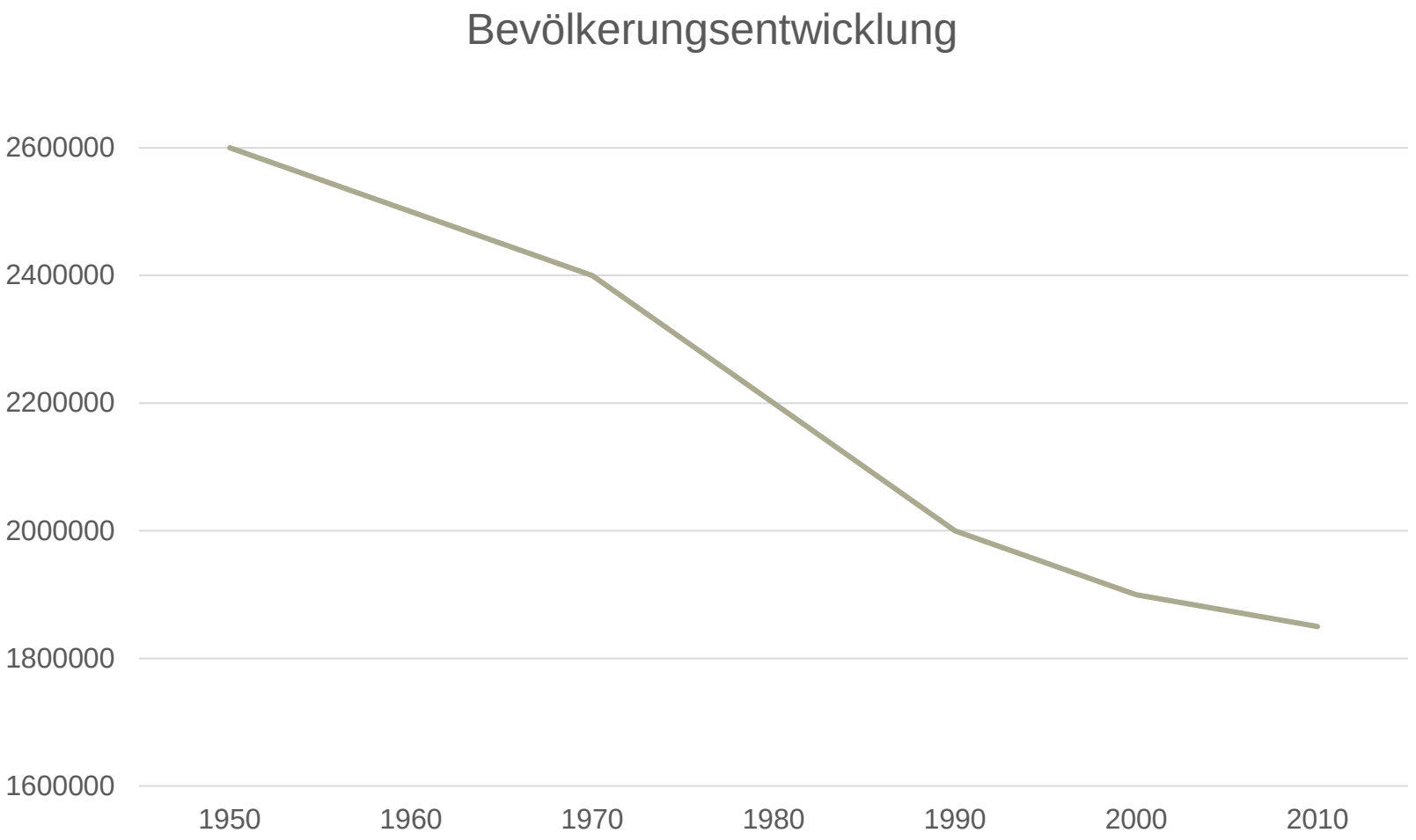
Warum man keine Icons verwenden sollte



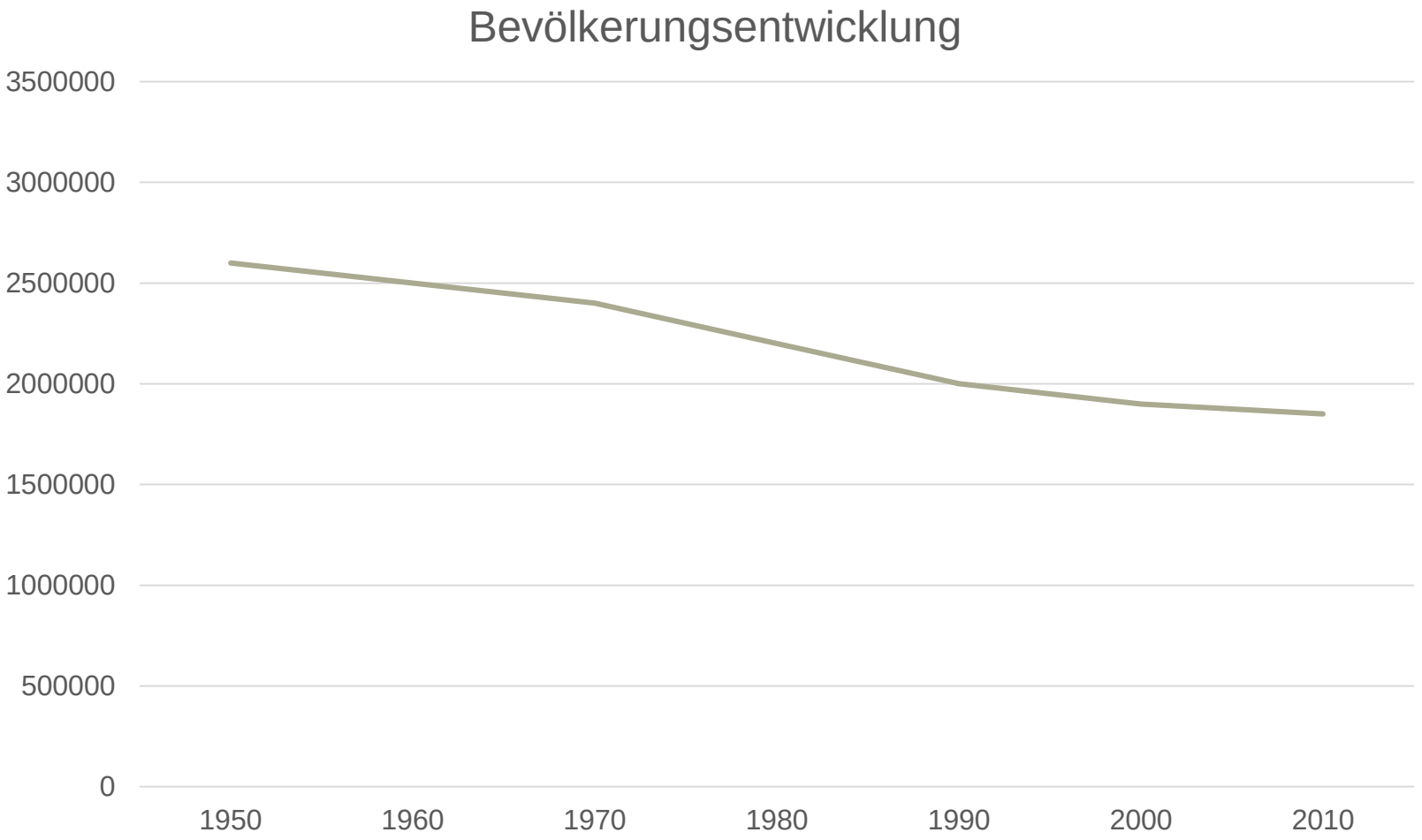
Warum man keine Icons verwenden sollte



Manchmal täuschen auch unverkürzte Achsen



Manchmal täuschen auch unverkürzte Achsen



Explorative Statistik

Umgang mit Ausreißern

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Einführung in die Ausreißeranalyse

- Bei einem Ausreißer handelt es sich um einen gemessenen oder erhobenen Wert, der nicht den Erwartungen entspricht bzw. der nicht zu den übrigen Werten passt
- Es existiert **keine eindeutige Definition** darüber, wann ein Wert als Ausreißer bezeichnet werden kann – beim Box-Plot werden z.B. alle Werte außerhalb des vierfachen IQR-Bereichs um den Median als Ausreißer klassifiziert

- Es gibt **drei mögliche Ursachen** für das Auftreten eines Ausreißers:

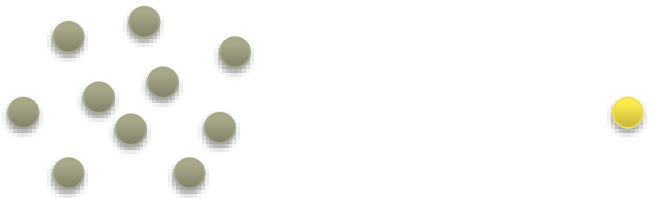
  - Der Ausreißer wurde durch einen **verfahrenstechnischen Fehler verursacht**, so etwa einen Fehler bei der Dateneingabe, beim Codieren der Daten oder einen technischen Ausfall bei der Datenerfassung bzw. -speicherung
  - Der Ausreißer kennzeichnet einen **außergewöhnlichen Wert**, etwa eine einzelne aus dem Rahmen fallende Beobachtung (der einzige Millionär), die sich jedoch erklären lässt – solche Ausreißer können mitunter ein Hinweis darauf sein, dass die Befragung falsch angelegt wurde
  - Der Ausreißer kennzeichnet einen korrekt erfassten Wert, für den es **keinerlei Erklärung** gibt

Einführung in die Ausreißeranalyse

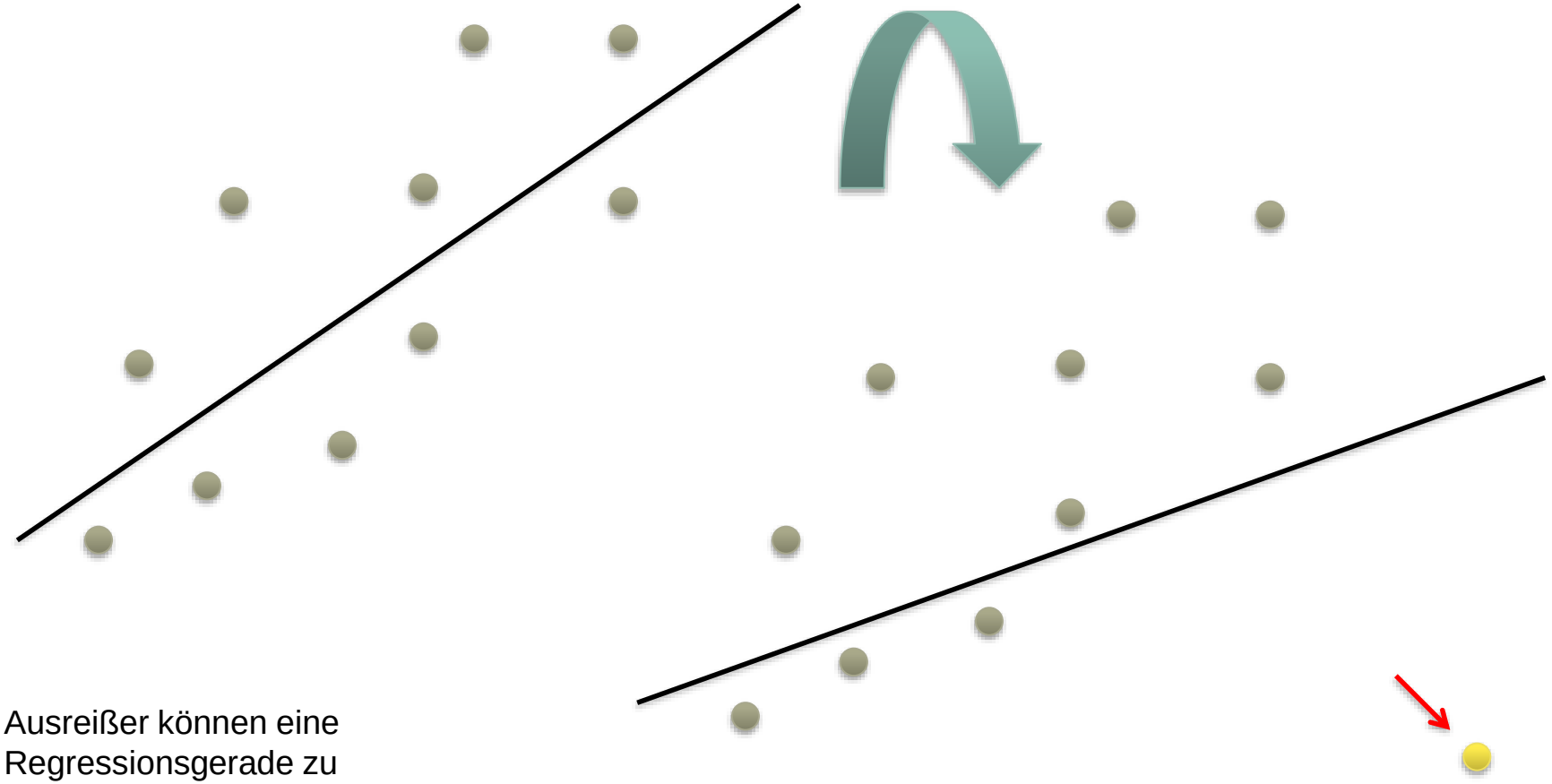
– Es ist zwischen normalen und multivariaten Ausreißern zu unterscheiden:

  – „Normaler“ Ausreißer = außergewöhnlich großer oder kleiner Wert (beispielsweise das persönliche Einkommen im Millionenbereich)
  – Multivariater Ausreißer = für sich betrachtet im normalen Bereich liegende Einzelwerte, die in ihrer Kombination quer durch die Variablen jedoch einen einzigartigen Fall ergeben (beispielsweise die 86jährige Frau mit Internetanschluss)

– Die entscheidende Frage jeder Ausreißeranalyse lautet: Werden die Ausreißer im Datensatz **beibehalten** oder können bzw. sollen sie **verworfen** werden?



Der Leverage-Effekt



Ausreißer können eine Regressionsgerade zu sich „hinziehen“ und so das Ergebnis stark beeinflussen

Wie ist mit Ausreißern umzugehen?

- Generell gibt es drei Möglichkeiten:
  - Eingang in die Analyse
  - Ausschluss aus der Analyse
  - Kennzeichnung als fehlender Wert

- Insbesondere drei Fragen sind für die Entscheidungsfindung von Bedeutung:
  - **Wie ist das Auftreten der Ausreißer zu erklären?**
  - Handelt es sich um Eingabefehler und ist es möglich, diese zu bereinigen?
  - Was sagen die Werte über Anlage und Durchführung der Erhebung aus?
  - **Welche Auswirkungen haben die Ausreißer auf die Ergebnisse der Datenanalyse?**
  - Beeinflussen sie beispielsweise den Verlauf einer Regressionsgraden? (Leverage-Effekt)
  - **Welcher Datenverlust entsteht, wenn die Ausreißer aus dem Datensatz entfernt werden?**

Explorative Statistik

Umgang mit fehlenden Werten

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Das Problem der fehlenden Daten

- Unter fehlenden Daten sind einzelne fehlende Werte zu verstehen

- Typische fehlende Werte bei Personenbefragungen:

  - Angaben zum Einkommen
  - Angaben zum eigenen Körper
  - Angaben zum Sexualverhalten

- Fehlende Werte sind dann ein Problem, wenn ein **Zusammenhang zwischen der Wahrscheinlichkeit des Fehlens und einem anderen Sachverhalt** zu vermuten ist, die Verteilung der fehlenden Werte also keine zufällige ist

  - Beispiel: Kommt es bei der Frage nach dem Einkommen tendenziell eher zu Auskunftsverweigerungen bei Personen mit niedrigem Einkommen, so wird dies das erhobene Durchschnittseinkommen nach oben verzerren

Das Problem der fehlenden Daten

- Bei der Untersuchung fehlender Daten ist daher vor allem zu klären:
  - Fehlen so viele Werte, dass eine sinnvolle Auswertung des Datensatzes unmöglich ist?
  - Sind die fehlenden Werte zufällig gestreut oder lässt sich ein Muster identifizieren?
- Generell bieten sich drei Möglichkeiten des Umgangs mit fehlenden Daten an:
  - Einzelne Fälle oder einzelne Variablen werden von der weiteren Auswertung ausgeschlossen
  - Es werden ausschließlich die vollständigen Fälle zur weiteren Auswertung zugelassen
  - Die fehlenden Werte werden induktiv oder statistisch ersetzt
- **Die richtige Entscheidung hängt von den Ursachen für das Fehlen der Werte ab**

Ursachen für fehlende Daten

- Das Fehlen von Daten kann auf vier Ursachen zurückgeführt werden:
  - Dateneingabefehler (z.B. Buchstaben in einem Zahlenfeld)
  - Codierungs- und Übertragungsfehler während der Eingabe oder der Speicherung von Daten
  - Ungenaue Datenfelder bei der Erhebung (z.B. „Studienrichtung“ bei einer Befragung von Nicht-Akademikern)
  - Aktionen des Befragten wie etwa das Vergessen von Angaben, widersinnige Angaben (höchster Schulabschluss ist die Mittlere Reife, trotzdem wurde eine Abiturnote eingetragen), Nichtauskunftsfähigkeit oder bewusste Entscheidung, eine Frage nicht zu beantworten (Einkommen, Körper, Sexualverhalten...)

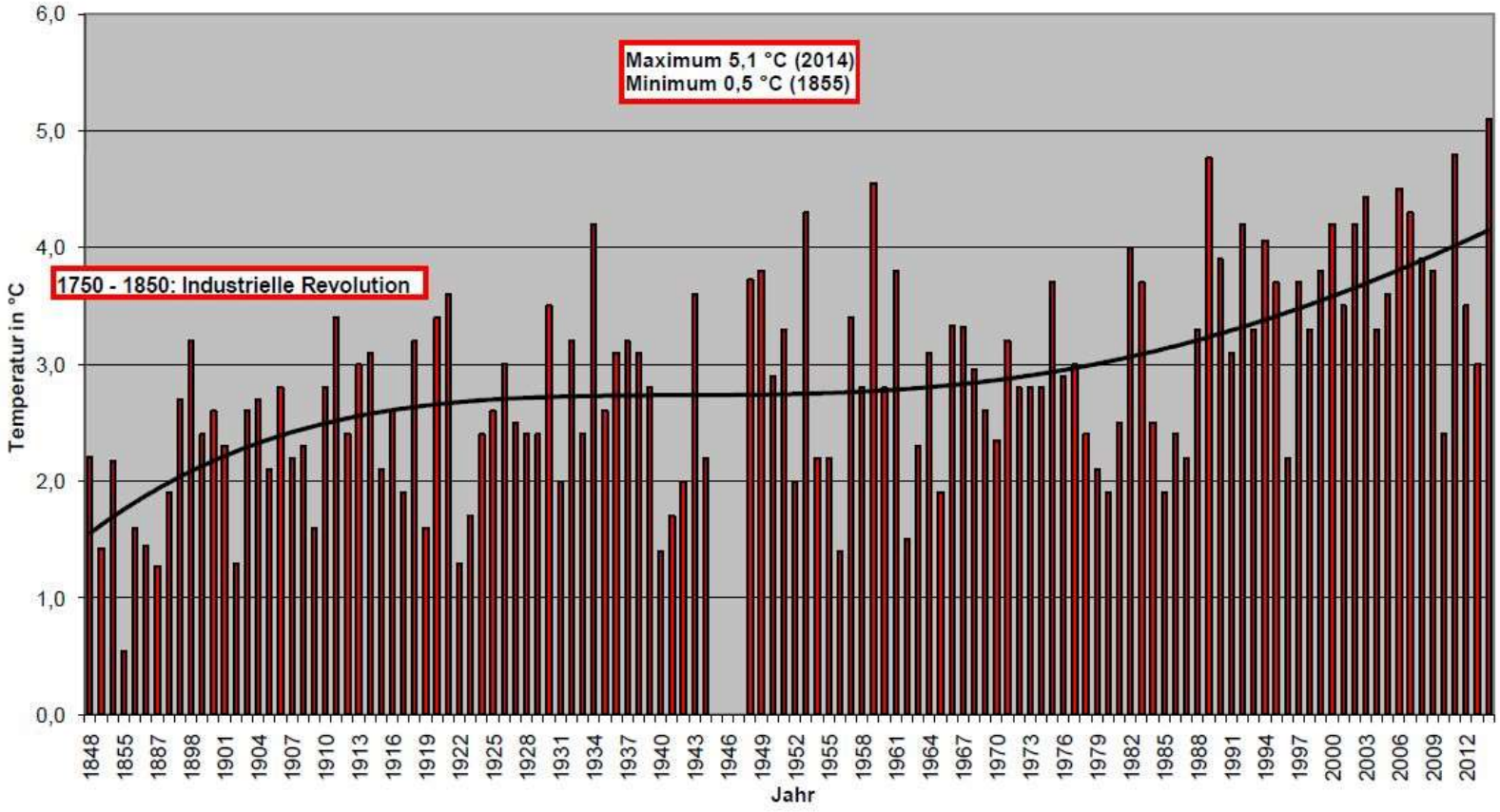
Ursachen für fehlende Daten

- Das Auftreten von fehlenden Werten ist bei der Arbeit mit realen Daten keinesfalls die Ausnahme, sondern vielmehr die Regel
- Die Wahrscheinlichkeit für das Auftreten fehlender Werte steigt erfahrungsgemäß mit der Größe des Datensatzes

- Bei der Analyse langer Zeitreihen, z.B. der Auswertung der Niederschlagsmengen der letzten 200 Jahre, werden aufgrund von Katastrophen, Krieg oder anderen Gründen immer wieder einzelne Werte nicht erfasst worden sein
- Gerade in der sozialwissenschaftlichen Forschung und bei der Marktforschung im Zuge der Befragung von hunderten oder tausenden Personen, kommt es aufgrund verschiedenster Ursachen häufig zu Einzelausfällen

- **Mit fehlenden Daten ist bei jeder marktforscherischen Untersuchung zu rechnen – ihr Auftreten sollte demzufolge keinesfalls ignoriert werden!**

Fehlende Werte in einer Zeitreihenbetrachtung



© Grafik: Nationalpark Harz, Daten: Deutscher Wetterdienst

Zufälligkeitsgrade

– Man unterscheidet in drei Zufälligkeitsgrade bezüglich des Auftretens fehlender Daten

MCAR,
MAR und
NRM

– Der Zufälligkeitsgrad entscheidet, wie mit fehlenden Werten umzugehen ist

– **MCAR = missing completely at random**

  – Fehlende Werte treten vollkommen zufällig auf
  – Die Wahrscheinlichkeit des Fehlens steht nicht in Zusammenhang mit anderen Größen
  – Es ist kein Zusammenhang zwischen dem Auftreten von fehlenden Werten der Variable Y und der Variable Y selbst (niedrige Einkommen werden tendenziell nicht angegeben) oder eine Korrelation mit einer anderen Variable X (Frauen sind tendenziell weniger bereit, Auskünfte über ihr Körpergewicht zu machen) feststellbar

Zufälligkeitsgrade

- **MAR = missing at random**

  - Das Auftreten von fehlenden Werten steht (teilweise) in Zusammenhang mit einer anderen erhobenen Variablen
  - Es ist kein Zusammenhang zwischen dem Auftreten von fehlenden Werten der Variable Y und der Variable Y selbst feststellbar, wohl aber eine (schwache) Korrelation des Auftretens von fehlenden Y-Werten mit einer anderen Variable X

- **NRM = nonrandom missing**

  - Das Auftreten von fehlenden Werten folgt klar erkennbaren Gesetzmäßigkeiten, eine Zufälligkeit ist sicher auszuschließen
  - Es kann entweder ein Zusammenhang zwischen dem Auftreten von fehlenden Werten der Variable Y und der Variable Y selbst oder mit einer anderen Variable X oder auch beides vorliegen, d.h. das Auftreten eines fehlenden Wertes kann vollständig durch eine andere Variable oder durch die Variable selbst erklärt werden

Umgang mit fehlenden Daten

- Der Umgang mit fehlenden Daten hängt wesentlich von deren Zufälligkeitsgrad ab

- **CCA = complete case approach**

  - Es werden ausschließlich vollständige Fälle für die weitere Analyse verwendet
  - Alle Fälle mit auch nur einem fehlenden Wert werden aus dem Datensatz entfernt
  - Die Methode kann nur bei zufällig fehlenden Daten (MCAR) angewendet werden
  - Günstig ist sie bei einer großen Stichprobe, in der Löschungen unkritisch sind

- **Ausschluss von Fällen oder Variablen**

  - Ziel des selektiven Ausschlusses ist die Verringerung des Gesamtanteils fehlender Werte
  - Abwägung zwischen Datenverlust und Reduktion der Probleme durch fehlende Werte
  - Günstigste Methode für nicht zufällig auftretende fehlende Werte (MAR, NRM)
  - Der Ausschluss von Fällen kann fallweise oder paarweise erfolgen

Umgang mit fehlenden Daten

- **Ersetzen fehlender Werte**

  - Grundidee: metrische Daten (und zwar ausschließlich diese) lassen sich ersetzen, wenn Regelmäßigkeiten erkennbar sind
  - Möglich ist der Ersatz von Werten über verschiedene induktive (nichtmathematische) und statistische (mathematische) Verfahren
  - Die wesentlichen Gefahren bei dieser Vorgehensweise bestehen darin, dass man den Datensatz **für vollständig hält** bzw. **durch Ersetzungen verzerrt**

Ausschlussverfahren

- **Fallweiser Ausschluss**
  - Fehlt ein einzelner Wert, wird der komplette Fall von der weiteren Analyse ausgeschlossen
  - Vorteil: Asymmetrien werden vermieden, da keine Teilfälle in die Analyse eingehen
  - Nachteil: Relevantes Datenmaterial geht verloren, der Stichprobenumfang sinkt
- **Paarweiser Ausschluss**
  - Fehlen einzelne Werte, wird mit den restlichen Werten des Falles weitergearbeitet
  - Vorteil: Alle Fälle bleiben erhalten, der Stichprobenumfang verändert sich nicht
  - Nachteil: Bei multivariaten Analysen u.U. unterschiedlich große Datensätze
- Um Fälle zu vermeiden, bei denen auf unterschiedlich große Datensätze zurückgegriffen und dabei verglichen wird, ist der fallweise Ausschluss das weitaus häufiger verwendete Ausschlussverfahren

Ersatzwertverfahren

- **Induktive Verfahren**

  - Die fehlenden Werte werden auf der Basis von Informationen ersetzt, die über die Stichprobe vorliegen
  - Nachbeobachtungen: Zusätzliche Beobachtungen oder Befragungen werden angestellt (Wie wirkt sich das auf die Repräsentativität aus?)
  - Externe Konstanten: Ein konstanter Wert aus einer externen Quelle oder aus einer früheren Studie wird ersatzweise verwendet

- **Statistische Verfahren**

  - Mittelwertersatz: Der fehlende Wert einer Variable wird durch das Mittel der Variablen ersetzt
  - Es existieren verschiedene Formen des Mittelwertersatzes: Ersatz durch das Mittel oder den Median der Nachbarpunkte, Ersatz durch einen Zeitreihen-Mittelwert und lineare Interpolation
  - Vorteil: Die Verfahren sind leicht anwendbar, benötigt werden nur die jeweiligen Mittelwerte
  - Nachteil: Die Varianz, die Verteilung der Daten und eventuelle Korrelationen werden verzerrt

Ersatzwertverfahren

- Linearer Trend: Ein fehlender Variablenwert wird durch einen linearen Trendwert ersetzt
- Voraussetzung: Für die gültigen Werte lässt sich ein aussagekräftiger linearer Trend identifizieren
- In diesem Fall können fehlende Werte durch die entsprechenden Werte der Trendgraden an der betreffenden Stelle ersetzt werden
- Nachteil: Der (durch zufällige Artefakte möglicherweise überschätzte) lineare Trend in den Variablen wird unbotmäßig verstärkt, die Varianz der Verteilung verringert sich

- Grundsätzlich ist bei allen Ersatzwertverfahren zu beachten, dass die Einbringung von Schätz- und Ersatzwerten in den Datensatz dokumentiert und im Datensatz so gut wie möglich kenntlich gemacht werden muss, damit im Rahmen einer sekundärstatistischen Analyse nicht von realen Werten ausgegangen wird

Was sollte man für die Klausur können?
(alle Angaben natürlich ohne Gewähr)

– Grundbegriffe (Skalenniveaus, Variablentypen etc.) werden über ein Multiple Choice-Quiz abgefragt

– Aufstellung von Häufigkeitstabellen und kumulierten Häufigkeitstabellen

– Berechnung von arithmetischem Mittel, getrimmtem arithmetischem Mittel, Median, Quartilen und Modus

– Berechnung von Varianz, Standard-abweichung, IQR und Spannweite

– Berechnung von Momentenkoeffizient, Quartilskoeffizient, Kurtosis und Exzeß

– Bei den Grafiken sind nur Box-Plots und Stem-and-Leaf-Plots zu zeichnen

– Von den drei Zusammenhangsmaßen (B-P-K, Spearman, Kendall) kommen mindestens zwei in der Klausur vor

Ressourcen für die Klausurvorbereitung

- Statistik-Wiki im Stud.IP
- Probeklausuren im Stud.IP
- Diskussionsforen im Stud.IP
- Multiple Choice-Quiz im Stud.IP

http://studip.hs-harz.de

- Übungsblätter zu Statistik I
- Aufgabenheft zu Statistik II
- Foliensätze zu Statistik I und II
- Links zu Open Source-Software

http://www.hs-harz.de/creinboth/

Zusätzliches Tutorium gewünscht?

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Statistik II

Christian Reinboth
M.Sc., Dipl.-Wi.Inf.(FH)

Sommersemester 2022
Berufsbegleitender Bachelorstudiengang Betriebswirtschaftslehre

Statistik
Wesentliche Kursinhalte (1)

- Kurzvorstellung
- Organisatorisches
- Bücher und Software

- Grundlagen

Statistik I

  - Einordnung
  - Grundbegriffe
  - Skalenniveaus
  - Variablentypen

- Qualitative und quantitative Forschung
  - Unterschiede
  - Vor- und Nachteile
  - Methoden der Datenerhebung
  - Methoden der Datenauswertung

- Erhebungsplanung und -durchführung
  - Erhebungsarten
    - Zufällige Auswahl
    - Klumpenstichprobe
    - Willkürliche Auswahl
    - Auswahl typischer Fälle
    - Konzentrationsverfahren
    - Mindeststichprobengröße
  - Gütekriterien
    - Bedeutung
    - Validität
    - Reliabilität
    - Objektivität
    - Repräsentativität
    - Sonstige Gütekriterien

- Gutes Fragebogendesign
  - Zieldefinition
  - Anschreiben
  - Incentivierung
  - Frageformulierung
  - Gängige Fragetypen

- Deskriptive Statistik
  - Häufigkeiten
    - Häufigkeiten
    - Häufigkeitstabellen
    - Bildung von Klassen
    - Verteilungsfunktion
    - Summenfunktion

Statistik
Wesentliche Kursinhalte (2)

- Statistische Lagemaße
  - Statistische Lagemaße
  - Arithmetisches Mittel
  - Median
  - Quartile
  - Modus
- Dispersionsparameter
  - Dispersionsparameter
  - Spannweite
  - Interquartilsabstand
  - Fünf-Werte-Zusammenfassung
  - Varianz
  - Standardabweichung
  - Variationskoeffizient
- Verteilungsmaße
  - Verteilungsmaße
  - Momentenkoeffizient
  - Quartilskoeffizient
  - Kurtosis / Exzeß
- Korrelationskoeffizienten
  - Korrelationskoeffizienten
  - Korrelation und Kausalität
  - Bravais-Pearson-Koeffizient
  - Rangkorrelationskoeffizienten
  - Spearman-Koeffizient
  - Kendall-Koeffizient
- Explorative Statistik
  - Grafische Darstellungen
    - Box-Whisker-Plot
    - Stem-and-Leaf-Plot
    - Objektivität von Grafiken
  - Ausreißer und fehlende Werte

Statistik II

Statistik
Wesentliche Kursinhalte (2)

Statistik II

- Induktive Statistik
  - Lineare Regression
    - Zielstellung
    - Voraussetzungen
    - Interdependenzproblem
    - Methode der kl. Quadrate
    - Ergebnisinterpretation
    - Bestimmtheitsmaß
  - Statistische Testverfahren
    - Statistische Tests
    - Chi-Quadrat-Test
    - Alpha-Fehlerinflation

- Mengenlehre
  - Mengenlehre
  - Logische Operatoren
  - Kommutativgesetz
  - Assoziativgesetz
  - Distributivgesetz
  - De Morgansche Regel
  - Venn-Diagramme
- Wahrscheinlichkeitslehre
  - Laplace-Wahrscheinlichkeit
  - Axiome von Kolmogoroff
  - Additionssatz
  - Multiplikationssatz
  - Pfaddiagramme
  - Kombinatorik
  - Satz von Bayes

- Konfidenzintervalle
- Statistische Software
  - Kostenlose Software
  - Einführung in R
- Klausurvorbereitung
  - Übungsaufgaben
  - Probeklausur
  - Fragestunde

Teil VI

Induktive Statistik

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Induktive Statistik

Lineare Regressionsanalyse

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Lineare Regression: Grundlagen

– Während im Rahmen der Korrelationsanalysen nur „inhaltsfreie“ Zusammenhänge zwischen Variablen untersucht wurden, unterstellt die lineare Regressionsanalyse eine **Wirkungsrichtung**: X beeinflusst Y (ausgedrückt als Gleichung Y = f(X))

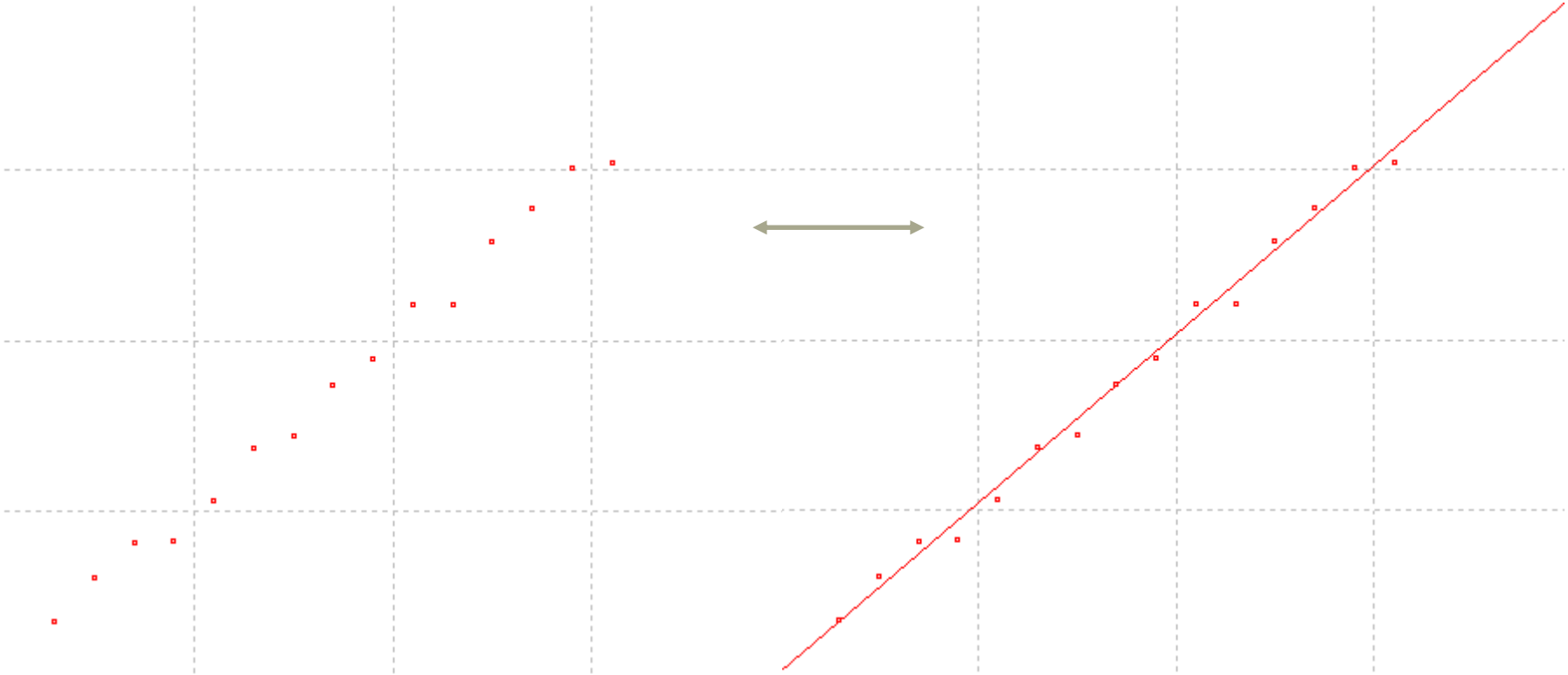
  – Wie wirkt sich der Preis (X) auf die Verkaufszahlen (Y) aus?
  – Wie wirkt sich die Zuckermenge (X) auf den Nährwert (Y) aus?
  – Wie wirkt sich die Lerndauer (X) auf den Punktestand (Y) aus?

$$Y = f(X)$$

– Wichtig: **Untersucht wird nur ein möglicher linearer Zusammenhang** – eine andere Form des Zusammenhangs (z.B. exponential, logarithmisch) wird dagegen nicht korrekt abgebildet

– Eine weitere Einschränkung: Im Rahmen dieser Vorlesung wird lediglich die Einfachregression (mit einer erklärenden Variablen), nicht jedoch die multiple Regression (mit mehreren erklärenden Variablen) betrachtet

Lineare Regression: Grundlagen



Lineare Regression: Grundlagen

- Die Regressionsanalyse ist das meistverwendete multivariate Analyseverfahren

- Im Rahmen einer (einfachen) linearen Regressionsanalyse wird die Beziehung zwischen einer abhängigen und einer unabhängigen Variablen untersucht, um

  - **Zusammenhänge quantitativ darzustellen** und zu erklären (Ursachenanalyse)
  - Werte der abhängigen Variablen zu **prognostizieren** (Wirkungsprognose)

  - Beispiel: Wie verändert sich die Absatzmenge (abhängige Variable) bei Veränderungen am Produktpreis, den Werbeausgaben oder der Anzahl der öffentlichen Verkaufsveranstaltungen (unabhängige Variablen)?

- Ergebnis des Verfahrens ist die **lineare Regressionsfunktion**

$$Y = f(X)$$

Lineare Regression: Interdependenz

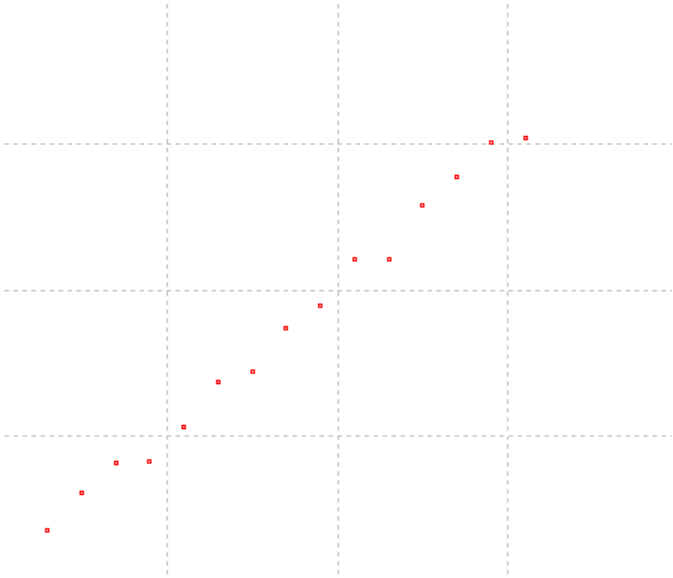
- Ein besonders Problem stellen **interdependente Beziehungen** dar
  - Beeinflusst der Bekanntheitsgrad eines Produkts die Absatzmenge oder beeinflusst die Absatzmenge den Bekanntheitsgrad eines Produkts?
  - Beeinflusst die Qualität einer Vorlesung die Motivation der Studierenden oder beeinflusst die Motivation der Studierenden die Qualität der Vorlesung?
- Ein solches interdependentes Beziehungssystem ist nicht in einer einzelnen Regressionsgleichung erfassbar, sondern nur in einem **Mehrgleichungsmodell**

Lineare Regression: Voraussetzungen

- Der linearen Regression liegen **drei wesentliche Annahmen** zugrunde:
  - Eine Variable X (die auch als **unabhängige Variable** bezeichnet wird) wirkt auf eine Variable Y (die wiederum als **abhängige Variable** bezeichnet wird), d.h. es gibt einen eindeutigen (und einseitigen) **Wirkungszusammenhang**
  - Der Zusammenhang zwischen X und Y ist **linear**
  - Sowohl X als auch Y sind **metrisch skaliert**
- Darüber hinaus wird angenommen, dass die Werte für Y Zufallsschwankungen unterliegen oder fehlerhaft gemessen werden können, während die Werte für X fehlerfrei vorliegen. Daraus ergibt sich, dass der Zusammenhang zwischen X und Y sich nicht fehlerfrei darstellen lässt, vielmehr muss nach dem Modell mit den wenigsten Fehlern (eben dem Regressionsmodell) gesucht werden.

Schätzung der Regressionsfunktion

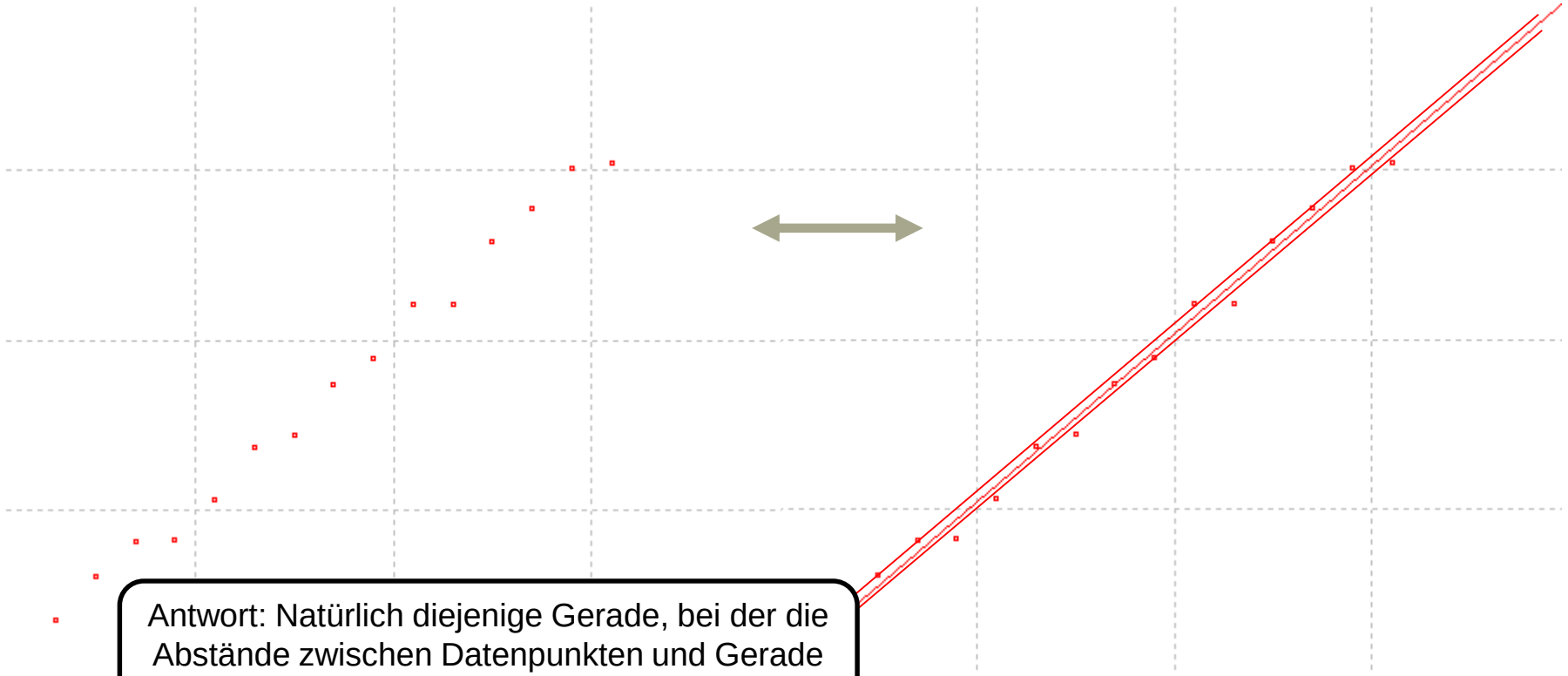
- Der Zusammenhang zwischen den beiden Variablen im Streudiagramm ist selten perfekt
- Beide Variablen bewegen sich hier im Beispiel jedoch tendenziell in die gleiche Richtung, ein linearer Trend ist klar erkennbar
- Es kommen nun theoretisch zahlreiche Geraden in Frage, um den Verlauf der Punkte nachzuzeichnen



Frage: Welche der möglichen Geraden beschreibt den Zusammenhang am besten?

Schätzung der Regressionsfunktion

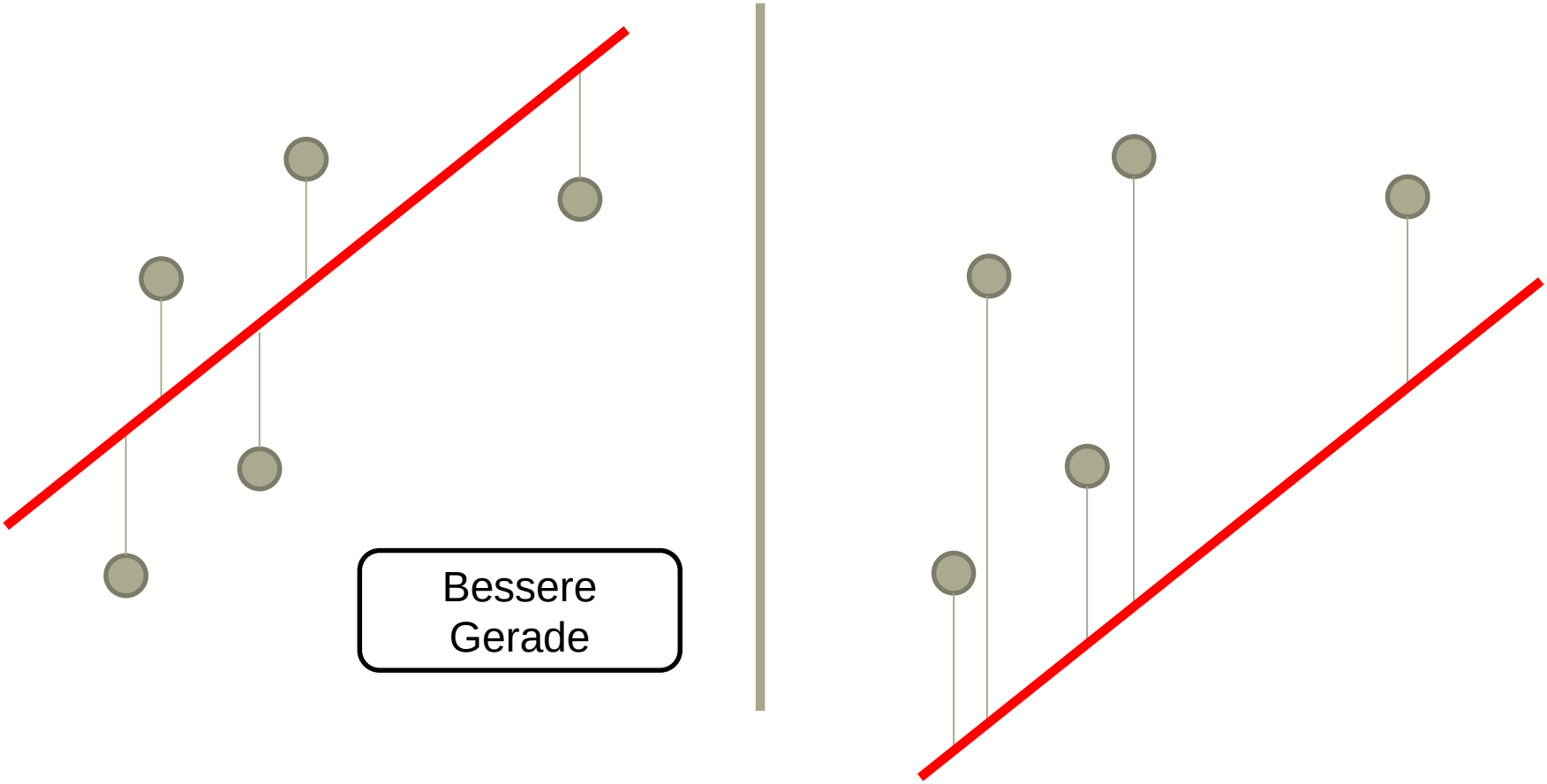
Frage: Welche der möglichen Geraden beschreibt den Zusammenhang am besten?



Antwort: Natürlich diejenige Gerade, bei der die Abstände zwischen Datenpunkten und Gerade minimal ausfallen. Aber wie findet man diese?

Schätzung der Regressionsfunktion

Frage: Welche der möglichen Geraden beschreibt den Zusammenhang am besten?



Methode der kleinsten Quadrate

– Lösungsansatz: **Minimierung der Summe der quadrierten Abweichungen** (der Geraden von den Werten) = **Methode der kleinsten Quadrate (MdkQ)**

– Die Methode der kleinsten Quadrate zielt – wie auch die intuitive Methode der simplen Abstandsminimierung – auf die **Minimierung der senkrechten Abstände der realen Werte von der Gerade** ab

– Die Abstände werden jedoch quadriert, so dass negative Vorzeichen wegfallen, wodurch die **Kompensation negativer und positiver Abstände** vermieden wird

– Schlussendlich wird diejenige Gerade selektiert, bei der die Summe der quadrierten Abstände minimal wird → sie ist die an die realen Werte **bestangepasste Gerade**

Methode der kleinsten Quadrate

- Regressionsfunktion: $Y = f(X)$
- Abgebildet über: $y = a + b * x$
- Berechnung von b: (Regressionskoeffizient)

$$b = \frac{\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y}}{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^2}$$

- Berechnung von a: (Konstantes Glied) $a = \bar{y} - b * \bar{x}$

Methode der kleinsten Quadrate

Nr.	x	y	x^2	(x * y)
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
Σ	…	…	…	…
Ø	…	…	//	//

$$b = \frac{\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y}}{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^2} \qquad a = \bar{y} - b * \bar{x} \qquad y = a + b * x$$

Übung: Methode der kleinsten Quadrate

Nr.	x	y
1	12	10000
2	15	15000
3	8	6000
4	11	11000
5	3	5000
6	17	23000
7	24	37000

Beispielfall mit bewusst gering gehaltener (Foliendarstellung…) Anzahl von Werten:

- x = Prozentualer Anteil des Werbebudgets eines Produkts am Gesamtbudget der Firma
- y = Verkaufte Einheiten des betrachteten Produkts in einem Untersuchungszeitraum
- Annahme: Das betrachtete Produkt, der Untersuchungszeitraum sowie das Gesamtbudget bleiben gleich

  (ceteris paribus)

Wie lautet die Regressionsgleichung?

Übung: Methode der kleinsten Quadrate

Nr.	x	y	x^2	(x * y)
1	12	10000	144	120000
2	15	15000	225	225000
3	8	6000	64	48000
4	11	11000	121	121000
5	3	5000	9	15000
6	17	23000	289	391000
7	24	37000	576	888000
Σ	90	107000	1428	1808000
Ø	12,86	15285,71	//	//

Übung: Methode der kleinsten Quadrate

$$b = \frac{\sum_{i=1}^{n}(x_i * y_i) - n * \bar{x} * \bar{y}}{\sum_{i=1}^{n}(x_i^2) - n * \bar{x}^2}$$

$$a = \bar{y} - b * \bar{x}$$

$$y = a + b * x$$

Beim Nachrechnen mit PSPP:
Rundungsfehler beachten

$$n = 7$$

$$\bar{x} = 12{,}86$$

$$\bar{y} = 15285{,}71$$

$$\sum_{i=1}^{n}(x_i^2) = 1428$$

$$\sum_{i=1}^{n}(x_i * y_i) = 1808000$$

$$b = \frac{1808000 - 7 * 12{,}86 * 15285{,}71}{1428 - 7 * 12{,}86^2}$$

$$= \frac{431980{,}39}{270{,}34} = 1597{,}92$$

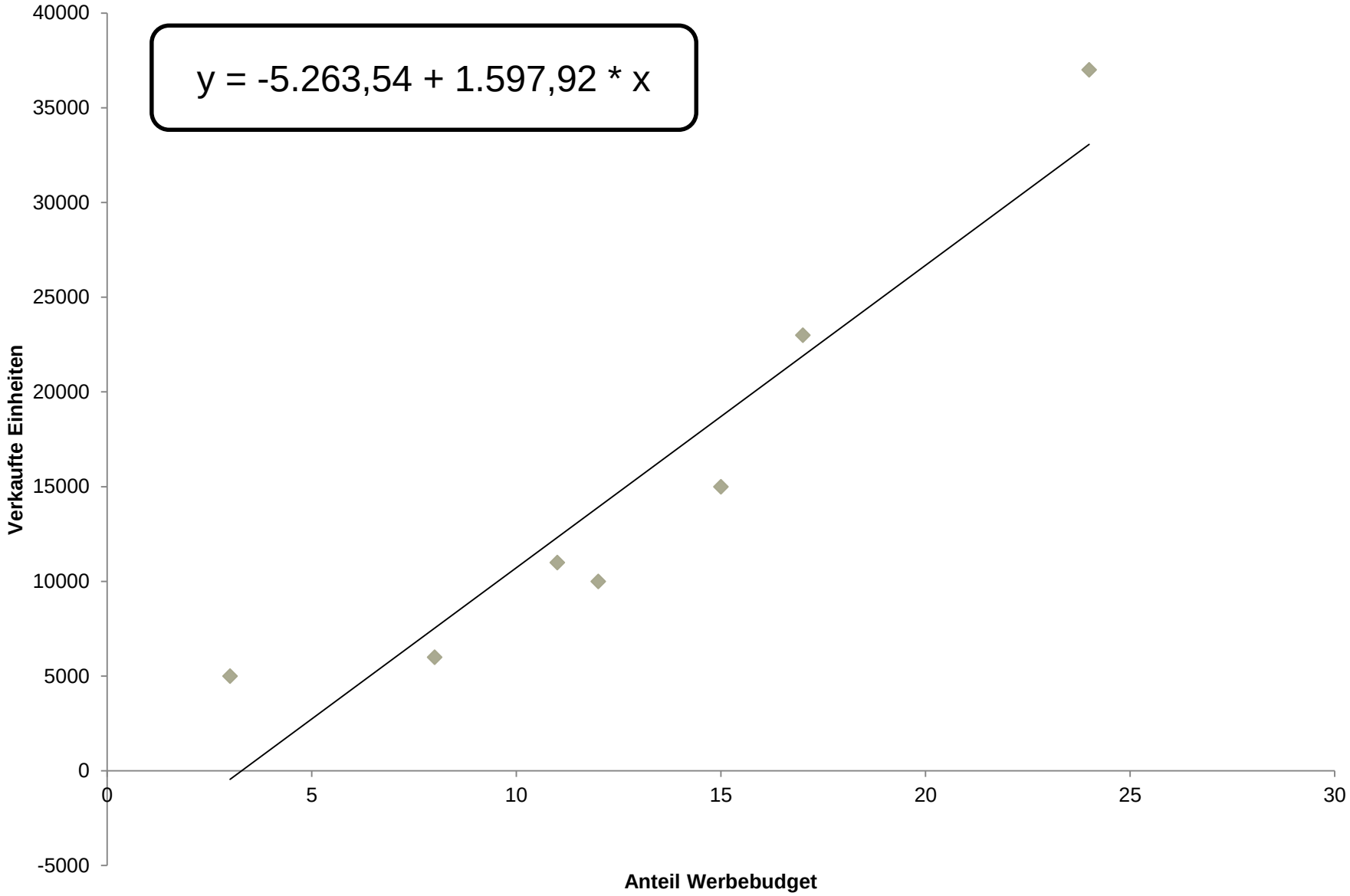
$$a = 15285{,}71 - 1597{,}92 * 12{,}86$$

$$= -5263{,}54$$

$$y = -5263{,}54 + 1597{,}92 * x$$

Interpretation der Regressionsgleichung

- Was lässt sich mit der Gleichung **y = -5.263,54 + 1.597,92 * x** anfangen?
  - Prognose unbekannter Werte: Bei einem Anteil am Werbebudget von 10% wären -5.263,54 + 1.597,92 * 10 = 10.715,66 verkaufte Einheiten zu erwarten
  - Aussage über den linearen Einfluss von X auf Y: Mit jedem Prozent, um den der Werbeetat angehoben wird, ist mit 1.597,92 zusätzlichen Verkäufen zu rechnen
  - Aber: Bei einem Werbeetat von 0% wären -5.263,54 verkaufte Einheiten zu erwarten – es stellt sich insofern die Frage, ob die Regressionsgleichung für große und kleine Werte von x noch gilt (klassisches Beispiel hierfür: Prognose der Geschmacksbewertung von Getränken auf Basis des zugegebenen Zuckers)
  - Bei der Konstruktion der Regressionsgeraden entspricht das konstante Glied **a** dem **Y-Achsenschnittpunkt**, der Regressionskoeffizient **b** der **Steigung**


y = -5.263,54 + 1.597,92 * x
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
Verkaufte Einheiten
0
5
10
15
20
25
30
Anteil Werbebudget

Das Bestimmtheitsmaß R^2

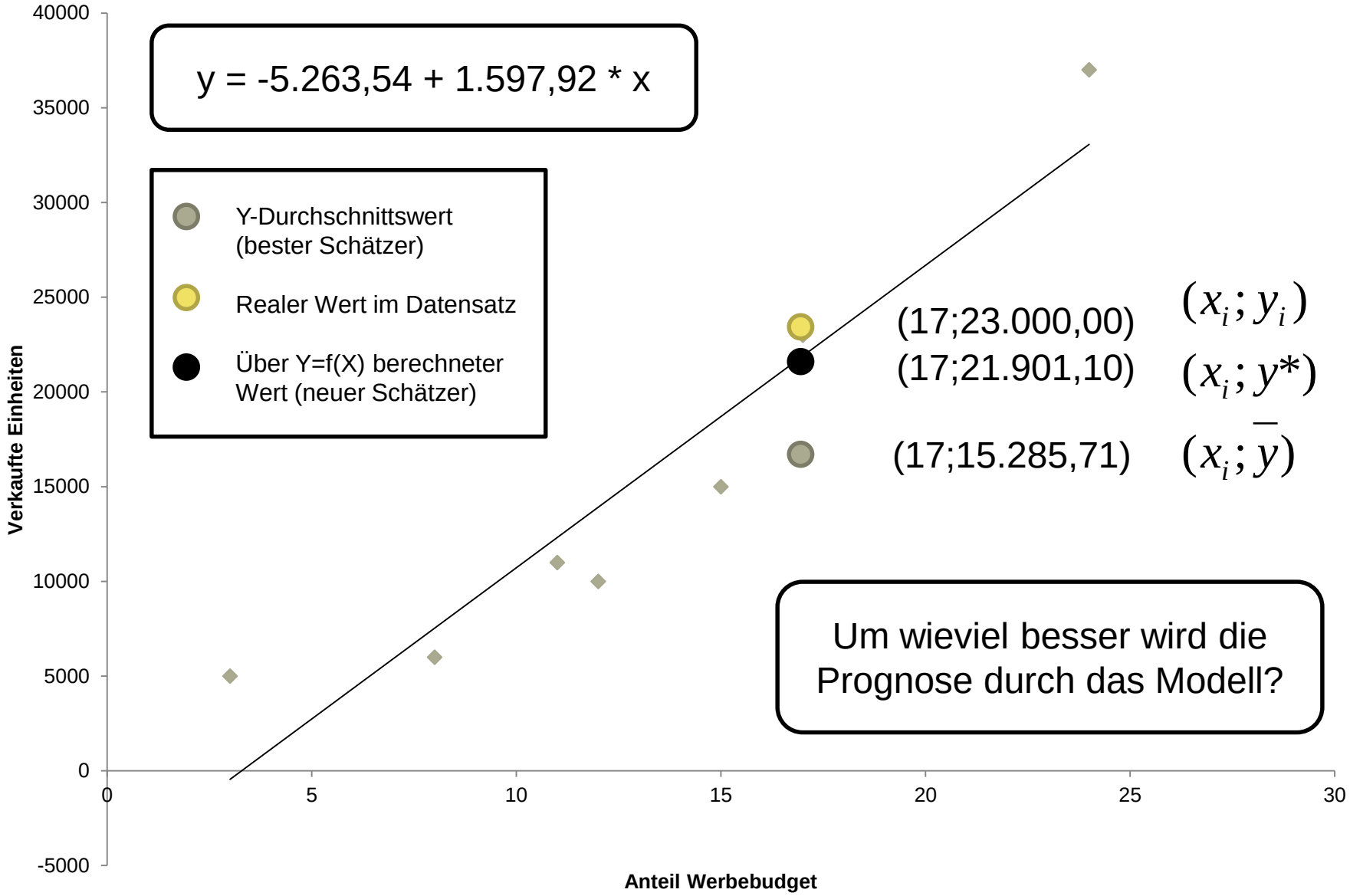
– Die Regressionsgerade gibt Zusammenhänge, die nicht perfekt linear sind (nicht alle Punkte liegen unmittelbar auf der Geraden), natürlich nur imperfekt wieder

– Es ist daher mit der Regressionsfunktion nur selten möglich, sämtliche Veränderungen in Y ausschließlich durch die Koeffizienten zu erklären

– In der Regel wird ein Teil der Veränderungen erklärt werden können, ein anderer Teil (die **Residuen**) wird dagegen unaufgeklärt bleiben

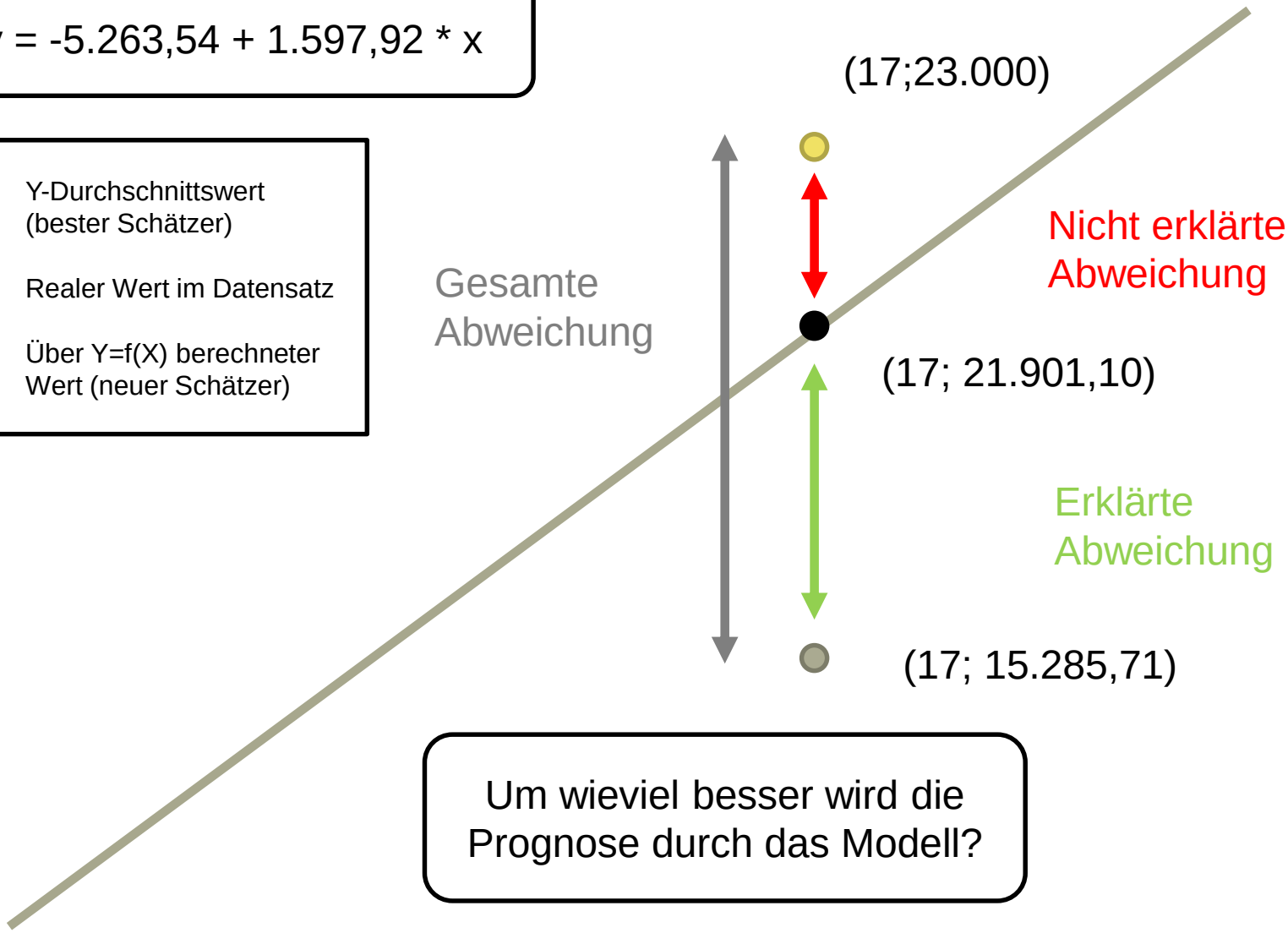
– Das **Verhältnis von erklärter Streuung zur Gesamtstreuung** ist ein gutes Maß für die **Güte des linearen Regressionsmodells**

– Die Residuen werden bei der Berechnung dieser Maßzahl quadriert, damit sich positive und negative Abweichungen nicht neutralisieren


y = -5.263,54 + 1.597,92 * x
Y-Durchschnittswert (bester Schätzer)
Realer Wert im Datensatz
Über Y=f(X) berechneter Wert (neuer Schätzer)
(17;23.000,00) $(x_i; y_i)$
(17;21.901,10) $(x_i; y^*)$
(17;15.285,71) $(x_i; \bar{y})$
Um wieviel besser wird die Prognose durch das Modell?
Verkaufte Einheiten
Anteil Werbebudget
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
0
5
10
15
20
25
30

$$y = -5.263{,}54 + 1.597{,}92 * x$$

- Y-Durchschnittswert (bester Schätzer)
- Realer Wert im Datensatz
- Über Y=f(X) berechneter Wert (neuer Schätzer)



Um wieviel besser wird die Prognose durch das Modell?

Das Bestimmtheitsmaß R²

– Die **Berechnung des Güßtemaßes R²** erfolgt mit:

$$R^2 = \frac{ESS}{TSS}$$

  – TSS = Total Sum of Squares = Summe aller quadrierten Abweichungen
  – ESS = Explained Sum of Squares = Summe aller erklärten quadrierten Abweichungen
  – RSS = Residual Sum of Squares = Summe aller nicht erklärten quadrierten Abweichungen
  – Das Verhältnis zwischen erklärter Streuung und Gesamtstreuung wird mit R² bezeichnet

– R² gibt den Anteil der erklärten Streuung an der Gesamtstreuung wieder
  > Güte der Anpassung und damit Güte des Regressionsmodells

  – R² ist als prozentualer Wert zu verstehen und liegt daher stets zwischen 0 und 1
  – R² = 1 → Gesamte Streuung wird erklärt, es besteht ein perfekter linearer Zusammenhang
  – Je kleiner R² ausfällt, desto mehr weicht der vorliegende Fall vom linearen Zusammenhang ab

– Beachte: R² ist ein Maß für den linearen – und nur für diesen – Zusammenhang

Das Bestimmtheitsmaß R²

Nr.	x	y	y*	$(y^* - \bar{y})^2$	$(y - y^*)^2$
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
Σ	//	…	//	…	…

ESS RSS

$$R^2 = \frac{ESS}{TSS}$$

Das Bestimmtheitsmaß entspricht übrigens dem quadrierten Bravais-Pearson-Korrelationskoeffizienten (lineare Korrelation)
[Vorsicht: Gilt nur für die lineare Einfachregression]

Übung: Bestimmtheitsmaß R²

Nr.	x	y	y*	$(y^* - \overline{y})^2$	$(y - y^*)^2$
1	12	10000	13911,50	1888453,12	15299832,25
2	15	15000	18705,26	11693322,20	13728951,67
3	8	6000	7519,82	60309047,49	2309852,83
4	11	11000	12313,58	8833556,74	1725492,42
5	3	5000	-469,78	248235465,14	29918493,25
6	17	23000	21901,10	43763384,85	1207581,21
7	24	37000	33086,54	316869548,69	15315169,17
Σ	//	15285,71	//	691592778,24	79505372,80

y = -5.263,54 + 1.597,92 * x

TSS = ESS + RSS = 771098151,03

Übung: Bestimmtheitsmaß R²

$$R^2 = \frac{ESS}{TSS} = \frac{691592778{,}24}{771098151{,}03} = 0{,}90$$

Hervorragender Wert! (max. +1)

Modellzusammenfassung (Verkauf)

R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
,95	,90	,88	3987,59

ANOVA (Verkauf)

	Quadratsumme	df	Mittel der Quadrate	F	Sig.
Regression	689924352,02	1	689924352,02	43,39	,001
Residual	79504219,41	5	15900843,88		
Gesamt	769428571,43	6			

Koeffizienten (Verkauf)

	Unstandardisierte Koeffizienten		Standardisierte Koeffizienten		
	B	Standardfehler	Beta	t	Sig.
(Konstante)	-5234,18	3460,63	,00	-1,51	,181
Werbung	1595,99	242,29	,95	6,59	,001

Komplexe Beispiele wie dieses lassen sich sehr gut in PSPP & Co. nachrechnen – man beachte aber die Rundungsfehler!

Induktive Statistik

Statistische Testverfahren

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Was sind statistische Testverfahren?

- Im Gegensatz zu den bereits bekannten Schätzverfahren, geht es bei den statistischen Testverfahren nicht mehr um die möglichst genaue Bestimmung unbekannter Parameter, sondern um die Prüfung vorab festgelegter Hypothesen

- Beispiele für mögliche Hypothesen / Forschungsfragen:

  - Im Harz gibt es während des Sommers mehr Regentage als in der Eiffel
  - BWL-Studierende geben mehr Geld für Literatur als Informatik-Studierende aus
  - Mit dem Alter von Mietern/innen steigt deren Wunsch nach barrierefreien Wohnungen
  - Weibliche Abiturientinnen schneiden in Mathematik besser als männliche Abiturienten ab

- Diese und andere Hypothesen können anhand von Daten aus Zufallsstichproben „getestet“ werden. Da Stichprobendaten zufälligen Schwankungen unterliegen, ist **kein** endgültiger Befund über die Richtigkeit der Hypothesen möglich - möglich ist lediglich eine Wahrscheinlichkeitsaussage.

Null- und Alternativhypothese

- Jeder Hypothesentest basiert auf einer **Nullhypothese H_0** (meistens: es existiert kein Effekt / kein Unterschied) sowie einer **Alternativhypothese H_1** (gegenteilige Aussage)

- Das Ergebnis des Tests **bezieht sich stets auf die Nullhypothese**, die entweder (mit einem gewissen Irrtumsrisiko) verworfen oder aber (dies ebenfalls einem gewissen Irrtumsrisiko) beibehalten werden kann

- Die Verwerfung geht weder mit einer Annahme der Alternativhypothese einher, noch ist sie ein Beweis dafür, dass die Nullhypothese nicht zutrifft

	H0 ist falsch	**H0 ist richtig**
Test verwirft H0	korrekt	Fehler 1. Art
Test verwirft H0 nicht	Fehler 2. Art	korrekt

Bedeutende statistische Hypothesentests

- Als Hypothesentest / Signifikanztest wird ein Verfahren bezeichnet, über das man auf der Basis vorliegender Beobachtungen (meist aus einer Stichprobe) zu einer begründeten Entscheidung über die Ungültigkeit einer Hypothese gelangen kann

- Im Rahmen dieser Vorlesung (kurz) angesprochene Testverfahren:
  - T-Test auf Gleichheit von Mittelwerten
  - Chi^2-Test auf Unabhängigkeit von Variablen
  - Kolmogoroff-Smirnov-Test auf Normalverteilung
  - Durbin-Watson-Test auf Autokorrelation von Residuen
  - Levene-Test auf Varianzgleichheit / Homoskedastizität

Wie lauten die H_0?

Wichtiger Hinweis: Um die zur Verfügung stehende Zeit optimal auszunutzen, werden wir nachfolgend nur den Chi^2-Test im Detail betrachten (alles weitere im Skript)

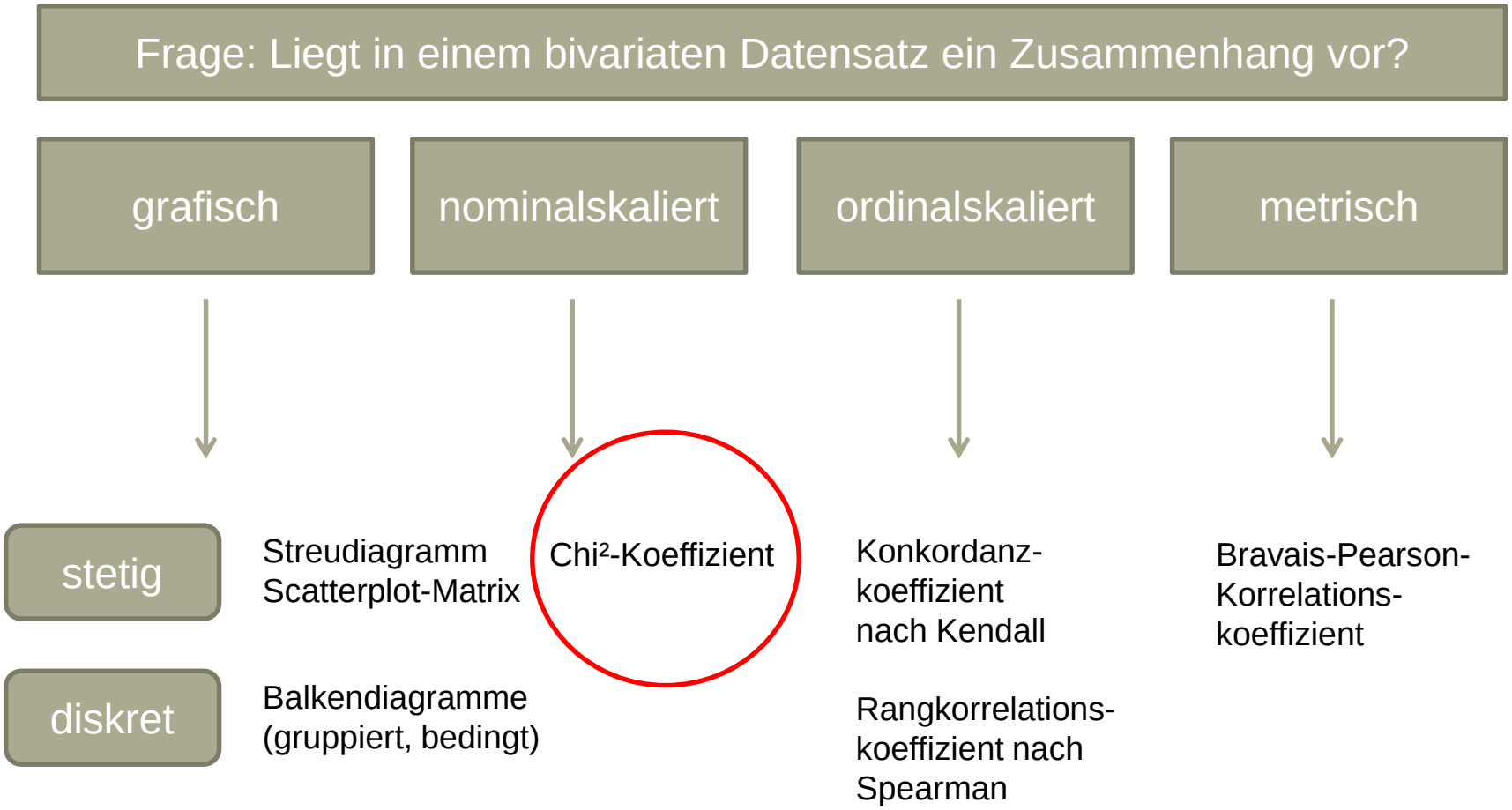
Induktive Statistik

Chi-Quadrat-Anpassungstest

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Erinnerung: Bivariate Zusammenhangsmaße



Chi²-Unabhängigkeitstest

- Beim Chi²-Unabhängigkeitstest (nachfolgend Chi²-Test) werden zwei nominal skalierte Merkmale auf stochastische Unabhängigkeit geprüft (Nullhypothese H_0: Die Merkmale X und Y sind stochastisch unabhängig voneinander)

- **Hierzu werden die real beobachteten Häufigkeiten mit den zu erwartenden Häufigkeiten bei völliger Unabhängigkeit der beiden Merkmale verglichen**

- Die bei Unabhängigkeit der Merkmale zu erwartende Verteilung lässt sich berechnen, indem man die sogenannten Randsummen multipliziert und durch die Anzahl der Gesamtwerte teilt

- Auf den folgenden Folien wird hierzu ein zusammenhängendes Beispiel betrachtet: 100 Personen wurden nach ihrem Schulabschluss sowie nach dem Schulabschluss ihrer Eltern befragt, um festzustellen, ob sich ein Zusammenhang finden lässt

Chi²-Unabhängigkeitstest

Bildungsabschluss/Eltern	Eltern haben Abitur	Eltern haben kein Abitur
Befragter hat Abitur	43	11
Befragter hat kein Abitur	12	34

– Zur Berechnung der im Fall völliger Unabhängigkeit zu erwartenden absoluten Häufigkeiten werden zunächst die Randsummen kalkuliert

Bildungsabschluss/Eltern	Eltern haben Abitur	Eltern haben kein Abitur	Rand
Befragter hat Abitur	43 [29,7]	11 [24,3]	54
Befragter hat kein Abitur	12 [25,3]	34 [20,7]	46
Rand	55	45	100

– Indem man die Randsummen multipliziert und durch die Gesamtsumme dividiert, erhält man den bei Unabhängigkeit zu erwartenden Wert, d.h. 55 * 54 / 100 = 29,7

Chi²-Unabhängigkeitstest

Bildungsabschluss/Eltern	Eltern haben Abitur	Eltern haben kein Abitur
Befragter hat Abitur	29,7	24,3
Befragter hat kein Abitur	25,3	20,7

- So würden sich also die 100 Befragten auf die vier Kategorien verteilen, gäbe es überhaupt keinen Zusammenhang zwischen dem eigenen Schulabschluss und dem Schulabschluss der Eltern

- Dass die tatsächlichen Werte von diesen Werten stark abweichen, ist bereits ein Indikator dafür, dass es einen Zusammenhang geben könnte

>> Mit Hilfe des Chi²-Tests soll nachfolgend festgestellt werden, ob die Abweichung so groß ist, dass ein Zusammenhang wahrscheinlich wird

Chi²-Unabhängigkeitstest

- Dazu werden die Differenzen zwischen erwarteten und tatsächlichen Werten quadriert und durch die zu erwartenden Werte dividiert, die Summe dieser Berechnungen ergibt dann den entscheidenden Chi^2-Wert

$(43 – 29{,}7)^2 / 29{,}7 = 5{,}955$
$(11 – 24{,}3)^2 / 24{,}3 = 7{,}279$
$(12 – 25{,}3)^2 / 25{,}3 = 6{,}991$
$(34 – 20{,}7)^2 / 20{,}7 = 8{,}545$
$= 28{,}77$

Warum werden die Differenzen quadriert?

- Es ergibt sich demnach ein Chi^2-Wert von 28,77

- Dieser ist dem Vergleichswert aus der tabellierten Chi^2-Verteilung gegenüberzustellen, wobei ein Fehlerniveau α von 5% (d.h. 1 – α = 0,950) bei einem Freiheitsgrad gewählt wurde (da sich unter Beibehaltung der Randsummen ein Wert frei festlegen lässt)

Chi²-Unabhängigkeitstest

– In der Tabelle der Chi²-Verteilung landet man bei dieser Vorgehensweise bei einem Vergleichswert von 3,84 („kritischer Wert“ des Testverfahrens)

	90%	95%	97,5%	99%	99,5%	99,9%
1	2,71	3,84	5,02	6,63	7,88	10,83
2	4,61	5,99	7,38	9,21	10,60	13,82
...	...	...	...	...	...	...

– Wird dieser durch den errechneten Wert überschritten, gilt die Nullhypothese, nach der die beiden Variablen „eigener Schulabschluss“ und „Schulabschluss der Eltern“ als voneinander völlig unabhängig einzustufen sind, als abgelehnt

– Da dies hier der Fall ist, lautet der Schluss, dass **mit hoher Wahrscheinlichkeit ein statistisch signifikanter Zusammenhang** zwischen den Variablen besteht

Chi²-Unabhängigkeitstest

– Der Chi^2-Unabhängigkeitstest läuft somit in vier Stufen ab:

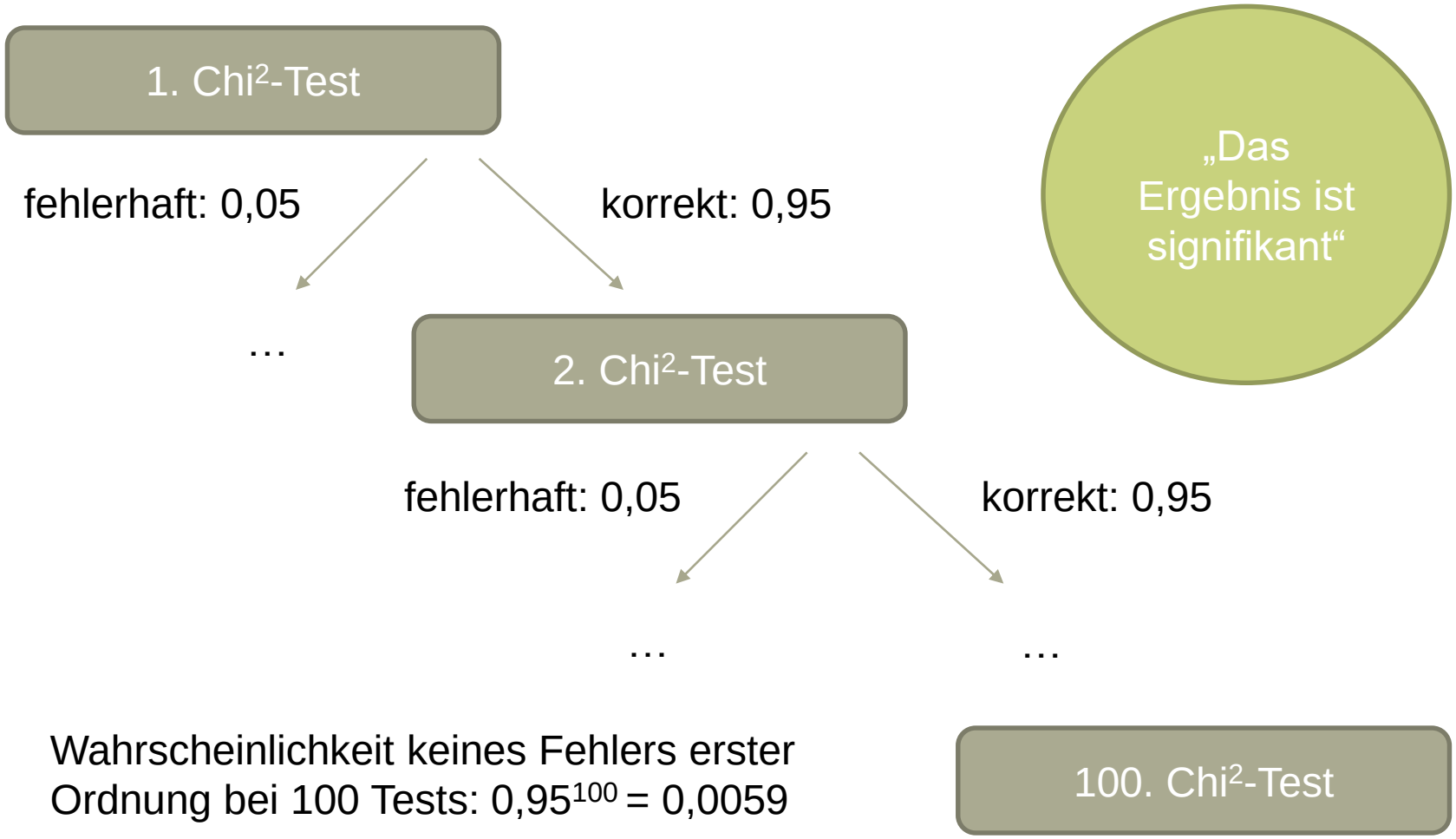
1. Berechnung der Randsummen für alle Zeilen und Spalten

2. Berechnung der zu erwartenden Häufigkeiten bei völliger Unabhängigkeit durch Multiplikation der jeweiligen Randsummen und Division durch die Gesamtsumme

3. Berechnung des Chi^2-Wertes durch Bildung der Summe der quadrierten Differenzen zwischen den tatsächlichen und den bei Unabhängigkeit zu erwartenden Häufigkeiten

4. Vergleich des Chi^2-Wertes mit dem kritischen Wert der Chi^2-Verteilung und Entscheidung über die Nullhypothese (Verwerfung oder Nicht-Verwerfung)

Das Problem der α-Fehlerinflation



Wahrscheinlichkeit keines Fehlers erster Ordnung bei 100 Tests: $0{,}95^{100} = 0{,}0059$

Das Problem der α-Fehlerinflation

- Führt man einen einzelnen Chi^2-Test (oder auch ein anderes statistisches Testverfahren) durch, muss a priori ein **Fehlerniveau α** festgelegt werden

- Liegt dieses Fehlerniveau z.B. bei 0,05, bedeutet dies, dass ein Fehler 1. Ordnung („false positives“) mit 5%iger Wahrscheinlichkeit auftritt, d.h. mit 5%iger Wahrscheinlichkeit wird eine falsche Signifikanz ausgewiesen

- Führt man nun aber eine Vielzahl von Tests an den gleichen Daten durch, ergeben sich fehlerhaft-signifikante Ergebnisse demnach mit steigender Wahrscheinlichkeit → dieser Effekt wird als **α-Fehler-Kumulierung / α-Fehlerinflation** bezeichnet

„Je mehr Hypothesen man auf einem Datensatz testet, desto höher wird die Wahrscheinlichkeit, dass eine davon (fehlerhaft) als zutreffend angenommen wird.“ (Definition der α-Fehlerinflation in der Wikipedia)

Übung: Chi²-Unabhängigkeitstest

– Eine an der Hochschule Harz durchgeführte Befragung, bei der unter anderem erhoben wurde, ob die Studierenden einem Nebenjob nachgehen, erbrachte folgendes – nach Geschlechtern aufgeteiltes – Ergebnis:

Geschlecht/Nebenjob	hat einen Nebenjob	hat keinen Nebenjob
Weibliche Studierende	35	26
Männliche Studierende	26	13

– Erinnerung: Der Chi^2-Unabhängigkeitstest erfolgt in vier Schritten:

1. Berechnung der Randsummen für alle Zeilen und Spalten
2. Berechnung der zu erwartenden Häufigkeiten bei völliger Unabhängigkeit
3. Berechnung des Chi^2-Wertes (über die Summe der quadrierten Differenzen)
4. Vergleich des Chi^2-Wertes mit dem kritischen Wert (bleibt hier gleich: 3,84)

Übung: Chi²-Unabhängigkeitstest

– Berechnung der Randsummen sowie der erwarteten Häufigkeiten bei Unabhängigkeit

Geschlecht/Nebenjob	hat einen Nebenjob	hat keinen Nebenjob	Rand
Weibliche Studierende	35 [37,21]	26 [23,79]	61
Männliche Studierende	26 [23,79]	13 [15,21]	39
Rand	61	39	100

– So würden sich also die 100 Befragten auf die vier Kategorien verteilen, gäbe es überhaupt keinen Zusammenhang zwischen dem Geschlecht der Befragten und der Wahrscheinlichkeit dafür, dass diese einen Nebenjob ausüben

– Dass die tatsächlichen Werte von diesen Werten kaum abweichen, ist bereits ein Indikator dafür, dass es keinen Zusammenhang geben dürfte

Übung: Chi²-Unabhängigkeitstest

- Im nächsten Schritt werden die Differenzen zwischen erwarteten und tatsächlichen Werten quadriert und durch die zu erwartenden Werte dividiert, die Summe dieser Berechnungen ergibt dann den entscheidenden Chi^2-Wert

$(35 – 37{,}21)^2 / 37{,}21 = 0{,}1313$
$(26 – 23{,}79)^2 / 23{,}79 = 0{,}2053$
$(26 – 23{,}79)^2 / 23{,}79 = 0{,}2053$
$(13 – 15{,}21)^2 / 15{,}21 = 0{,}3211$
$= 0{,}8630$

- Es ergibt sich demnach ein Chi^2-Wert von 0,8630

- Dieser ist dem Vergleichswert aus der tabellierten Chi^2-Verteilung gegenüberzustellen, wobei ein Fehlerniveau α von 5% (d.h. 1 – a = 0,950) bei einem Freiheitsgrad gewählt wurde (da sich unter Beibehaltung der Randsummen ein Wert frei festlegen lässt)

Übung: Chi²-Unabhängigkeitstest

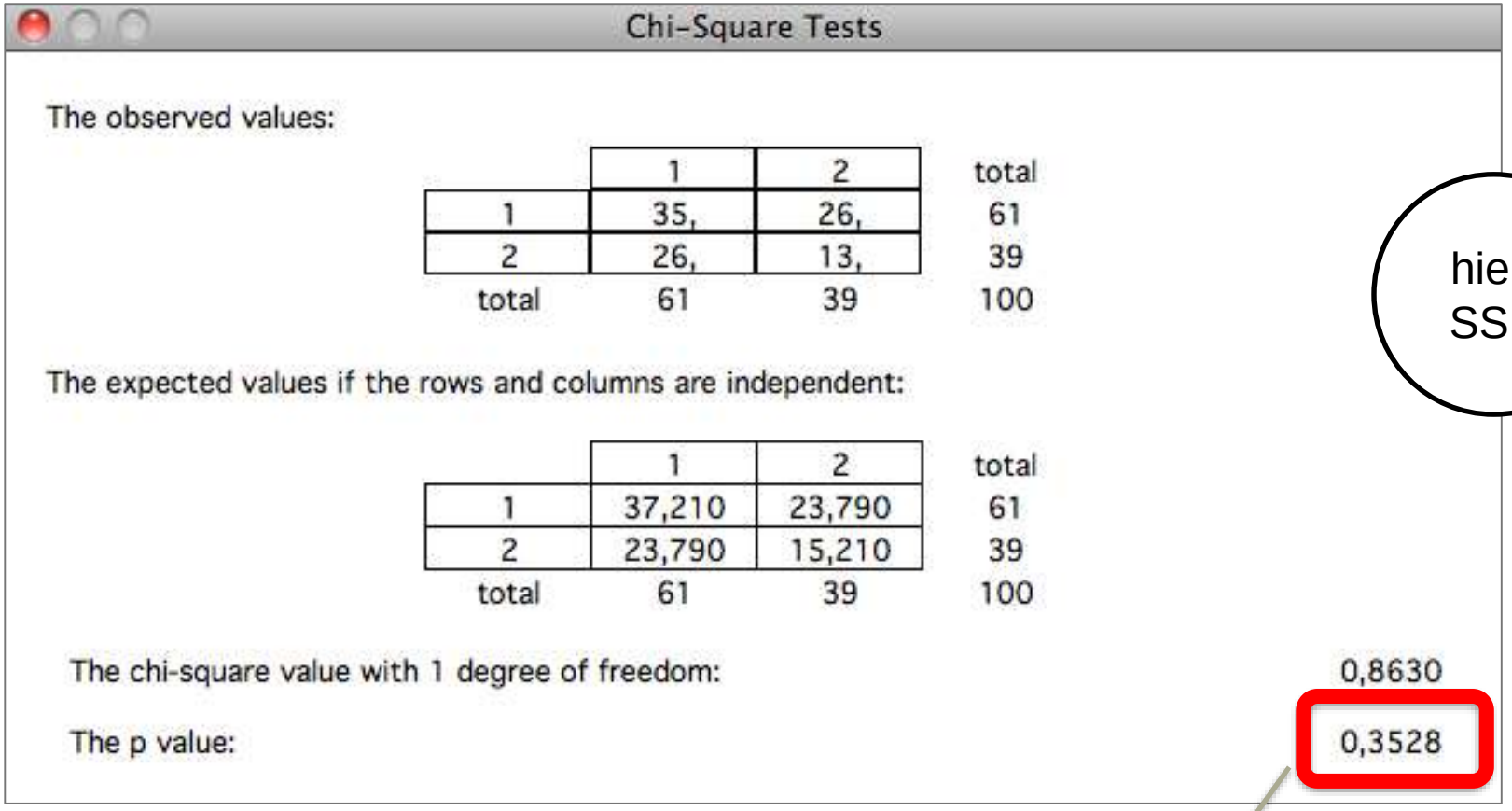
– In der Tabelle der Chi²-Verteilung landet man bei dieser Vorgehensweise bei einem Vergleichswert von 3,84 („kritischer Wert“ des Testverfahrens)

Chi²	90%	95%	97,5%	99%	99,5%	99,9%
1	2,71	3,84	5,02	6,63	7,88	10,83
2	4,61	5,99	7,38	9,21	10,60	13,82
...	...	...	...	...	...	...

– Wird dieser durch den errechneten Wert überschritten, gilt die Nullhypothese, nach der die beiden Variablen „Geschlecht“ und „Nebenjob“ als voneinander völlig unabhängig einzustufen sind, als abgelehnt

– Da dies hier nicht der Fall ist, lautet der Schluss, dass die Nullhypothese (Variablen sind unabhängig) nicht verworfen werden kann (aber: kein Beweis für ihre Gültigkeit)

Wie laufen Testverfahren mit Software ab?



Chi-Square Tests

The observed values:

	1	2	total
1	35,	26,	61
2	26,	13,	39
total	61	39	100

The expected values if the rows and columns are independent:

	1	2	total
1	37,210	23,790	61
2	23,790	15,210	39
total	61	39	100

The chi-square value with 1 degree of freedom: 0,8630

The p value: 0,3528

hier: SSP

„Signifikanzwert“ – was ist das?

Interpretation des Signifikanzwertes

- **Der p-Wert / Signifikanzwert gibt die Wahrscheinlichkeit dafür an, dass die real beobachteten Werte / Abweichungen auftreten, wenn die Nullhypothese zutrifft**

- Am Beispiel des Chi^2-Unabhängigkeitstests:

  - Nullhypothese: Die betrachteten Merkmale x und y sind stochastisch unabhängig
  - Großer p-Wert: Es ist wahrscheinlich, dass die realen Werte bei Gültigkeit der Nullhypothese erreicht werden konnten → Beibehaltung der Nullhypothese
  - Kleiner p-Wert: Es ist unwahrscheinlich, dass die realen Werte bei Gültigkeit der Nullhypothese erreicht werden konnten → Verwerfung der Nullhypothese

- Der p-Wert wird oft (leicht falsch) als Wahrscheinlichkeit dafür interpretiert, dass das Zurückweisen einer Nullhypothese H_0 falsch ist (Irrtumswahrscheinlichkeit)

Großer Signifikanzwert	=	Nullhypothese beibehalten
Kleiner Signifikanzwert	=	Nullhypothese zurückweisen

Teil VII

Mengenlehre

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Grundbegriffe der Wahrscheinlichkeitslehre

- **Zufallsvorgang**: Ein Zufallsvorgang ist ein Vorgang, der in einem von mehreren möglichen Ergebnissen mündet, die sich wiederum gegenseitig ausschließen
- Welches Ereignis eintritt, kann vorab nicht mit Sicherheit ausgesagt werden

- **Zufallsexperiment**: Ein Zufallsexperiment ist die (beliebig häufige) Wiederholung eines Zufallsvorgangs unter kontrollierten, gleich bleibenden Rahmenbedingungen

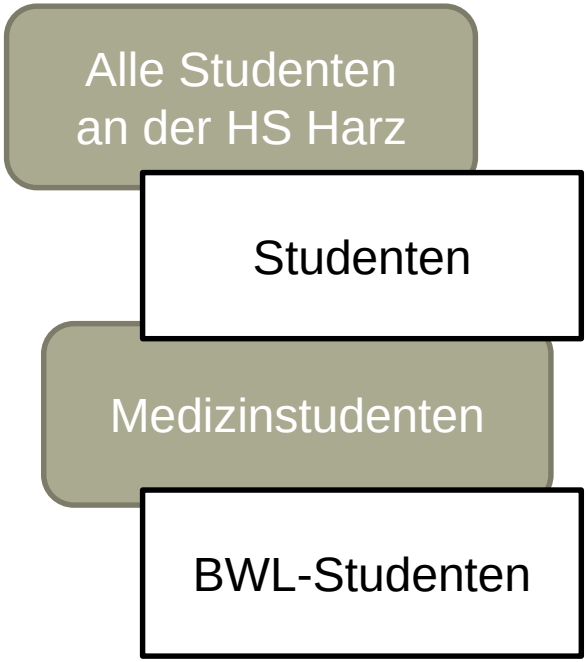
- Typische Beispiele für Zufallsexperimente
  - „Kopf oder Zahl“-Spiel mit einer fairen Münze
  - Würfeln mit einem (oder mehreren) fairen Würfeln
  - Lauf einer Kugel durch den Kessel beim Roulettespiel
  - Ziehung von Lottozahlen (ohne Zurücklegen) aus einer Trommel
  - Ziehen von Karten (mit oder ohne Zurücklegen) aus einem Kartenstapel
  - Ziehen von schwarzen/weißen Kugeln (mit oder ohne Zurücklegen) aus einer Urne

Ist die „zufällige“ Auswahl von Passanten ebenfalls ein Zufallsexperiment?

Grundbegriffe der Mengenlehre

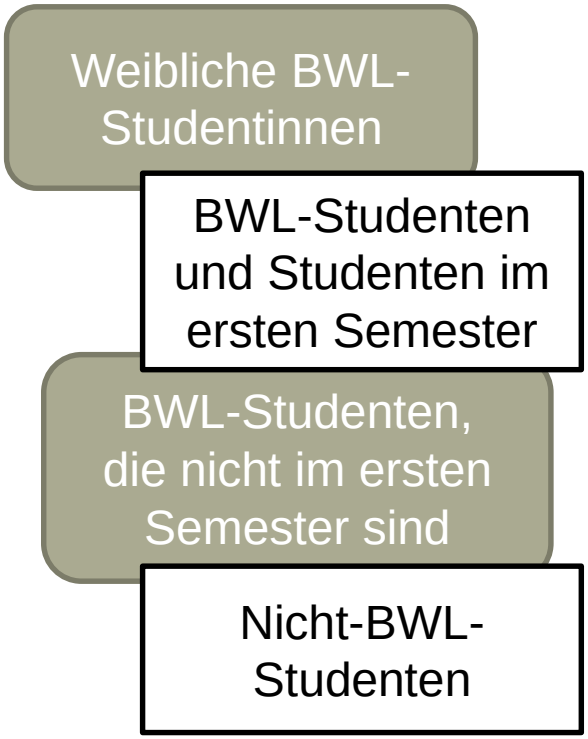
– Um die **Ergebnisse von Zufallsexperimenten** beschreiben zu können, wird nachfolgend auf das Vokabular der **Mengenlehre** zurückgegriffen

- **Menge**
  = Eine Gruppe von Elementen (Ω)
- **Elemente**
  = Einzelne Mitglieder einer Menge
  (nicht teilbare Elementarereignisse)
- **Leere Menge**
  = Eine Menge ohne ein Element (Ø)
- **Teilmenge**
  = Eine Untermenge einer anderen Menge
  (z.B. A ist eine Teilmenge von Ω: $A \subseteq \Omega$)



Grundbegriffe der Mengenlehre

- **Schnittmenge**
  = Eine Menge aller Elemente, die zugleich in zwei Mengen (A und B) enthalten sind
- **Vereinigungsmenge**
  = Eine Menge aller Elemente, die entweder in A oder B (oder in A und B) enthalten sind
- **Differenzmenge**
  = Eine Menge aller Elemente, die zwar in einer Menge (A), zugleich aber nicht in einer anderen Menge (B) enthalten sind
- **Komplementärmenge**
  = Eine Menge aller Elemente, die nicht zu einer anderen Menge (A) gehören (d.h. der Rest des Ereignisraums G)



Logische Operatoren und Mengen

– **Logisches UND (Konjunktion, A∩B)**

Menge A	Menge B	UND
W	W	W
W	F	F
F	W	F
F	F	F

Wahrheitstabelle

– **Logisches ODER (Disjunktion, A∪B)**

Menge A	Menge B	ODER
W	W	W
W	F	W
F	W	W
F	F	F

Logische Operatoren und Mengen

– **Logisches NICHT (Negation, Ā)**

Menge A	NICHT
W	F
F	W

– Wie lassen sich zentrale Begriffe mit Operatoren ausdrücken?

  – Schnittmenge von A und B: $A \cap B$
  – Vereinigungsmenge von A und B: $A \cup B$
  – Differenzmenge von A und B: $A \setminus B$
  – Komplementärmenge von A: $\bar{A}$

Regeln für das Rechnen mit Mengen

– **Kommutativgesetz**

Die Argumente einer kommutativen Operation können vertauscht werden, ohne dass sich das Ergebnis ändert

Beispiel: $1 + 2 = 2 + 1$

$1 * 2 = 2 * 1$

– Das Kommutativgesetz in der Mengenlehre:

$A \cap B = B \cap A$

$A \cup B = B \cup A$

Regeln für das Rechnen mit Mengen

– **Assoziativgesetz**

Eine zweistellige Verknüpfung ist assoziativ, wenn die Reihenfolge der Ausführung keine Rolle spielt (die Klammersetzung ist somit beliebig)

Beispiel: $(1 + 2) + 3 = 1 + (2 + 3)$

$(1 * 2) * 3 = 1 * (2 * 3)$

– Das Assoziativgesetz in der Mengenlehre:

$(A \cap B) \cap C = A \cap (B \cap C)$

$(A \cup B) \cup C = A \cup (B \cup C)$

Regeln für das Rechnen mit Mengen

– **Distributivgesetz**

Das Distributivgesetz regelt die Auflösung
von Klammern (z.B. durch Ausmultiplikation)

Beispiel: $(1 + 2) * 3 = (1 * 3) + (2 * 3)$
$(1 - 2) * 3 = (1 * 3) - (2 * 3)$

– Das Distributivgesetz in der Mengenlehre:

$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Regeln für das Rechnen mit Mengen

– **De Morgansche Regel**

…müsste eigentlich Ockhamsche Regel heißen, da sie bereits William von Ockham („Ockhams Rasiermesser“ / „Occam's razor“) bekannt war

„Von mehreren möglichen Erklärungen für ein und denselben Sachverhalt ist die einfachste Theorie allen anderen vorzuziehen.“

Augustus de Morgan (1806 – 1871)
(Quelle: WikiMedia; Lizenz: gemeinfrei)

William von Ockham (1288 – 1347)
(Quelle: WikiMedia; Lizenz: gemeinfrei)

– Die De Morgansche Regel lautet:

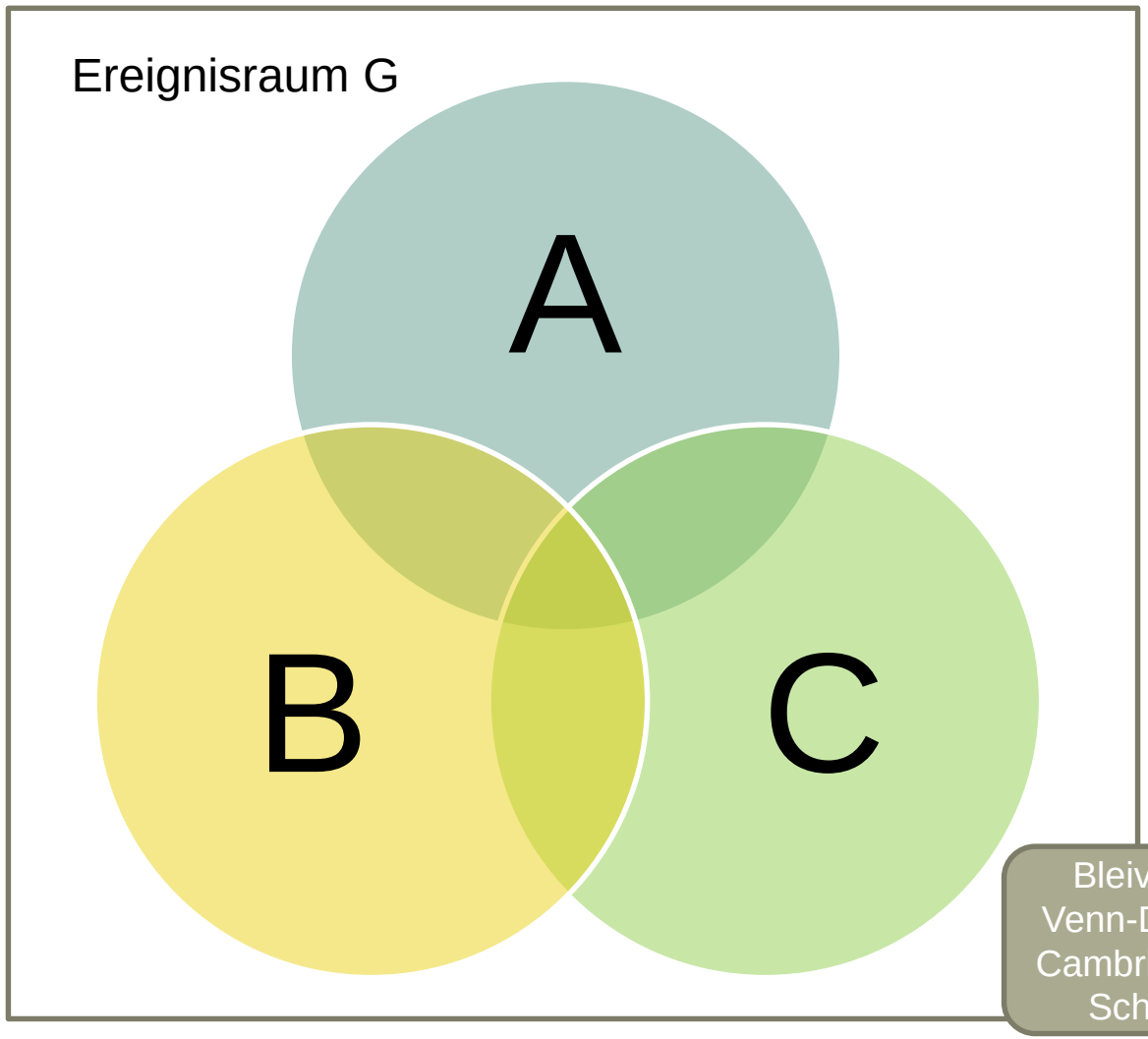
$$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$$

$$\overline{(A \cap B)} = \overline{A} \cup \overline{B}$$

Übung: Logische Operatoren und Mengen

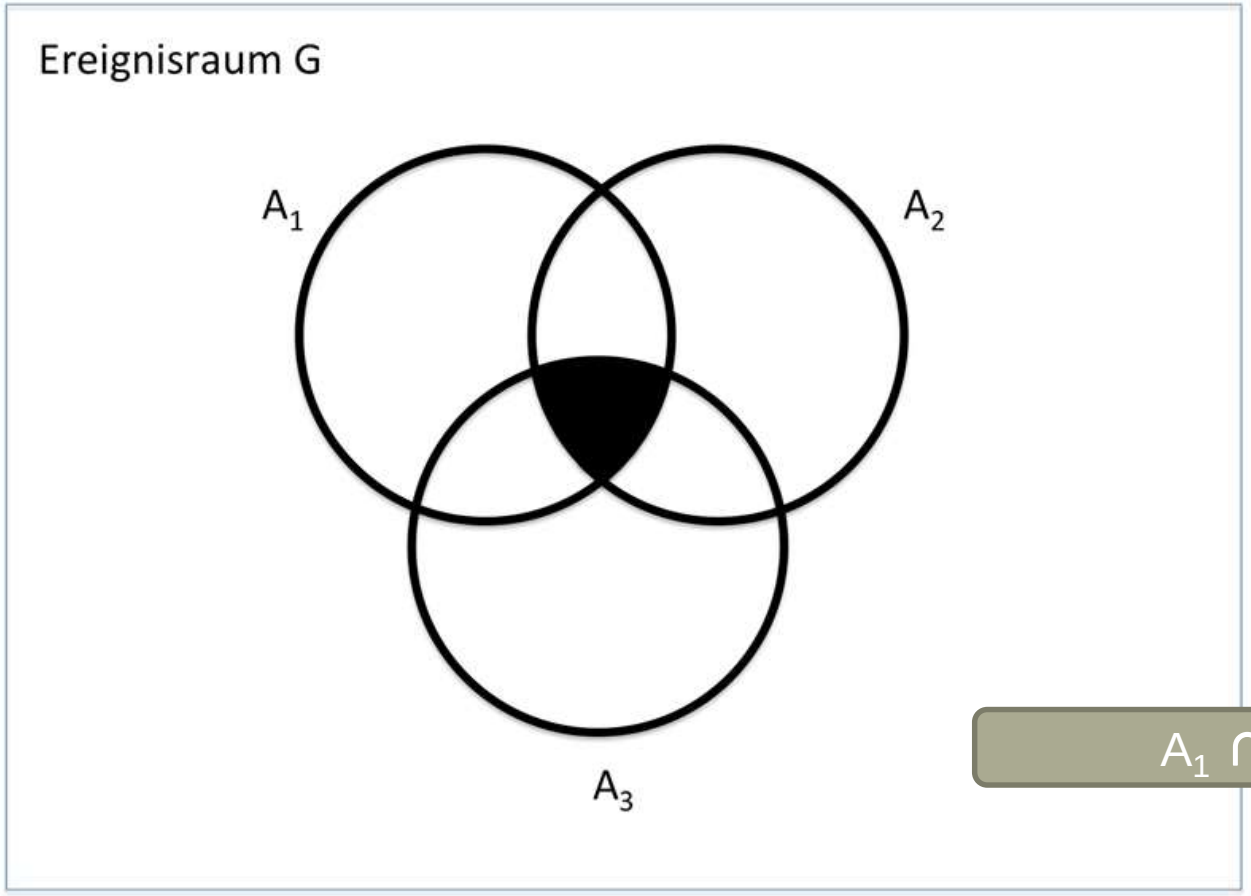
- Die Menge Ω = [1; 2; 3; 4; 5; 6; 7; 8; 9; 10] verfügt über drei Teilmengen
  - Menge der geraden Zahlen A = [2; 4; 6; 8; 10]
  - Menge der ungeraden Zahlen B = [1; 3; 5; 7; 9;]
  - Menge der zweistelligen Zahlen C = [10]
- Die nachfolgenden Beispiele verdeutlichen die Anwendung der Operatoren
  - $A \cap B = B \cap A = \emptyset$
  - $B \cap C = C \cap B = \emptyset$
  - $A \cap C = C \cap A = [10]$
  - $(A \cap B) \cap C = A \cap (B \cap C) = \emptyset$
  - $(A \cap B) \cup C = (A \cup C) \cap (B \cup C) = \emptyset$
  - $(A \cup B) \cap C = (A \cap C) \cup (B \cap C) = [10]$
  - $A \cup B = B \cup A = [1; 2; 3; 4; 5; 6; 7; 8; 9; 10]$
  - $(A \cup B) \cup C = A \cup (B \cup C) = [1; 2; 3; 4; 5; 6; 7; 8; 9; 10]$

Mengenvisualisierung mit Venn-Diagrammen



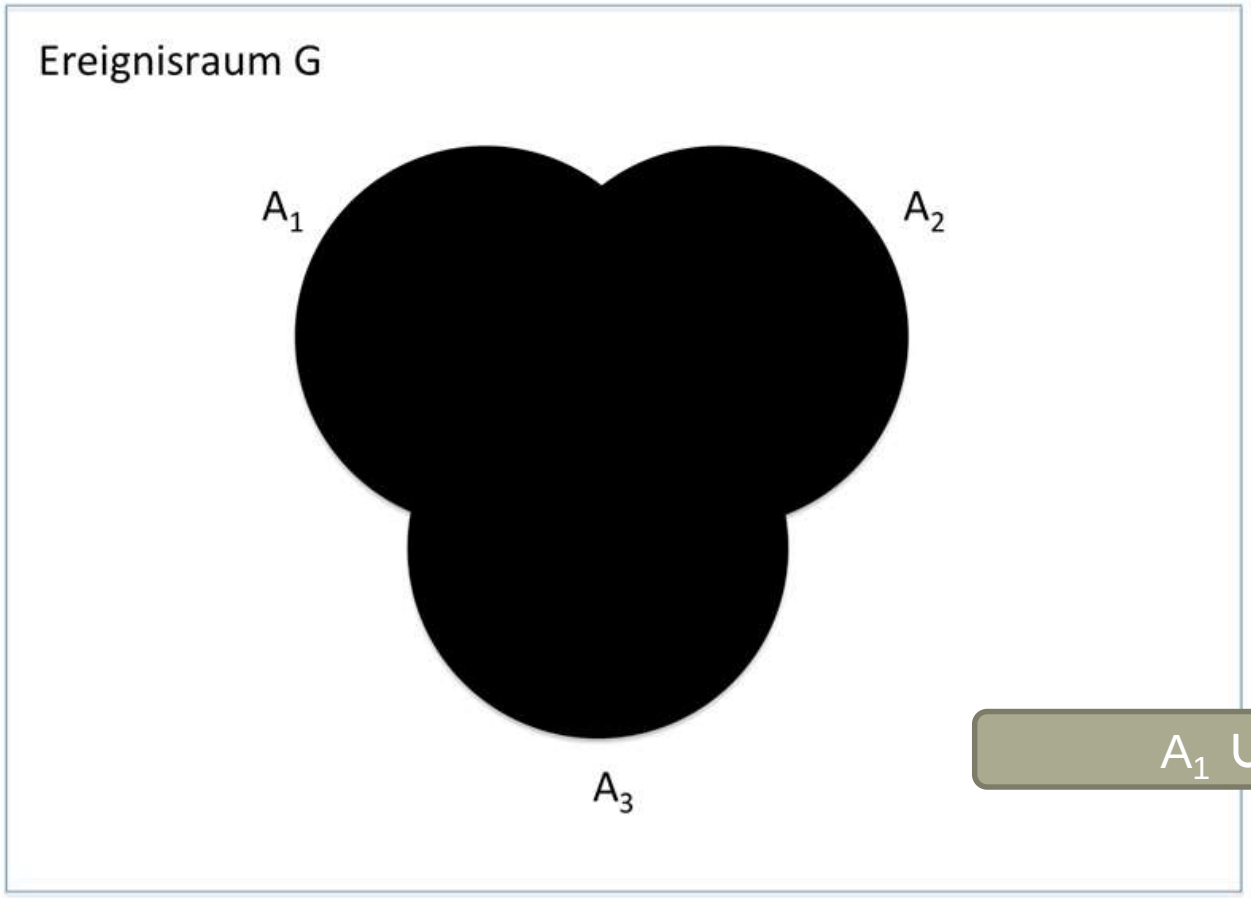
Bleiverglastes Fenster mit einem Venn-Diagramm in Venns Studienort Cambridge (Quelle: WikiMedia; User: Schutz; Lizenz: CC BY-SA 2.5)

Beispiel: Konstruktion von Venn-Diagrammen



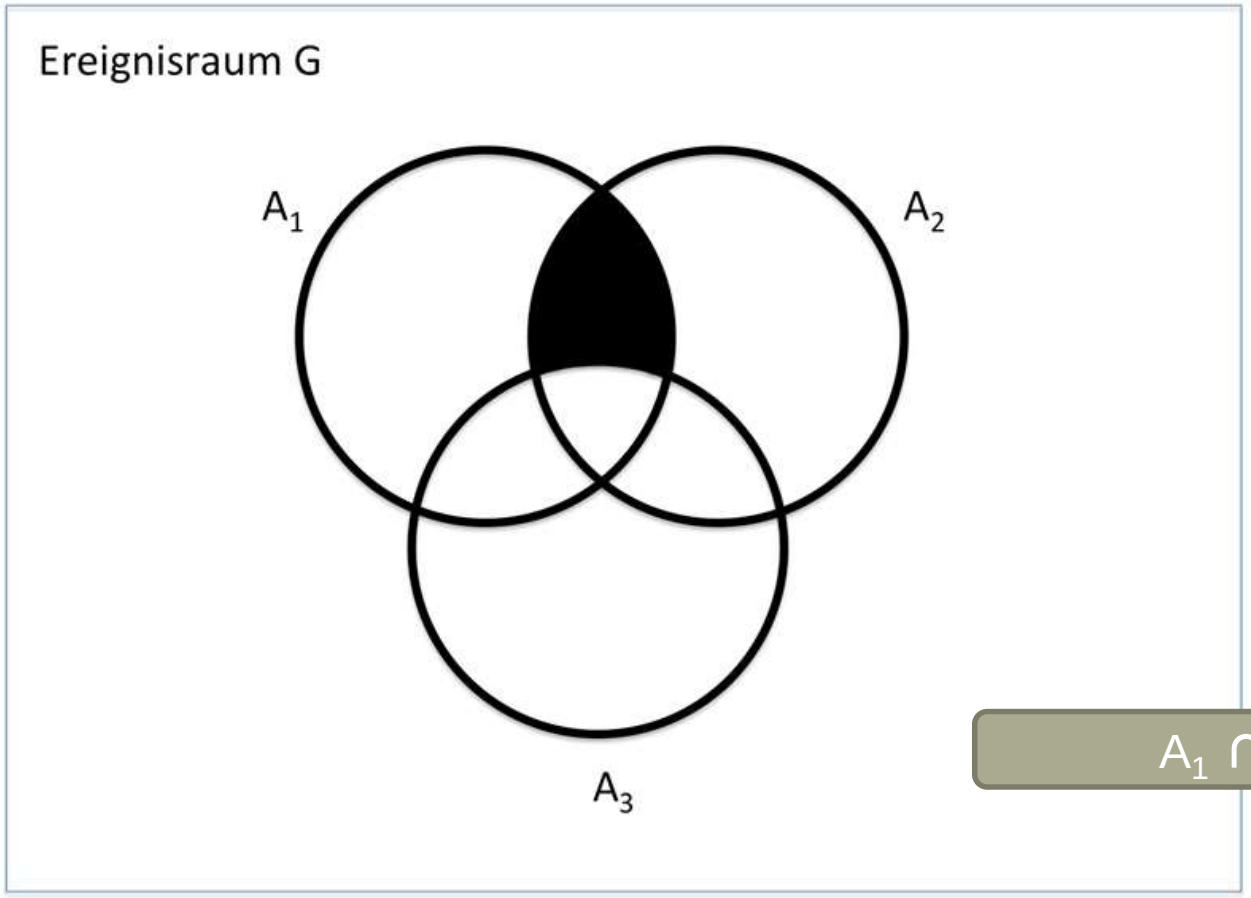
$A_1 \cap A_2 \cap A_3$

Beispiel: Konstruktion von Venn-Diagrammen



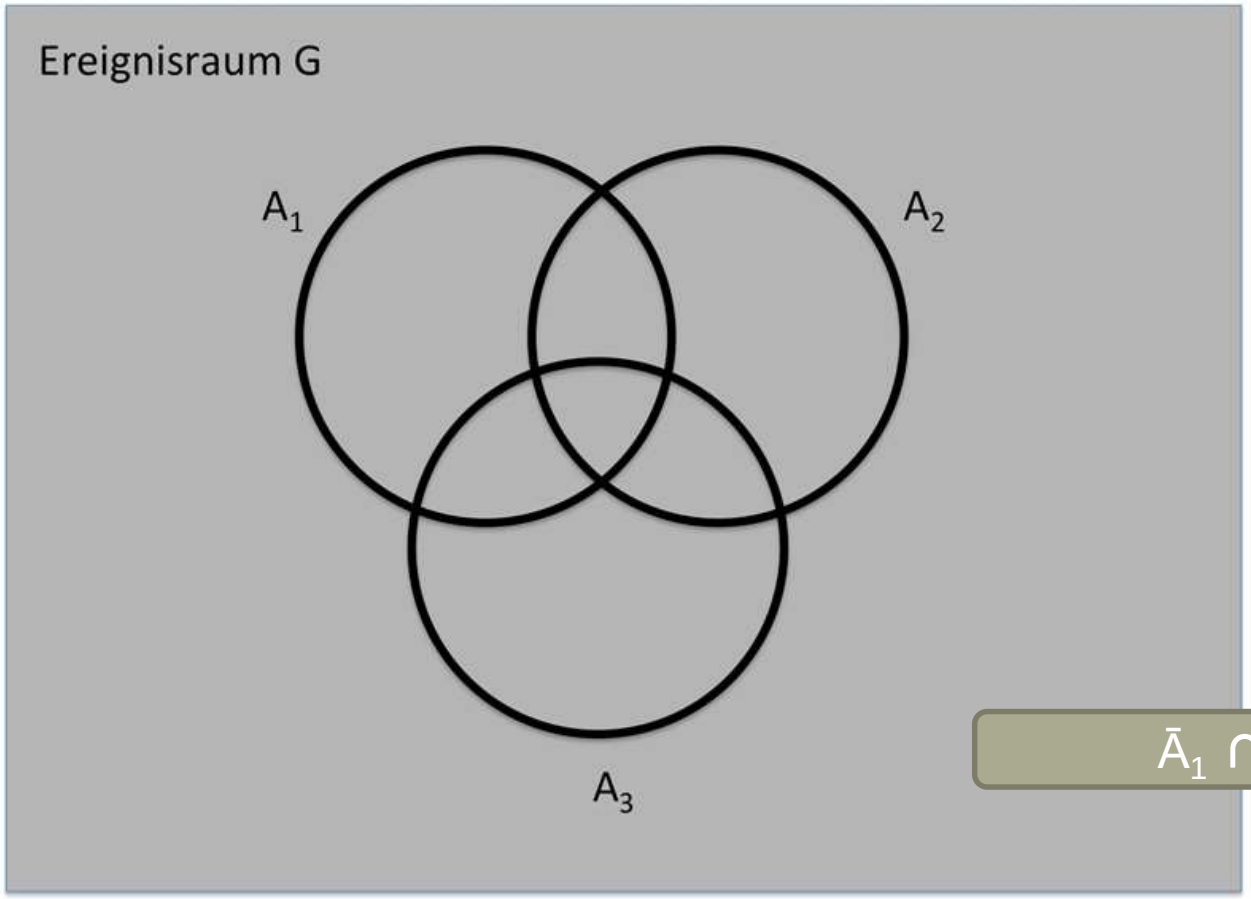
$A_1 \cup A_2 \cup A_3$

Beispiel: Konstruktion von Venn-Diagrammen



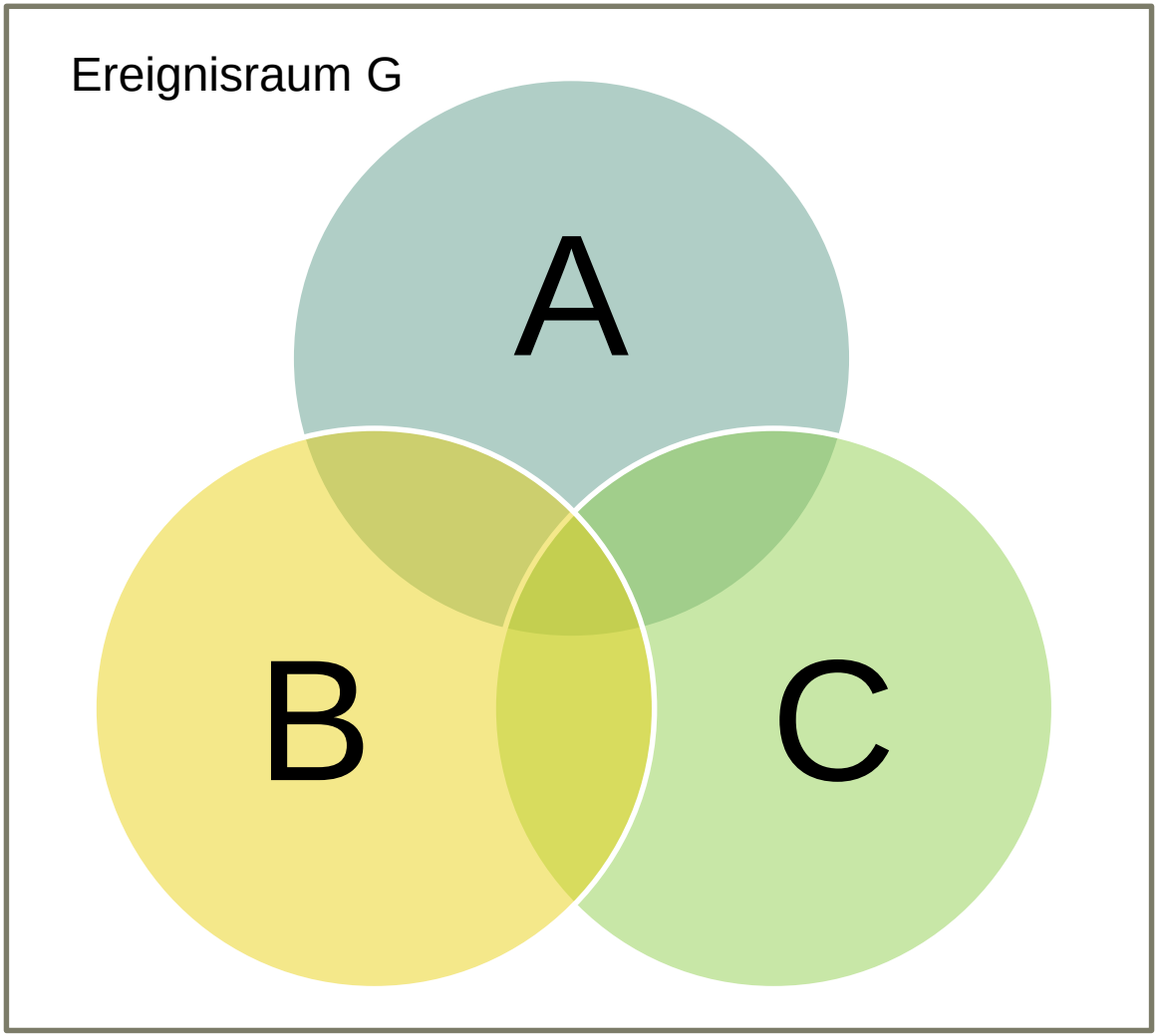
$$A_1 \cap A_2 \cap \bar{A}_3$$

Beispiel: Konstruktion von Venn-Diagrammen



$\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3$

Mengenvisualisierung mit Venn-Diagrammen



Welche Fläche entspricht…?

$A \cap B$

$A \cap C$

$A \cup B$

$A \cup B \cup C$

$A \cap B \cap C$

$\bar{A}$

$\bar{A} \cap B$

Teil VIII

Wahrscheinlichkeitslehre

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Der klassische Wahrscheinlichkeitsbegriff

– Besitzt ein Zufallsvorgang A endlich viele Elementarereignisse und verfügt jedes dieser Ereignisse über die gleiche Eintrittschance, berechnet man die **Wahrscheinlichkeit für das Eintreten eines bestimmten Ereignisses P(A)** (das aus mehreren Elementarereignissen bestehen kann) nach Laplace wie folgt:

  P (A) = Σ für A günstiger Elementarereignisse / Σ möglicher Elementarereignisse

– Die Wahrscheinlichkeit auf eine 3 beim einmaligen Würfeln liegt daher bei:

  P(3) = [3] / [1; 2; 3; 4; 5; 6] = 1 / 6 = 0,167 = 16,7%

– Die Wahrscheinlichkeit auf eine gerade Zahl beim Würfen liegt dagegen bei:

  P (gerade Zahl) = [2; 4; 6] / [1; 2; 3; 4; 5; 6] = 3 / 6 = 0,5 = 50%

Einige Laplace-Wahrscheinlichkeiten

- Wahrscheinlichkeit für „Kopf“ beim Münzwurf: $\frac{[K]}{[Z,K]} = \frac{1}{2}$
- Wahrscheinlichkeit für eine ungerade Zahl beim Würfeln: $\frac{[1,3,5]}{[1,2,3,4,5,6]} = \frac{3}{6} = \frac{1}{2}$
- Wahrscheinlichkeit für eine gerade Zahl beim Würfeln: $\frac{[2,4,6]}{[1,2,3,4,5,6]} = \frac{3}{6} = \frac{1}{2}$
- Wahrscheinlichkeit für eine Summe > 4 beim Würfeln: $\frac{[5,6]}{[1,2,3,4,5,6]} = \frac{2}{6} = \frac{1}{3}$
- Wahrscheinlichkeit für sechs Richtige in der Lotterie: $\frac{1}{?}$

Woher nehmen wir den Nenner?

Einige Ereignisse und Gegenereignisse

- Wahrscheinlichkeit für eine 3 beim Würfelwurf: $\frac{[3]}{[1,2,3,4,5,6]} = \frac{1}{6}$
- Gegenereignis zu einer 3 beim Würfelwurf: $\frac{[1,2,4,5,6]}{[1,2,3,4,5,6]} = \frac{5}{6}$
- Wahrscheinlichkeit für mindestens eine 3 beim Würfelwurf: $\frac{[3,4,5,6]}{[1,2,3,4,5,6]} = \frac{4}{6} = \frac{2}{3}$
- Gegenereignis zu mindestens einer 3 beim Würfelwurf: $\frac{[1,2]}{[1,2,3,4,5,6]} = \frac{2}{6} = \frac{1}{3}$

Warum ist das Gegenereignis zu „mindestens 3“ nicht „höchstens 3“, sondern „höchstens 2“?

Weitere Wahrscheinlichkeitsbegriffe

- **Frequentistischer Wahrscheinlichkeitsbegriff:** Ableitung von a priori nicht bekannten Wahrscheinlichkeiten aus vergangenen Erfahrungen

  - Beispiel: Wenn 8 der letzten 10 neu auf den Markt gebrachten Digitalkameras einen Produktlebenszyklus von unter 6 Monaten hatten, kann mit 80% Wahrscheinlichkeit davon ausgegangen werden, dass sich dies bei einem neuen Modell ebenso verhält

    (nur möglich, wenn sich die Vorgänge nicht gegenseitig beeinflussen)

- **Subjektiver Wahrscheinlichkeitsbegriff:** Subjektiv durch Personen (auf Basis von (Teil-) Daten oder „Bauchgefühl“) vorgenommene Wahrscheinlichkeitsschätzungen

- **Im Rahmen dieser Vorlesung wird nachfolgend nur noch der klassische Wahrscheinlichkeitsbegriff nach Pierre de Laplace von Bedeutung sein**

Die drei Axiome von Kolmogorov

- **Axiom 1: Die Wahrscheinlichkeit eines Ereignisses A eines Zufallsvorgangs ist eine nichtnegative reelle Zahl**

  $$P(A) \geq 0$$

  (Die Wahrscheinlichkeit eines Ereignisses darf nicht < 0 sein)

- **Axiom 2: Die Wahrscheinlichkeiten aller möglichen Elementarereignisse eines Zufallsvorgangs ergeben zusammen den Wert 1**

  $$P(\Omega) = 1$$

  (Die Wahrscheinlichkeit aller Ereignisse darf nicht > 1 sein)

- **Axiom 3: Die Wahrscheinlichkeit der Vereinigungsmenge zweier oder mehrerer Ereignisse eines Zufallsvorgangs berechnet sich aus der Summe der Einzelwahrscheinlichkeiten der Ereignisse, wenn diese paarweise disjunkt sind**

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

  falls

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

Was verraten uns die drei Axiome?

- **Axiom 1: Die Wahrscheinlichkeit eines Ereignisses A eines Zufallsvorgangs ist eine nichtnegative reelle Zahl**

$$P(A) \geq 0$$

„Die Wahrscheinlichkeit, eine 6 zu würfeln, liegt bei -16,7 %“

„Die Wahrscheinlichkeit, eine 6 zu würfeln, liegt bei 16,7%“

- **Axiom 2: Die Wahrscheinlichkeiten aller möglichen Elementarereignisse eines Zufallsvorgangs ergeben zusammen den Wert 1**

$$P(\Omega) = 1$$

„Die Wahrscheinlichkeit, eine gerade Zahl zu würfeln, liegt bei 120%“

„Die Wahrscheinlichkeit, eine gerade Zahl zu würfeln, liegt bei 50%“

„Die Wahrscheinlichkeit, eine Zahl zwischen 1 und 6 zu würfeln, liegt bei 100%“

Was verraten uns die drei Axiome?

– **Axiom 3: Die Wahrscheinlichkeit der Vereinigungsmenge zweier oder mehrerer Ereignisse eines Zufallsvorgangs berechnet sich aus der Summe der Einzelwahrscheinlichkeiten der Ereignisse, wenn diese paarweise disjunkt sind**

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

falls

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

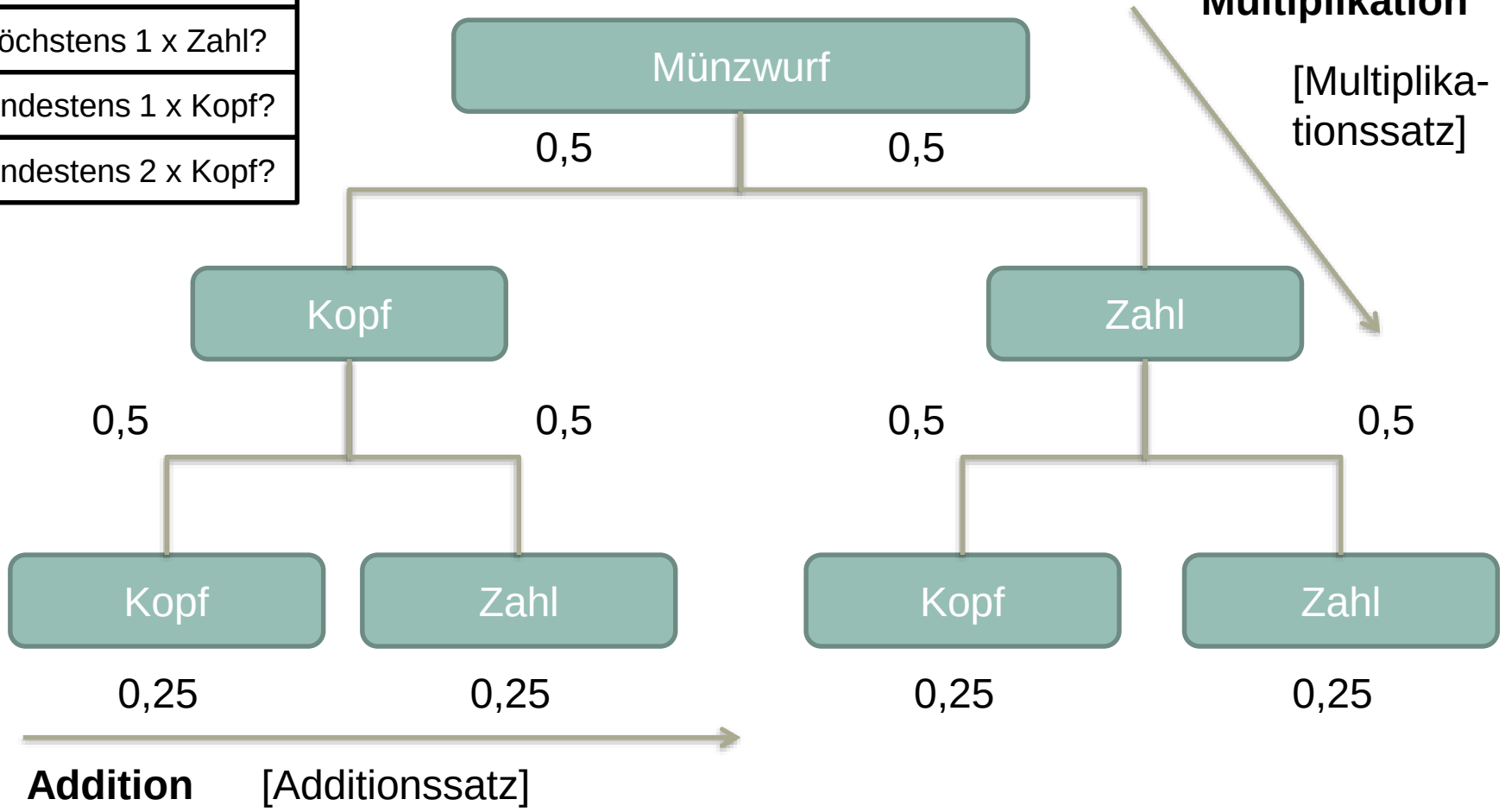
(auch bekannt als: Additivität bei disjunkten Ereignissen)

„Die Wahrscheinlichkeit, eine Zahl kleiner 3 oder eine Zahl kleiner 2 zu würfeln, liegt bei [P(2) + P(1)] + [P(1)] = [1/6 + 1/6] + [1/6] = 3/6 = 1/2 = 50%“

„Die Wahrscheinlichkeit, eine gerade Zahl zu Würfeln, liegt bei P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2 = 50%“

Pfaddiagramme von Zufallsexperimenten

Genau 1 x Zahl?
Höchstens 1 x Zahl?
Mindestens 1 x Kopf?
Mindestens 2 x Kopf?



Auch im Pfaddiagramm findet sich Laplace

– Klassische Wahrscheinlichkeitsdefinition nach Laplace:

$$P(A) = \frac{\sum \text{für } A \text{ günstiger Elementarereignisse}}{\sum \text{möglicher Elementarereignisse}}$$

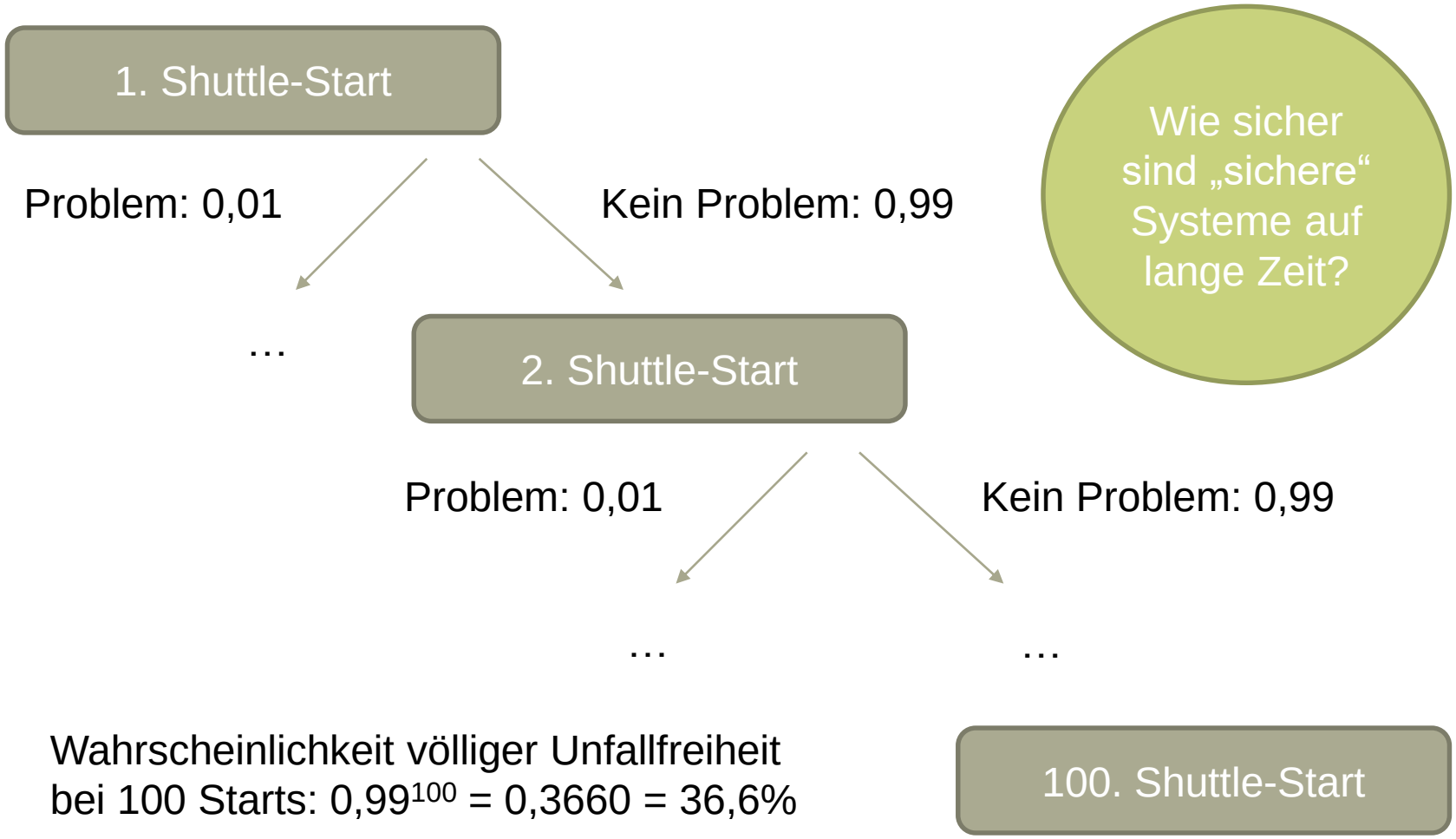
– Wahrscheinlichkeit für mindestens 1 x Zahl beim zweifachen Münzwurf:

$$P(A) = \frac{(Z;K);(K;Z);(Z;Z)}{(Z;K);(K;Z);(Z;Z);(K;K)} = \frac{3}{4} = 0{,}75 = 75\%$$

Absturz der Challenger am 28.01.1986
(Quelle: WikiMedia; Lizenz: gemeinfrei)

„The Challenger Disaster“ (BBC, 2013)
über die Arbeit der Rogers-Kommission

„Die Chance auf ein Versagen liegt bei nur 1%“



"The probability of a train derailment was infinitesimal. That meant it was only a matter of time."

N. K. Jemisin

Additions- und Multiplikationssätze

- Sind zwei Ereignisse A und B miteinander unvereinbar (disjunkt, d.h. ohne eine Schnittmenge), so gilt für sie der **Additionssatz für unvereinbare Ereignisse**:

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

- Können zwei Ereignisse A und B auch über eine Schnittmenge verfügen (nicht disjunkt), so gilt für sie der **Additionssatz für beliebige Ereignisse**:

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

Warum der Abzug?

- Sind zwei Ereignisse stochastisch unabhängig, d.h. beeinflusst das Eintreten eines Ereignisses nicht die Wahrscheinlichkeit des Eintretens des anderen Ereignisses, so gilt für sie der **Multiplikationssatz bei stochastischer Unabhängigkeit**:

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

- Liegt keine stochastische Unabhängigkeit vor, spricht man von einer **bedingten Wahrscheinlichkeit** (z.B. der Wahrscheinlichkeit von B unter der Bedingung, dass zuvor A eintritt) – den Umgang damit lernen wir im Kurs noch kennen

Übung: Rechnen mit den A- und M-Sätzen

- Zwei Sachbearbeiter suchen unabhängig voneinander nach Belegen für eine (unstrittige) Steuerhinterziehung in den gleichen Unterlagen, wobei jeder von ihnen mit einer Trefferquote von 0,4 arbeitet. Wie groß ist die Chance dafür, dass mindestens einer der beiden den erforderlichen Beweis findet?

- Zur Lösung dieser Aufgabe werden der **Additionssatz für beliebige Ereignisse** (es kann ja der Fall eintreten, dass beide Sachbearbeiter fündig werden) und der **Multiplikationssatz bei stochastischer Unabhängigkeit** (die Sachbearbeiter beeinflussen sich bei ihrer Suche nicht gegenseitig) benötigt

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

Additionssatz

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

Multiplikationssatz

(alternativ ist die Lösung natürlich auch über ein Pfaddiagramm möglich)

Übung: Rechnen mit den A- und M-Sätzen

- Zwei Sachbearbeiter suchen unabhängig voneinander nach Belegen für eine (unstrittige) Steuerhinterziehung in den gleichen Unterlagen, wobei jeder von ihnen mit einer Trefferquote von 0,4 arbeitet. Wie groß ist die Chance dafür, dass mindestens einer der beiden den erforderlichen Beweis findet?

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

$$P(A \cup B) = 0{,}4 + 0{,}4 - P(A \cap B)$$

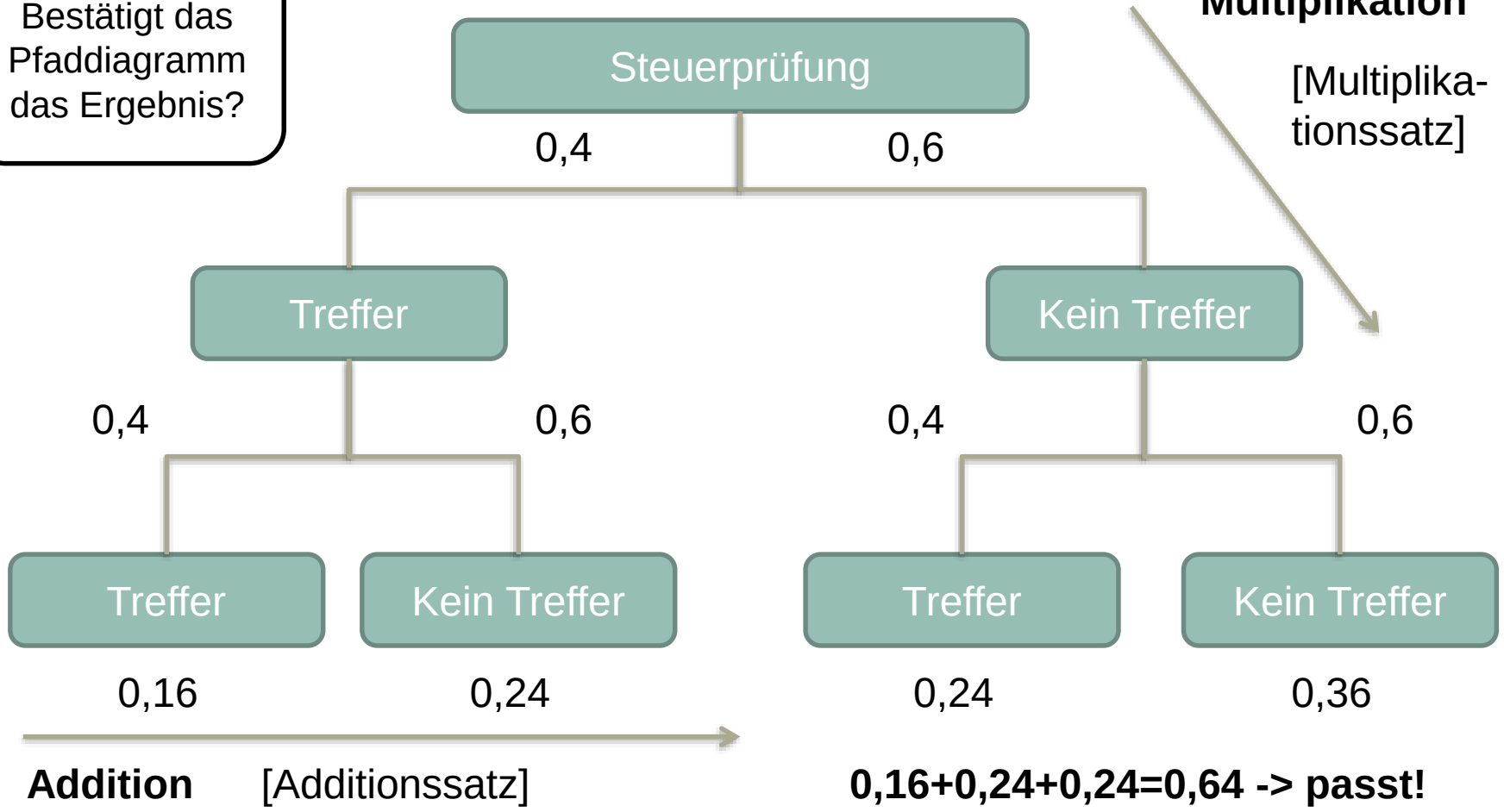
$$P(A \cap B) = P(A) * P(B)$$

$$P(A \cap B) = 0{,}4 * 0{,}4 = 0{,}16$$

$$P(A \cup B) = 0{,}4 + 0{,}4 - 0{,}16 = 0{,}64$$

Übung: Rechnen mit den A- und M-Sätzen

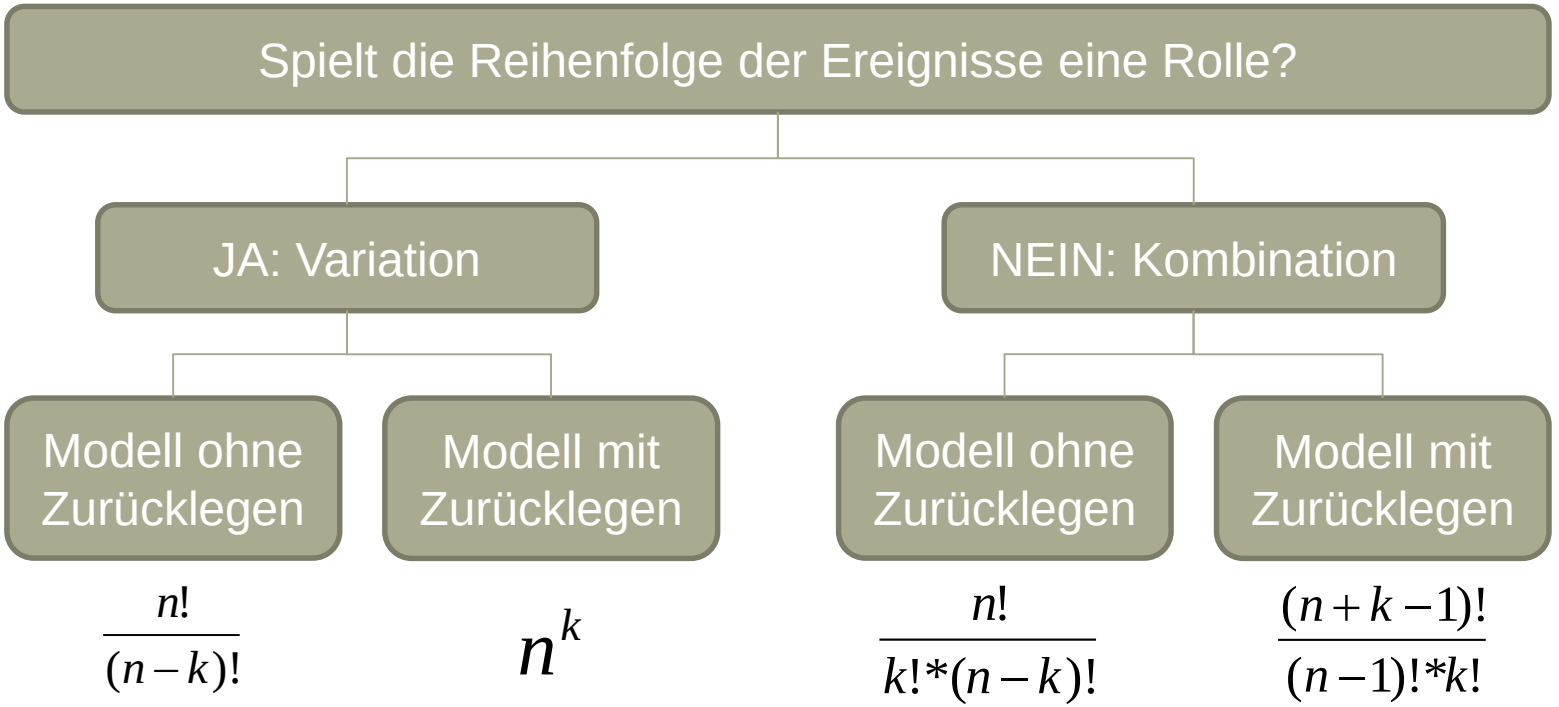
Bestätigt das Pfaddiagramm das Ergebnis?



Kombinatorik: Wie viele Möglichkeiten gibt es?

Kernproblem: Um mit der Laplace-Wahrscheinlichkeit rechnen zu können, muss die Anzahl der günstigen sowie die Anzahl der möglichen Ereignisse bekannt sein – wie berechnen sich diese unter verschiedenen Rahmenbedingungen?

(Beispiel: Wie viele Möglichkeiten gibt es, um einen Lotto-Schein auszufüllen?)



Variation – Modell ohne Zurücklegen

- Wann spricht man von einer **Variation – Modell ohne Zurücklegen**?
  - Auswahl von Objekten (Ereignissen) in einer bestimmten **Reihenfolge**
  - Jedes Objekt (Ereignis) kann dabei **nur ein Mal auftreten** (eintreten)

- Beispiel: Berechnung der Anzahl möglicher 4-stelliger PIN-Kombinationen (k) aus 10 Ziffern (n), wenn jede Ziffer pro PIN maximal ein Mal auftreten kann

$$\frac{n!}{(n-k)!} = \frac{10!}{(10-4)!} = \frac{3628800}{720} = 5040$$

Wie viele Reihenfolgen gibt es, in denen k aus n Elementen angeordnet werden können, wenn jedes Element nur ein Mal gezogen werden kann?

Kurze Wiederholung: 6! (gesprochen „6 Fakultät“) = 6 * 5 * 4 * 3 * 2 * 1 = 720

10*9*8*7 = 5040 Warum?

Variation – Modell ohne Zurücklegen

– Einen Sonderfall stellt die **Permutation** bei Auswahl aller Objekte (n = k) dar:

$$\frac{n!}{(n-k)!} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n!$$

Wie viele Reihenfolgen gibt es, in denen n Elemente angeordnet werden können?

– Rechenlogik im Sonderfall (PIN mit 10 aus 10 Ziffern ohne Zurücklegen)

  – Für die erste Stelle der PIN kommen insgesamt 10 Ziffern in Frage
  – Für die zweite Stelle der PIN kommen nun noch 9 Ziffern in Frage
  – Für die dritte Stelle der PIN kommen nun noch 8 Ziffern in Frage
  – Für die vierte Stelle der PIN kommen nun noch 7 Ziffern in Frage
  – Für die fünfte Stelle der PIN kommen nun noch 6 Ziffern in Frage…

  – 10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1 = 10! = 3.628.800

Variation – Modell ohne Zurücklegen

– Aus einer Urne mit 3 Kugeln (A, B, C) werden 2 Kugeln gezogen

Nummer	Anordnung	Wird die Anordnung gezählt?
1	A, B	JA
2	A, C	JA
3	B, A	JA
4	B, C	JA
5	C, A	JA
6	C, B	JA

$$\frac{n!}{(n-k)!} = \frac{3!}{(3-2)!} = \frac{6}{1} = 6$$

Variation – Modell mit Zurücklegen

- Wann spricht man von einer **Variation – Modell mit Zurücklegen**?
  - Auswahl von Objekten (Ereignissen) in einer bestimmten **Reihenfolge**
  - Jedes Objekt (Ereignis) kann dabei **mehrere Male auftreten** (eintreten)

- Beispiel: Berechnung der Anzahl möglicher 4-stelliger PIN-Kombinationen (k) aus 10 Ziffern (n), wenn jede Ziffer pro PIN beliebig häufig auftreten kann

$$n^k = 10^4 = 10000$$

Wie viele Reihenfolgen gibt es, in denen k aus n Elementen angeordnet werden können, wenn jedes Element beliebig oft (bzw. maximal k-mal) gezogen werden kann?

  - Für die erste Stelle der PIN kommen insgesamt 10 Ziffern in Frage
  - Für alle weiteren Stellen kommen ebenfalls noch 10 Ziffern in Frage

  - $10 * 10 * 10 * 10 = 10^4$

Variation – Modell mit Zurücklegen

– Aus einer Urne mit 3 Kugeln (A, B, C) werden 2 Kugeln gezogen

Nummer	Anordnung	Wird die Anordnung gezählt?
1	A, B	JA
2	A, C	JA
3	B, A	JA
4	B, C	JA
5	C, A	JA
6	C, B	JA
7	A, A	JA
8	B, B	JA
9	C, C	JA

$$n^k = 3^2 = 9$$

Kombination – Modell ohne Zurücklegen

- Wann spricht man von einer **Kombination – Modell ohne Zurücklegen**?
  - Auswahl von Objekten (Ereignissen) ohne Beachtung der **Reihenfolge**
  - Jedes Objekt (Ereignis) kann dabei **nur ein Mal auftreten** (eintreten)

- Beispiel: Berechnung der möglichen Kombinationen beim Lotto (6 aus 49, Ziehen ohne Zurücklegen, die Reihenfolge spielt beim Gewinn keine Rolle)

$$\frac{n!}{k!*(n-k)!} = \frac{49!}{6!*(49-6)!} = 13983816$$

Dieser Term wird auch als Bionomialkoeffizient bezeichnet (nCr-Taste auf vielen Taschenrechnern)

- Die Wahrscheinlichkeit auf einen Hauptgewinn in der Lotterie liegt nach der klassischen Definition von Laplace also bei 1 / 13.983.816 = 0,000000715%

Kombination – Modell ohne Zurücklegen

– Aus einer Urne mit 3 Kugeln (A, B, C) werden 2 Kugeln gezogen

Nummer	Anordnung	Wird die Anordnung gezählt?
1	A, B	JA
2	A, C	JA
3	B, A	NEIN (bereits in 1 gezählt)
4	B, C	JA
5	C, A	NEIN (bereits in 2 gezählt)
6	C, B	NEIN (bereits in 4 gezählt)

$$\frac{n!}{k!*(n-k)!} = \frac{3!}{2!*(3-2)!} = \frac{6}{2} = 3$$

Kombination – Modell mit Zurücklegen

– Wann spricht man von einer **Kombination – Modell mit Zurücklegen**?

  – Auswahl von Objekten (Ereignissen) ohne Beachtung der **Reihenfolge**
  – Jedes Objekt (Ereignis) kann dabei **mehrere Male auftreten** (eintreten)

– Beispiel: Aus einer Urne mit 10 nummerierten Kugeln wird 3 Mal eine Kugel gezogen, wobei die gezogene Kugel jedes Mal wieder zurückgelegt wird. Wie viele Kombinationsmöglichkeiten für Kugeln ergeben sich?

$$\frac{(n+k-1)!}{(n-1)!*k!} = \frac{(10+3-1)!}{(10-1)!*3!} = \frac{479001600}{362880*6} = \frac{479001600}{2177280} = 220$$

Wie viele Möglichkeiten gibt es, k aus n Elementen zu kombinieren, wenn die Elemente immer wieder neu gezogen werden können?

Kombination – Modell mit Zurücklegen

– Aus einer Urne mit 3 Kugeln (A, B, C) werden 2 Kugeln gezogen

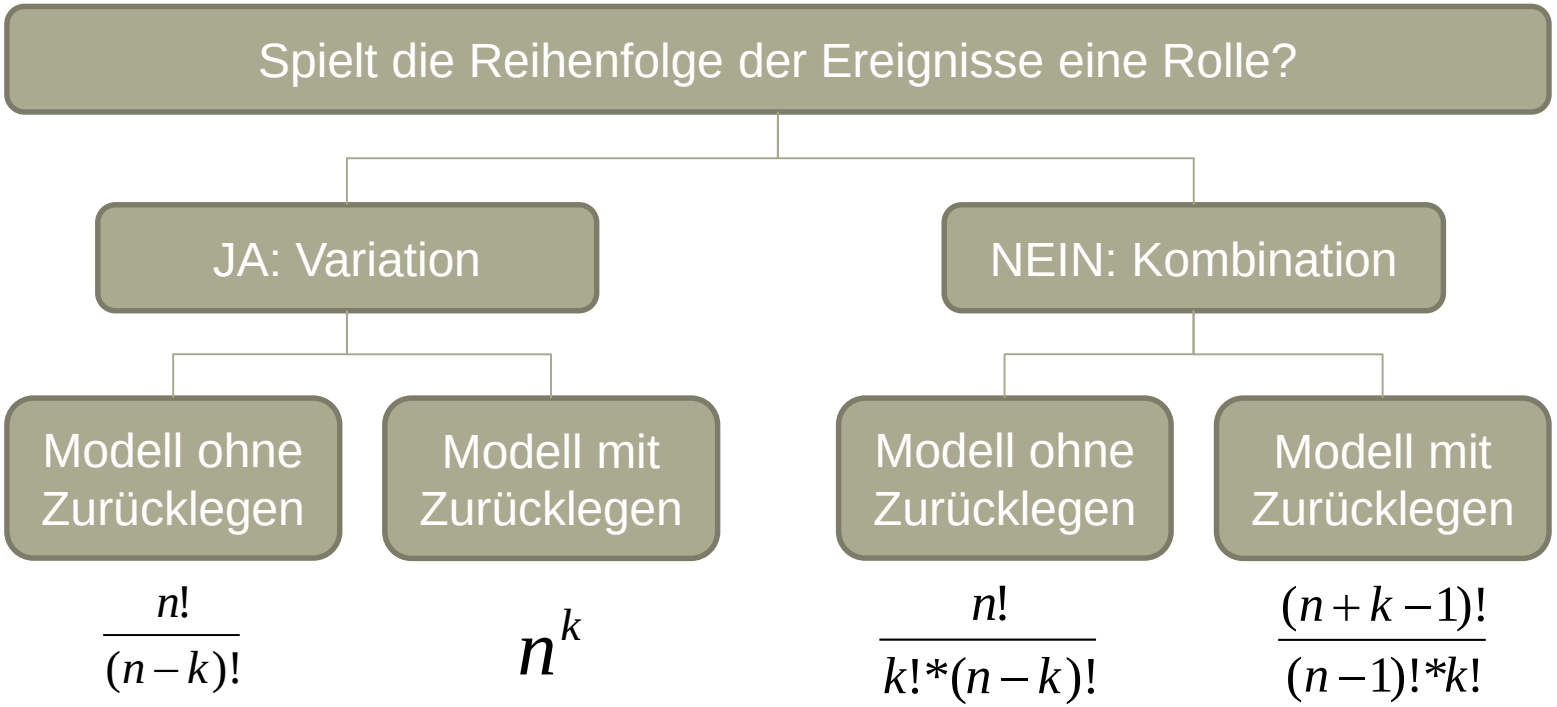
Nummer	Anordnung	Wird die Anordnung gezählt?
1	A, B	JA
2	A, C	JA
3	B, A	NEIN (bereits in 1 gezählt)
4	B, C	JA
5	C, A	NEIN (bereits in 2 gezählt)
6	C, B	NEIN (bereits in 4 gezählt)
7	A, A	JA
8	B, B	JA
9	C, C	JA

$$\frac{(n+k-1)!}{(n-1)!*k!} = \frac{(3+2-1)!}{(3-1)!*2!} = \frac{24}{2*2} = \frac{24}{4} = 6$$

Kombinatorik: Wie viele Möglichkeiten gibt es?

Kernproblem: Um mit der Laplace-Wahrscheinlichkeit rechnen zu können, muss die Anzahl der günstigen sowie die Anzahl der möglichen Ereignisse bekannt sein – wie berechnen sich diese unter verschiedenen Rahmenbedingungen?

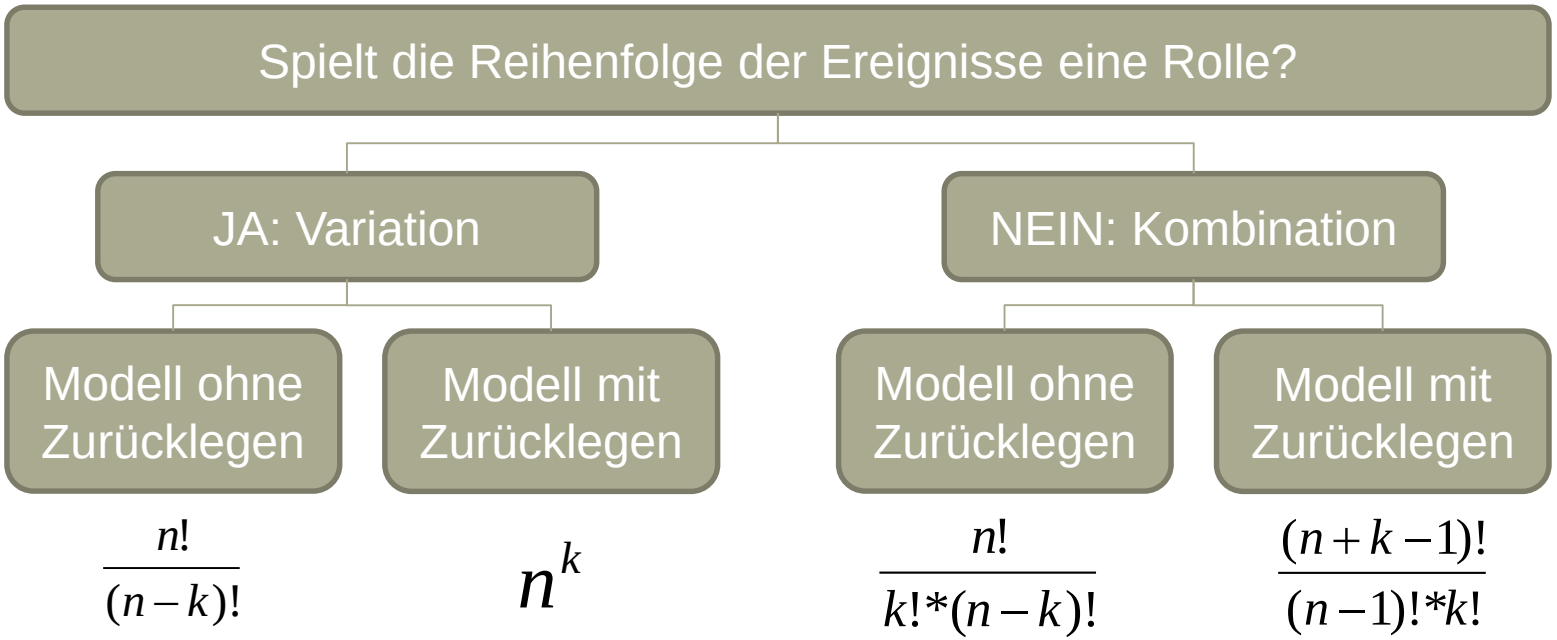
(Beispiel: Wie viele Möglichkeiten gibt es, um einen Lotto-Schein auszufüllen?)



Übung: Wie viele Möglichkeiten gibt es?

– Wie viele Möglichkeiten für eine vierstellige PIN existieren, wenn...

  – ...keine der vier Ziffern bekannt ist?
  – ...bekannt ist, dass eine der vier Ziffern eine 6 ist?
  – ...bekannt ist, dass die Ziffer 6 an erster Stelle steht?



Übung: Wie viele Möglichkeiten gibt es?

- Wie viele Möglichkeiten für eine vierstellige PIN existieren, wenn...
  - ...keine der vier Ziffern bekannt ist?
  - ...bekannt ist, dass eine der vier Ziffern eine 6 ist?
  - ...bekannt ist, dass die Ziffer 6 an erster Stelle steht?

Erste Annahme: Es müssten immer weniger Möglichkeiten werden...

- In diesem Fall liegt eine Variation (die Reihenfolge der Ziffern spielt bei Eingabe der PIN eine Rolle) mit Zurücklegen (alle Ziffern können mehrfach auftreten) vor
  - Wenn keine Ziffer bekannt ist: $n^k = 10^4 = 10000$
  - Wenn bekannt ist, dass die PIN eine 6 enthält: $4 * n^k = 4 * 10^3 = 4000$
  - Wenn bekannt ist, dass die 6 an erster Stelle steht: $n^k = 10^3 = 1000$

Rechnen mit bedingten Wahrscheinlichkeiten

– Bisherige Grundannahme: Ereignisse treten unabhängig voneinander ein – d.h. welche Zahl gewürfelt wurde, wirkt sich nicht auf den nächsten Würfelwurf aus

– Neue Grundannahme: Die Wahrscheinlichkeit des Eintretens eines Ereignisses A hängt von der Wahrscheinlichkeit des Eintretens eines vorherigen Ereignisses B ab

– Die bedingte Wahrscheinlichkeit von A unter der Bedingung B ist definiert als

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$ Was wiederum umgeformt werden kann zu

$$P(A \cap B) = P(A \mid B) * P(B) \quad \text{für} \quad P(B) > 0$$

Sind A und B stochastisch unabhängig voneinander, so wird vereinfacht zu

$$P(A \mid B) = P(A) \quad \text{und} \quad P(A \cap B) = P(A) * P(B)$$

Übung: Würfeln mit zwei Würfeln

- Wie groß ist (nach Laplace) die Wahrscheinlichkeit, beim gleichzeitigen Würfeln mit zwei Würfeln eine Gesamtzahl größer als 8 zu erzielen?
  - Von 36 Kombinationen (6 * 6) erfüllen nur 10 diese Bedingung
  - Die Wahrscheinlichkeit liegt also bei 10 / 36 = 0,278 = 27,8%
- Würfelt man nacheinander, kennt man das Ergebnis des ersten Wurfs bereits. Handelt es sich um eine 4, stellt sich die Frage, wie groß die Chance auf eine Augenzahl größer 8 nun unter dieser Bedingung ist
  - Dies wäre der Fall, wenn der zweite Würfel mindestens eine 5 zeigt

Woher kommen die 2/6?

$$P(S>8 \mid W_1=4)\frac{P(S>8 \cap W_1=4)}{P(W_1=4)} = \frac{\frac{2}{6} * \frac{1}{6}}{\frac{1}{6}} = \frac{1}{3} = 33{,}3\%$$

Satz der totalen Wahrscheinlichkeit

– Bilden die Ereignisse A_1, A_2, … A_k überschneidungsfrei (disjunkt) einen vollständigen Ereignisraum Ω, so gilt für ein Ereignis B ⊃ Ω der Satz der totalen Wahrscheinlichkeit

$$P(B) = \sum_{i=1}^{k} P(B \mid A_i) * P(A_i)$$

– Anwendungsbeispiel: Drei Maschinen (A_1, A_2, A_3) stellen Bauteile mit einer Fehlerrate von A_1 = 0,02, A_2 = 0,04 und A_3 = 0,03 her. Aus Kapazitätsgründen werden mit A_1 50%, mit A_2 30% und mit A_3 20% der Bauteile produziert. Wie groß ist die Wahrscheinlichkeit, ein fehlerhaftes Bauteil zu erhalten?

$$P(Fehler) = \sum_{i=1}^{3} P(Fehler \mid Maschine) * P(Maschine)$$

$$P(Fehler) = (0{,}02 * 0{,}5) + (0{,}04 * 0{,}3) + (0{,}03 * 0{,}2) = 0{,}028 = 2{,}8\%$$

Rechnen mit dem Satz von Bayes

- Das berühmte „Taxi-Problem“ wurde erstmalig von Arthur Engel formuliert

  - In einer Stadt existieren zwei Taxi-Firmen: Green Cab und Blue Cab
  - Der Marktanteil von Green Cab (mit grünen Fahrzeugen) liegt bei 85%
  - Der Marktanteil von Blue Cab (mit blauen Fahrzeugen) liegt bei 15%

  - Es kommt zu einem Unfall mit Fahrerflucht und einem einzigen Zeugen
  - Der Zeuge hat (unstrittig) ein Taxi gesehen und glaubt (strittig), dass es ein blaues Taxi war – **aber wie hoch ist die Zuverlässigkeit dieser Aussage?**

  - Das Gericht ordnet einen Sehtest an, bei dem sich herausstellt, dass der Zeuge die Farbe von Fahrzeugen bei Nacht mit 80%iger Wahrscheinlichkeit korrekt erkennt – **war der Unfallwagen also mit 80%iger Sicherheit blau?**

Rechnen mit dem Satz von Bayes

– Viele Probanden antworten so – aber warum ist diese Annahme falsch?

  – Es bleibt unberücksichtigt, dass die meisten Taxen grün und nicht blau sind
  – Die Wahrscheinlichkeit, dass der Zeuge ein blaues Taxi gesehen hat, ist also nicht besonders groß – die Farbwahrnehmung ist dann erst der zweite Schritt

– In diesem Fall muss mit dem **Satz von Bayes** gerechnet werden

$$P(A_i \mid B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B \mid A_i) * P(A_i)}{P(B)} = \frac{P(B \mid A_i) * P(A_j)}{\sum_{j=1}^{k} P(B \mid A_i) * P(A_j)}$$

(Die Formel sehen wir uns nach einigen Vorüberlegungen gleich noch genauer an)

Rechnen mit dem Satz von Bayes

- Bevor wir uns der Formel zuwenden also noch ein paar Vorüberlegungen...

- Wären insgesamt nur 100 Taxen in der Stadt unterwegs…

  - …wären von diesen 85 grün (85% Marktanteil)
  - ...wären von diesen 15 blau (15% Marktanteil)

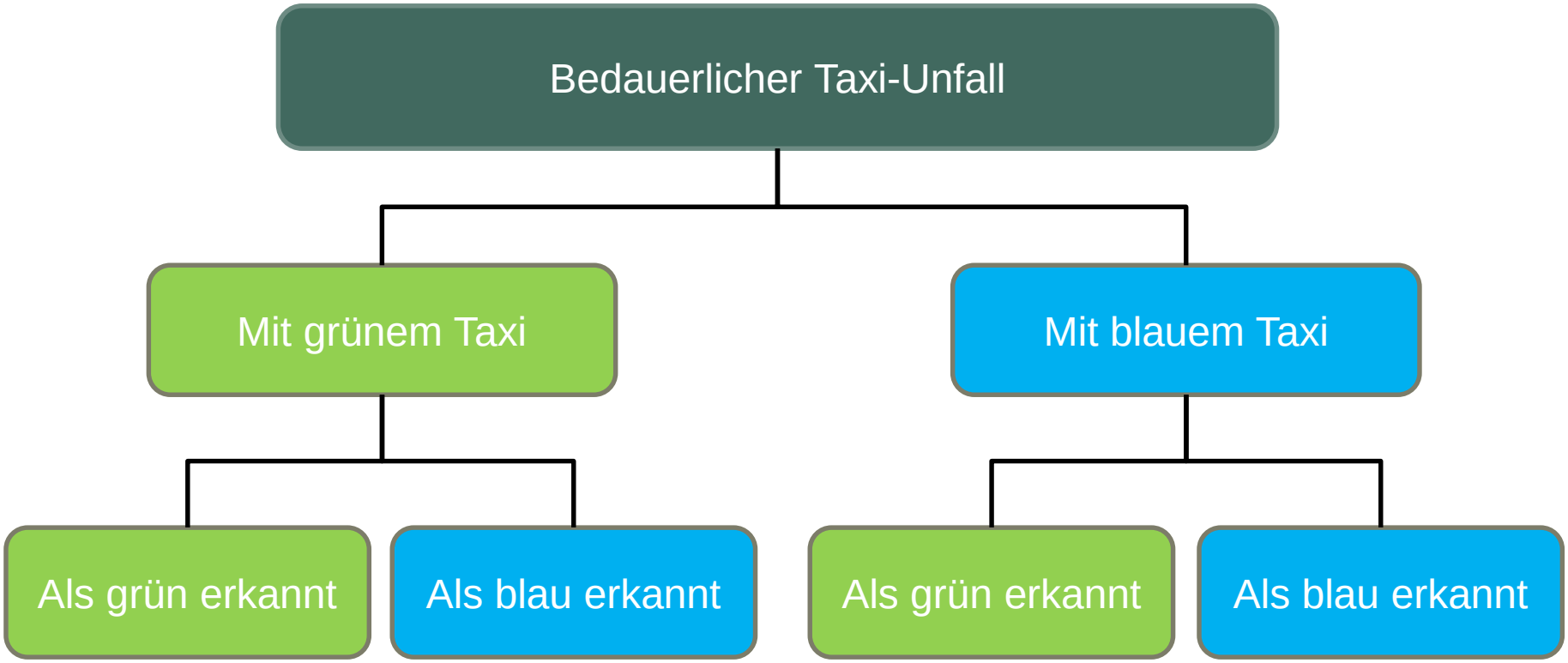
- Da der Zeuge Farben mit 80%iger Sicherheit korrekt erkennt…

  - …würde er 68 grüne Taxen als grün erkennen – und 17 als blau
  - …würde er 12 blaue Taxen als blau erkennen – und 3 als grün

- Diese Rahmenbedingungen müssen beachtet werden, will man wissen, wie groß die Chance für eine korrekte Aussage des Zeugen wirklich ist

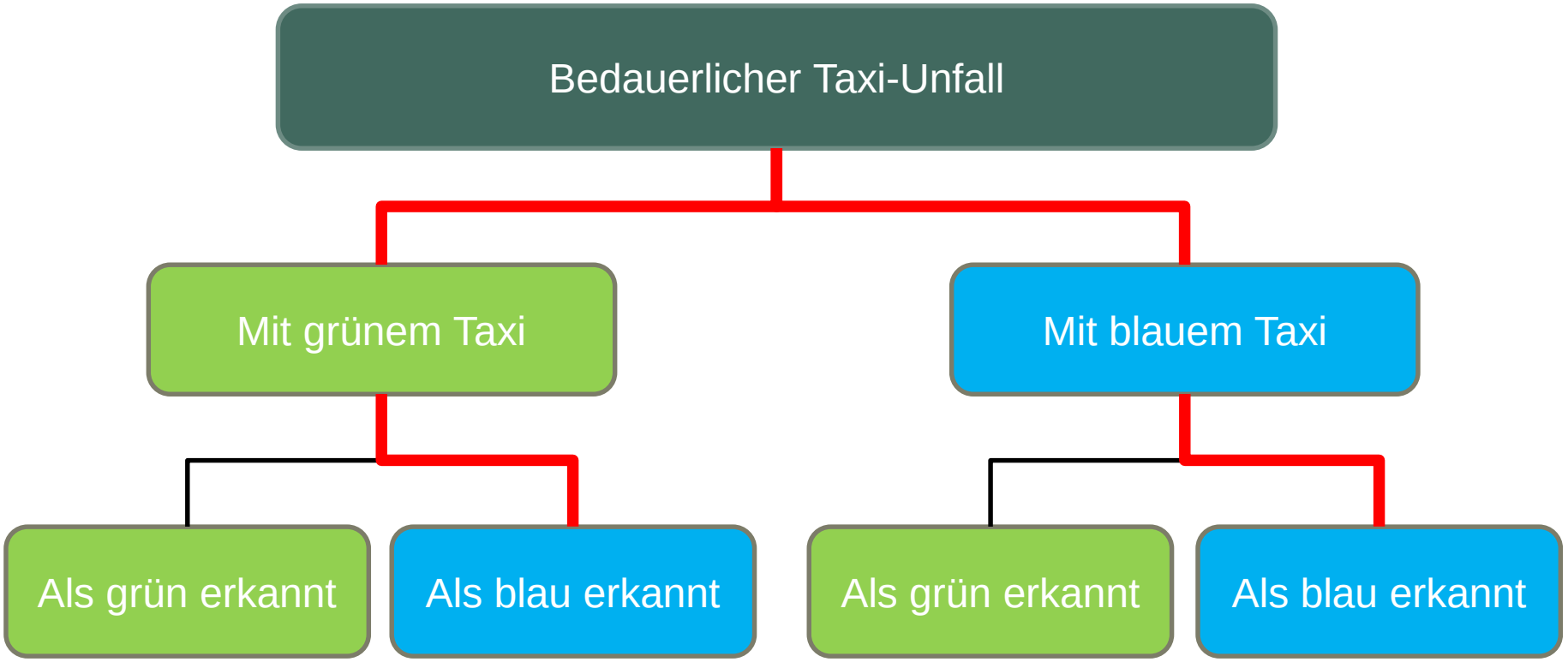
Rechnen mit dem Satz von Bayes

Welche Möglichkeiten gibt es insgesamt?



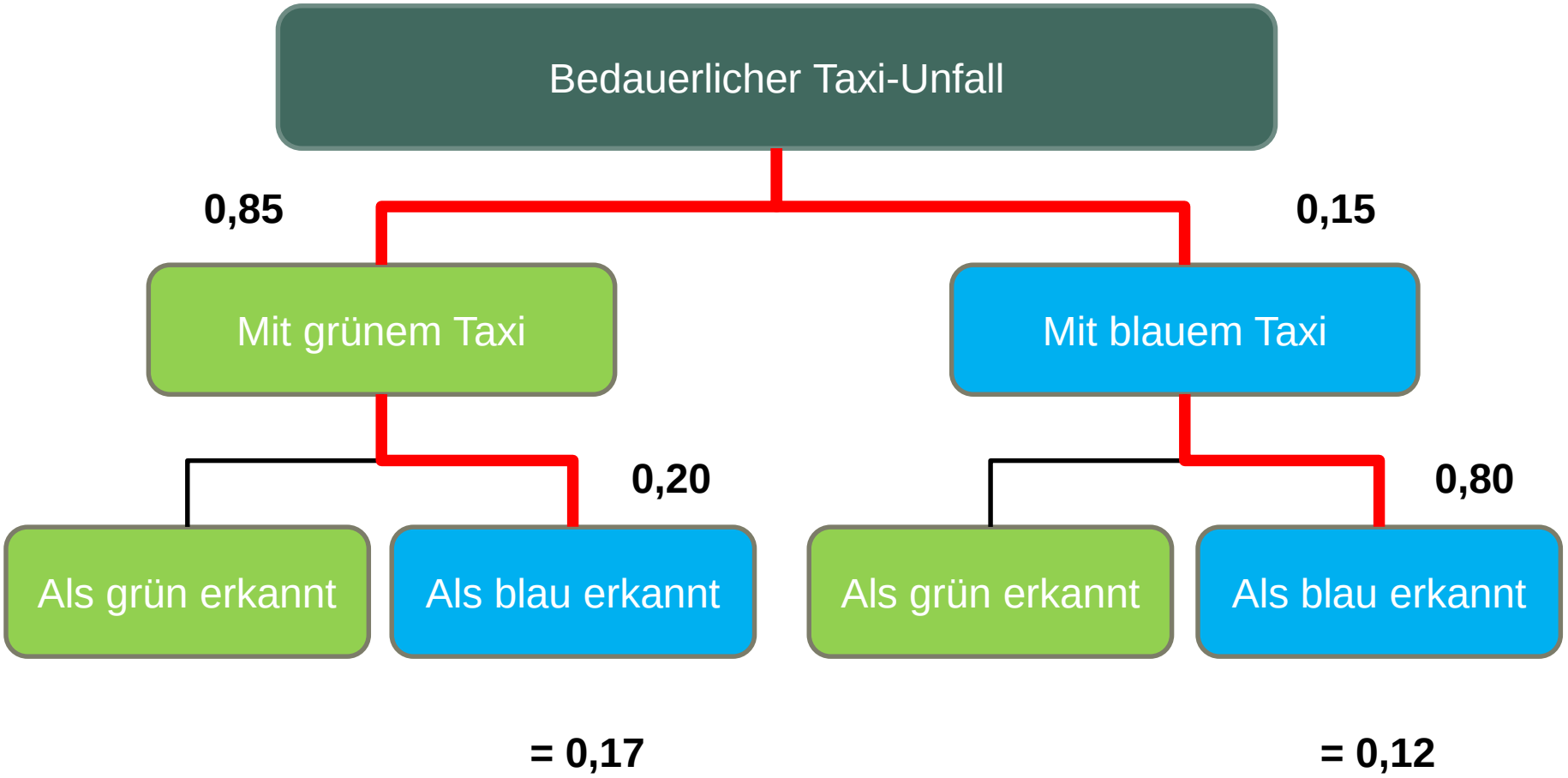
Rechnen mit dem Satz von Bayes

Welche Möglichkeiten sind von Bedeutung?



Rechnen mit dem Satz von Bayes

Welche Möglichkeiten sind von Bedeutung?



Rechnen mit dem Satz von Bayes

- Da der Zeuge das Taxi als blau identifiziert, sind zwei Pfade von Bedeutung
  - Das Unfalltaxi war grün (85%) und wird als blau erkannt (20%) -> 0,17
  - Das Unfalltaxi war blau (15%) und wird als blau erkannt (80%) -> 0,12
- Unter Berücksichtigung des klassischen Wahrscheinlichkeitsbegriffs nach Laplace würde man an der Stelle intuitiv – hoffentlich – wie folgt vorgehen:
  - P (A) = Σ günstiger Elementarereignisse / Σ möglicher Elementarereignisse
  - P (das Unfalltaxi war blau) = 0,12 / (0,17 + 0,12) = 0,12 / 0,29 = 0,41 = 41%
- Auch wenn diese Vorgehensweise eher intuitiv als formelgeleitet ist, führt sie letztlich zum korrekten Ergebnis – die Vorgehensweise unter Berücksichtigung des Satz von Bayes bzw. des Bayes-Theorem findet sich auf der nächsten Folie

Rechnen mit dem Satz von Bayes

Wahrscheinlichkeit für B unter der Bedingung, dass A_i eingetreten ist (Zeuge hält ein blaues Taxi für blau)

Wahrscheinlichkeit für den Eintritt des Ereignisses A_i (Taxi war blau)

$$P(A_i \mid B) = \frac{P(B \mid A_i) * P(A_i)}{P(B)}$$

Wahrscheinlichkeit für A_i unter der Bedingung, dass B eingetreten ist

(Taxi war wirklich blau (A_i) wenn der Zeuge es für blau hält (B))

Wahrscheinlichkeit dafür, dass B eintritt (die Summe aller Pfade, bei denen der Zeuge das Taxi am Ende für blau hält)

Rechnen mit dem Satz von Bayes

– Welche Größen sind für die formelgestützte Berechnung erforderlich?

TG = Taxi ist grün TB = Taxi ist blau

ZG = Zeuge hält das Taxi für grün ZB = Zeuge hält das Taxi für blau

Die Basisrate für TG liegt bei 0,85, die Basisrate für TB liegt bei 0,15

Als bedingte Wahrscheinlichkeiten für die Zeugenaussagen ergeben sich

P(ZG|TG) = 0,8 P(ZG|TB) = 0,2 P(ZB|TG) = 0,2 P(ZB|TB) = 0,8

$$P(A_i \mid B) = \frac{P(B \mid A_i) * P(A_i)}{P(B)} = \frac{P(ZB \mid TB) * P(TB)}{P(ZB \mid TB) * P(TB) + P(ZB \mid TG) * P(TG)}$$

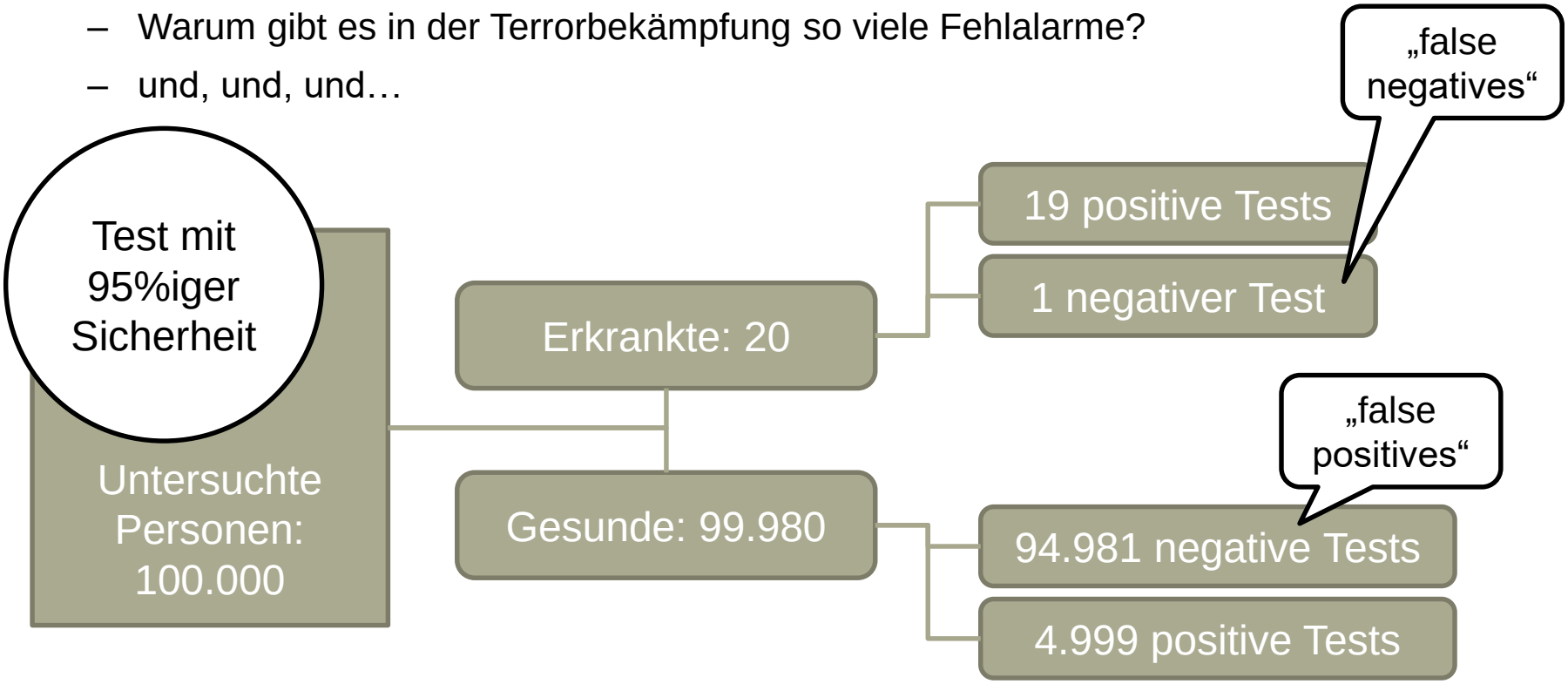
$$= \frac{0{,}80 * 0{,}15}{(0{,}80 * 0{,}15) + (0{,}20 * 0{,}85)} = 0{,}41$$

← Deutlich geringer als 0,8…

Rechnen mit dem Satz von Bayes

– Für welche „Alltagsphänomene“ ist der Satz von Bayes von Bedeutung?

  – Warum werde keine flächendeckenden HIV-Tests durchgeführt?
  – Warum gibt es in der Terrorbekämpfung so viele Fehlalarme?
  – und, und, und…



Übung: Rechnen mit dem Satz von Bayes

- Ein Unternehmen stellt Spritzgussteile auf zwei verschiedenen Maschinen her, wobei 70% der Teile auf Maschine X und 30% der Teile auf Maschine Y produziert werden. Die Wahrscheinlichkeit für einen Fertigungsfehler liegt bei Maschine X bei 10%, bei Maschine Y dagegen bei 20%

  - Wie groß ist die Wahrscheinlichkeit für einen Produktionsfehler?
  - Wie groß ist die Wahrscheinlichkeit, dass sich ein entdeckter Produktionsfehler auf Maschine Y zurückführen lässt?

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

Additionssatz

$$P(A_i \mid B) = \frac{P(B \mid A_i) * P(A_i)}{P(B)}$$

Satz von Bayes

Übung: Rechnen mit dem Satz von Bayes

- Ein Unternehmen stellt Spritzgussteile auf zwei verschiedenen Maschinen her, wobei 70% der Teile auf Maschine X und 30% der Teile auf Maschine Y produziert werden. Die Wahrscheinlichkeit für einen Fertigungsfehler liegt bei Maschine X bei 10%, bei Maschine Y dagegen bei 20%

  - Wie groß ist die Wahrscheinlichkeit für einen Produktionsfehler?

$$P(A \cup B) = P(A) + P(B) = (0{,}7 * 0{,}1) + (0{,}3 * 0{,}2) = 0{,}13$$

  - Wie groß ist die Wahrscheinlichkeit, dass sich ein entdeckter Produktionsfehler auf Maschine Y zurückführen lässt?

$$P(A_i \mid B) = \frac{P(B \mid A_i) * P(A_i)}{P(B)} = \frac{(0{,}3 * 0{,}2)}{(0{,}7 * 0{,}1) + (0{,}3 * 0{,}2)} = 0{,}4615$$

Denksport: Anlasslose Massenüberwachung

Eine Behörde überwacht mit Hilfe einer Software die unverschlüsselte E-Mail-Kommunikation deutscher Internetnutzer*innen. Die Software, die E-Mails auf eine Reihe von Schlüsselbegriffen und Phrasen filtert, die auf illegale und / oder terroristische Aktivitäten hinweisen könnten, stuft eine tatsächlich sicherheitsrelevante Kommunikation mit einer sehr hohen Wahrscheinlichkeit von 99,5% als potentielle Bedrohung ein. Die Wahrscheinlichkeit dafür, dass eine harmlose E-Mail fälschlicherweise als potentielle Bedrohung klassifiziert wird, liegt dagegen nur bei 0,5%.

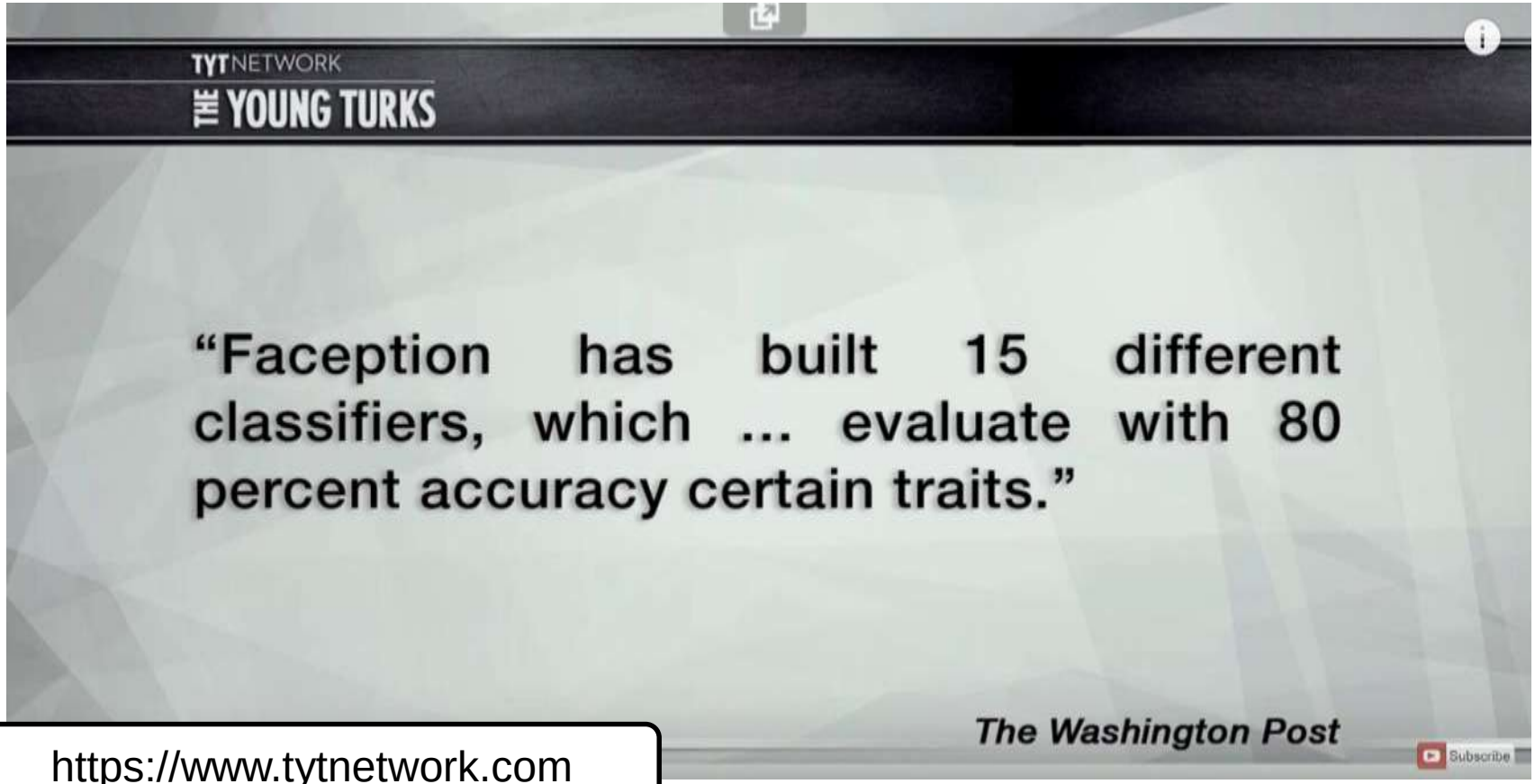
In Deutschland gibt es 71.000.000 Internetnutzer*innen. Nachfolgend gehen wir davon aus,

– dass jeder Nutzer täglich 10 unverschlüsselte Mails verschickt, die von der Software gesichtet werden,
– dass 10.000 Nutzer das Internet für die Vorbereitung illegaler oder terroristischer Aktivitäten nutzen
– und dass jede vierte Mail, die von einem dieser 10.000 Nutzer verschickt wird, einen auffindbaren Hinweis auf eine solche Aktivität enthält.

Wie groß ist die Wahrscheinlichkeit dafür, dass eine an einem beliebigen Tag durch die Bedrohungen zu 99,5% korrekt klassifizierende Software als potentielle Bedrohung eingestufte E-Mail auch tatsächlich auf eine reale Bedrohungslage hinweist?

Auflösung unter: http://scienceblogs.de/frischer-wind/2017/05/30/anlasslose-massenueberwachung-und-der-satz-von-bayes/

Wie viele „false positives“ generiert eine Anti-Terror-Software mit 80% Treffergenauigkeit?



https://www.tytnetwork.com

Teil IX

Konfidenzintervalle

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Was sind Konfidenzintervalle?

– Da Vollerhebungen selten sind, steht man häufig vor der Aufgabe, Parameter aus der Grundgesamtheit (etwa die Lage des arithmetischen Mittels) aus Stichprobendaten heraus schätzen zu müssen. Hierfür bieten sich zwei Vorgehensweise an:

  – **Punktschätzung:** Der Parameter wird als einzelner Wert geschätzt – z.B. das arithmetische Mittel der Grundgesamtheit aus dem arithmetischen Mittel der Stichprobe. Das Problem: Die Wahrscheinlichkeit, genau den richtigen Wert zu treffen, ist äußerst gering. Gleichzeitig kann man aber auch nicht wissen, wie weit man vom realen Wert entfernt liegt.

    „Der geschätzte arithmetische Mittelwert liegt bei 5 g. Wir wissen aber nicht, wie weit das vom realen arithmetischen Mittelwert entfernt ist.“

    Aussagekraft?

  – **Intervallschätzung:** Mehr Aussagekraft hat eine Intervallschätzung, d.h. die Abgrenzung eines Intervalls, in dem sich der wahre Parameter der mit einer gewissen Sicherheit befindet.

    „Mit 95%iger Sicherheit liegt der reale arithmetische Mittelwert zwischen 4,8 g und 5,6 g.“

Was sind Konfidenzintervalle?

- Wie kommt nun aber ein solches Konfidenzintervall zustande?
  (vom lateinischen confidere = vertrauen, d.h. Vertrauensintervall)

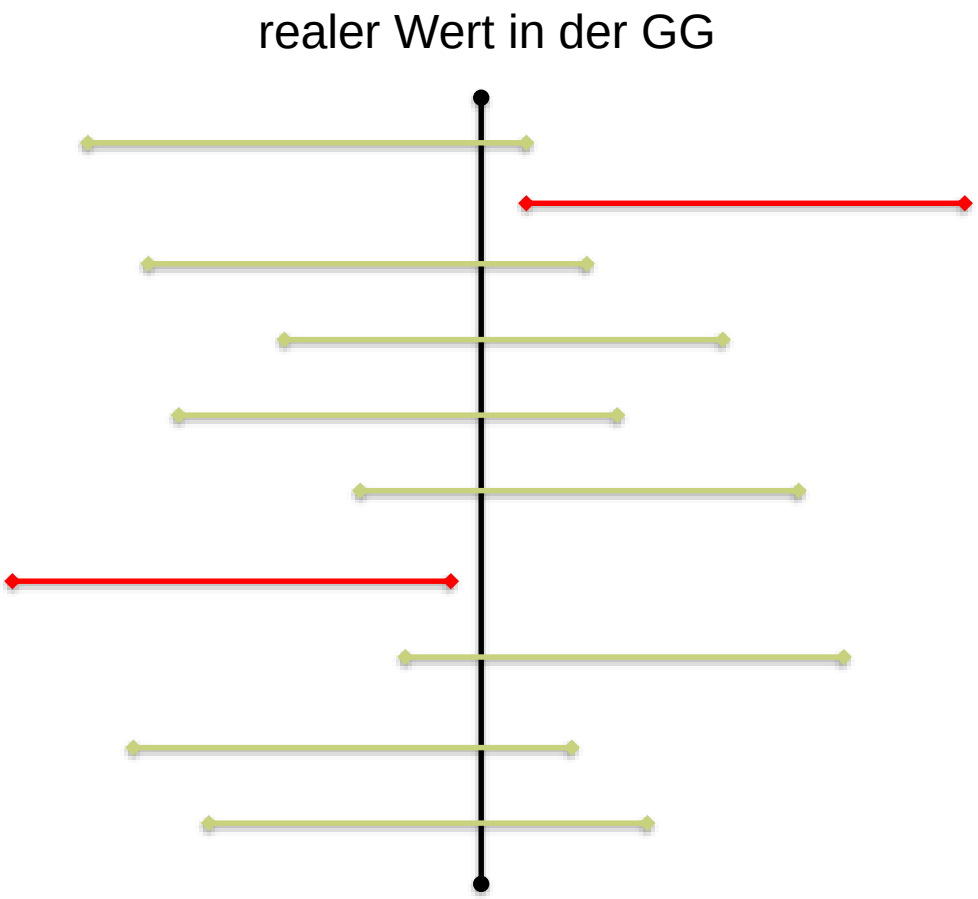
- Folgende Ausgangssituation ist gegeben:

  - Es ist bekannt, dass eine zu untersuchende Variable normalverteilt ist
  - Erwartungswert und/oder Standardabweichung sind aber unbekannt
  - Daten einer (repräsentativen) Stichprobe liegen für Schätzungen vor

- Auf Basis der Daten aus der Stichprobe soll nun versucht werden, den Bereich zu bestimmen, in dem sich der Wert (z.B. Erwartungswert) der Grundgesamtheit befindet

- Je breiter dieses Intervall ist, umso größer ist die Sicherheit, dass der gesuchte Wert auch tatsächlich in dem Intervall liegt – umso geringer ist aber auch der Aussagewert des Intervalls -> dies wird über das Vertrauensniveau / Konfidenzniveau γ reguliert

Was sagt das Konfidenzniveau aus?



Bei einem Konfidenzniveau von 95% schließen 95% der Konfidenzintervalle dieser Breite bei unendlicher Wiederholung der Stichprobenziehung den realen Wert in der Grundgesamtheit ein.

>>> Ein beliebiges Konfidenzintervall auf diesem Konfidenzniveau gehört also mit 95%iger Wahrscheinlichkeit zu der Gruppe von Konfidenz-intervallen, welche den realen Wert einschließen.

Alternativ: Die Wahrscheinlichkeit, dass der reale Wert in keinem der 95%-Intervalle liegt, beträgt 5%.

Konfidenzniveau und Konfidenzbreite

- Wie man sich leicht vorstellen kann, hängt die Breite eines Konfidenzintervalls wesentlich vom jeweils gewählten Konfidenzniveau bzw. Vertrauensniveau ab

- Dies lässt sich logisch wie folgt herleiten:

  - Je breiter ein Konfidenzintervall ausfällt, desto wahrscheinlicher ist, dass es den realen Wert in der Grundgesamtheit einschließt
  - Je größer das Konfidenzniveau eines Konfidenzintervalls ist, umso wahrscheinlicher ist, dass es den realen Wert in der Grundgesamtheit einschließt
  - Daraus folgt: Je größer das Vertrauensniveau, desto breiter das Konfidenzintervall

- Wichtig: Das Konfidenzniveau muss immer vor der Aufstellung eines Intervalls festgelegt und darf keinesfalls im Nachhinein so „angepasst“ werden, dass ein gewünschtes Ergebnis erreicht wird

Einige bedeutende Konfidenzintervalle

- Konfidenzintervall um den Erwartungswert
  - ...bei normalverteilter Grundgesamtheit und bekannter Standardabweichung der Merkmalsverteilung
  - ...bei normalverteilter Grundgesamtheit und unbekannter Standardabweichung der Merkmalsverteilung
  - ...bei unbekannter Merkmalsverteilung in der Grundgesamtheit
- Konfidenzintervall um die Varianz
- Konfidenzintervall um die Standardabweichung
- Konfidenzintervall um den Stichprobenanteilswert

Wichtiger Hinweis: Um die uns zur Verfügung stehende Zeit optimal auszunutzen, werden wir nachfolgend nur das Konfidenzintervall um den Erwartungswert μ bei bekannter Standardabweichung σ betrachten

Konfidenzintervall um μ bei bekanntem σ

– Beispiel: Das Gewicht von Spritzgussbauteilen sei normalverteilt bei einer Standardabweichung σ von 10 g und unbekanntem Erwartungswert μ. Eine Stichprobe vom Umfang 100 erbringt einen Mittelwert von 20 g.

– Bestimmt werden soll das Konfidenzintervall um den Erwartungswert μ mit einem Konfidenzniveau von 95%

$$P(\bar{x} - z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$z_{(1-\frac{\alpha}{2})}$ Entsprechendes Quantil aus der Standardnormalverteilung (in diesem Fall: $z_{(0,975)} = 1{,}96$)

$\bar{x}$ = arithmetisches Mittel (Stichprobe)
σ = Standardabweichung (Grundges.)
n = Stichprobenumfang

Konfidenzintervall um μ bei bekanntem σ

$$P(\bar{x} - z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$z_{(1-\frac{0,05}{2})} = z_{(0,975)} = 1,96$ (Wert aus der Tabelle der Z-Verteilung)

$$P(20 - 1,96 * \frac{10}{\sqrt{100}} \leq \bar{x} \leq 20 + 1,96 * \frac{10}{\sqrt{100}}) = 1 - 0,05$$

$$P(20 - 1,96 * 1 \leq \bar{x} \leq 20 + 1,96 * 1) = 0,95$$

$$P(18,04 \leq \bar{x} \leq 21,96) = 0,95$$

Übung: Konfidenzintervall um μ (bei σ bek.)

– Das Gewicht von Studierenden ist – aller Wahrscheinlichkeit nach – normalverteilt bei einer Standardabweichung σ von 520 g und unbekanntem Erwartungswert μ. Eine Untersuchung von 20 Studierenden erbringt einen Mittelwert von 67,3 kg.

– Bestimmt werden soll das Konfidenzintervall um den Erwartungswert μ mit einem Konfidenzniveau von 99%

$$P(\bar{x} - z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$z_{(1-\frac{\alpha}{2})}$ Entsprechendes Quantil aus der Standardnormalverteilung (in diesem Fall: $z_{(0,995)} = 2,58$)

$\bar{x}$ = arithmetisches Mittel (Stichprobe)
σ = Standardabweichung (Grundges.)
n = Stichprobenumfang

Übung: Konfidenzintervall um μ (bei σ bek.)

$$P(\bar{x} - z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{(1-\frac{\alpha}{2})} * \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$z_{(1-\frac{0,01}{2})} = z_{(0,995)} = 2,58$ (Wert aus der Tabelle der Z-Verteilung)

$$P(67,3 - 2,58 * \frac{0,52}{\sqrt{20}} \leq \bar{x} \leq 67,3 + 2,58 * \frac{0,52}{\sqrt{20}}) = 1 - 0,01$$

$$P(67,3 - 2,58 * 0,12 \leq \bar{x} \leq 67,3 + 2,58 * 0,12) = 0,99$$

$$P(66,99 \leq \bar{x} \leq 67,61) = 0,99$$

Beispiele für weitere Konfidenzintervalle

$$P(\bar{x} - t_{(1-\frac{\alpha}{2};n-1)} * \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{(1-\frac{\alpha}{2};n-1)} * \frac{s}{\sqrt{n-1}}) = 1-\alpha$$

(Konfidenzintervall um den Erwartungswert bei unbekannter Standardabweichung)

$$P(\hat{p} - z_{(1-\frac{\alpha}{2})} * \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{(1-\frac{\alpha}{2})} * \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}) = 1-\alpha$$

(Konfidenzintervall um den Stichprobenanteilswert)

Für den rechnerischen Part der Klausur wird aus Zeitgründen nur das Konfidenzintervall um den Erwartungswert μ bei bekannter Standardabweichung σ von Relevanz sein.

Teil X

Statistische Software

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

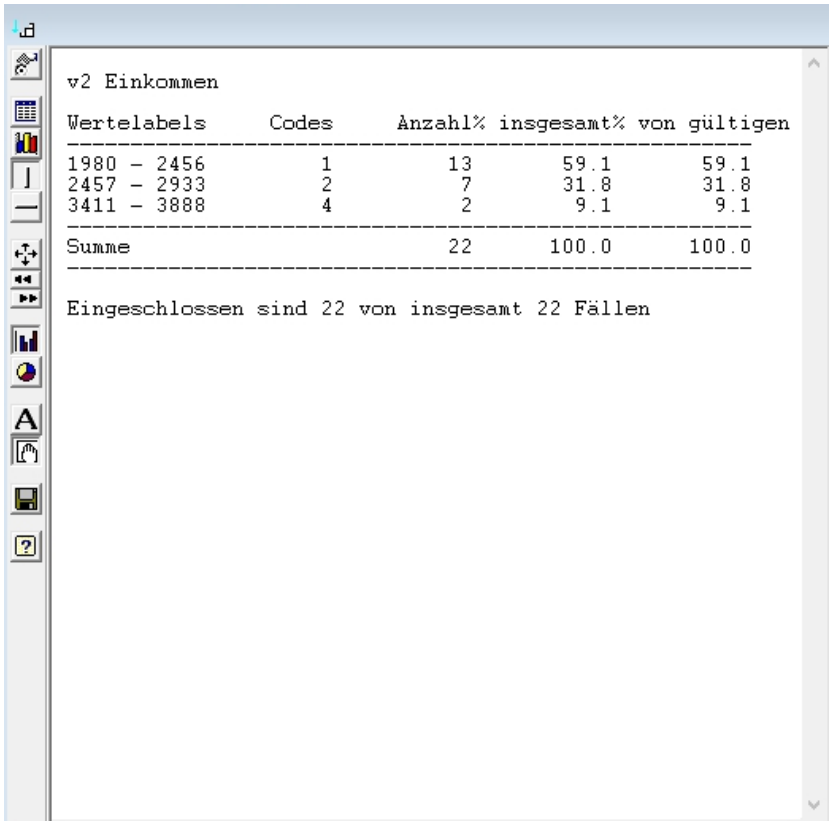
Statistische Software

Kostenlose Statistik-Software

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Warum eine gesonderte Software-Einführung?
(Nur weil wir nicht per Hand rechnen wollen?)



v2 Einkommen

Wertelabels	Codes	Anzahl	% insgesamt	% von gültigen
1980 – 2456	1	13	59.1	59.1
2457 – 2933	2	7	31.8	31.8
3411 – 3888	4	2	9.1	9.1
Summe		22	100.0	100.0

Eingeschlossen sind 22 von insgesamt 22 Fällen

– Praxisnah: In keinem Betrieb würde eine lineare Regressionsanalyse noch „per Hand“ durchgeführt
– Vorbereitung: Wer im Rahmen der BA empirisch arbeiten möchte, wird hierfür Software einsetzen wollen

Und warum freie Software?

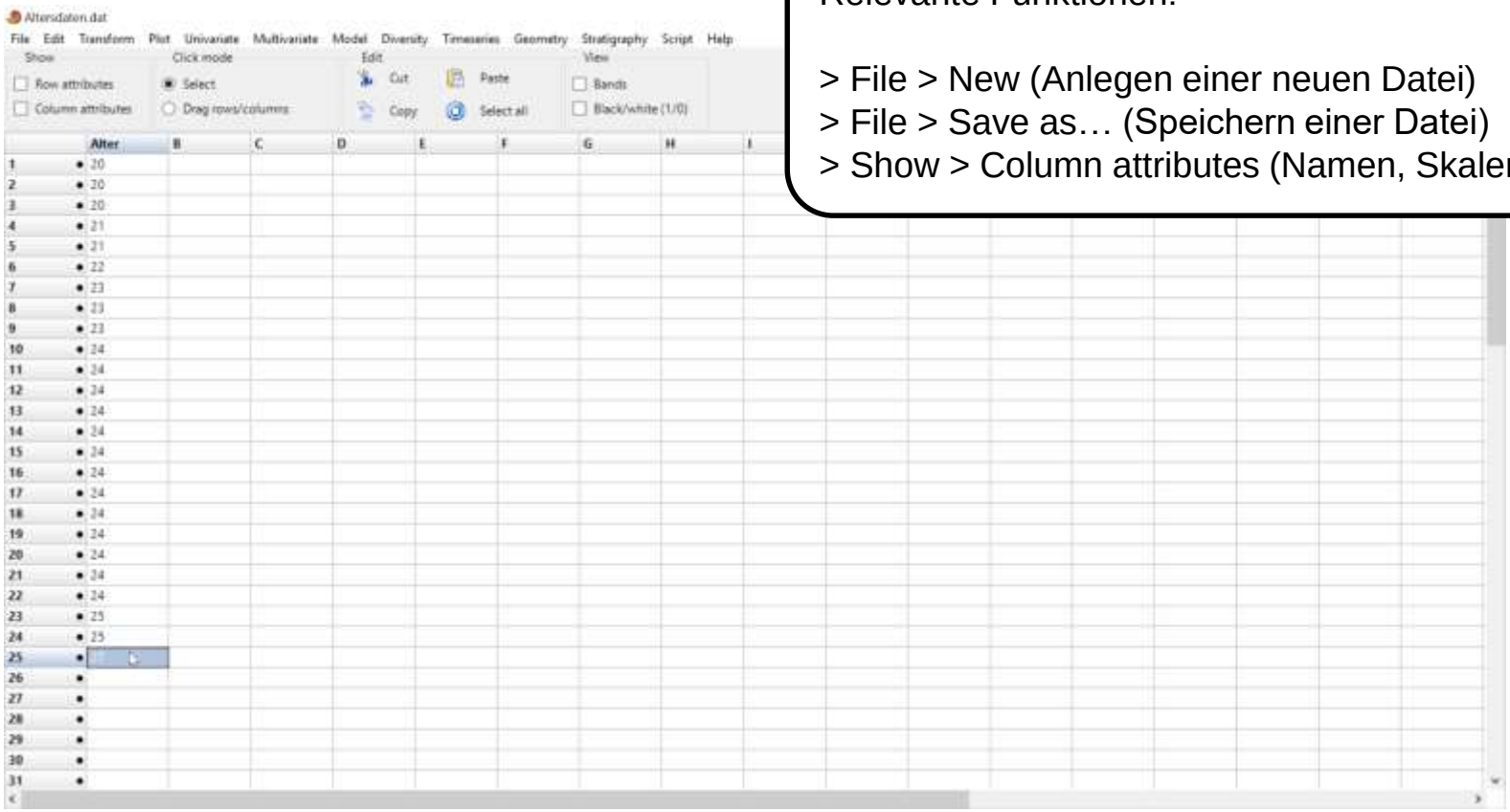
– Eine einfache SPSS-Lizenz kostet 1.168,00 EUR pro User und Jahr
– Freie Software ist ohne Kosten in Studium und Beruf einsetzbar

Unser zentraler Beispieldatensatz
(bereits aus der Hauptvorlesung bekannt)

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

Wie bekommen wir diese Daten nun in PAST?

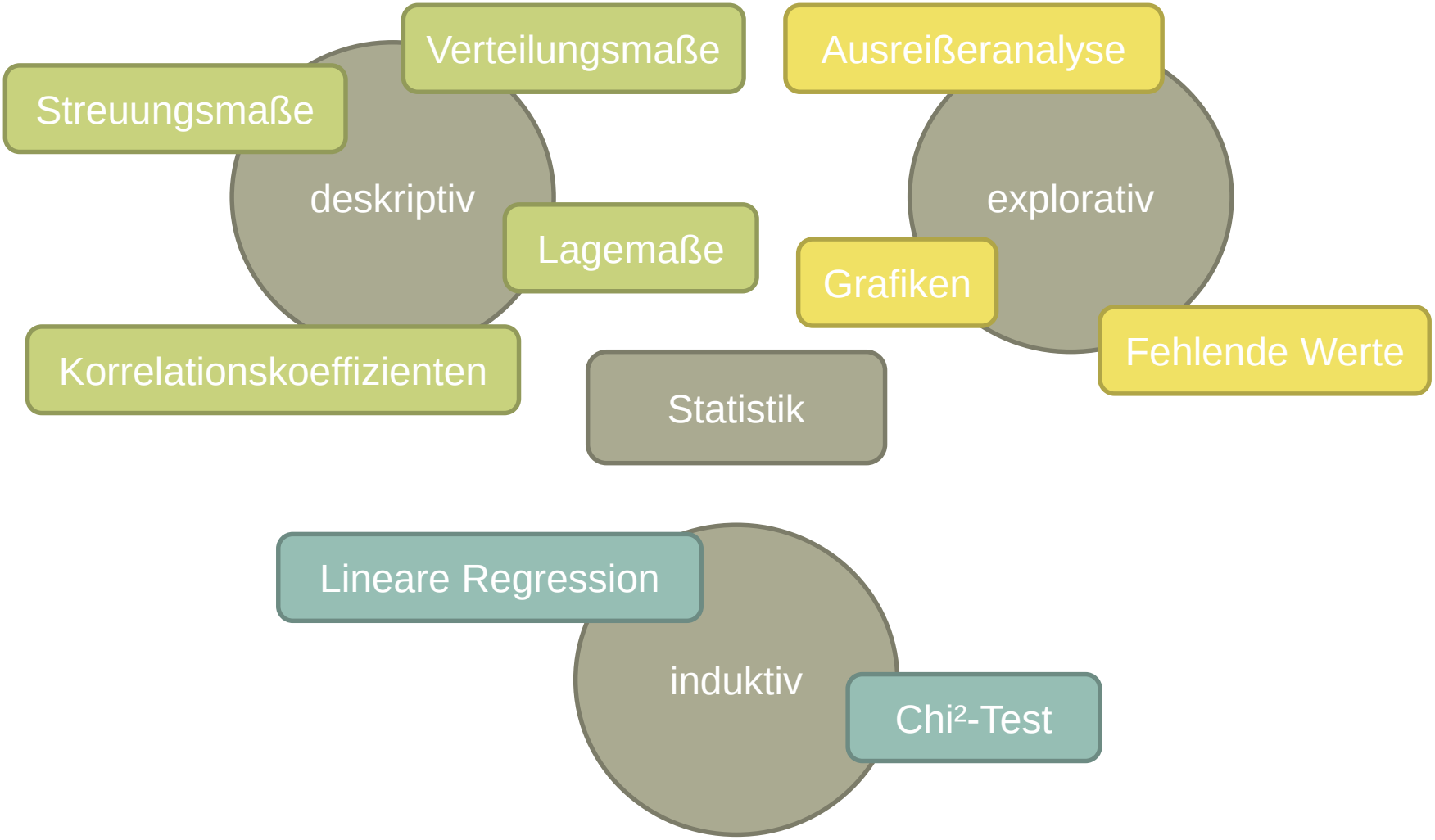
Eingabe von Daten in PAST



Relevante Funktionen:

> File > New (Anlegen einer neuen Datei)
> File > Save as… (Speichern einer Datei)
> Show > Column attributes (Namen, Skalen)

Wo befinden wir uns?



Lagemaße und Streuungsmaße

> Univariate > Summary statistics

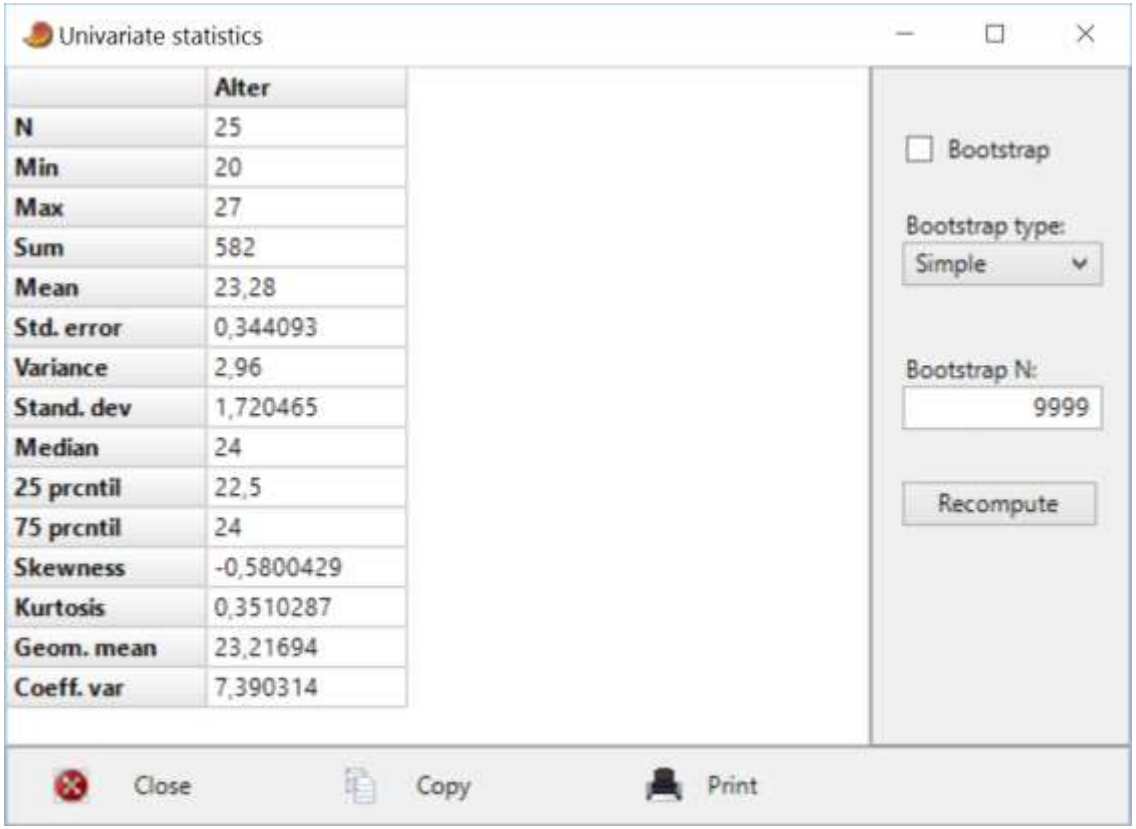
Was ist hier was?

N = Anzahl der Werte
Min = kleinster Wert
Max = größter Wert

Mean = arithmetisches Mittel
Geom. mean = Geometrisches Mittel

25 prcentil = Unteres Perzentil
Median = Mittleres Perzentil
75 prcentil = Oberes Perzentil

Variance = Varianz
Stand dev. = Standardabweichung



Das „SPSS-Analyseproblem“

- **Software führt <u>JEDE</u> Analyse unabhängig von den Voraussetzungen durch!**

- ...also auch die Berechnung des arithmetischen Mittels

  - ... aus Schulnoten
  - ... aus Geschlechtern
  - ... aus Kontonummern
  - ... aus Telefonnummern
  - ... aus Präferenzrängen

- Bei komplexen Verfahren sind noch weit schlimmere „Vergehen“ denkbar
- Die fachlichen Kenntnisse der Anwender/innen sind daher entscheidend

- **Darum: <u>KEINE</u> Analyse ohne vorherige Prüfung der Voraussetzungen!**

Warum ergeben sich andere Streuungsmaße?

– In der Vorlesung haben wir die Standardvarianz als Durchschnitt der quadrierten Abweichungen berechnet:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  – $s^2 = 2{,}8416$ | $s = 1{,}6875$

– Mit Hilfe von PAST berechnen wir die sog. Stichprobenvarianz mit den Freiheitsgraden (n-1) im Vorfaktor:

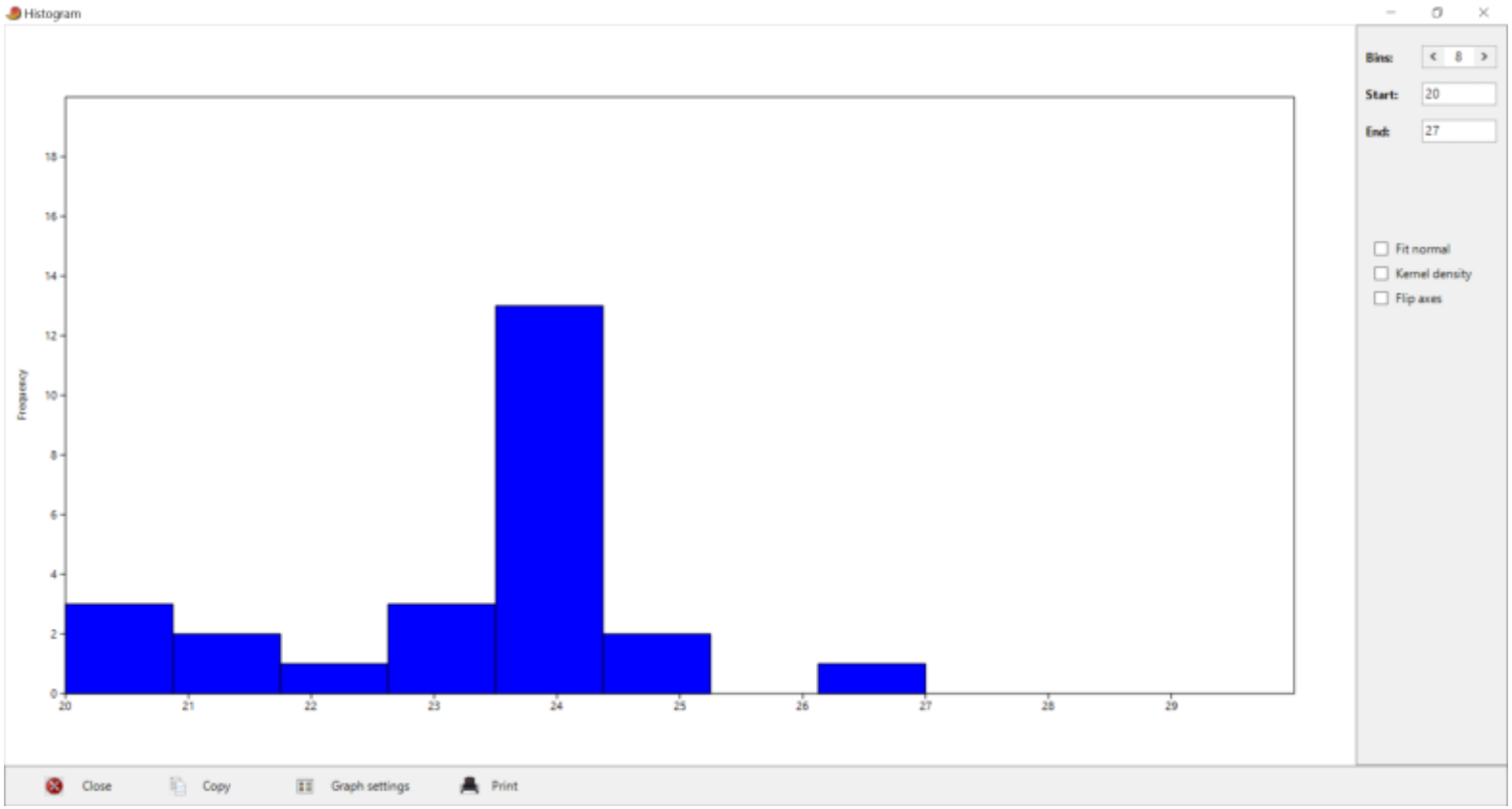
  – $s^2 = 2{,}96$ | $s = 1{,}72$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Ist die Wahl der Formel eher für große oder eher für kleine Datensätze relevant?

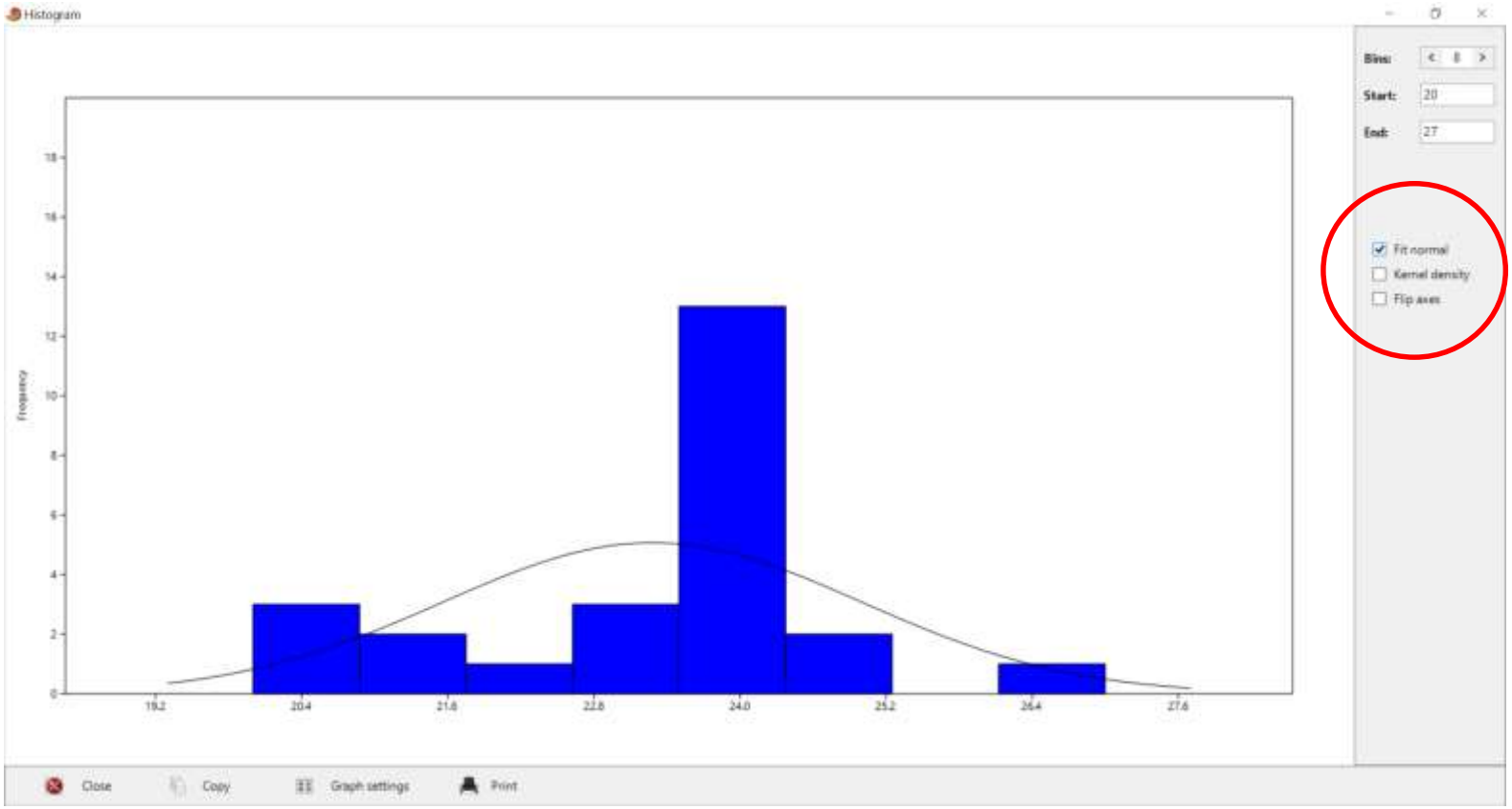
Gibt es einen Modus?

> Plot > Histogram



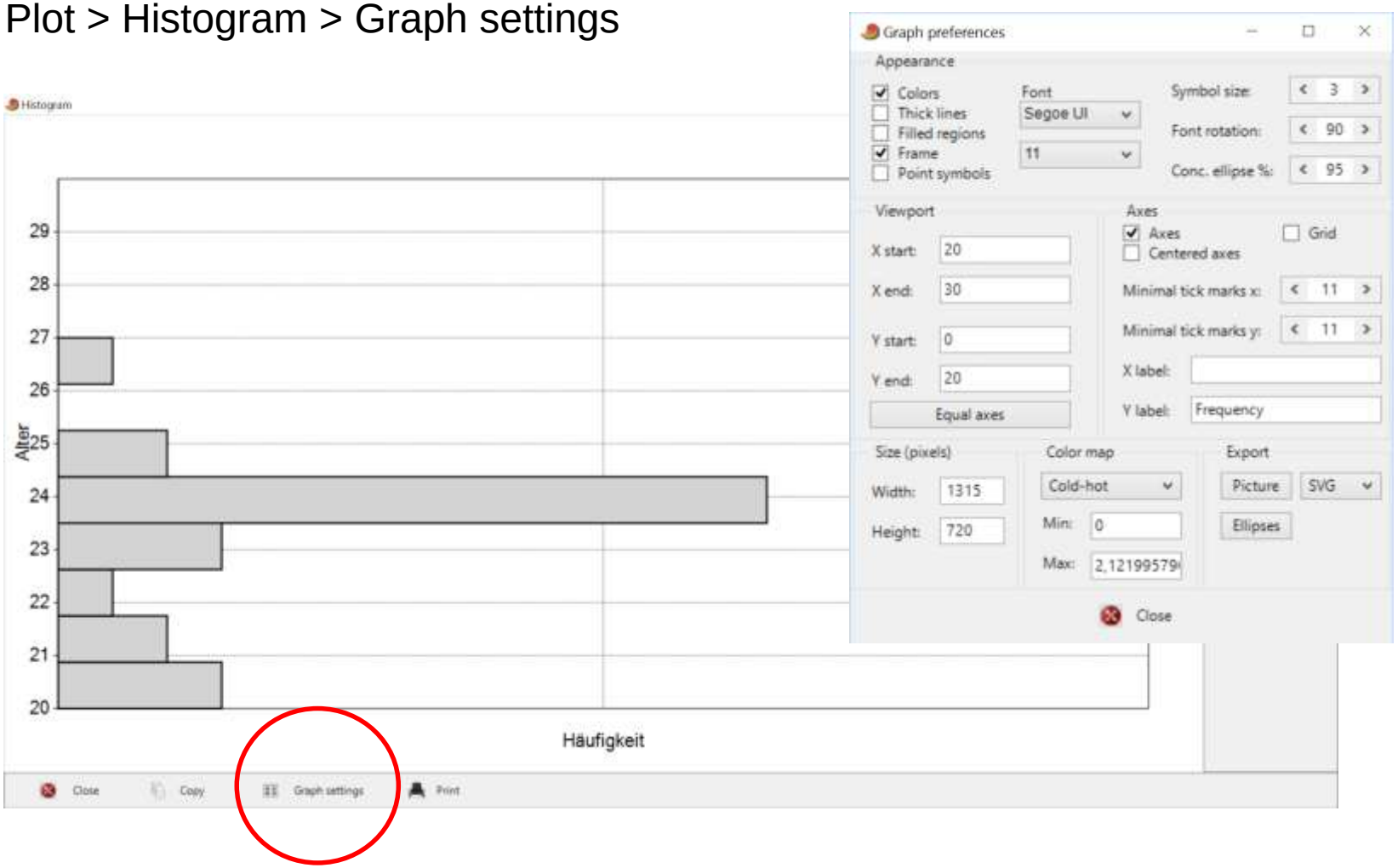
Folgt die Verteilung einer Normalverteilung?

> Plot > Histogram



Lässt sich die Grafik noch individualisieren?

> Plot > Histogram > Graph settings

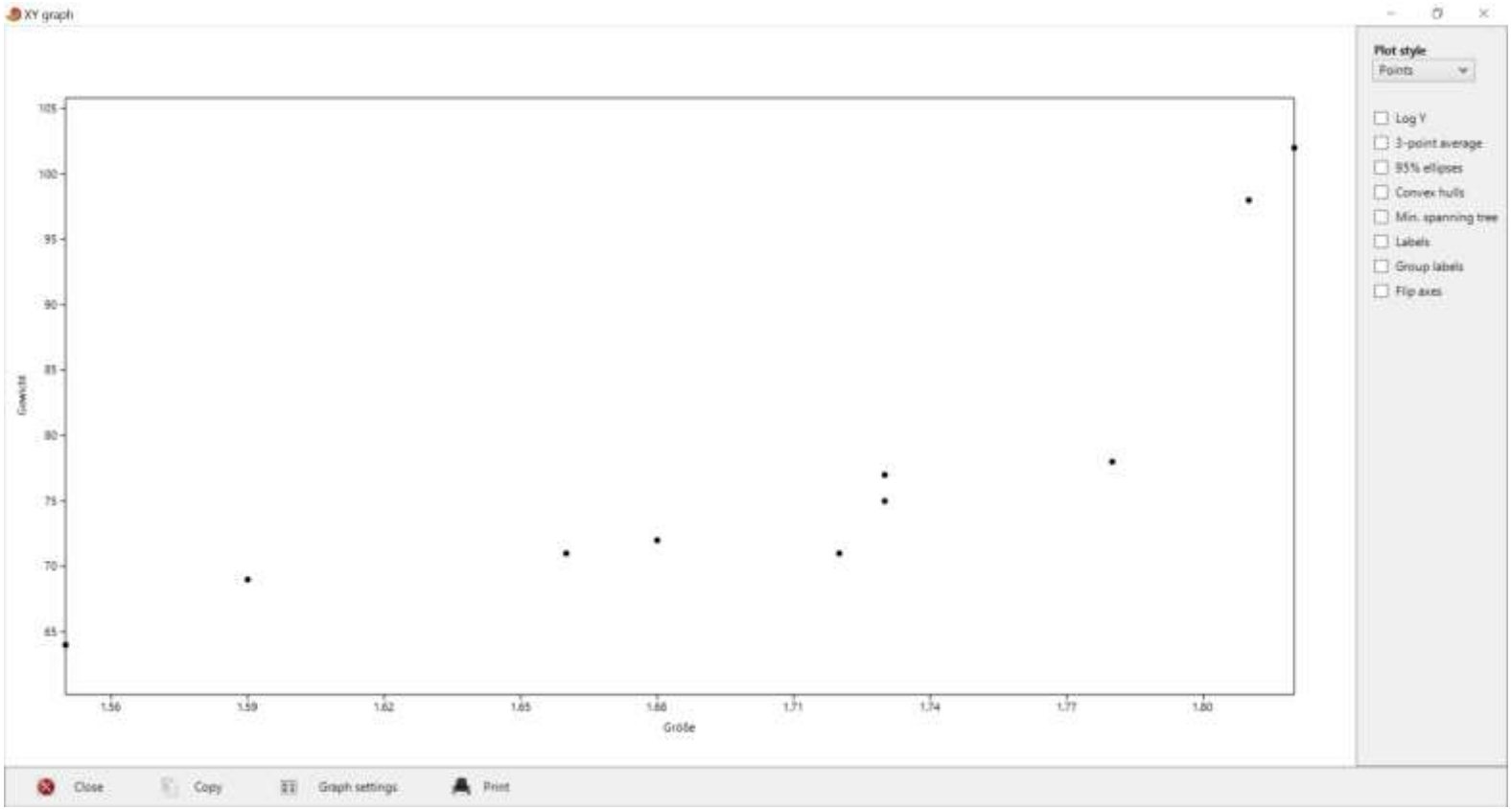


Bivariater Datensatz für Korrelationsanalysen
(ebenfalls aus der Hauptvorlesung bekannt)

Befragte/r	Größe (m)	Gewicht (kg)
1	1,55	64
2	1,68	72
3	1,72	71
4	1,73	75
5	1,82	102
6	1,81	98
7	1,66	71
8	1,78	78
9	1,73	77
10	1,59	69

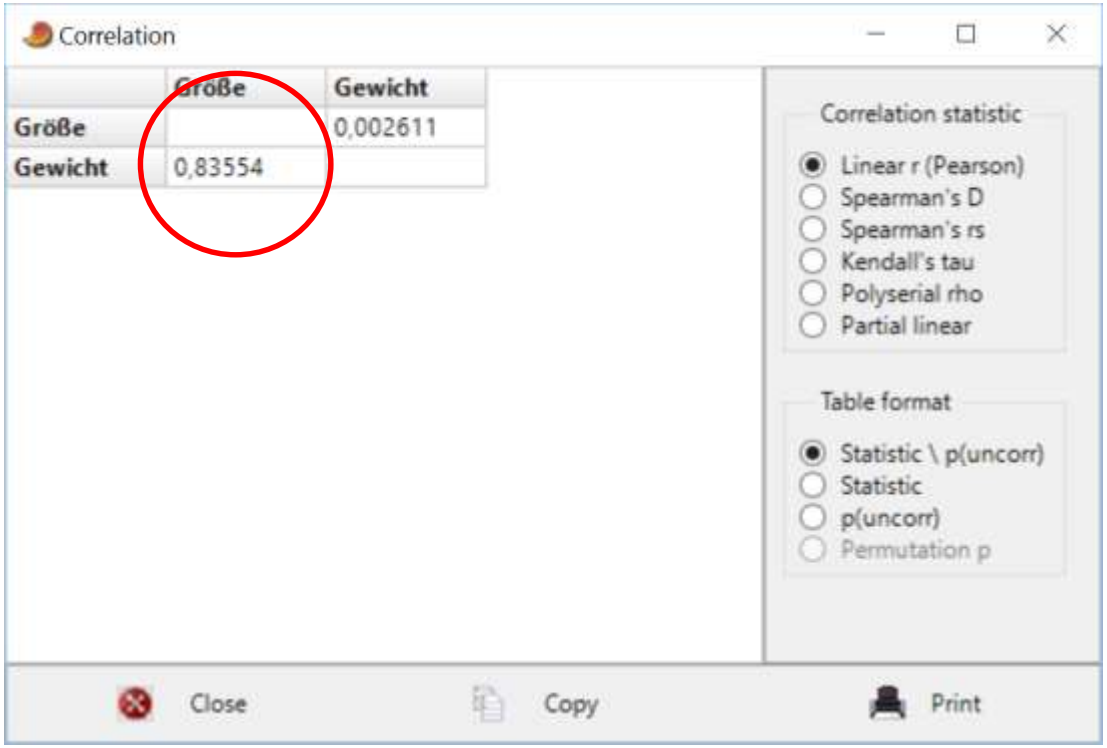
Ist ein Zusammenhang grafisch plausibel?

> Plot > XY graph



Berechnung von Korrelationskoeffizienten

> Univariate > Correlation



Was ist hier was?

Kendall‘s tau =
Konkordanzkoeffizient nach Kendall

Linear r (Pearson) =
Bravais-Pearson-Korrelationskoeffizient

Spearman‘s rs =
Rangkorrelationskoeffizient nach Spearman

Interpretation des Betrags von x

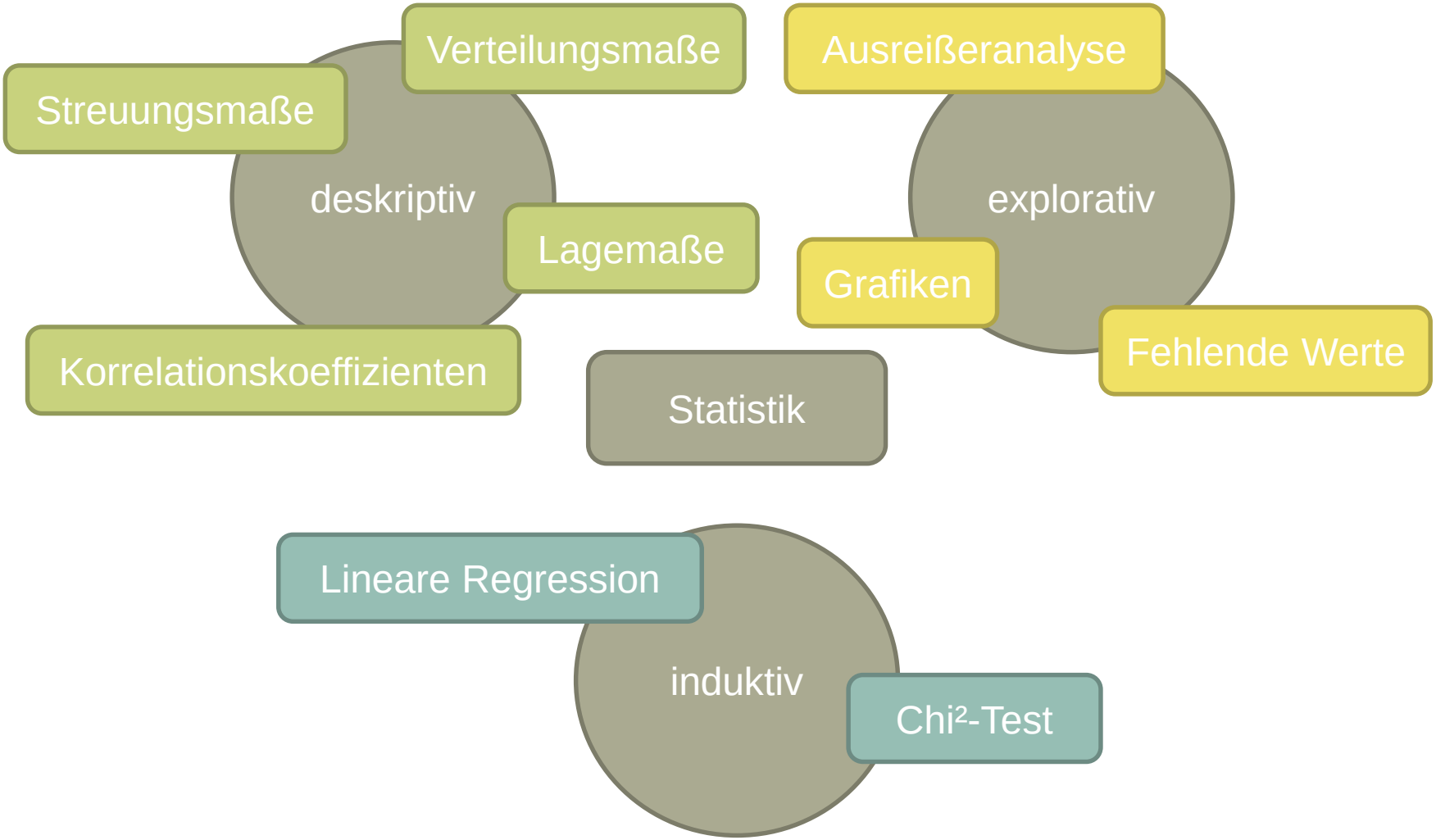
x = 0 = keine Korrelation
0 < x < 0,5 = schwache Korrelation
0,5 <= x < 0,8 = mittlere Korrelation
0,8 <= x < 1 = starke Korrelation
x = 1 = perfekte Korrelation

Korrelation ist nicht gleich Kausalität

- Eine über einen Korrelationskoeffizienten identifizierte Korrelation sollte näher untersucht, dabei jedoch **niemals inhaltlich interpretiert werden**
- Grund dafür ist, dass eine Korrelation nicht notwendigerweise auf einem Ursache-Wirkungs-Zusammenhang beruht – auch wenn es in vielen Fällen leider äußerst verführerisch ist, diese Annahme zu treffen

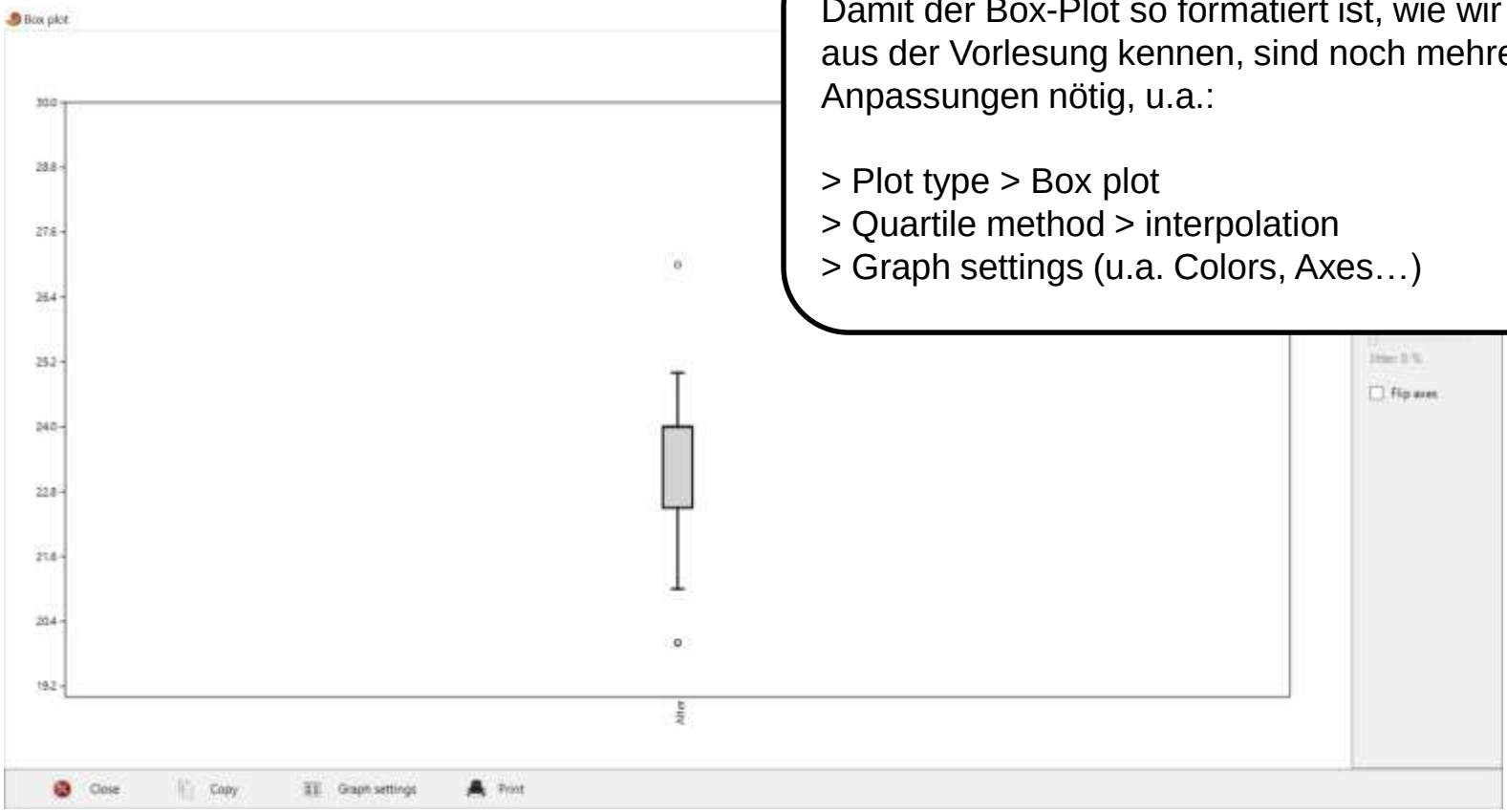
- Tatsächlich kann es verschiedene Erklärungen für Korrelationen geben
  - Einseitiger Zusammenhang: X beeinflusst Y bzw. Y beeinflusst X
  - Beidseitiger Zusammenhang: X und Y beeinflussen sich gegenseitig
  - Es handelt sich um einen reinen Zufallseffekt in den Daten (Scheinkorrelation)
  - Eine dritte Variable (Z) beeinflusst X und Y gleichermaßen (Scheinkorrelation)
- Ein klassisches Beispiel für eine Scheinkorrelation ist die Korrelation zwischen Storchenzahl und Geburtenquote (verbunden über die Variable „Urbanisierung“)

Wo befinden wir uns?



Erstellung eines Box-Plots

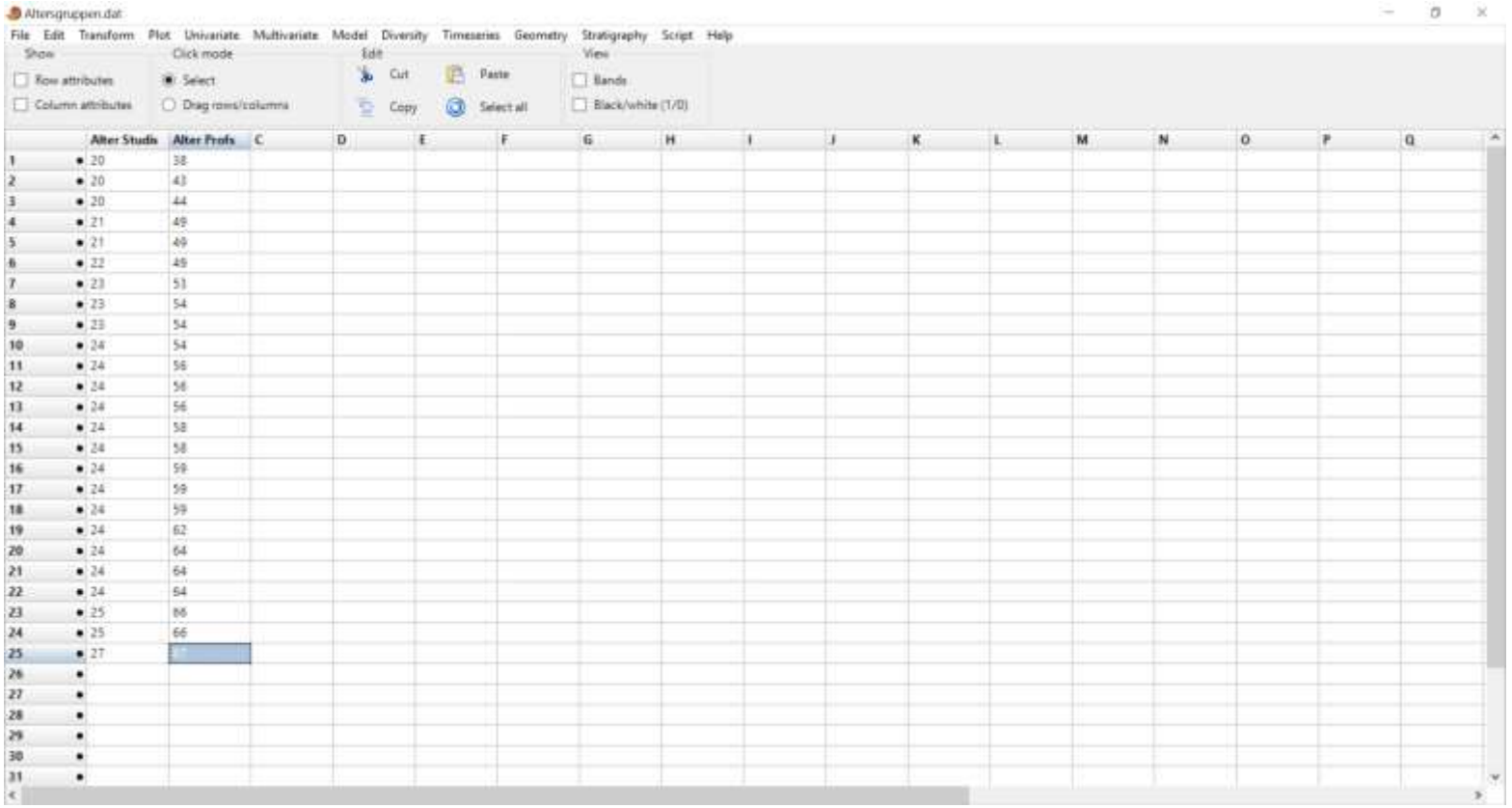
> Plot > Barchart/Boxplot



Damit der Box-Plot so formatiert ist, wie wir es aus der Vorlesung kennen, sind noch mehrere Anpassungen nötig, u.a.:

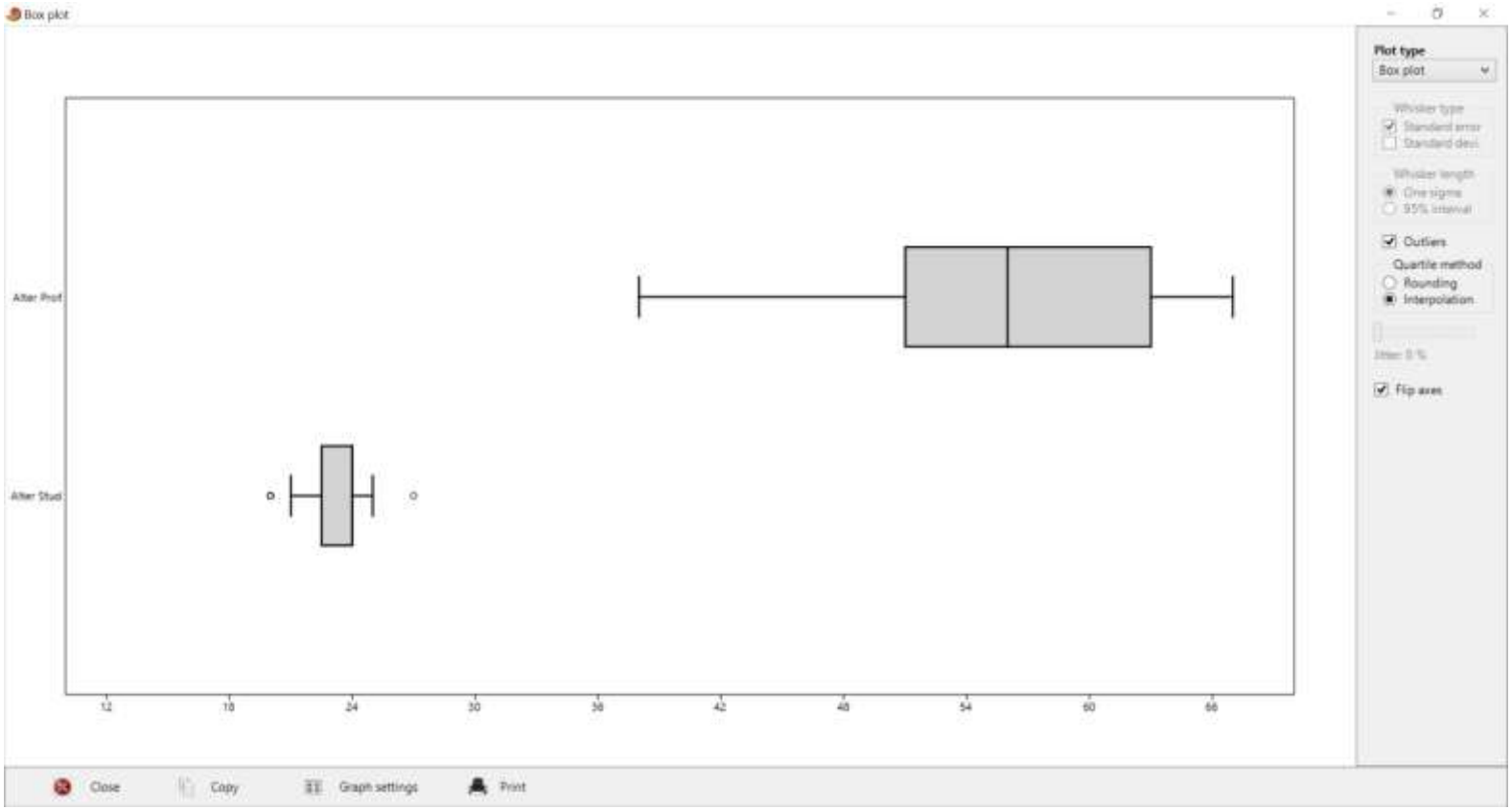
> Plot type > Box plot
> Quartile method > interpolation
> Graph settings (u.a. Colors, Axes…)

Erstellung vergleichender Box-Plots
(nach Erweiterung des Datensatzes)

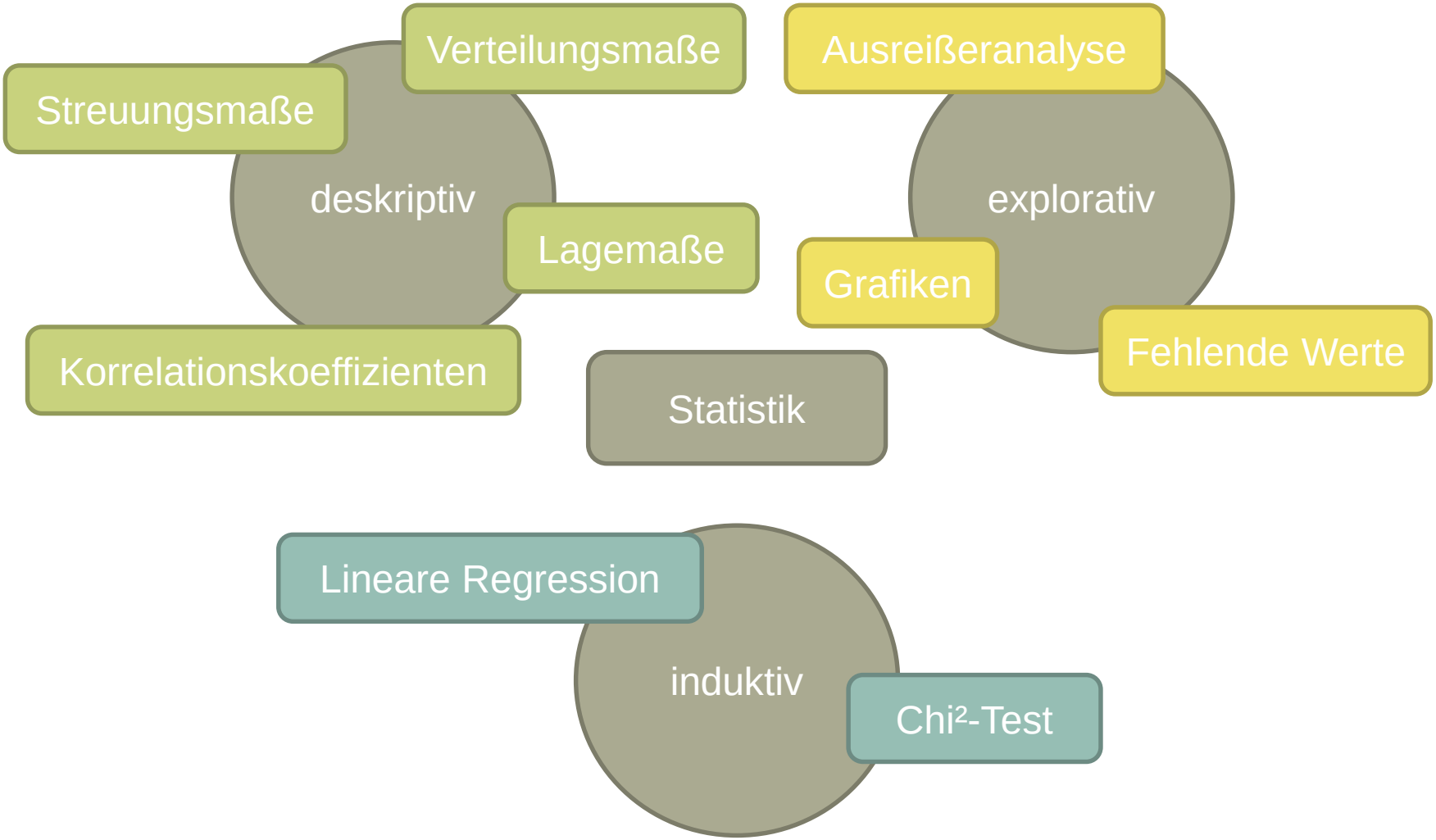


Erstellung vergleichender Box-Plots

> Plot > Barchart/Boxplot



Wo befinden wir uns?



Beispieldatensatz zur linearen Regression

Nr.	x	y
1	12	10000
2	15	15000
3	8	6000
4	11	11000
5	3	5000
6	17	23000
7	24	37000

Beispielfall mit bewusst gering gehaltener (Foliendarstellung…) Anzahl von Werten:

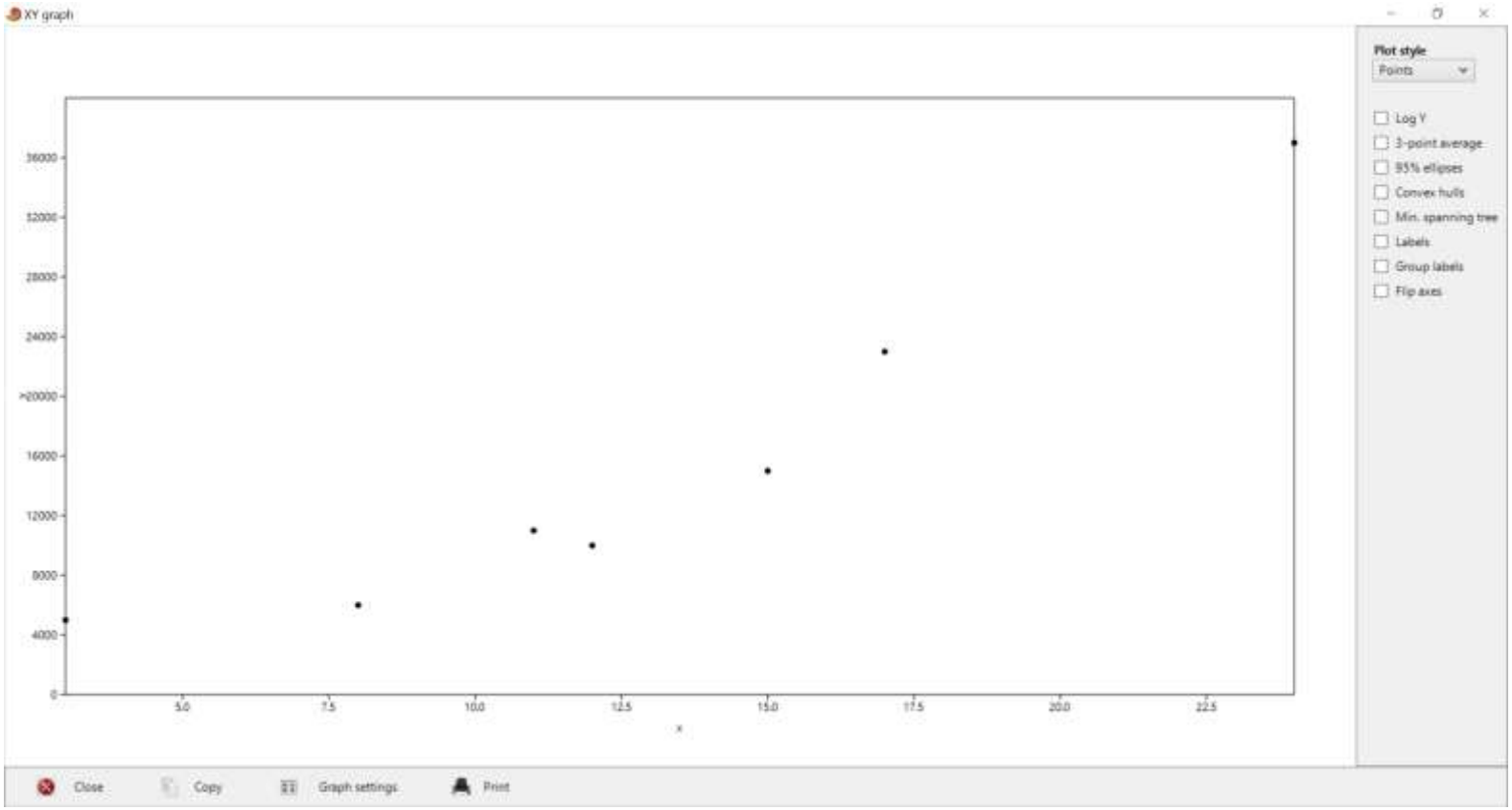
- x = Prozentualer Anteil des Werbebudgets eines Produkts am Gesamtbudget der Firma
- y = Verkaufte Einheiten des betrachteten Produkts in einem Untersuchungszeitraum
- Annahme: Das betrachtete Produkt, der Untersuchungszeitraum sowie das Gesamtbudget bleiben gleich

  (ceteris paribus)

Wie lautet die Regressionsgleichung?

Ist ein Zusammenhang grafisch plausibel?

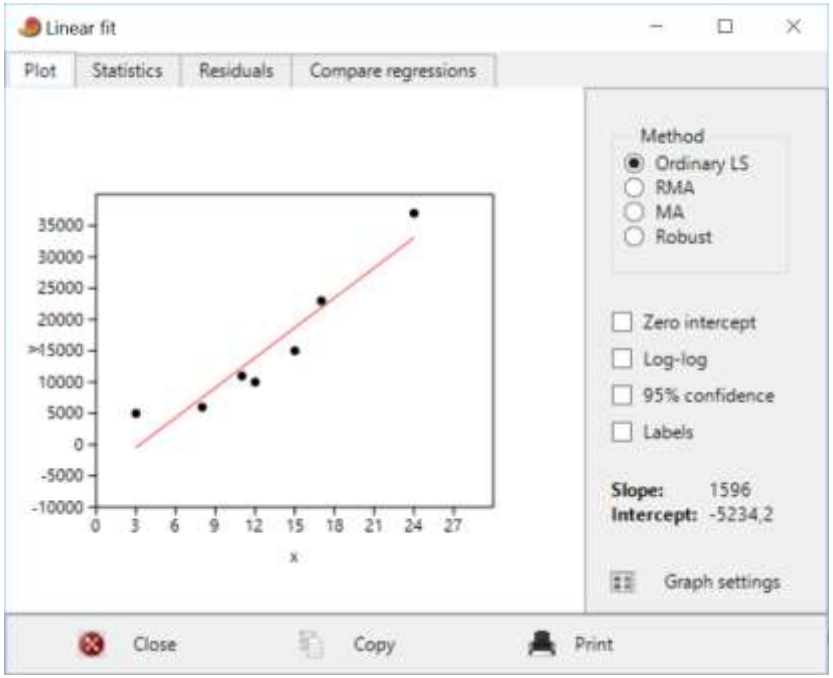
> Plot > XY graph

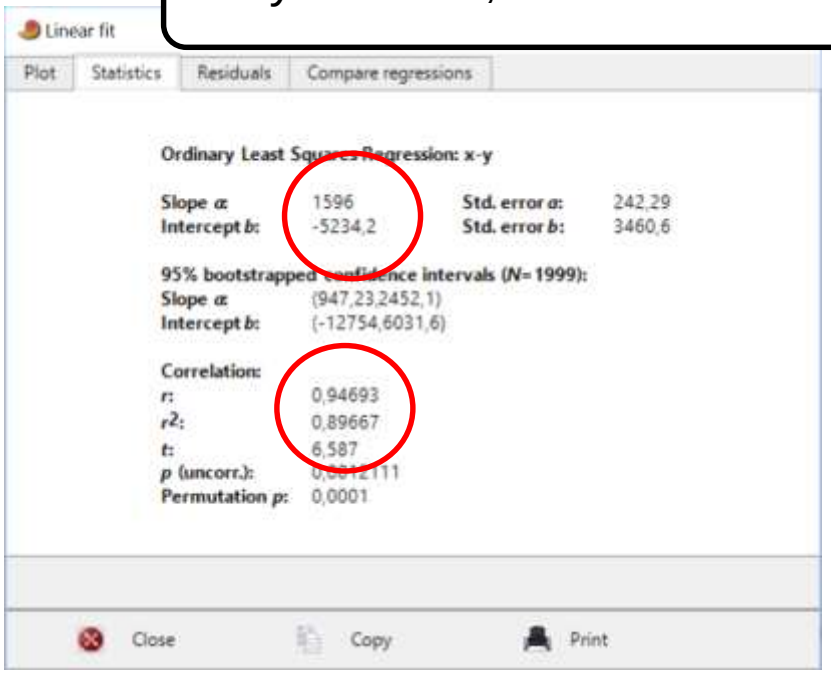


Erstellung und Bewertung des LR-Modells

> Model > Linear > Bivariate

$$y = -5.234{,}2 + 1.596 * x$$

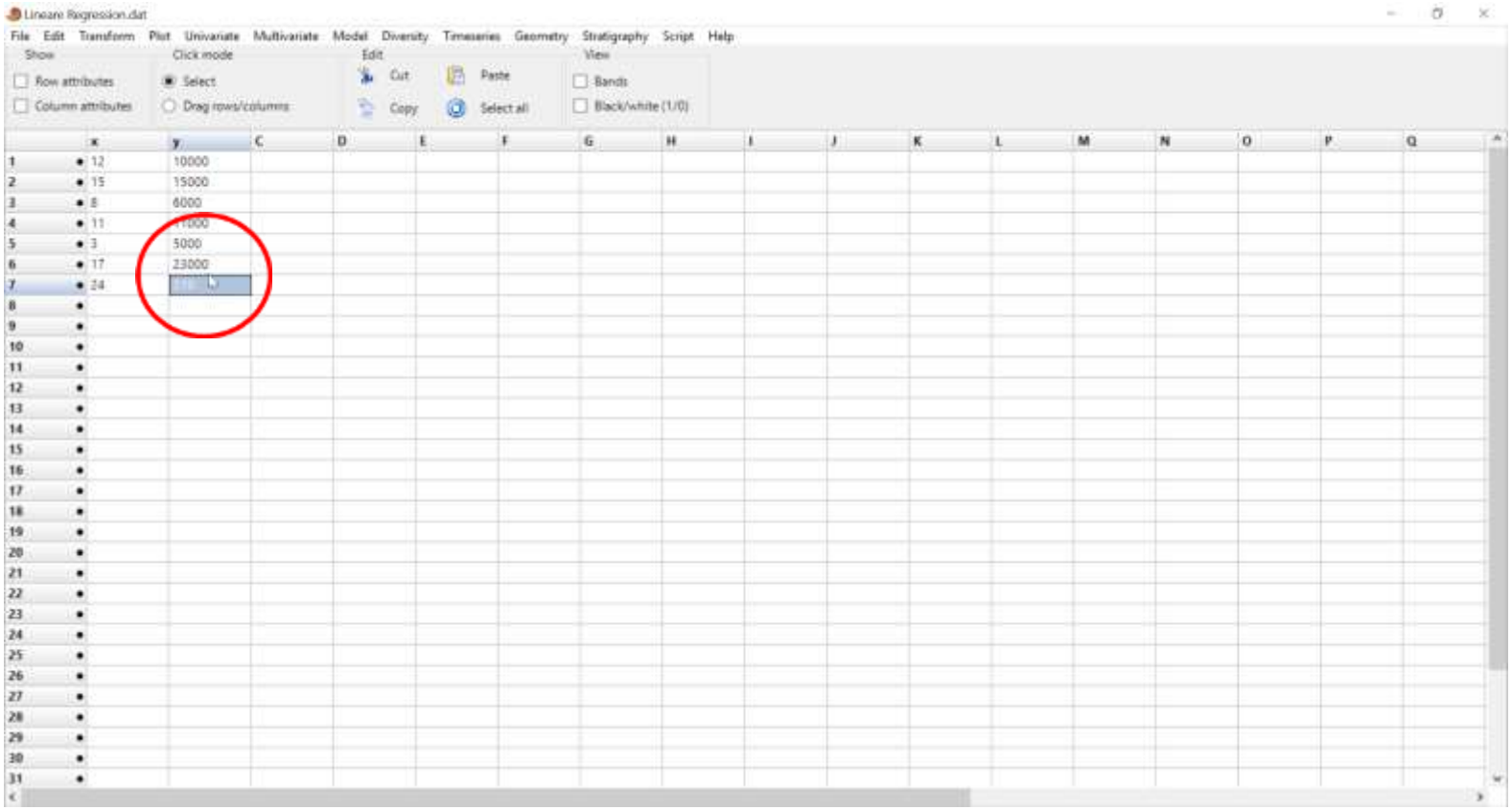




Was ist hier was?

Slope = Konstantes Glied
Intercept = Regressionskoeffizient
r^2 = Bestimmtheitsmaß / Gütekriterium

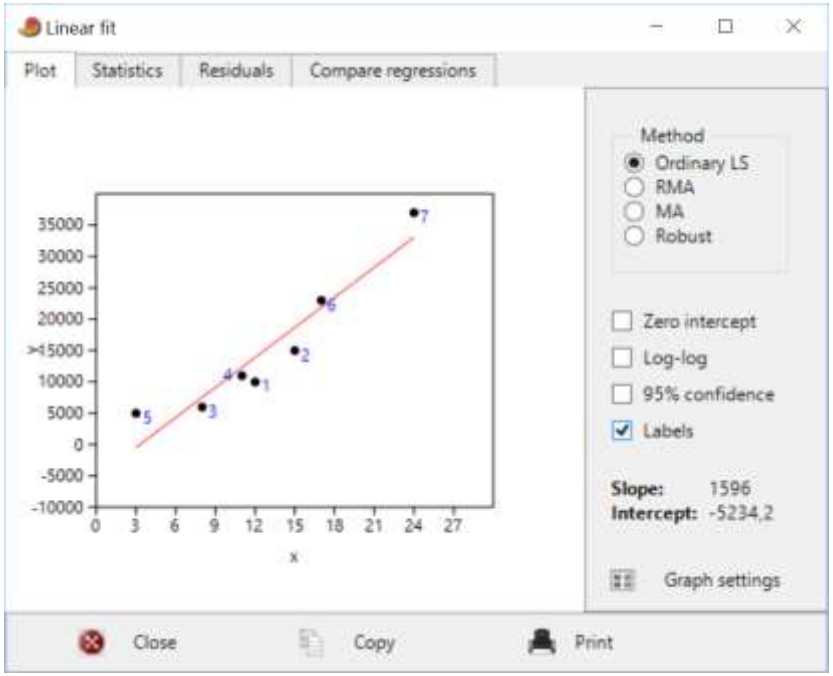
Sichtbarmachung des Leverage-Effekts
(Was eine kleine Änderung bewirken kann…)

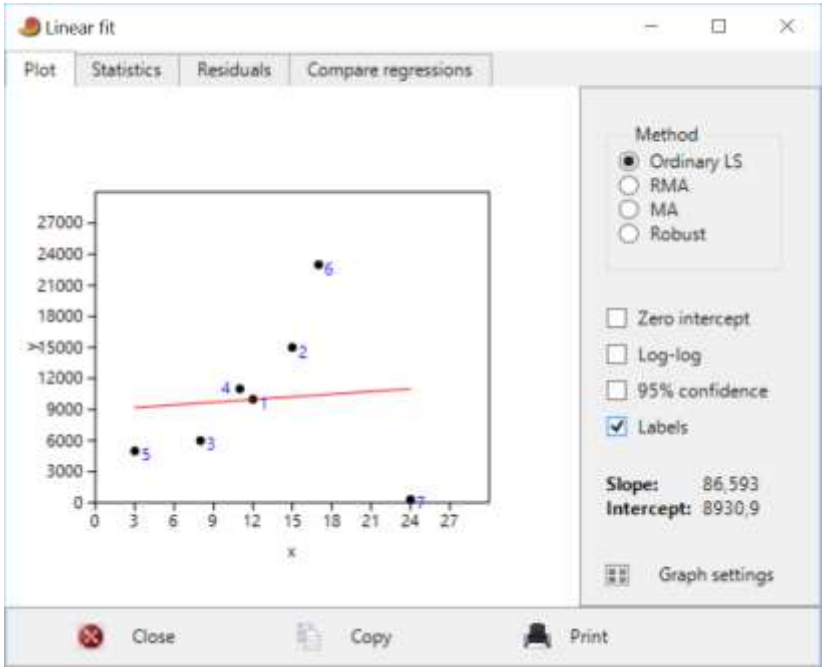


Sichtbarmachung des Leverage-Effekts

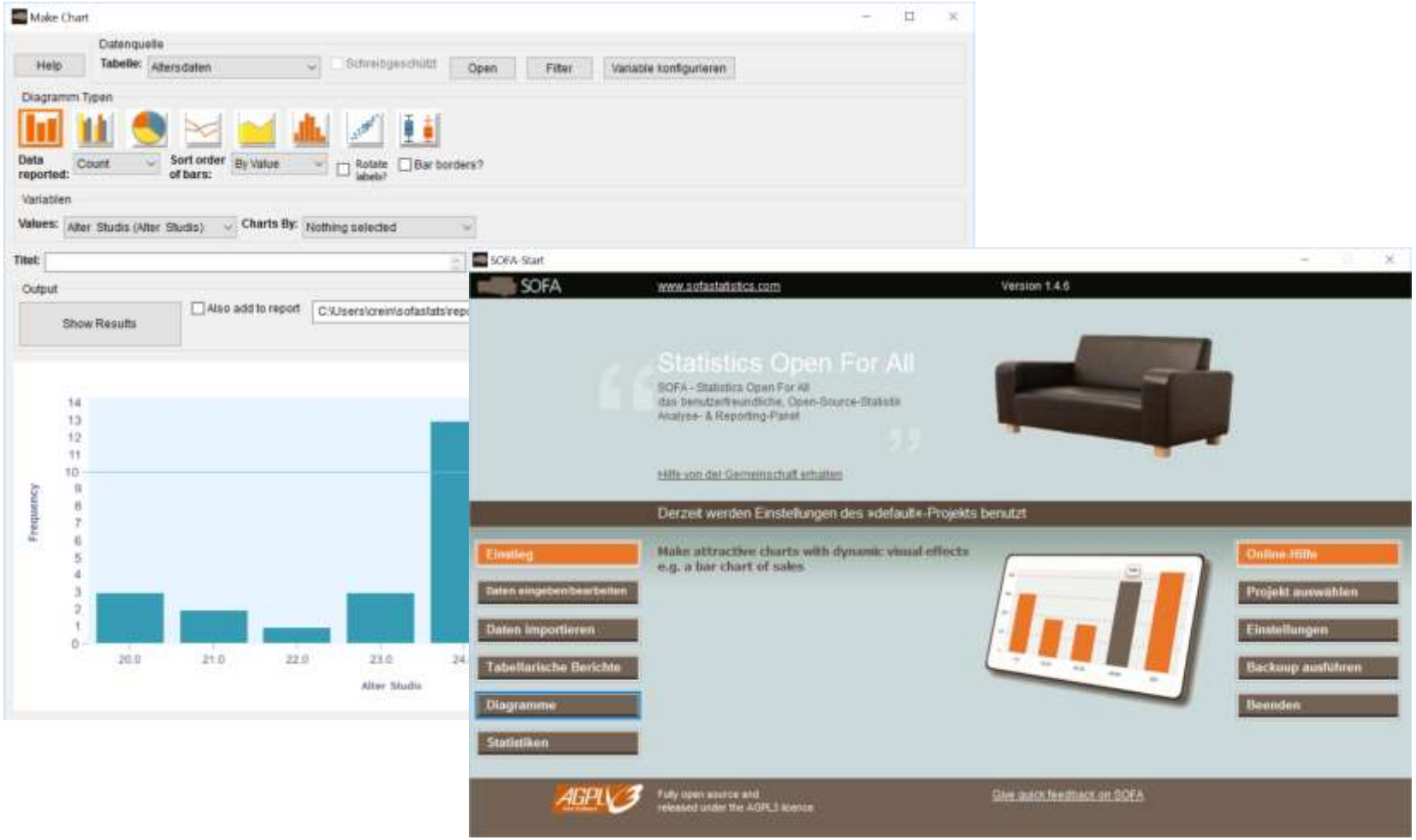
> Model > Linear > Bivariate

Wie deutlich verschlechtert sich hier r^2?



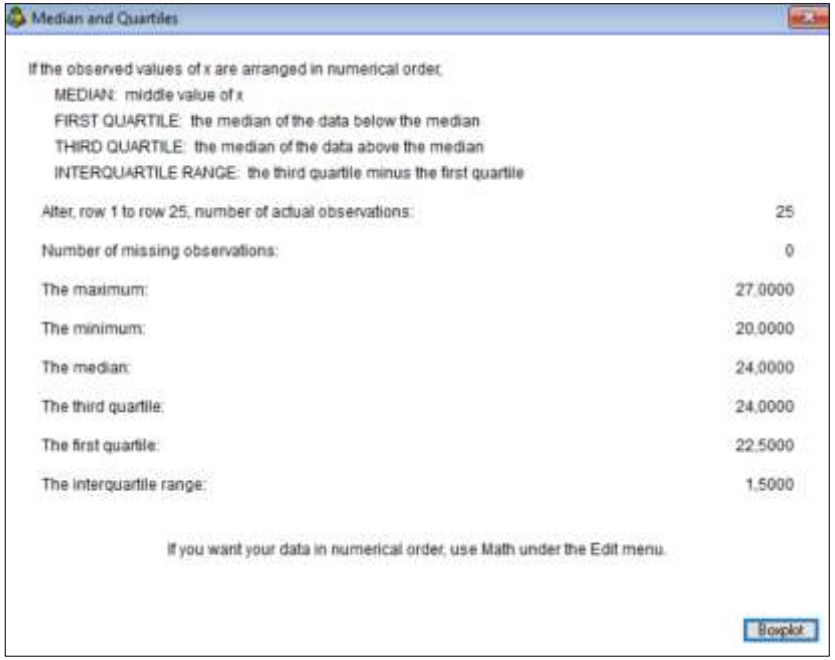


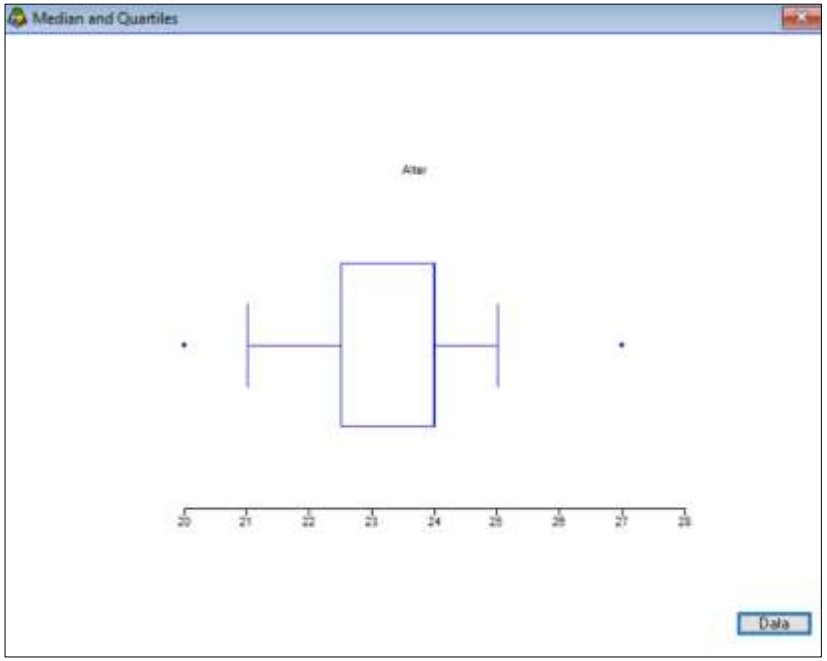
Was kann andere (freie) Software (besser)?



Erstellung von Box-Plots mit SSP

> Describing Data > Median, Quartiles > Box-Plot

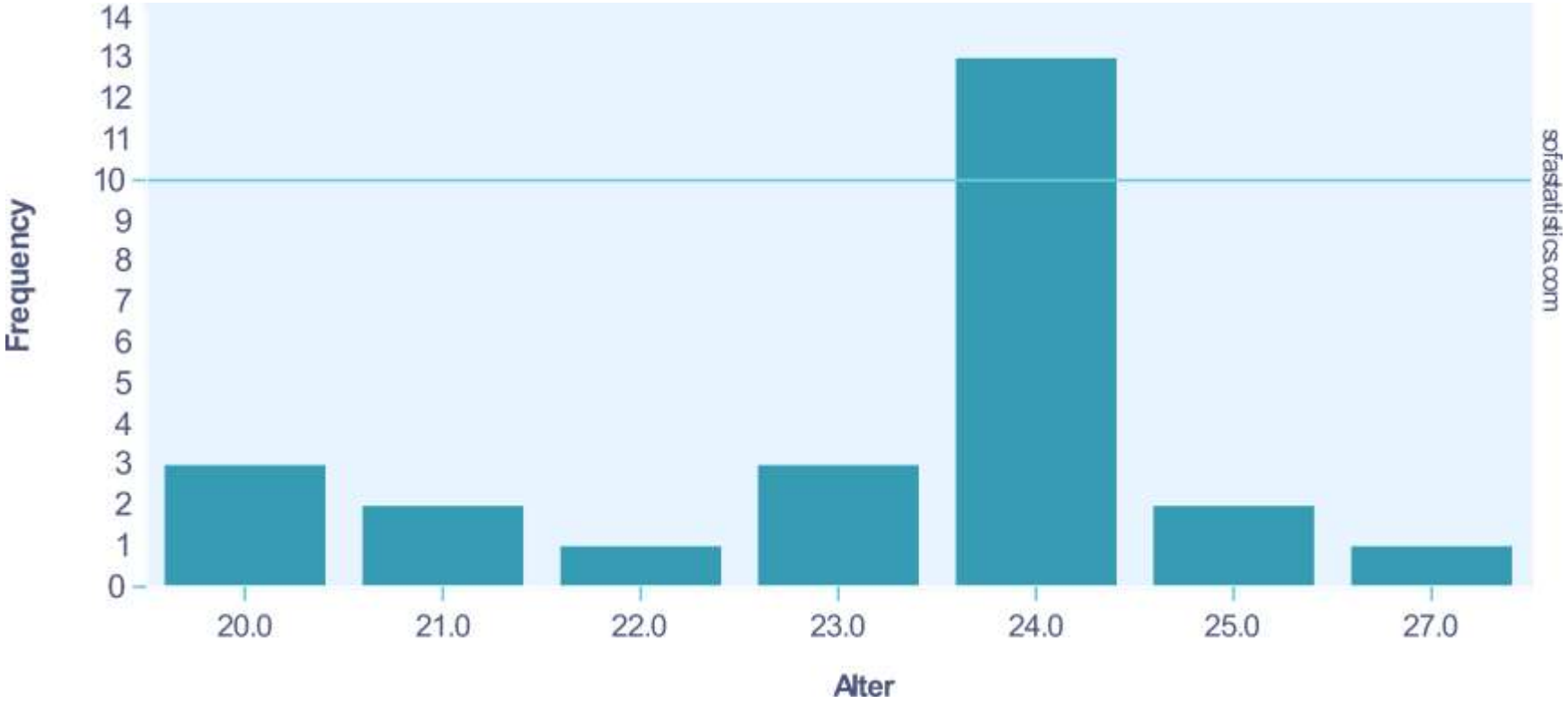




Schöne Übersicht der Konstruktionsgrößen – weniger ansehnlicher Box-Plot

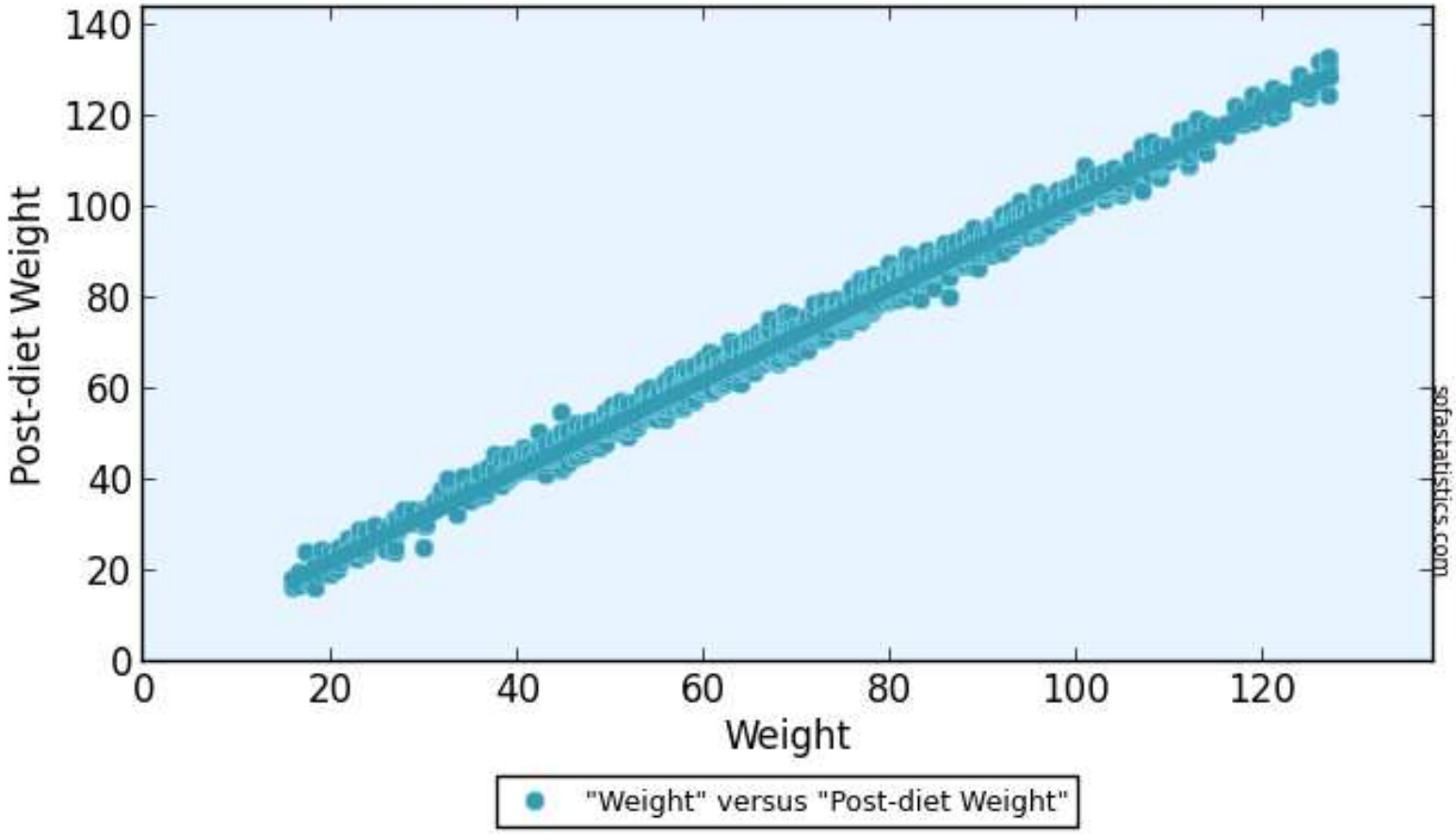
Erstellung „schöner“ Grafiken mit SOFA

> Diagramme > Balkendiagramm erstellen



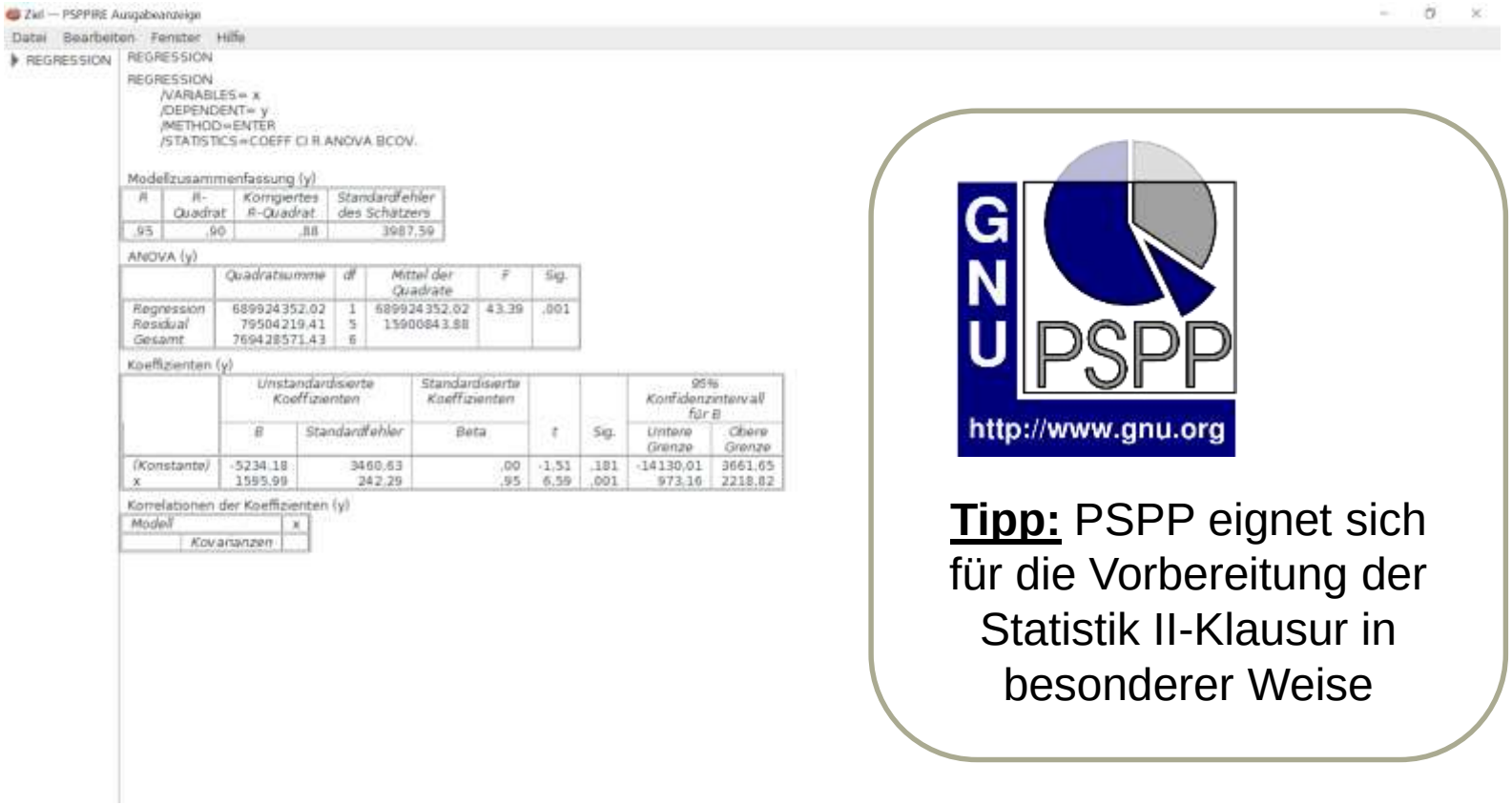
Erstellung „schöner“ Grafiken mit SOFA

> Diagramme > Scatterplot erstellen



Detailergebnisse der Regression in PSPP

> Analysieren > Regression > Linear



Zwi — PSPPIRE Ausgabeanzeige

Datei Bearbeiten Fenster Hilfe

REGRESSION

REGRESSION

```
REGRESSION
  /VARIABLES= x
  /DEPENDENT= y
  /METHOD=ENTER
  /STATISTICS=COEFF CI R ANOVA BCOV.
```

Modellzusammenfassung (y)

R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
,95	,90	,88	3987,59

ANOVA (y)

	Quadratsumme	df	Mittel der Quadrate	F	Sig.
Regression	689924352,02	1	689924352,02	43,39	,001
Residual	79504219,41	5	15900843,88		
Gesamt	769428571,43	6			

Koeffizienten (y)

	Unstandardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.	95% Konfidenzintervall für B	
	B	Standardfehler	Beta			Untere Grenze	Obere Grenze
(Konstante)	-5234,18	3460,63	,00	-1,51	,181	-14130,01	3661,65
x	1595,99	242,29	,95	6,59	,001	973,16	2218,82

Korrelationen der Koeffizienten (y)

Modell		x
	Kovarianzen	

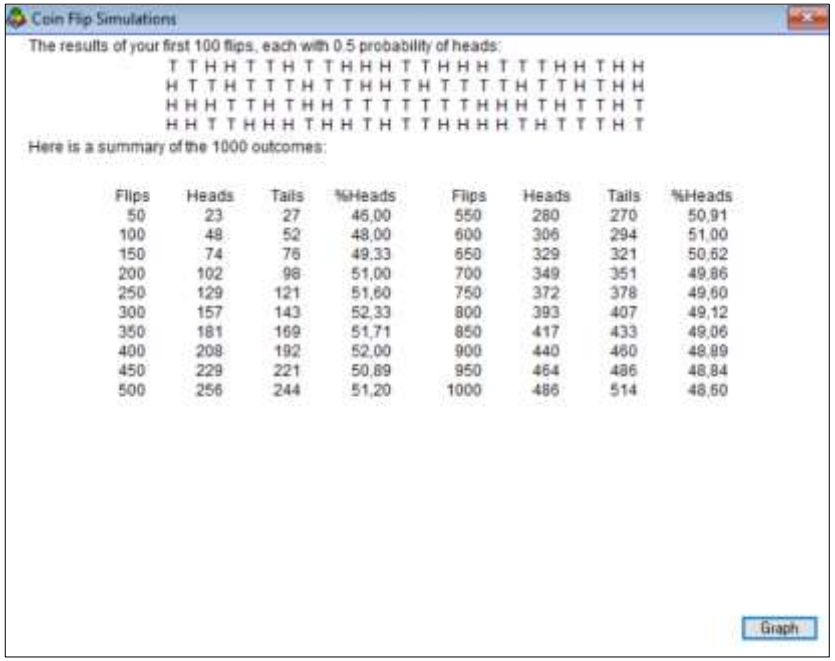
GNU PSPP

http://www.gnu.org

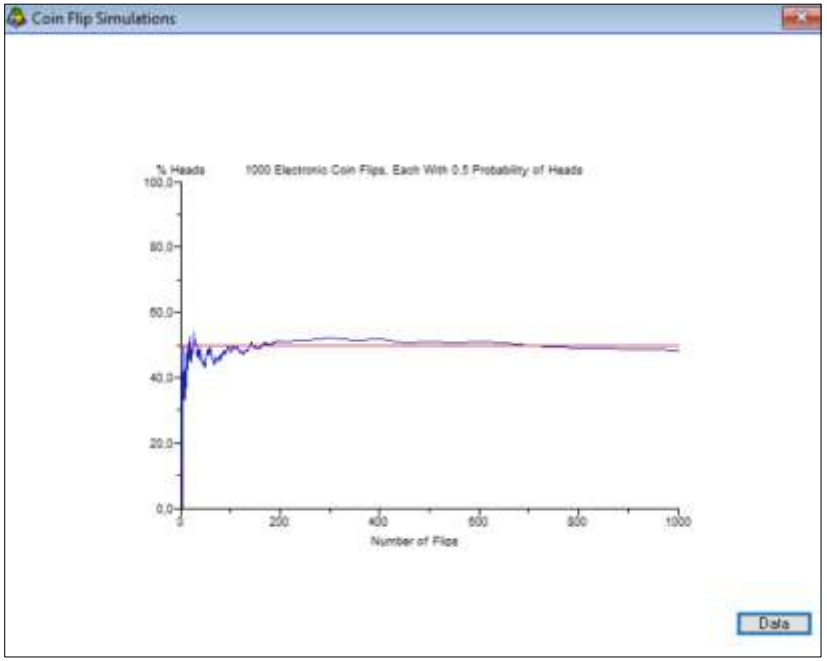
Tipp: PSPP eignet sich für die Vorbereitung der Statistik II-Klausur in besonderer Weise

Simulation von Münzwürfen in SSP

> Uncertainty > Coin Flip Simulation



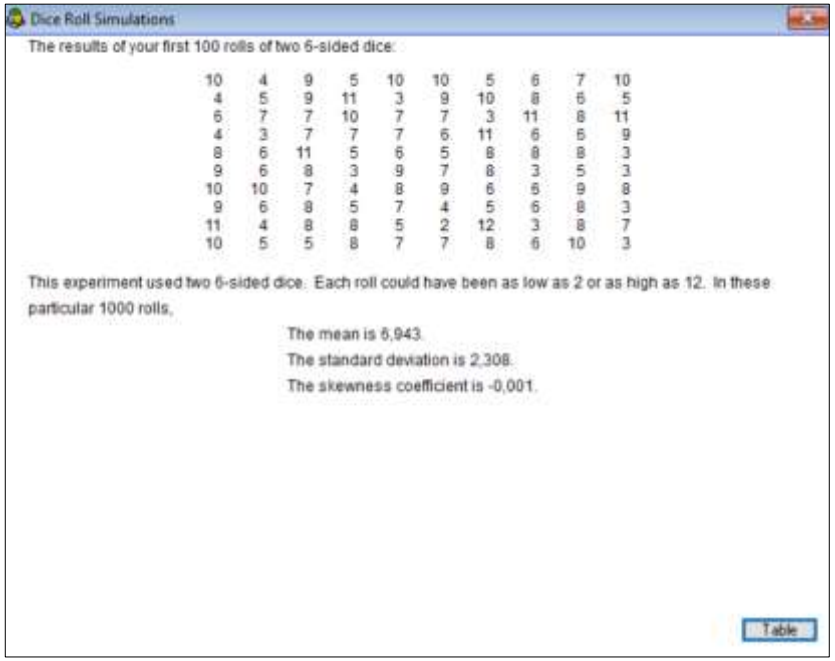
Flips	Heads	Tails	%Heads	Flips	Heads	Tails	%Heads
50	23	27	46,00	550	280	270	50,91
100	48	52	48,00	600	306	294	51,00
150	74	76	49,33	650	329	321	50,62
200	102	98	51,00	700	349	351	49,86
250	129	121	51,60	750	372	378	49,60
300	157	143	52,33	800	393	407	49,12
350	181	169	51,71	850	417	433	49,06
400	208	192	52,00	900	440	460	48,89
450	229	221	50,89	950	464	486	48,84
500	256	244	51,20	1000	486	514	48,60

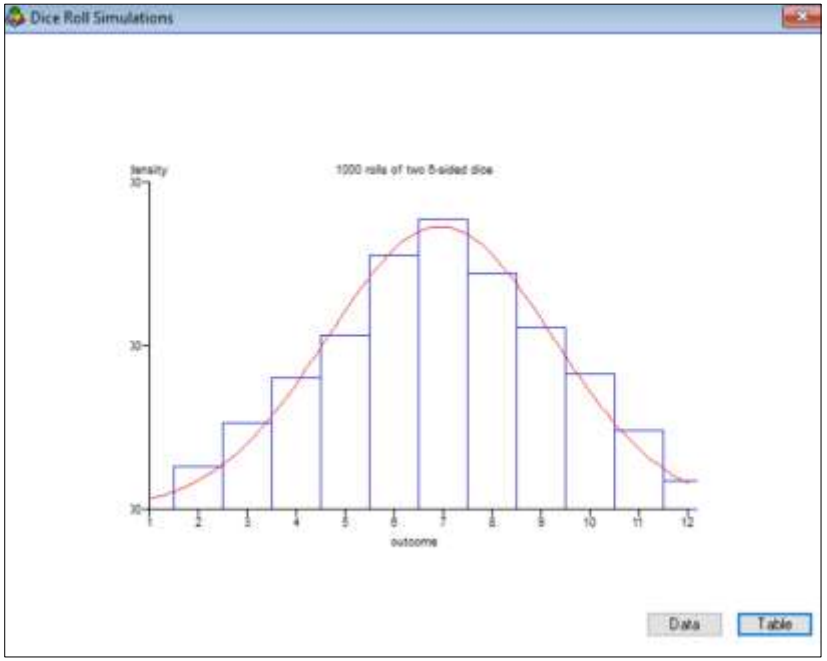


<u>Gesetz der Großen Zahlen:</u> Die relative Häufigkeit eines Zufallsergebnisses stabilisiert sich um die theoretische Wahrscheinlichkeit eines Zufallsergebnisses, wenn das zu Grunde liegende Zufallsexperiment immer wieder unter denselben Voraussetzungen durchgeführt wird.

Simulation von Würfelwürfen in SSP

> Uncertainty > Dice Roll Simulation



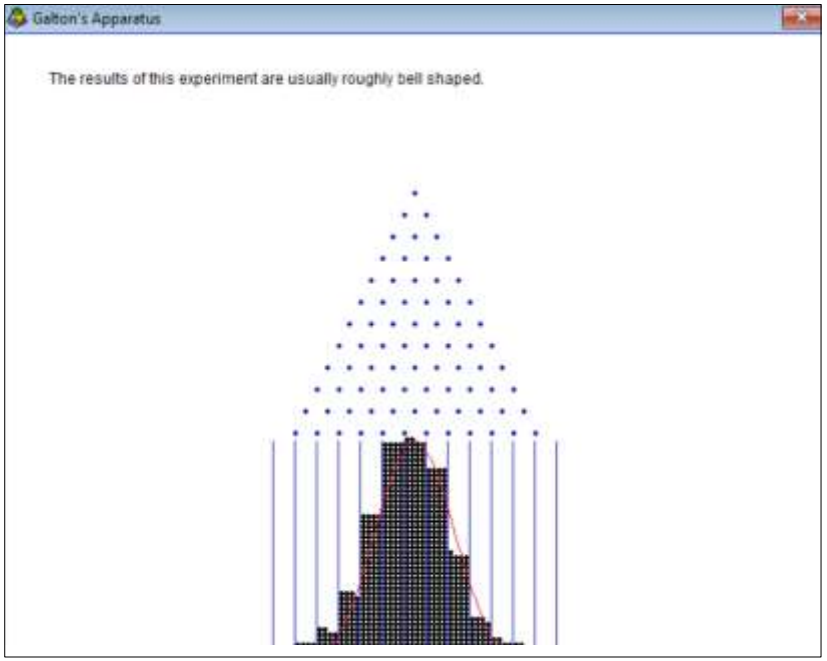


<u>Gesetz der Großen Zahlen:</u> Die relative Häufigkeit eines Zufallsergebnisses stabilisiert sich um die theoretische Wahrscheinlichkeit eines Zufallsergebnisses, wenn das zu Grunde liegende Zufallsexperiment immer wieder unter denselben Voraussetzungen durchgeführt wird.

Simulation eines Galtonbretts in SSP

> Uncertainty > Galton‘s Apparatus

Foto: Klaus-Dieter Keller; Lizenz: gemeinfrei; Quelle: Wikimedia



Mit Hilfe eines Galtonbretts lässt sich visuell demonstrieren, warum viele Zufallsvariablen der Bionomialverteilung folgen.

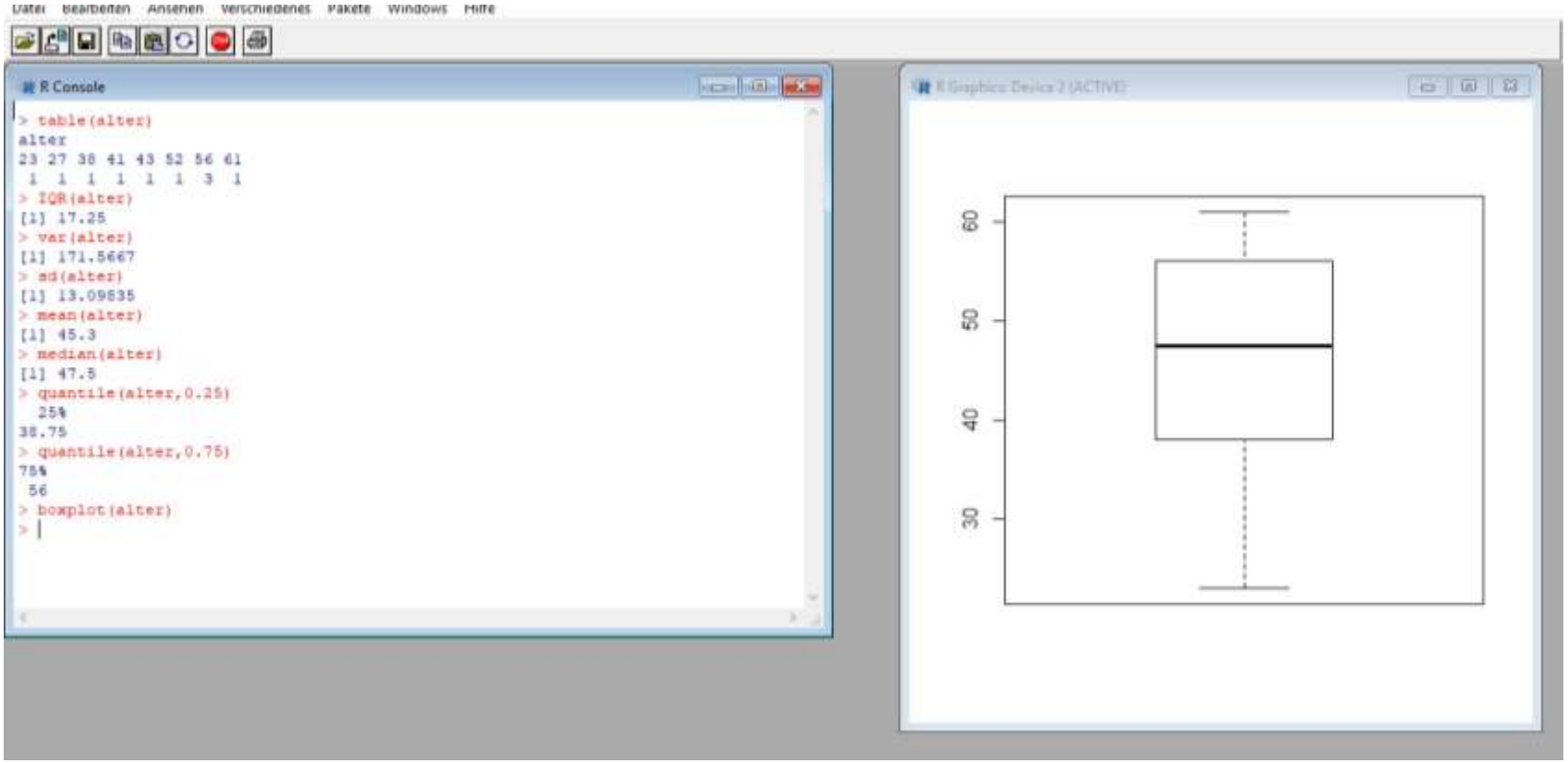
Bestimmung der optimalen Stichprobengröße

$$n = \frac{\frac{Z^2 * p * q}{e^2}}{1 + \frac{\frac{Z^2 * p * q}{e^2} - 1}{N}}$$

- Was passiert bei....
  - größerer Grundgesamtheit?
  - kleinerer Grundgesamtheit?
  - bekannten Anteilswerten?
  - weniger Sicherheit?
  - mehr Sicherheit?

Kostenloser Download unter:
http://www.statistikberatung.eu/SampleSizer.zip

…und das Beste kommt zum Schluss: R



Statistische Software

Einführung in die Nutzung von R

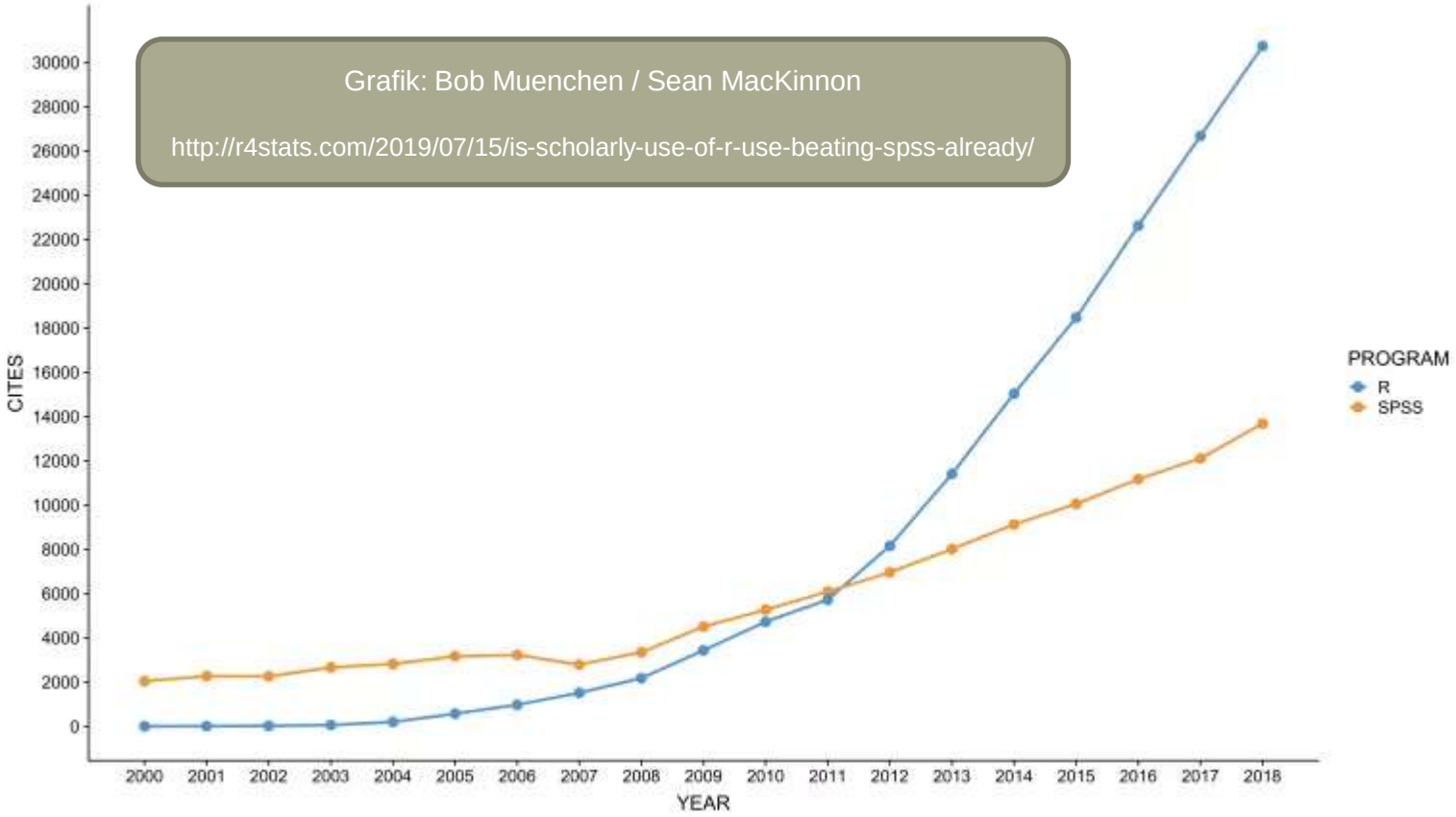
▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

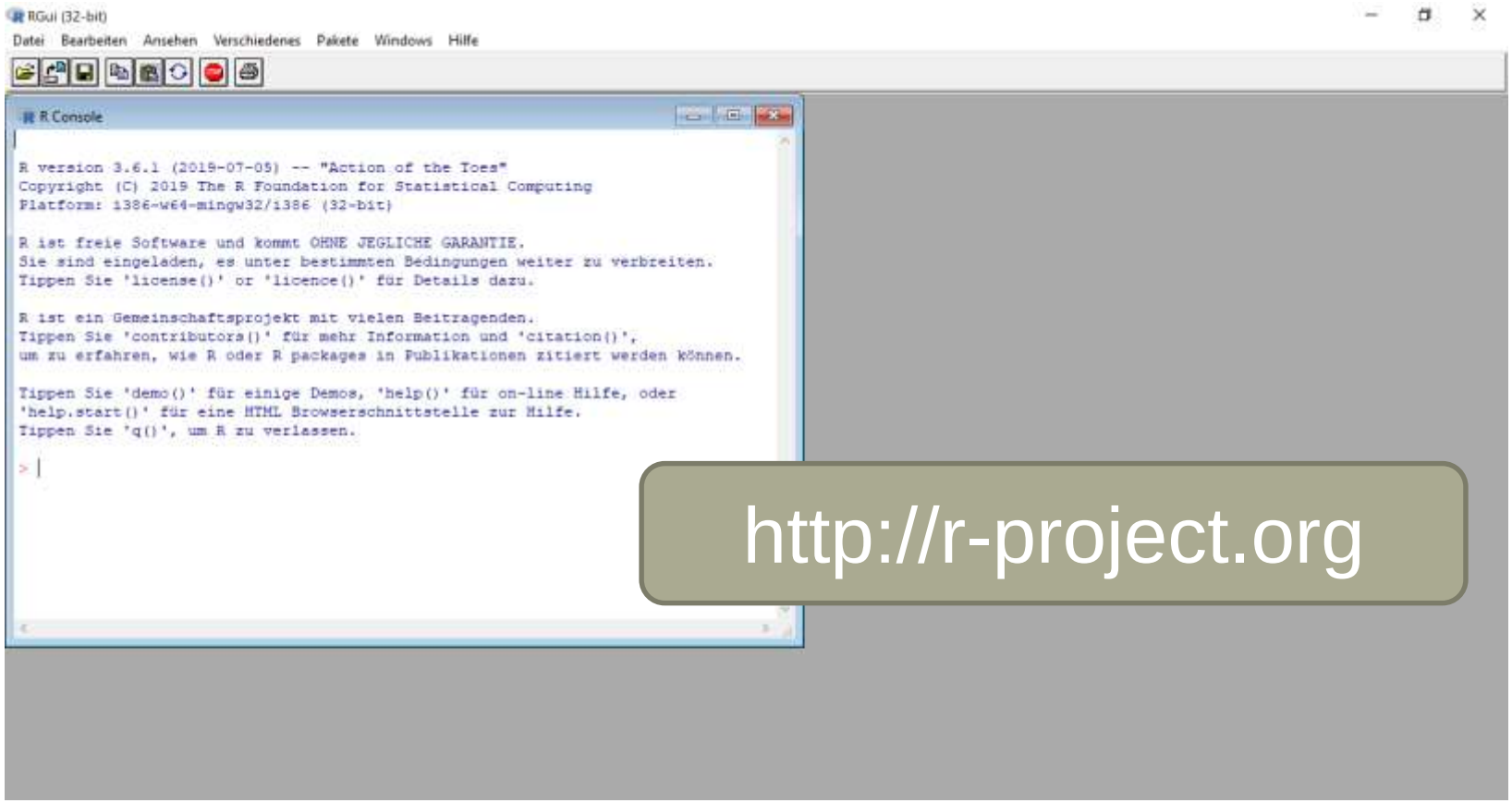
Was ist R?

- R ist eine Programmiersprache, entwickelt 1992 von Ross Ihaka und Robert Gentleman (Auckland)
- R ist Open Source-Software und somit nicht nur frei verfügbar sondern auch frei erweiterbar
- Mittlerweile stehen schon mehr als 12.000 dieser Erweiterungen (sog. Packages) zur Verfügung, viele davon aus der Statistik
- R wird immer populärer und lässt SPSS & Co. allmählich hinter sich

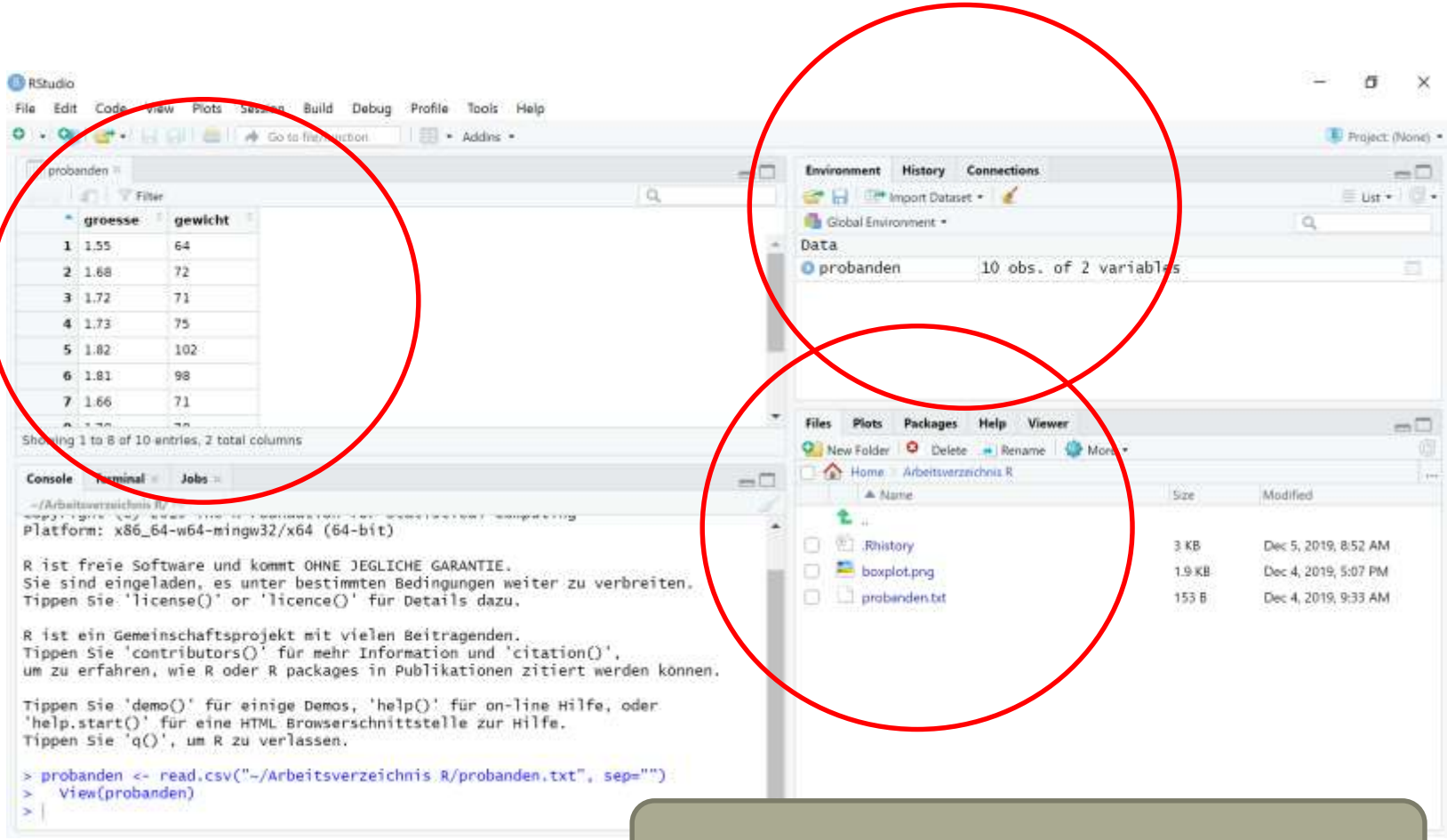
Nennung von R in Fachpublikationen



Download von R



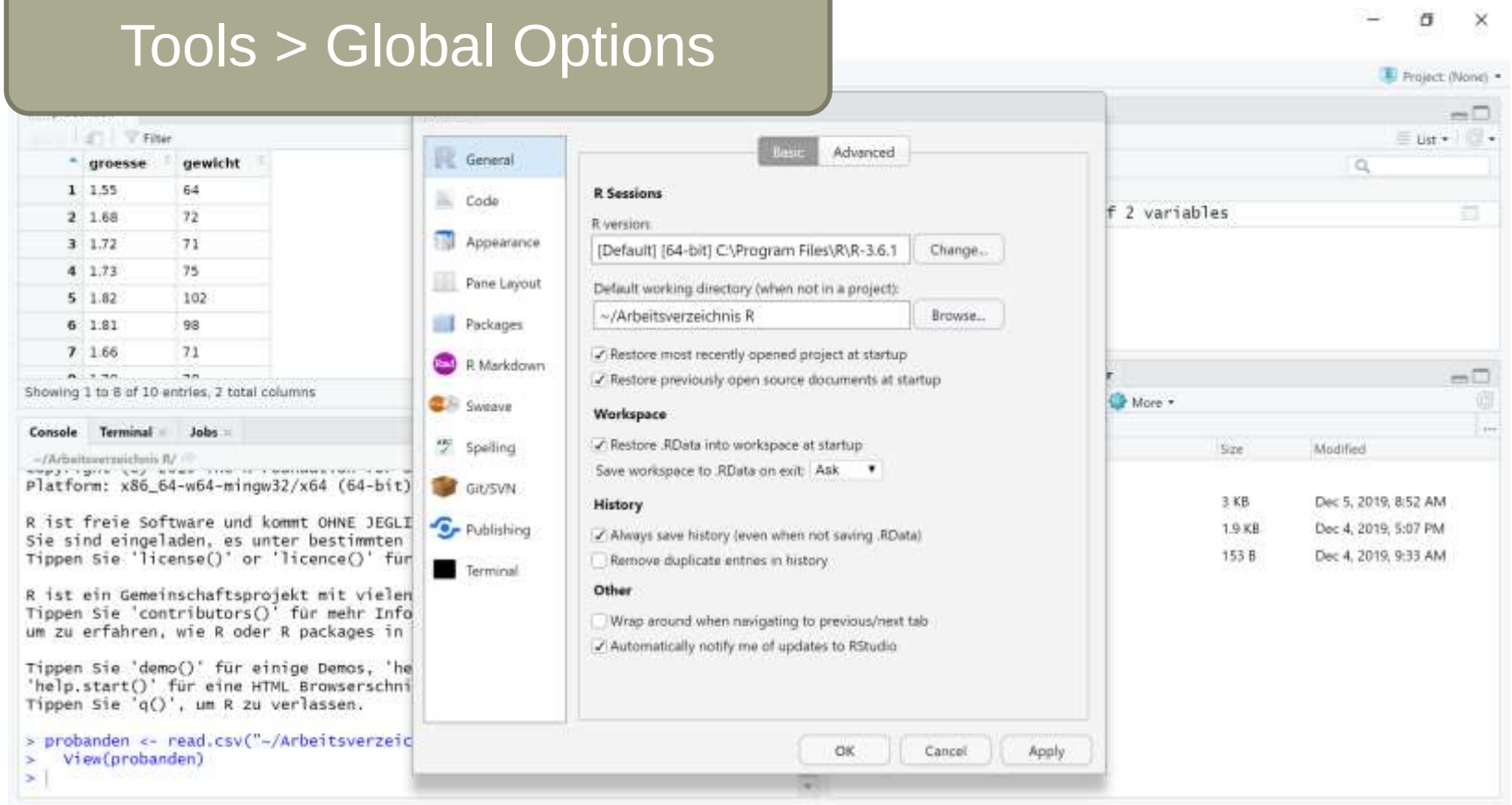
Download von RStudio



https://rstudio.com

Einrichtung des Arbeitsverzeichnisses

Tools > Global Options



Nutzung von R als „Taschenrechner“

Testen wir einmal folgende Eingaben…

1+5

5-1

2*3

3/2

sqrt(4) -> Square Root = Quadratwurzel

Warum diese Form?

– Bei sqrt() handelt es sich um eine **Funktion**, die als Ergebnis die Quadratwurzel einer Zahl liefert, die der Funktion beim Aufruf als **Argument** übergeben wird.

Unser Übungsdatensatz (aus der Vorlesung)

Befragte/r	Größe (m)	Gewicht (kg)
1	1,55	64
2	1,68	72
3	1,72	71
4	1,73	75
5	1,82	102
6	1,81	98
7	1,66	71
8	1,78	78
9	1,73	77
10	1,59	69

Anlegen eines Datensatzes in R

Eingabe von zwei Datenreihen in Form von Vektoren

```
gewicht<-c(64,72,71,75,102,98,71,78,77,69)
groesse<-c(1.55,1.68,1.72,1.73,1.82,1.81,1.66,1.78,1.73,1.59)
```

(c = combine)

Zusammenführen der Datenreihen zu einem Datensatz

```
probanden<-data.frame(groesse,gewicht)
```

Ausgabe der eingegebenen Daten

```
gewicht
probanden
```

Was passiert im Fenster „Global Environment“?

Anlegen eines Datensatzes in R

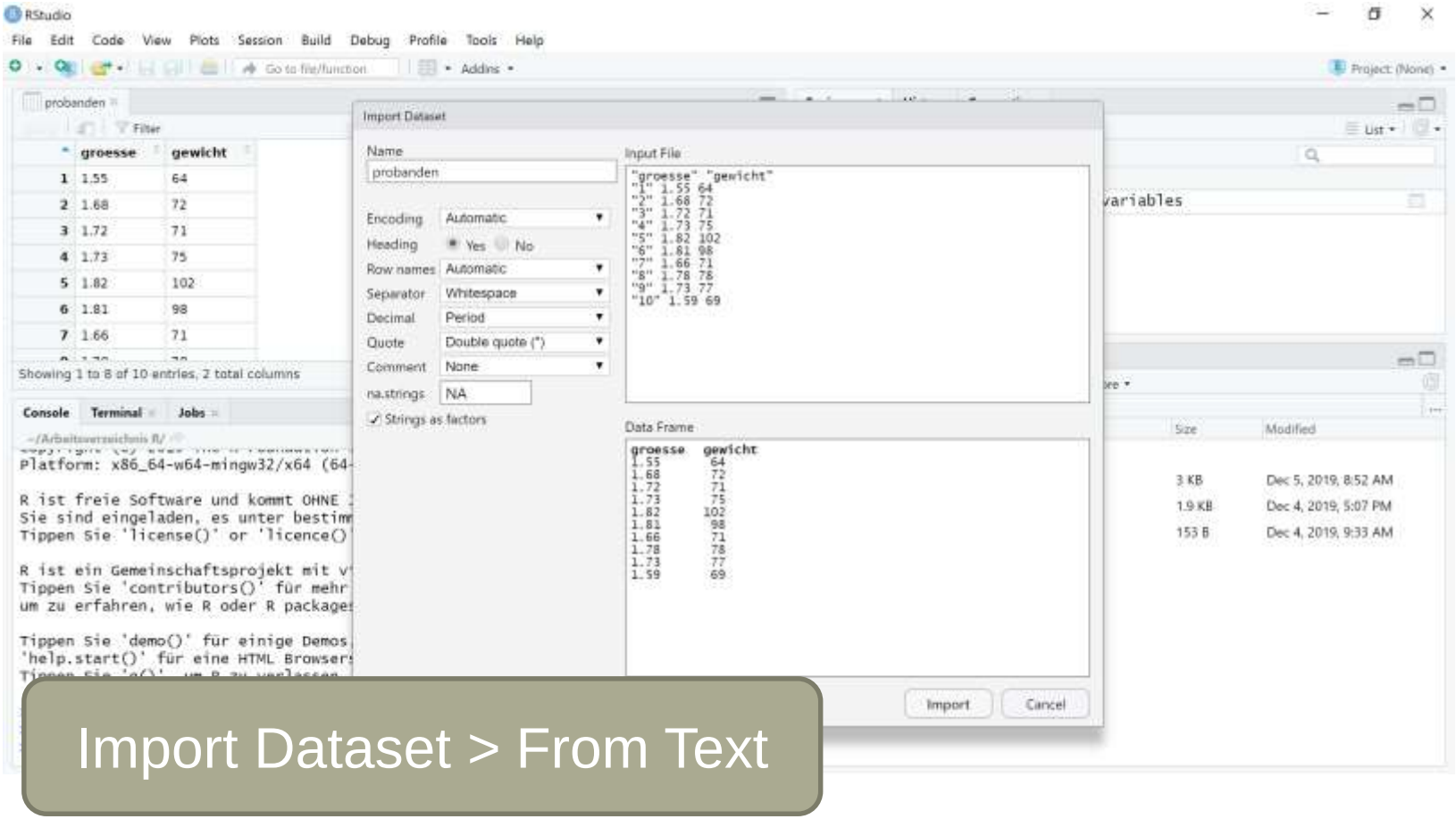
Ausgabe der eingegebenen Daten

probanden$gewicht
length(probanden$gewicht)
ls() -> Anzeigen aller Objekte

Abspeichern und Laden dieses Datensatzes

write.table(probanden,"probanden.txt")
rm(probanden,gewicht,groesse) (rm = remove)
read.table(„probanden.txt“) -> generiert nur eine Ausgabe
probanden<-read.table("probanden.txt") -> Zuweisung zu einem Frame

Geht das nicht auch viel einfacher…?



Import Dataset > From Text

Einige einfache statistische Auswertungen

Grundlegende Angaben

sum(probanden$gewicht)	Summe aller Werte
min(probanden$gewicht)	Kleinster Wert im Datensatz
max(probanden$gewicht)	Größter Wert im Datensatz

Statistische Lagemaße

mean(probanden$gewicht)	Arithmetisches Mittel
median(probanden$gewicht)	Median / 50%-Perzentil
quantile(probanden$gewicht,0.25)	25%-Perzentil (frei änderbar)
summary(probanden$gewicht)	Sechs-Werte-Zusammenfassung

Einige einfache statistische Auswertungen

Statistische Streuungsmaße

IQR(probanden$gewicht)	Interquartilsabstand
var(probanden$gewicht)	Varianz
sd(probanden$gewicht)	Standardabweichung

Spannweite und Variationskoeffizient sind nicht drin – und nun?

max(probanden$gewicht) - min(probanden$gewicht)	-> Spannweite
sd(probanden$gewicht)/mean(probanden$gewicht)	-> Variationskoeffizient

Learning: Was es in R (noch) nicht gibt, kann man sich selbst zusammenstellen…

Erstellung einfacher Grafiken in R

boxplot(probanden$gewicht)

Wie lässt sich dieser Plot noch modifizieren?

boxplot(probanden$gewicht, col="lightblue")

Lassen sich auch mehr als zwei Argumente ergänzen?

boxplot(probanden$gewicht, col="lightblue", horizontal=TRUE)

Spielt die Reihenfolge der Argumente eine Rolle?

boxplot(probanden$gewicht, horizontal=TRUE, col="lightblue")

Erstellung einfacher Grafiken in R

Wie lässt sich der Boxplot noch modifizieren?

```
boxplot(probanden$gewicht)
boxplot(probanden$gewicht, range=0)
boxplot(probanden$gewicht, plot=FALSE)
boxplot(probanden$gewicht, border="green")
boxplot(probanden$gewicht, sub="Box-Whisker-Plot")
boxplot(probanden$gewicht, main="Gewichtsverteilung")
boxplot(probanden$gewicht, ylab="kg", xlab="Stichprobe")
```

Welche Farben kennt R denn?

```
colors()
```

Erstellung einfacher Grafiken in R

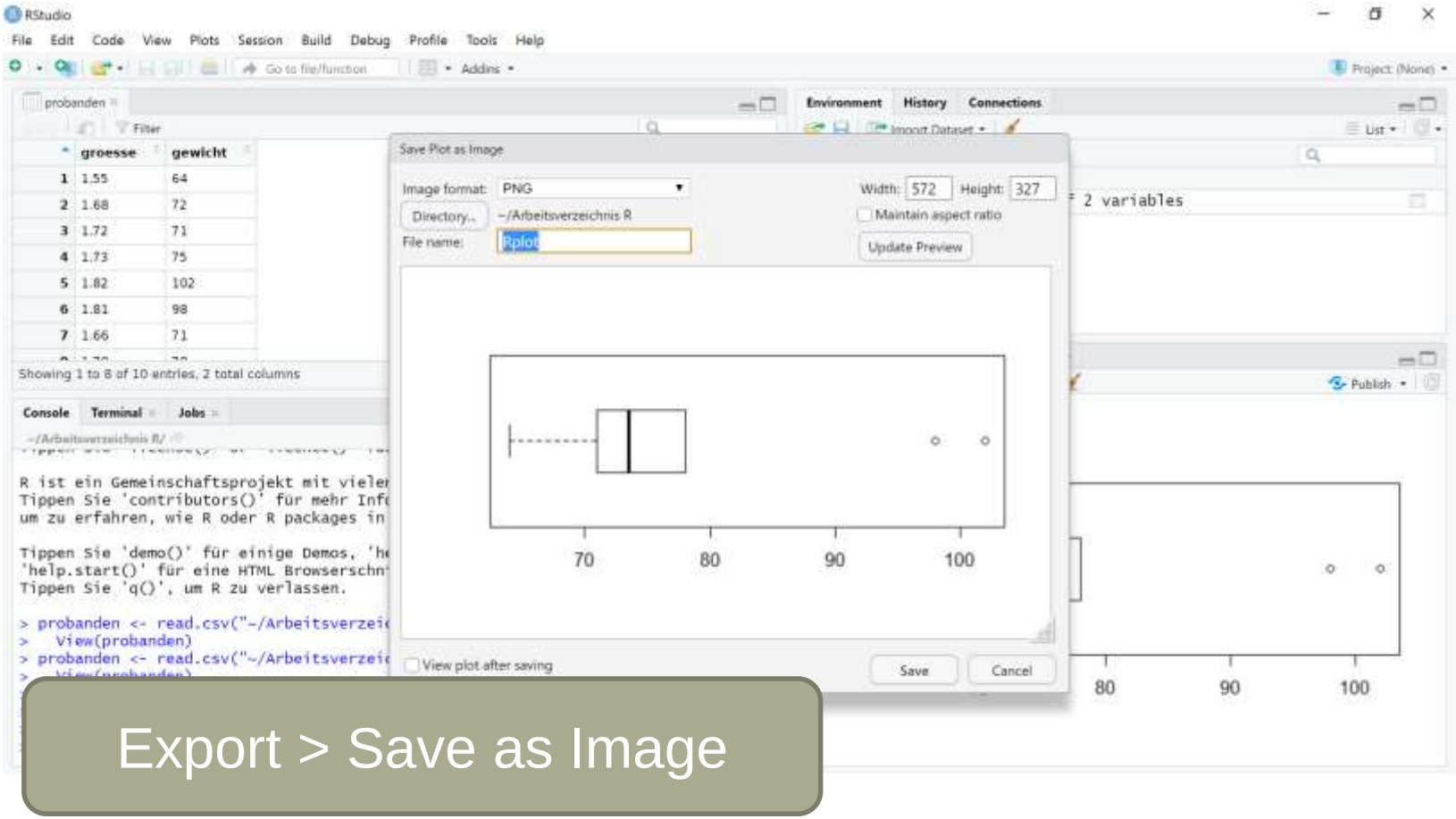
Was für Grafiktypen existieren noch?

hist()	Histogramm
pie()	Kreisdiagramm
barplot()	Balkendiagramm
plot()	Streudiagramm

Wie lässt sich diese Grafik nun abspeichern?

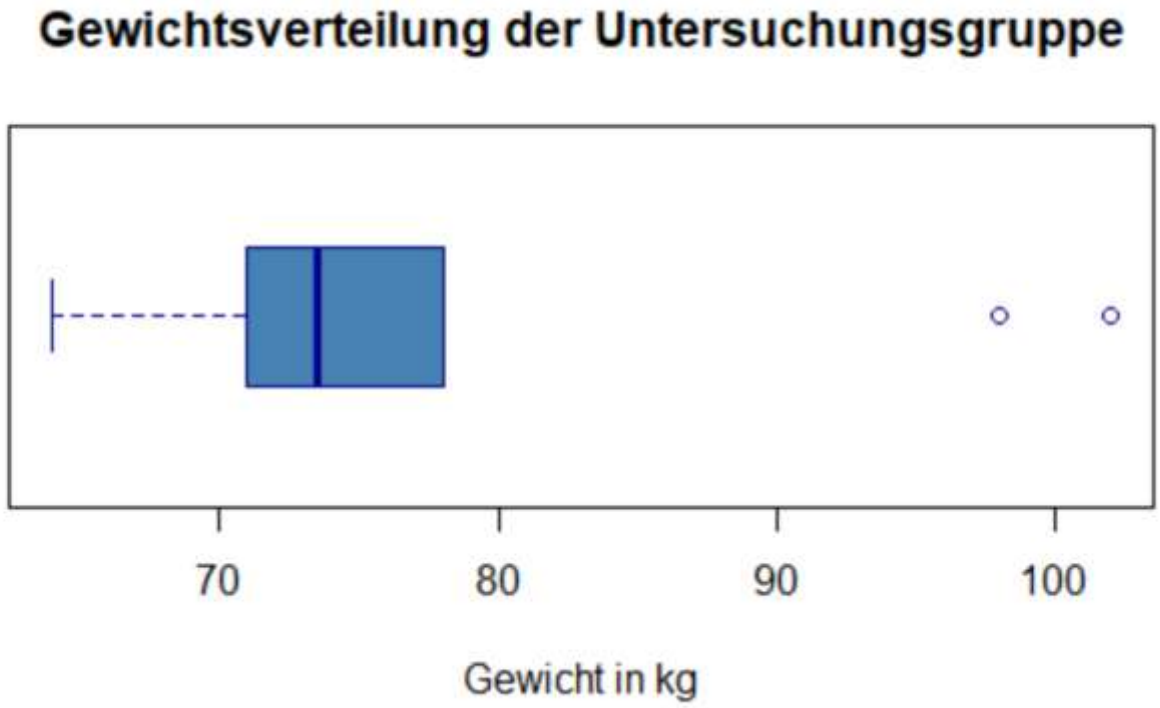
```
png("boxplot.png")
boxplot(probanden$gewicht)
dev.off()
```

Geht das nicht auch viel einfacher…?



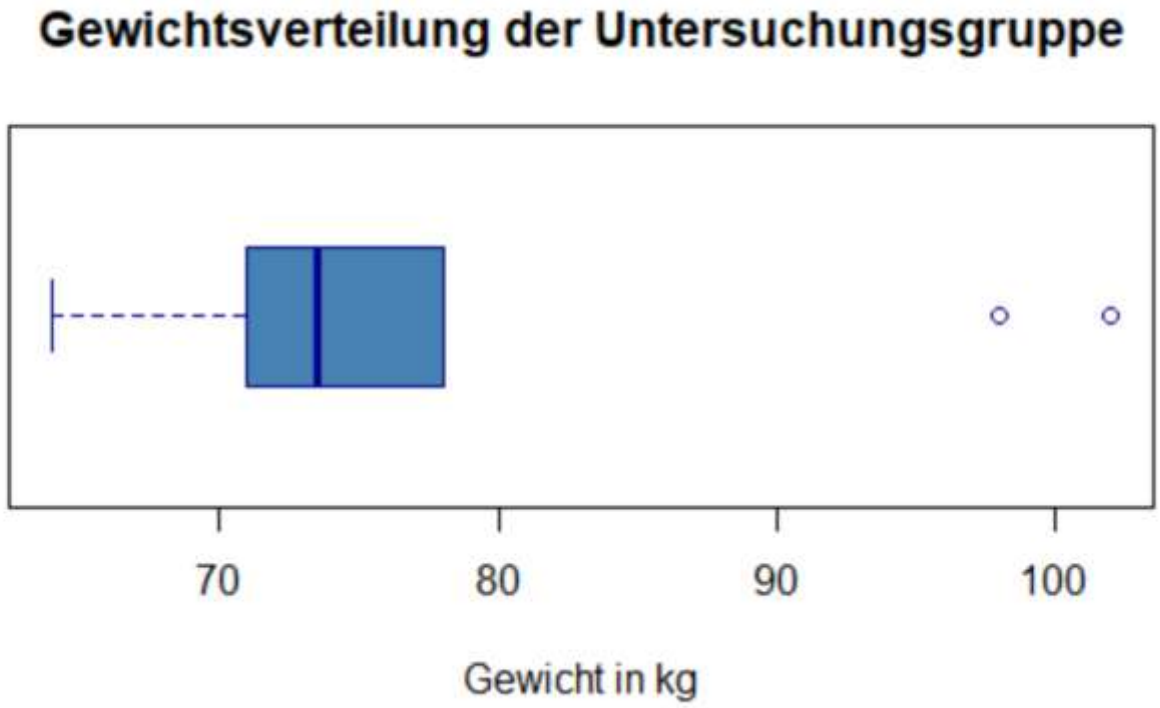
Export > Save as Image

Wie lässt sich dieses Diagramm erstellen?



?

Wie lässt sich dieses Diagramm erstellen?




```
boxplot(probanden$gewicht, horizontal = TRUE, border="darkblue",
col="steelblue",main="Gewichtsverteilung der Untersuchungsgruppe",
xlab="Gewicht in kg")
```

Eine kleine Datensatz-Erweiterung…

Befragte/r	Größe (m)	Gewicht (kg)	Geschlecht
1	1,55	64	M
2	1,68	72	M
3	1,72	71	M
4	1,73	75	W
5	1,82	102	W
6	1,81	98	M
7	1,66	71	W
8	1,78	78	W
9	1,73	77	M
10	1,59	69	W

Erstellung eines gruppierten Box-Plots

Was fällt bei der Ausgabe der Daten in „probanden$geschlecht“ auf?

Vergleichen wir also mal die beiden Teilstichproben in einem Box-Plot

```
boxplot(probanden$gewicht ~ probanden$geschlecht, horizontal = TRUE)
```

Funktioniert! An den Beschriftungen müssen wir aber noch arbeiten…

```
boxplot(probanden$gewicht ~ probanden$geschlecht, horizontal = TRUE,
main="Gewichtsverteilung nach Geschlechtern", xlab="Gewicht in kg",
ylab="Geschlecht")
```

Was ist mit anderen Grafiken?

Erstellen wir mal ein einfaches Balken- oder Kreisdiagramm:

```
alter<-c(21,21,21,23,23,26,27,27,27,29)
barplot(alter)
pie(alter)
```

So soll das aber nicht aussehen!

Lösung: Statt der Datenreihe ist die Tabelle als Argument zu übergeben

```
table(alter)
barplot(table(alter))
pie(table(alter))
```

Suche nach Zusammenhängen mit R

Rangkorrelationskoeffizienten (Spearman, Kendall)

```
cor.test(probanden$gewicht,probanden$groesse,method="spearman")
cor.test(probanden$gewicht,probanden$groesse,method="kendall")
```

Bravais-Pearson-Korrelationskoeffizient

```
cor.test(probanden$gewicht,probanden$groesse,method="pearson")
```

Chi-Quadrat-Test auf Unabhängigkeit

```
chisq.test(probanden$gewicht,probanden$groesse)
```

Was verraten die Warn-meldungen?

Sehen wir uns noch das Streudiagramm an

Erstellung und Konfiguration eines Streudiagramms


```
plot(probanden$groesse, probanden$gewicht)
plot(probanden$groesse, probanden$gewicht, pch=2)
plot(probanden$groesse, probanden$gewicht, pch=16)
plot(probanden$groesse, probanden$gewicht, col="red")
plot(probanden$groesse, probanden$gewicht, col.lab="blue")
```

Bekommen wir die Regressionsgrade in das Diagramm?

```
abline(lm(probanden$gewicht ~ probanden$groesse), col="red")
summary(lm(probanden$gewicht ~ probanden$groesse))
```

-> lm = Linear Model

Beispieldatensatz zur linearen Regression

Nr.	x	y
1	12	10000
2	15	15000
3	8	6000
4	11	11000
5	3	5000
6	17	23000
7	24	37000

Beispielfall mit bewusst gering gehaltener (Foliendarstellung…) Anzahl von Werten:

- x = Prozentualer Anteil des Werbebudgets eines Produkts am Gesamtbudget der Firma
- y = Verkaufte Einheiten des betrachteten Produkts in einem Untersuchungszeitraum
- Annahme: Das betrachtete Produkt, der Untersuchungszeitraum sowie das Gesamtbudget bleiben gleich

  (ceteris paribus)

Wie lautet die Regressionsgleichung?

Lineare Regressionsanalyse mit R

Anlage des Datensatzes und Generierung des Streudiagramms

```
x<-c(12,15,8,11,3,17,24)
y<-c(10000,15000,6000,11000,5000,23000,37000)
plot(x,y)
plot(x,y,xlab="Anteil am Werbebudget",ylab="Umsatz")
```

Durchführung einer einfachen linearen Regressionsanalyse

```
summary(lm(y~x))              -> Ausgabe der Werte
abline(lm(y ~ x), col="red")  -> Einfügen der Regressionsgeraden
```

Wie interpretiert man das Ergebnis?

```
Residuals:
    1     2     3     4     5     6     7
-3918 -3706 -1534 -1322  5446  1102  3930

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5234.2     3460.6  -1.512  0.19082
x             1596.0      242.3   6.587  0.00121 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3988 on 5 degrees of freedom
Multiple R-squared:  0.8967,    Adjusted R-squared:  0.876
F-statistic: 43.39 on 1 and 5 DF,  p-value: 0.001211
```

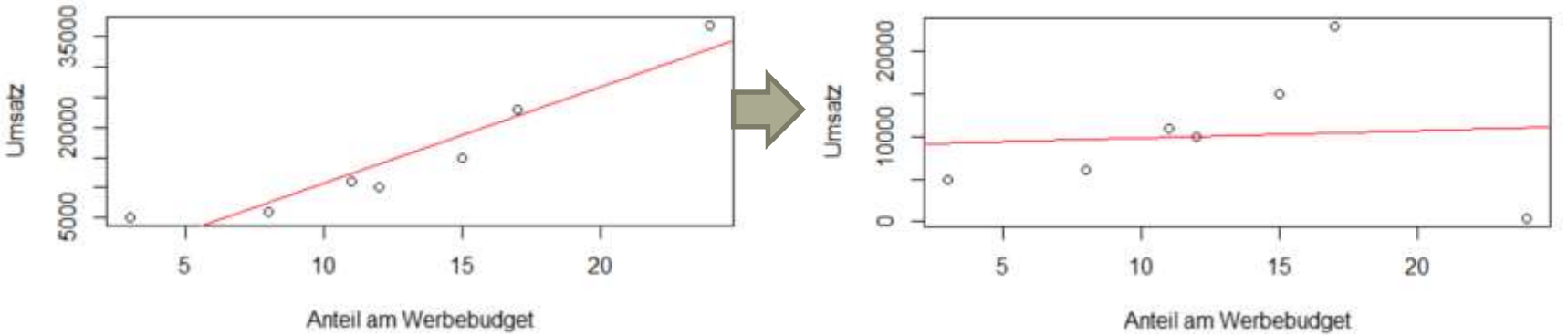
y = Regressionskoeffizient
Intercept = Konstantes Glied
R-squared = Bestimmtheitsmaß / Gütekriterium

Also:

y (Umsatz) = - 5.234,2 + 1.596 x (Werbekostenanteil)
bei einer Streuungsaufklärung von 89,67% (sehr gut)

Demonstration des Leverage-Effekts

```
y<-c(10000,15000,6000,11000,5000,23000,370)
plot(x,y,xlab="Werbebudget",ylab="Umsatz")
abline(lm(y ~ x), col="red")
summary(lm(y~x))
```



Nützliche Hinweise für die Arbeit mit R

ls()	Welche Objekte existieren?
rm(list=ls())	Alle Objekte im Speicher löschen
# Kommentar	Kommentare in Skripten hinterlegen
	mean(probanden$groesse) Mittelwertsberechnung
	mean(probanden$groesse) # Mittelwertsberechnung
?mean()	Aufruf der Hilfefunktion (in diesem Fall zu mean())
z<-c(1:40)	Generierung einer Zahlenreihe von 1 bis 40
z <- rnorm(100,0,1)	Generierung von 100 normalverteilten Zufallszahlen
useNA = "ifany"	Erstellung einer Kategorie für fehlende Werte
fix()	Öffnen des Editors zur Veränderung von Daten

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Ein eigenständiger R-Kurs ist derzeit in Vorbereitung…

…wer hätte denn grundsätzlich Interesse an einem solchen Kurs?

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

Was sollte man für die Klausur können?
(alle Angaben natürlich ohne Gewähr)

– Grundbegriffe (Skalenniveaus, Variablentypen etc.) werden über ein Multiple Choice-Quiz abgefragt

– Aufstellung von Häufigkeitstabellen und kumulierten Häufigkeitstabellen

– Berechnung von arithmetischem Mittel, getrimmtem arithmetischem Mittel, Median, Quartilen und Modus

– Berechnung von Varianz, Standard-abweichung, IQR und Spannweite

– Berechnung von Momentenkoeffizient, Quartilskoeffizient, Kurtosis und Exzeß

– Bei den Grafiken sind nur Box-Plots und Stem-and-Leaf-Plots zu zeichnen

– Von den drei Zusammenhangsmaßen (B-P-K, Spearman, Kendall) kommen mindestens zwei in der Klausur vor

Was sollte man für die Klausur können?
(alle Angaben natürlich ohne Gewähr)

- Berechnung und Interpretation einer einfachen linearen Regressionsfunktion (einschließlich des Bestimmtheitsmaßes)
- Interpretation von Venn-Diagrammen
- Mehrstufige Zufallsexperimente
  - Additionssätze
  - Multiplikationssätze
  - Baum-/Pfaddiagramme
- Variationen und Kombinationen
  - Variation mit Zurücklegen
  - Variation ohne Zurücklegen
  - Kombination mit Zurücklegen
  - Kombination ohne Zurücklegen
- Bedingte Wahrscheinlichkeiten
  - Insbesondere Satz von Bayes
- Konfidenzintervall um μ
- Chi^2-Unabhängigkeitstest

Ressourcen für die Klausurvorbereitung

- Statistik-Wiki im Stud.IP
- Probeklausuren im Stud.IP
- Diskussionsforen im Stud.IP
- Multiple Choice-Quiz im Stud.IP

http://studip.hs-harz.de

- Übungsblätter zu Statistik I
- Aufgabenheft zu Statistik II
- Foliensätze zu Statistik I und II
- Links zu Open Source-Software

http://www.hs-harz.de/creinboth/

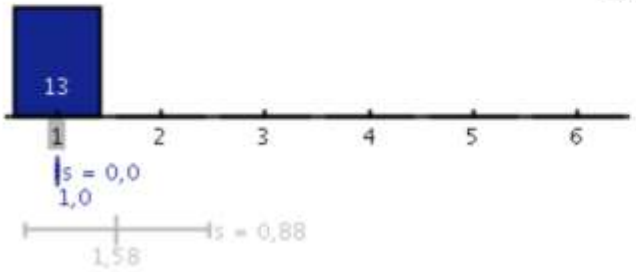
Zusätzliches Tutorium gewünscht?

Bitte die Stud.IP-Evaluation nicht vergessen
(eine Rücklaufquote von > 70% wäre gut…)

Evaluation der Präsenzveranstaltung

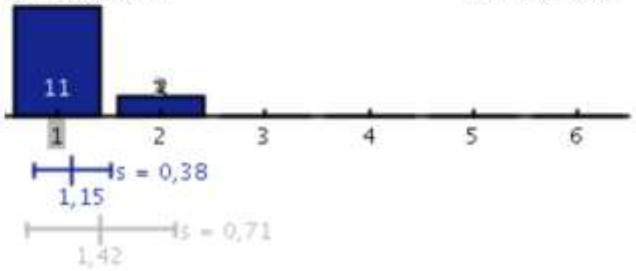
Wie sorgfältig ist der Dozent auf die Veranstaltung vorbereitet

sehr gut — sehr schlecht



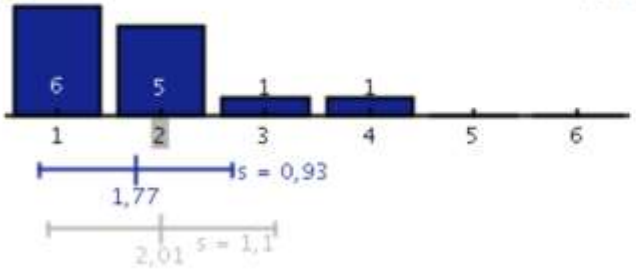
Wie souverän beherrscht der Dozent den zu vermittelnden Stoff?

sehr souverän — sehr unsicher



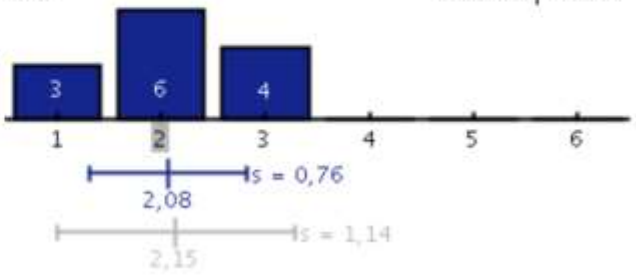
In Anbetracht der Schwierigkeit der Lehrinhalte, wie stimulierend ist der Vortragsstil des Dozenten?

sehr anregend — sehr langweilig



In welchem Umfang gelang es dem Dozenten, Ihr Interesse an dem behandelten Lehrstoff zu wecken oder zu vertiefen?

sehr — überhaupt nicht



Vielen Dank für die Aufmerksamkeit...

...und maximalen Erfolg bei der Abschlussklausur!

▲ Hochschule Harz
Hochschule für angewandte Wissenschaften

Sommersemester 2022
Christian Reinboth, M.Sc.
Fachbereich Wirtschaftswissenschaften

▲ Hochschule Harz

Hochschule für angewandte Wissenschaften

Christian Reinboth

Telefon +49 3943 – 896

Telefax +49 3943 – 5896

E-Mail creinboth@hs-harz.de

Friedrichstraße 57 – 59

38855 Wernigerode