Statistik für Journalistinnen und Journalisten

Vorlesung im Sommersemester 2017
an der TU Dortmund

Jörg Rahnenführer

SoSe 2017, Fakultät Statistik, TU Dortmund

1.1 Dozent

Prof. Dr. Jörg Rahnenführer

- Mathegeäude, Raum 720
- Email: rahnenfuehrer@statistik.tu-dortmund.de

Wissenschaft

- Studium der Mathematik mit Nebenfach Psychologie in Düsseldorf
- Forschung an WU Wien, UC Berkeley, MPI Saarbrücken
- Professor für Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, Technische Universität Dortmund
- Forschungsgebiete: Statistische Methoden für Medizin, Genetik und Bioinformatik
- Wissenschaftliche Kooperationen mit Prof. Henrik Müller und Prof. Holger Wormer vom Institut für Journalistik

1.1 Dozent

Wissenschaftliche Kooperationen

- Julia Serong, Lars Koppers, Edith Luschmann, Alejandro Molina Ramirez, Kristian Kersting, Jörg Rahnenführer, Holger Wormer (2017): **Öffentlichkeitsorientierung von Wissenschaftsinstitutionen und Wissenschaftsdisziplinen**. Publizistik, 1-26.
- Elena Erdmann, Karin Boczek, Lars Koppers, Gerret von Nordheim, Christian Pölitz, Alejandro Molina, Katharina Morik, Henrik Müller, Jörg Rahnenführer, Kristian Kersting (2016): **Machine Learning meets Data-Driven Journalism: Boosting International Understanding and Transparency in News Coverage**. arXiv:1606.05110 [cs, stat].
- Karin Boczek, Gerret von Nordheim, Lars Koppers (2016): **Update eines Klassikers: Beispiele für die Kombination manueller Inhaltsanalysen mit Latent Dirichlet Allocation**. Jahrestagung der DGPuK 2016.

1.1 Organisatorisches

Vorlesung Statistik für Journalisten

- Zeit: Dienstag, 10:15-11:45 Uhr
- Raum: Hörsaal E5 (Universitätsbibliothek)
- Homepage: https://www.statistik.tu-dortmund.de/genetik-sj17.html

Hauptklausur

- Zeit: Donnerstag, 03.08.2017, 12:00-14:00 Uhr
- Raum: SRG I / Hörsaal
- Anmeldung bis spätestens Mittwoch, 26.07.2017, per Email im Prüfungsamt bei Frau Kitsche (`iris.kitsche@tu-dortmund.de`)

Nachklausur

- Zeit: Montag, 18.09.2017, 10:00-12:00 Uhr
- Raum: 3.406 in EF 50
- Anmeldung bis spätestens Sonntag, 10.09.2017, per Email im Prüfungsamt bei Frau Kitsche (`iris.kitsche@tu-dortmund.de`)

1.1 Inhalt

Themen der Vorlesung:

- Merkmale und Datentypen
- Univariate Daten
  - Statistische Kennzahlen für die Lage
  - Statistische Kennzahlen für die Streuung
- Bivariate Daten
  - Zusammenhangsmaße
  - Lineare Regression
- Wahrscheinlichkeitsrechnung
  - Zufallsvariablen
  - Verteilungen
- Unabhängigkeit
  - Bedingte Wahrscheinlichkeiten
  - Stochastische Unabhängigkeit
  - Korrelation versus Kausalität
- Statistische Tests und Signifikanz

1.2 Literaturempfehlungen

- Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard: Statistik: Der Weg zur Datenanalyse, Springer.
- Krämer, Walter: Statistik verstehen: Eine Gebrauchsanweisung, Piper.
- Ligges, Uwe: Programmieren mit R, Springer.
- Müller, Christine; Denecke, Liesa: Stochastik in den Ingenieurwissenschaften: Eine Einführung mit R, Springer.

1.3 Statistik

- There are three kinds of lies - lies, damned lies, and statistics. (Leonard Henry Courtney, 1832-1918)
- Raten ist billig - falsch raten ist teuer. (Chinesisches Sprichwort)

1.3 Statistik

- There are three kinds of lies - lies, damned lies, and statistics. (Leonard Henry Courtney, 1832-1918)
- Raten ist billig - falsch raten ist teuer. (Chinesisches Sprichwort)

- Statistik ist (auch) Spaß
- Statistik ist (auch) Intuition
- Statistik ist (auch) Überraschung

- Wie uns Statistik hilft: Der Zufall folgt kontrollierten Regeln!

1.3 Statistik in den Medien

Die Rolle von Zahlen, Daten und Statistiken im Journalismus

- Zunehmende Bedeutung und Verfügbarkeit von Daten
- Beliebte Kategorie: Faktencheck
- Beliebt: "Empirische Studien belegen, dass ..."
- (Statistische) Einordnung der Relevanz und der Generalisierbarkeit von Studienergebnissen wichtig
- Daten sind in der Regel mit Rauschen und Fehlern behaftet!
- Statistische Methoden werden benötigt um Signal und Rauschen zu unterscheiden.
- Heiße Themen: Datenjournalismus und Visualisierung

1.3 Statistik in den Medien

Zusammenhang von Rauchverbot und sinkender Anzahl an Herzinfarkten?
Statistisch fragwürdige Aussagen in den Ruhr Nachrichten:

15.11.2010 18:59 Uhr

Ruhr Nachrichten.de

Kardiologen einig: Rauchverbot senkt Zahl der Herzinfarkte

NRW Das Rauchverbot in Kneipen und öffentlichen Gebäuden zeigt eine größere Wirkung als bislang angenommen. Das bestätigen Herzspezialisten in ganz NRW. *Von Florian Habersack*

So wurden an der Uni-Klinik Münster im Jahr 2007 541 Herzinfarkt-Fälle registriert. Für dieses Jahr rechnet Professor Holger Reinecke, kommissarischer Leiter der Kardiologie der Uni-Klinik, nur noch mit circa 300 Fällen – ein Rückgang um 44,5 Prozent. „Dieses Resultat hat uns selbst überrascht", sagt Reinecke. Aber der Zusammenhang ist für ihn klar: „Das hängt mit dem Gesetz zum Nichtraucherschutz zusammen."

- ForumSollte das Rauchverbot in NRW noch konsequenter sein?
- ArtikelWarum Nichtraucherschutz Leben retten kann

Sowohl der Direktor der Kardiologischen Klinik im Herz- und Diabeteszentrum Bad Oeynhausen, Dieter Horstkotte, als auch Heinrich Klues, Chefarzt der Kardiologie des Helios Klinikums Krefeld bestätigen den Trend. In Bad Oeynhausen liegt der Rückgang der Herzinfarktzahlen bei 15 Prozent, in Krefeld bei etwa 18 Prozent. „Ich bin mir ganz sicher, dass das Rauchverbot ein wesentlicher Faktor dafür ist", sagt Klues.

Schneller Effekt

Gegen diesen Trend entwickeln sich allerdings die Zahlen der Uni-Klinik Essen. Hier stiegen die Herzinfarkt-Fälle von 215 (2008) auf voraussichtlich 360 in diesem Jahr. Aber: Vor allem im Ruhrgebiet werde das Rauchverbot auch „viel zu lax gehandhabt", wie Professor Raimund Erbel, Chef des Westdeutschen Herzzentrums der Uniklinik Essen, bemerkt.

Für viele Kardiologen steht fest: Ein Rauchverbot bringt – vor allem Passivrauchern – einen schnellen Effekt. Dieser Trend entspricht auch einer internationalen Studie der amerikanischen Herzgesellschaft AHA. Sie untersuchte die Entwicklung der Herzinfarkt-Fälle nach der Verhängung eines Rauchverbots in zwölf Ländern. Heraus kam ein Rückgang um 36 Prozent.

Abb. 1: Artikel aus den Ruhr Nachrichten vom 15.11.2010

1.3 Statistik in den Medien

Unstatistik des Monats (www.unstatistik.de):

"Der Berliner Psychologe Gerd Gigerenzer, der Bochumer Ökonom Thomas Bauer und der Dortmunder Statistiker Walter Krämer haben im Jahr 2012 die Aktion „Unstatistik des Monats“ ins Leben gerufen. Sie hinterfragen jeden Monat sowohl jüngst publizierte Zahlen als auch deren Interpretationen. Die Aktion will so dazu beitragen, mit Daten und Fakten vernünftig umzugehen, in Zahlen gefasste Abbilder der Wirklichkeit korrekt zu interpretieren und eine immer komplexere Welt und Umwelt sinnvoller zu beschreiben."

1.3 Statistik in den Medien

Aktuelle Unstatistiken:

30.03.2017

Armes Deutschland

„Die Armut in Deutschland ist auf einen neuen Höchststand von 15,7 Prozent angestiegen." Diese Aussage in der Pressemitteilung zum Armutsbericht 2017 des Paritätischen Gesamtverbands ist die Unstatistik des Monats März. Nach Lesart des Paritätischen Wohlfahrtsverbandes beginnt Armut in Deutschland unterhalb eines Einkommens von 60 Prozent des Medians - unabhängig davon, wie hoch der Median eigentlich liegt. › **Mehr**

24.02.2017

Postfaktische US-Terrorismus-Politik

US-Präsident Donald Trump unterzeichnete einen Erlass zum "Schutz der Nation vor der Einreise ausländischer Terroristen", laut dem Bürgern aus sieben Ländern mit vorwiegend muslimischer Bevölkerung die Einreise in die USA 90 Tage lang nicht erlaubt sein sollte. Die Unstatistik des Monats Februar betrachtet die Faktenlage für diesen Erlass Trumps. › **Mehr**

31.01.2017

Feinstaub durch Silvesterknaller - Schall und Rauch statt akute Gesundheitsgefahr

Als vermeintliche Umweltsünder ins neue Jahr: Mit Sätzen wie „Die Deutschen haben ihre Luft an Silvester mit Lust und Vorsatz vergiftet" geißelten Medien wie die Süddeutsche Zeitung oder Focus Online in den ersten Januartagen die private Böllerei. Was für viele Menschen in Deutschland als gute Tradition zum Jahreswechsel gilt, gehöre angesichts der Feinstaubbelastung als „nicht mehr zeitgemäßes Vergiften der Luft" verboten, wird etwa der Bundesgeschäftsführer der Deutschen Umwelthilfe (DUH) Jürgen Resch zitiert. › **Mehr**

20.12.2016

Big Data knackt Ihre Psyche

„Zeige mir deine ‚Facebook-Likes' und ich sage dir, wer du bist." Wie ein Lauffeuer verbreitete sich im Dezember die Botschaft, dass ein derartiger Facebook-Algorithmus Donald Trump zum Wahlsieg

1.4 Statistiksoftware – Excel

Excel

- **keine** geeignete Statistiksoftware – Beispiel Varianz
- gut geeignet für Dateneingabe und Datenansicht in der Tabelle

1.4 Statistiksoftware – R

R

- `http://www.R-project.org`
- Entwicklung seit 1992
- Inspiriert von S / S-PLUS
- Einsatz an Universitäten und Forschungsinstituten sowie immer mehr in der Industrie
- Lizenz: GPL 2 (freie Software, Open Source Software)
- nur behelfsmäßige grafische Benutzeroberflächen
- wird teilweise auch im Datenjournalismus verwendet

1.4 Statistiksoftware – SAS

SAS

- `http://www.sas.com`
- Name:
  - zunächst: „Statistical Analysis Systems“
  - heute: „sas“
- SAS Institute, gegründet 1976, 2 Mrd. US$ Jahresumsatz.
- Einsatz vor allem in Medizin, Pharmaindustrie, im CRM, beim Scoring
- sehr schlechte grafische Benutzeroberflächen

CRM: Customer Relationship Management

1.4 Statistiksoftware – SPSS

SPSS

- http://www.spss.com
- Name:
  - zunächst „Statistical Package for the Social Sciences“
  - dann „Superior Performing Software System“
  - heute „SPSS“
- gegründet 1968 als eigenständige Firma
- 2009 an IBM verkauft
- Einsatz vor allem in den Sozial- und Geisteswissenschaften, im CRM, beim Scoring
- ordentliche grafische Benutzeroberfläche

1.4 Statistiksoftware – Anforderungen

Anforderungen an Statistiksoftware:

- Interaktive Arbeit mit Daten für die Datenanalyse
- Erstellung statistischer Grafik
- Hohe numerische Genauigkeit
- Hohe Rechengeschwindigkeit
- Verarbeitung großer Datenmengen
- Automatisierbarkeit von Methoden und sich wiederholender Abläufe
- Einfache Bedienbarkeit / Programmierbarkeit
- Nebenbedingungen: Preis, unterstützte Hardware, Parallelisierung, ...

Diese Anforderungen widersprechen sich teilweise. Je nach Gewichtung der Schwerpunkte erfolgt die Wahl einer entsprechenden Software.

Software kann niemals besser sein als ihr Benutzer!

2.1 Merkmale

Definition 1

Ein **Merkmal** *ist eine abstrahierende Eigenschaft von einem oder mehreren verschiedenen Beobachtungen (Merkmalsausprägungen = Merkmalswerten), die pro Merkmalsträger erfasst wird.*
Ein Merkmal „besteht“ also aus Merkmalsausprägungen, die an Merkmalsträgern einer Gesamtheit erfasst werden.

2.2 Merkmalstypen

Definition 2

- *Ein Merkmal heißt* **qualitativ**, *wenn es eine Eigenschaft, („Qualität“) eines Merkmalsträgers bezeichnet.*
- *Ein Merkmal heißt* **quantitativ**, *wenn seine Ausprägungen „echte“ Messwerte sind. (Solche sind addier-, subtrahier- und multiplizierbar.)*
- *Quantitative Merkmale heißen auch* **metrisch** *oder* **kardinal**.

2.2 Merkmalstypen

Definition 3

- *Bei qualitativen Merkmalen werden zwei Arten unterschieden:*
  - **Ordinale** *Merkmale lassen zwar kein Addieren oder Multiplizieren zu, aber eine Anordnung ihrer Ausprägungen.*
  - **Nominale** *Merkmale erlauben auch kein Sortieren ihrer Ausprägungen.*

  Man sagt, Merkmalsausprägungen liegen auf einer Nominalskala, Ordinalskala bzw.- Kardinalskala.
- *Merkmale auf der Kardinalskala werden als* **metrisch** *bezeichnet.*
- *Bei der Kardinalskala unterscheidet man zwischen Intervallskala und Verhältnisskala (bei dieser gibt es zusätzlich einen definierten Nullpunkt, so dass auch Verhältnisse berechnet werden dürfen).*

2.2 Merkmalstypen

Definition 3

- **Diskrete** *Merkmale haben höchstens abzählbar unendlich viele Ausprägungen, d.h. ihre Ausprägungen lassen sich mit Hilfe der natürlichen Zahlen abzählen.*
- **Stetige** *Merkmale können überabzählbar viele Werte annehmen, z.B. jede reelle Zahl in einem Intervall.*

2.2 Merkmalstypen

Beispiele

- *Metrische Merkmale*: Größe, Alter (beide Verhältnisskala), Datum (Intervallskala)
- *Ordinale Merkmale*: Frage nach Gesundheitszustand, Schulnoten
- *Nominale Merkmale*: Geschlecht, Name, Postleitzahl
- *Diskrete Merkmale*: Geschlecht, Einwohnerzahl
- *Stetige Merkmale*: Temperatur, Größe

Stetige Merkmale werden häufig nur diskret beobachtet wegen Messgenauigkeit (Temperatur) oder Übereinkunft (Alter).

Merkmalstypen werden in Statistiksystemen durch enstprechende Datenformate abgebildet.

2.2 Ein Datenbeispiel für Kardinalskala

Beispiel 1

Patientendaten:
a) Gewicht (in kg); `NA`: fehlender Wert (Not Available)

Zufällige Auswahl des Gewichts von 200 Patienten:

```
 85  70  75  70  92  88  68 101  74  80  87  68  95  33  75 117 105  88  76
 82 107  92  87  91  83  80  85  95  75  60  85  75  73  58  93  70 100  94
100  75  80  85  87  43  90  92  89  NA 100  96  58  72  77  83  48  74  90
 58  78  75  56  70  75  70  67  95  74  88  70  68  66 102  72  74 113  72
 81  75  55  60  75  90  71  93  NA  94  75  89  90  80  52  90 105  90  82
 80  83  80  89  70  67  92 108  58  75  75 110  85  58  74  93  97  65  83
110  87  81  64 103 120  65  85  79  95 110  70  90  85  94  88  88 130  70
 69  78 100  88  86  85  76  60  79  90  88 104  69  96  59  75  NA  75  66
 70  86  80  65  94  72  62  75 105  91  79  88  80  85  69  87  54  96  70
 82  70  95  78  95  95  84  70  90  65  67  85  NA  92  87  63 120  65  55
 65  81  NA  54  81  63  64  77  70  75
```

2.3 Übungsaufgaben

- Betrachten Sie journalistische Artikel, in denen Aussagen über Merkmale (auf Zahlenbasis) gemacht werden.
  Bestimmen Sie die Skalenniveaus der jeweiligen Merkmale.
- Gegeben sei ein Datensatz mit folgenden individuellen Angaben von Patienten: Name, Geburtsdatum, Augenfarbe, Gewicht, Größe, Bewertung des Gesundheitszustands (1-10), Körpertemperatur.
  Welche Skalenniveaus liegen für diese Merkmale vor?

2.3 Übungsaufgaben

RuhrNachrichten.de

Suchbegriff SUCHEN

Städte Nachrichten Sport Lifestyle Bilder Videos Meine Themen Stellen Immo Reise RNshopping Anzeigen Abo

Nachrichten Vermischtes Aktuelles & Berichte Zahl der Versuchstiere in NRW gestiegen

zur eZeitung | Einloggen | Registrieren

Aktionstag am 24. April

Zahl der Versuchstiere in NRW gestiegen

NRW Nordrhein-Westfalen liegt weit vorne - im negativen Sinn: Die Zahl der eingesetzten Versuchstiere ist 2015 gestiegen, mit 432.000 liegt NRW im Ländervergleich auf dem zweiten Platz. Heute ist der Internationale Tag des Versuchstiers. Wir wollten wissen, wo in der Region Tierversuche stattfinden und haben dazu auch die Universitäten befragt.

Tierversuche – bei diesem Wort entstehen Bilder im Kopf, von Affen mit Elektroden im Hirn und Kaninchen mit verätzten Augen. Weil das Thema aus ethischer Sicht so umstritten ist und Bilder von Labortieren regelmäßig für emotionale Reaktionen sorgen, gehen Wissenschaftler in den seltensten Fällen von sich aus mit Informationen über Tierversuche an die Öffentlichkeit.

Die Fakten: In Deutschland wurden nach aktuellsten Zahlen des Bundesministeriums für Ernährung und Landwirtschaft im Jahr 2015 2.028.072 Tiere für Tierversuche eingesetzt. Hinzu kamen 771.889 Tiere, die zu wissenschaftlichen Zwecken getötet wurden, ohne dass an ihnen Behandlungen durchgeführt wurden. Auf Nordrhein-Westfalen entfiel dabei nach Angaben des Vereins „Ärzte gegen Tierversuche“ mit 432.006 Tieren ein Anteil von 15,4 Prozent. Im Ländervergleich liegt NRW damit auf dem zweiten Platz hinter Spitzenreiter Baden-Württemberg. Im Jahr 2014 sind hier noch 25.000 Versuchstiere weniger eingesetzt worden.

2.3 Übungsaufgaben

Gegeben sei ein Datensatz mit folgenden individuellen Angaben von Patienten: Name, Geburtsdatum, Augenfarbe, Gewicht, Größe, Bewertung des Gesundheitszustands (1-10), Körpertemperatur.
Welche Skalenniveaus liegen für diese Merkmale vor?

Name:	Nominalskala
Geburtsdatum:	Kardinalskala (Intervallskala) - metrisch
Augenfarbe:	Nominalskala
Gewicht:	Kardinalskala (Verhältnisskala) - metrisch
Größe:	Kardinalskala (Verhältnisskala) - metrisch
Bewertung Gesundheit:	Ordinalskala
Körpertemperatur:	Kardinalskala (Intervallskala) - metrisch (für Kelvin statt Celsius: Verhätnisskala)

3.1.1 Histogramm: Absolute Häufigkeitsverteilung

Durch Verwendung von „Kästen“ werden oft absolute Häufigkeiten für Intervalle dargestellt und als **Histogramm** bezeichnet:

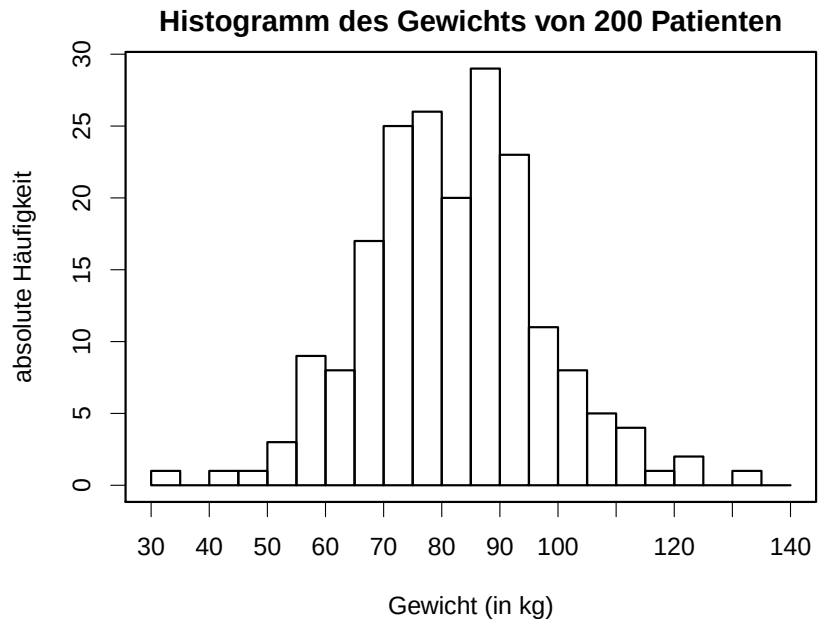


Abb. 2: Histogramm des Gewichts von 200 Patienten aus Beispiel 1 (absolute Häufigkeiten, Klassenbreite 5 kg)

3.1.1 Absolute Häufigkeitsverteilung

Interpretation

- Jeder Kasten entspricht einer Werteklasse, hier links offene und rechts abgeschlossene Intervalle, also:
  $A_1 := (30, 35], A_2 := (35, 40], A3 := (40, 45], ...$
- x-Achse: Merkmalsausprägungen
- y-Achse: absolute Häufigkeiten $H(A_k)$ der Klassen

3.1.1 Absolute Häufigkeitsverteilung

Häufig spielt die gewählte Klassenbreite eine entscheidende Rolle, Strukturen in den Daten zu finden. Bei Einteilung in 1 kg Klassen sieht man, dass Zahlen, die auf 0 oder 5 enden, häufiger auftreten als andere:

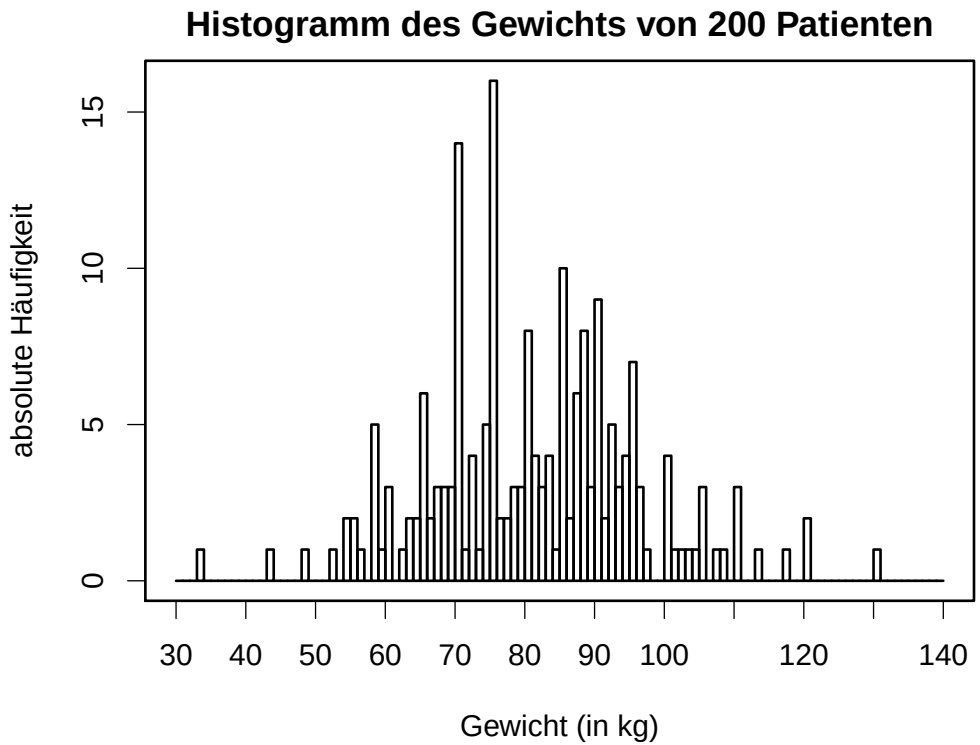


Abb. 3: Histogramm des Gewichts von 200 Patienten aus Beispiel 1 (absolute Häufigkeiten, Klassenbreite 1 kg)

3.1.1 Absolute Häufigkeitsverteilung

Zu eng gewählte Klassen führen aber zu einem unruhigen Bild, das die globale Struktur der Verteilung schlecht erkennen lässt, dafür eignen sich breitere Klassen, wie etwa hier in 10 kg Schritten:

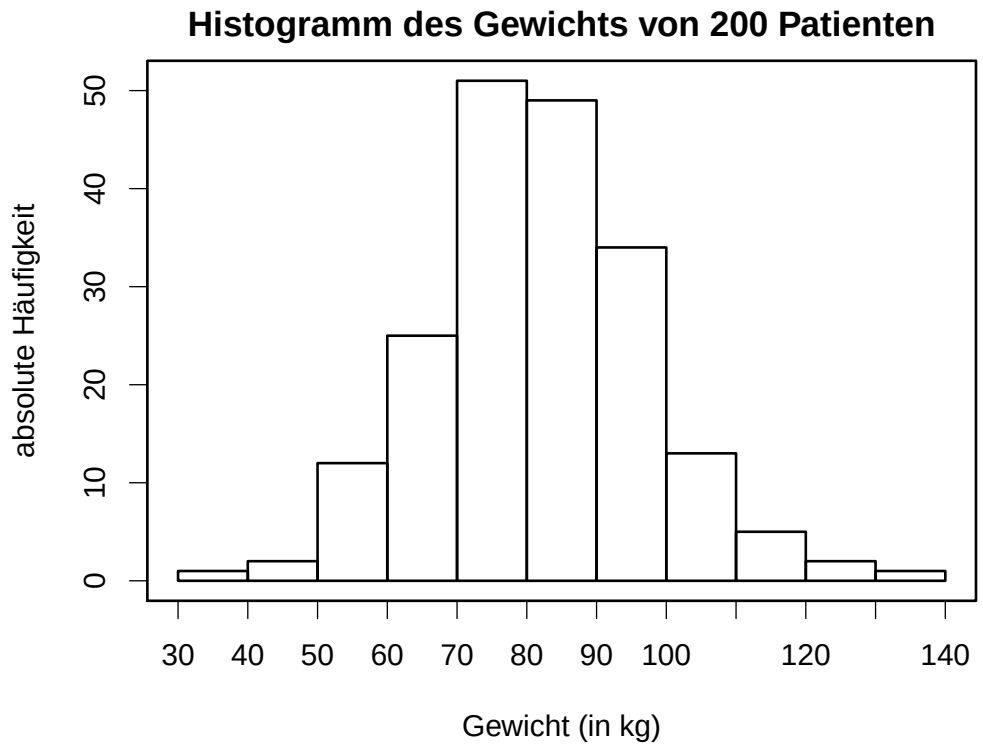


Abb. 4: Histogramm des Gewichts von 200 Patienten aus Beispiel 1 (absolute Häufigkeiten, Klassenbreite 10 kg)

3.1.2 Relative Häufigkeitsverteilung

Von der Darstellung absoluter zur Darstellung relativer Häufigkeiten

- Bisher wurde das Histogramm verwendet, um die absolute Häufigkeitsverteilung darzustellen.
- In der Literatur spricht man häufig nur bei der *Darstellung der relativen Häufigkeitsverteilung* von einem Histogramm.

3.1.2 Relative Häufigkeitsverteilung

Definition 4

Ein **Histogramm** *ist eine grafische Darstellung der relativen Häufigkeitsverteilung der n Werte eines stetigen metrischen Merkmals X auf einer Gesamtheit.*

- *Dabei werden die Werte in* **Klassen** $A_1, A_2, \ldots, A_K$ *eingeteilt.*
- *Die Klassengrenzen werden auf der x-Achse aufgetragen.*
- *Für jede Klasse wird ein Kasten gezeichnet, der auf der x-Achse durch die Klassengrenzen begrenzt wird.*
- *Die Fläche (!) eines Kastens beschreibt die relative Häufigkeit einer Klasse.*

3.1.2 Relative Häufigkeitsverteilung

Eigenschaften

- Die Kastenhöhe der k-ten Klasse ist $r_k = {}^{h_k}/_{b_k}$, wobei h_k die relative Häufigkeit und b_k die Klassenbreite sind.
- Bei der Klassenbreite $b_k = 1$ ist also die Kastenhöhe $r_k = h_k$, d.h. die relative Häufigkeit der Klasse.
- **Die Gesamtfläche der Kästen ist gleich 1.**
- Mit dieser Definition lässt sich das Histogramm zur Veranschaulichung der empirischen Dichte (dazu später mehr) verwenden.

3.1.2 Relative Häufigkeitsverteilung

Die Darstellung aus Abbildung 2 mit relativen Häufigkeiten sieht also nahezu identisch aus, einzig die y-Achse ist in diesem Fall verändert:

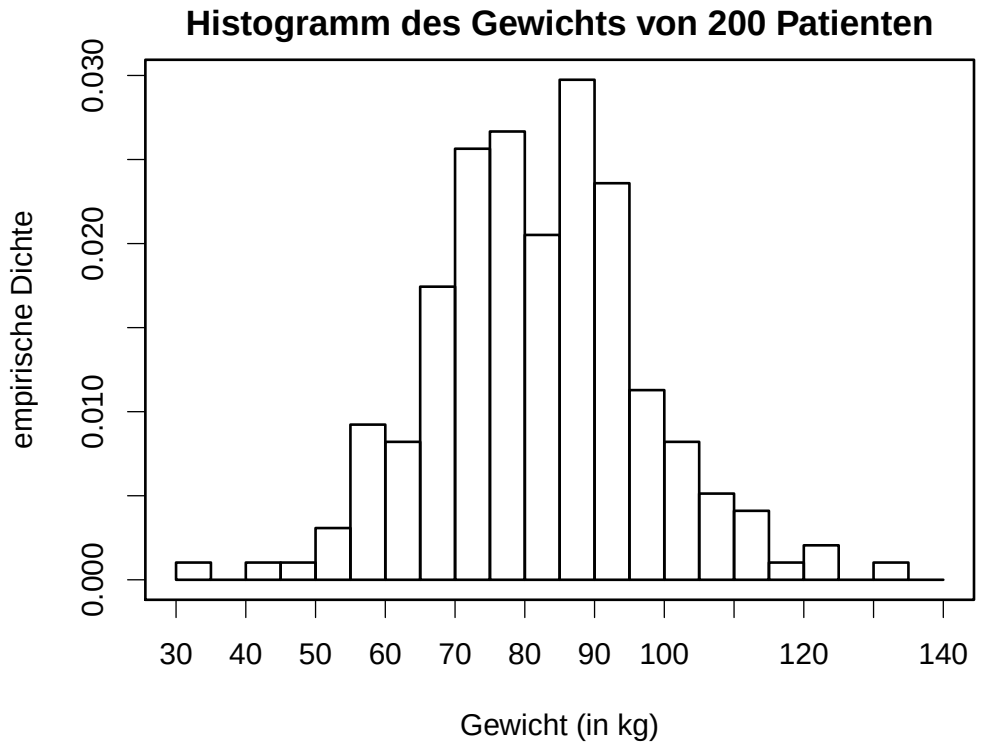


Abb. 5: Histogramm des Gewichts von 200 Patienten aus Beispiel 1 (empirische Dichte, Klassenbreite 5 kg)

3.1.2 Einschub: Beschriftung einer Grafik

Die **Beschriftung einer Grafik** ist von zentraler Bedeutung für gute Datenanalyse und Präsentationen.

Zu einer geeigneten Beschriftung zählen:

- Beschriftungen an den Achsen (mit Einheit!)
- Titel und evtl. Untertitel
- Beschriftung muss gut lesbar sein (Größe und Schriftart)
- Inhalt einer Grafik soll verstanden werden können, ohne den Fließtext zu lesen
- unterschiedliche Farben, Symbole und Linienarten
  - sollen so gewählt werden, dass sie gut unterscheidbar sind
  - sollen aber auch nicht von anderen Farben, Symbole und Linienarten ablenken
  - müssen in einer Legende erklärt werden

3.1.3 Stabdiagramm

Idee

Bei *qualitativen Merkmalen* hat sich eingebürgert, ein sogenanntes **Stabdiagramm** (**Balkendiagramm**) zu verwenden, indem pro Merkmalsausprägung ein schmaler Stab (Balken) mit der Höhe H_k oder h_k (aber natürlich einheitlich) über dem Merkmalswert gezeichnet wird. Es bezeichnen H_k die absolute und h_k die relative Häufigkeit.

Bemerkungen

- Bei Stabdiagrammen werden die Merkmalsausprägungen für qualitative Merkmale gleichabständig auf der x-Achse gezeichnet.
- **Stäbe sind immer (im Gegensatz zu Kästen beim Histogramm) voneinander separiert!**

3.1.3 Stabdiagramm

Beispiel 2

Strahlentherapie bei Patienten

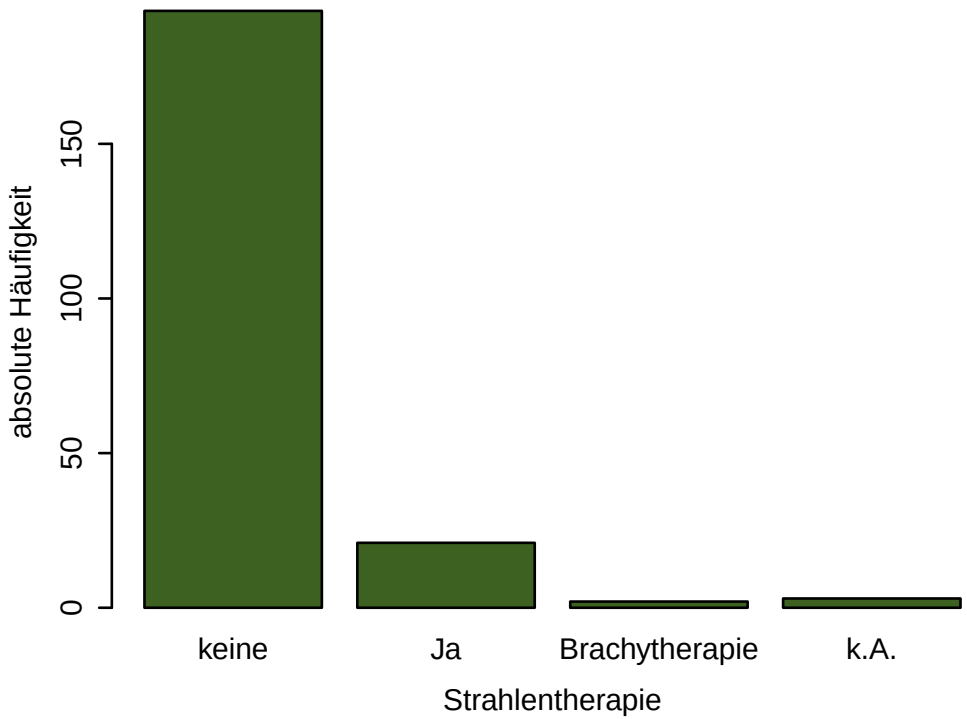


Abb. 6: Strahlentherapie bei Patienten

3.1.3 Stabdiagramm

Stabdiagramme von *ordinalen* oder *metrisch diskreten* Merkmalen:

- Da die Merkmalsausprägungen von ordinalen oder metrisch diskreten Merkmalen eine „natürliche“ Reihenfolge haben, wird selbige verwendet.

Stabdiagramme von *nominalen* Merkmalen:

- Da die Merkmalsausprägungen von nominalen Merkmalen *keine* „natürliche“ Reihenfolge haben, ist es erlaubt, die Merkmalswerte beliebig anzuordnen.
- Häufig ist es interessant, die Werte nach ihrer Häufigkeit anzuordnen, das häufigste zuerst, dann das zweithäufigste, usw.

3.2 Empirische Verteilungsfunktion

Idee

Die **empirische Verteilungsfunktion** (oder relative Summenhäufigkeitsfunktion) $F(x) = S^{rel}(x)$ kann für klassierte stetige, ordinale und metrische Merkmale durch die Folge der Summenhäufigkeiten S_k, $k = 1, \ldots, K$, der verwendeten Merkmalsklassen festgelegt werden und in ein Histogramm eingezeichnet werden.

3.2 Empirische Verteilungsfunktion

Definition 5 (Empirische Verteilungsfunktion)

Sei X ein Merkmal mit reellen Zahlen als Ausprägungen $x_1, \ldots, x_n$. Dann heißt die Funktion

$$F(x) := \sum_{j=1}^{k} h_j$$

die den Anteil der Werte x_i angibt, die kleiner oder gleich x sind, **empirische Verteilungsfunktion**, *wobei*

a) $x \in$ *[Ende von Klasse k, Ende von Klasse $k+1$) bzw.*

b) $x \in$ *[Mitte von Klasse k, Mitte von Klasse $k+1$), $k < K$,*

und
$F(x) = 0$ *für alle $x <$ Ende bzw. Mitte von Klasse 1 sowie*
$F(x) = 1$ *für alle $x >$ Ende bzw. Mitte von Klasse K.*

Typischerweise wird hier pro einzigartiger Ausprägung x_i eine eigene Klasse gewählt.

3.2 Empirische Verteilungsfunktion

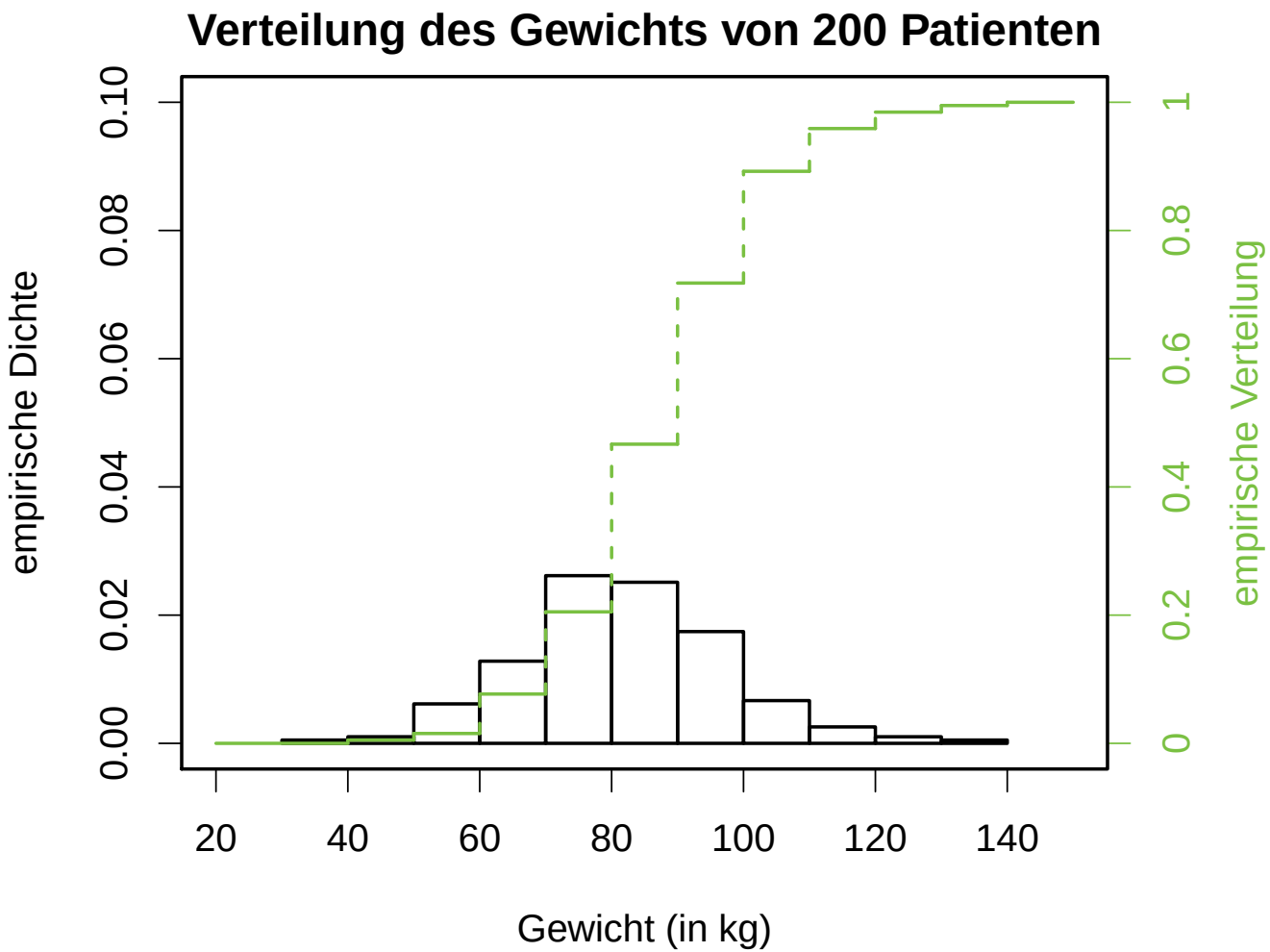
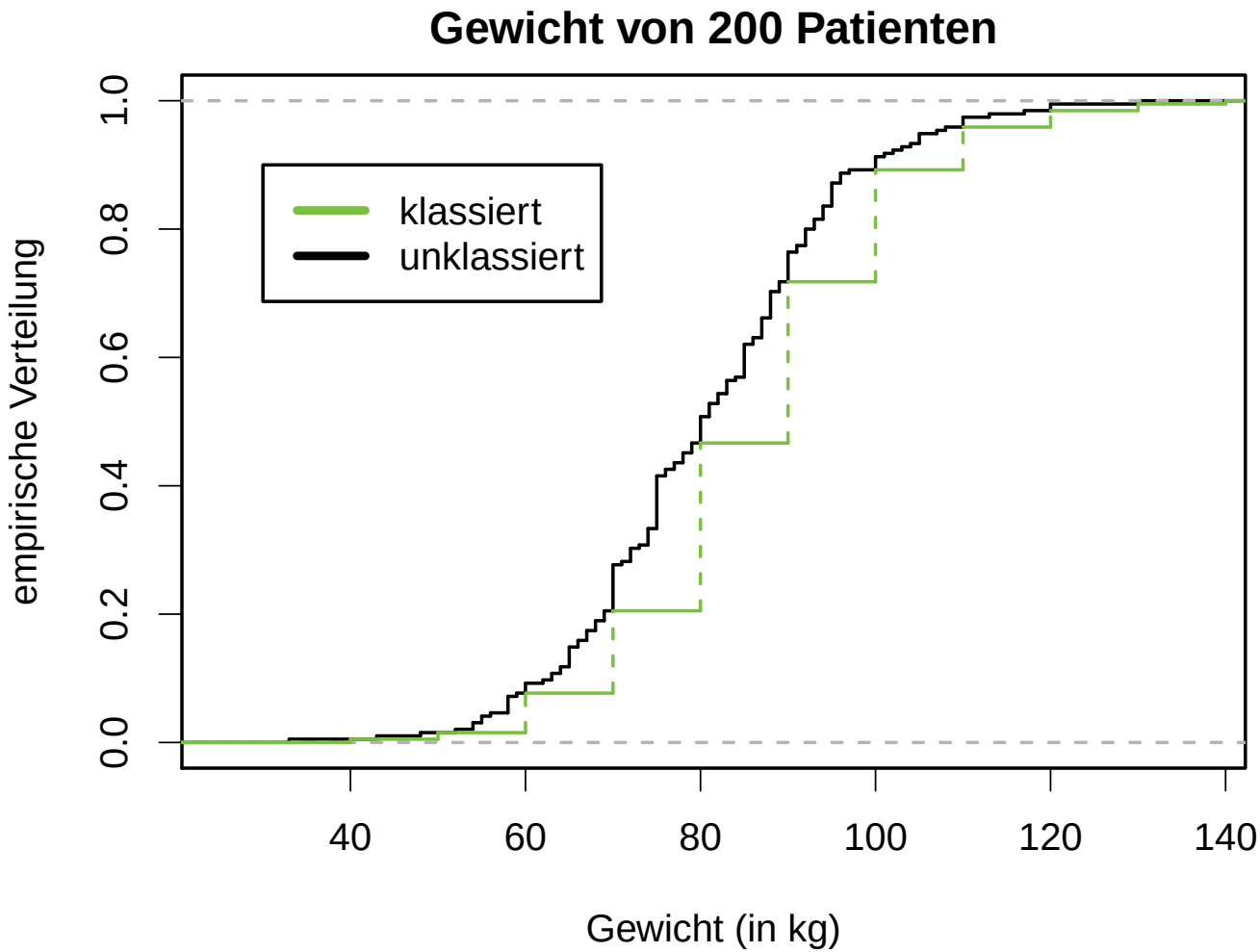


Abb. 7: empirische Verteilungsfunktion und Histogramm zu Beispiel 1: Gewicht von 200 Patienten, Klassenbreite 10 kg

3.2 Empirische Verteilungsfunktion

Die empirische Verteilungsfunktion (ohne Histogramm) der *nicht* klassierten Daten:



3.3 Kuchen- oder Kreis-Diagramm

Idee

- Zur Visualisierung von Klassenanteilen an einer Gesamtheit wird häufig ein **Kuchen- bzw. Kreis-Diagramm** verwendet.
- Dabei wird ein Kreis so in Sektoren aufgeteilt, dass die **Sektorflächen** (!) proportional zu den absoluten (bzw. relativen) Klassenhäufigkeiten sind.
- Die Kreissektoren erinnern an Kuchenstücke.
- Unterschiedlich große Gesamtheiten können durch proportional zur Größe der Gesamtheit gewählte Kreisflächen (!) bei mehreren Kreis-Diagrammen angedeutet werden.

3.3 Kuchen- oder Kreis-Diagramm

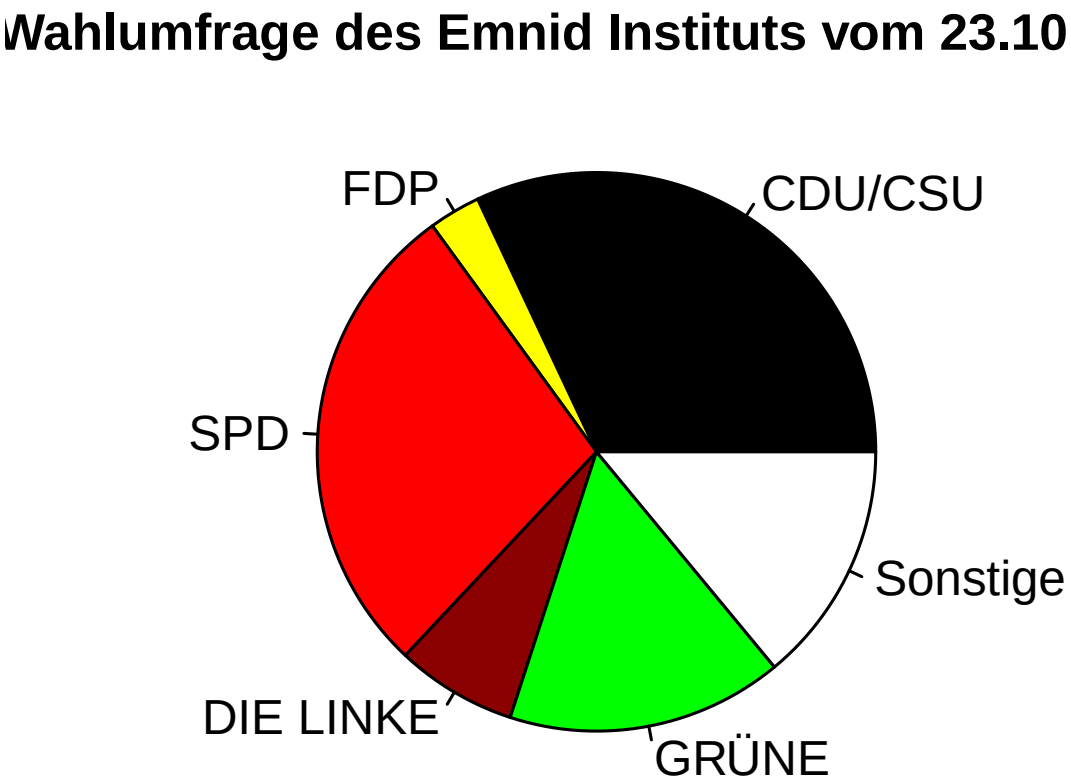


Abb. 9: Wahlumfrage des Emnid Instituts vom 23.10.2011 – Versuch 1

3.3 Kuchen- oder Kreis-Diagramm

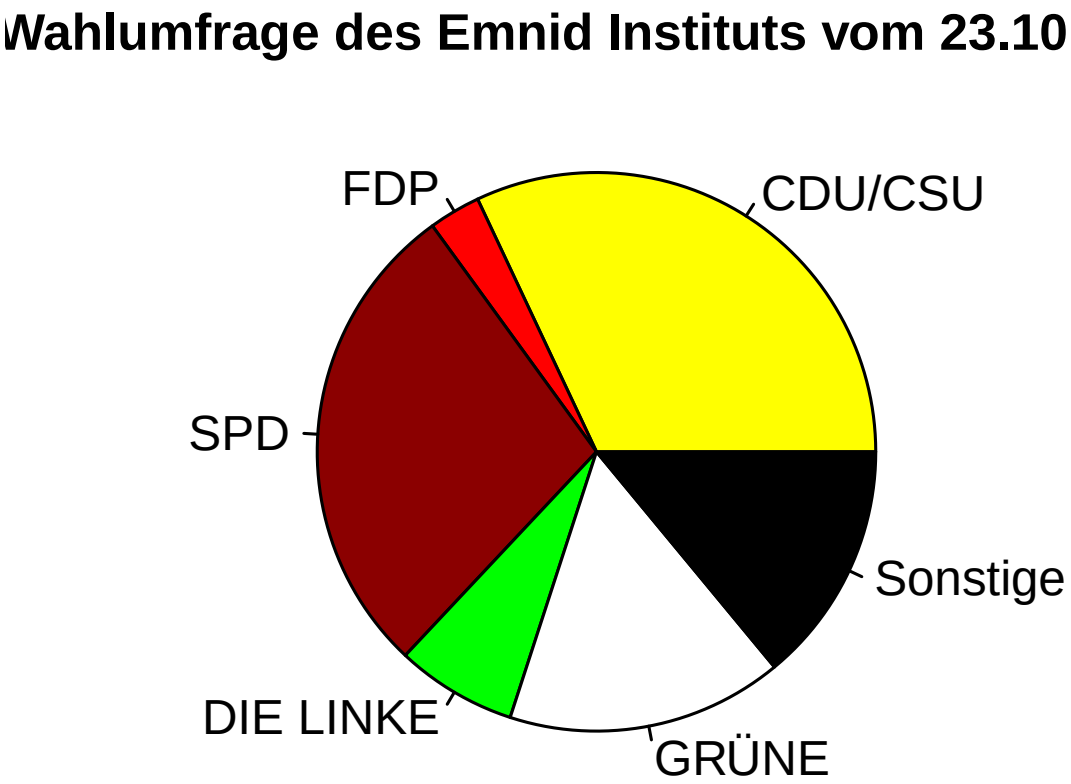


Abb. 10: Wahlumfrage des Emnid Instituts vom 23.10.2011 – Versuch 2

3.3 Kuchen- oder Kreis-Diagramm

Ganz schlimm: Niemals 3D-Kuchendiagramme!!!

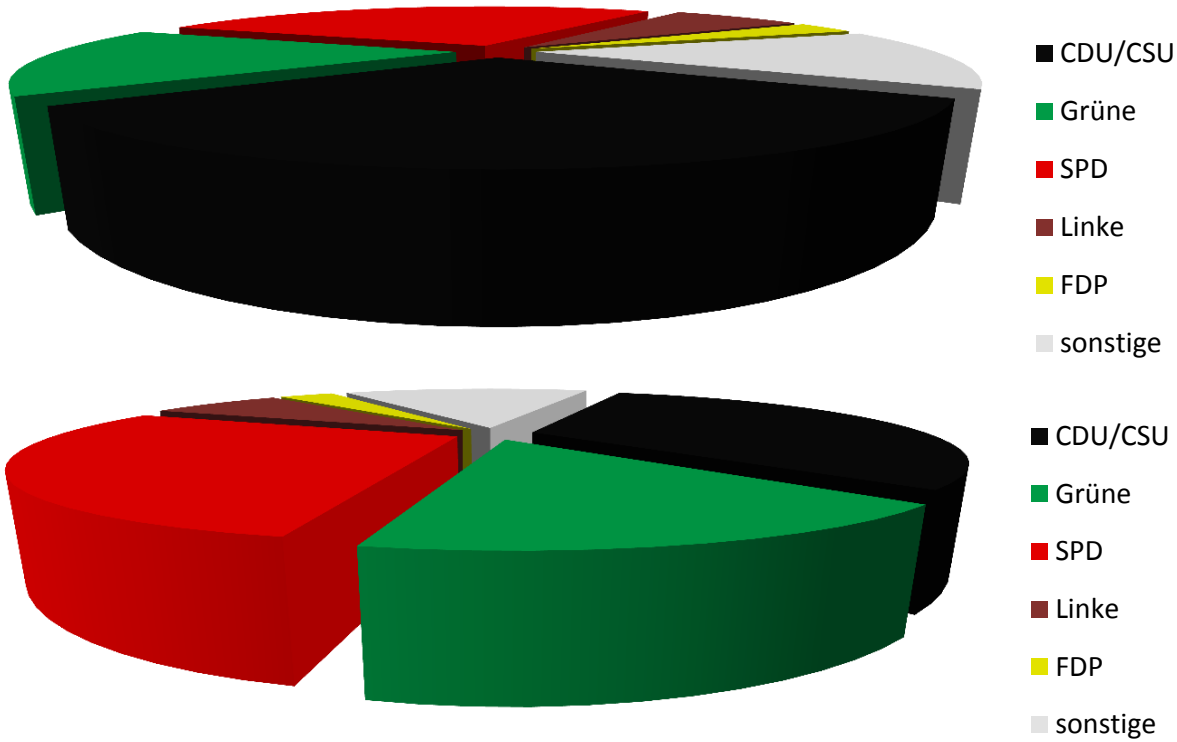


Abb. 11: Wahlumfrage des Emnid Instituts vom 23.10.2011 –Versuche 3 und 4

3.3 Kuchen- oder Kreis-Diagramm

Was sollte verbessert werden?

- 3D: Flächen hinten wirken kleiner als die Flächen vorne
- Farben: Dunkle Farben wirken mächtiger als helle
- Farben: Leuchtende Farben ziehen die Aufmerksamkeit an, graue Töne werden weniger wahrgenommen.
- Kreissegmente (Winkel) viel schlechter vergleichbar als Stäbe / Balken

- $\longrightarrow$ Niemals 3D-Kuchendiagramme verwenden!
- $\longrightarrow$ Kreisdiagramme vermeiden!

3.3 Kuchen- oder Kreis-Diagramm

Viel besser:

- **Balkendiagramme verwenden!**
- Dabei an geeignete Farbwahl denken
- Im folgenden Beispiel ist die Farbwahl entgegen der Konvention getroffen worden, wobei aber die Farbtöne zumindest ähnlich gehalten wurden und die Einfärbung letztendlich weniger ablenkend wirken sollte.

3.3 Kuchen- oder Kreis-Diagramm

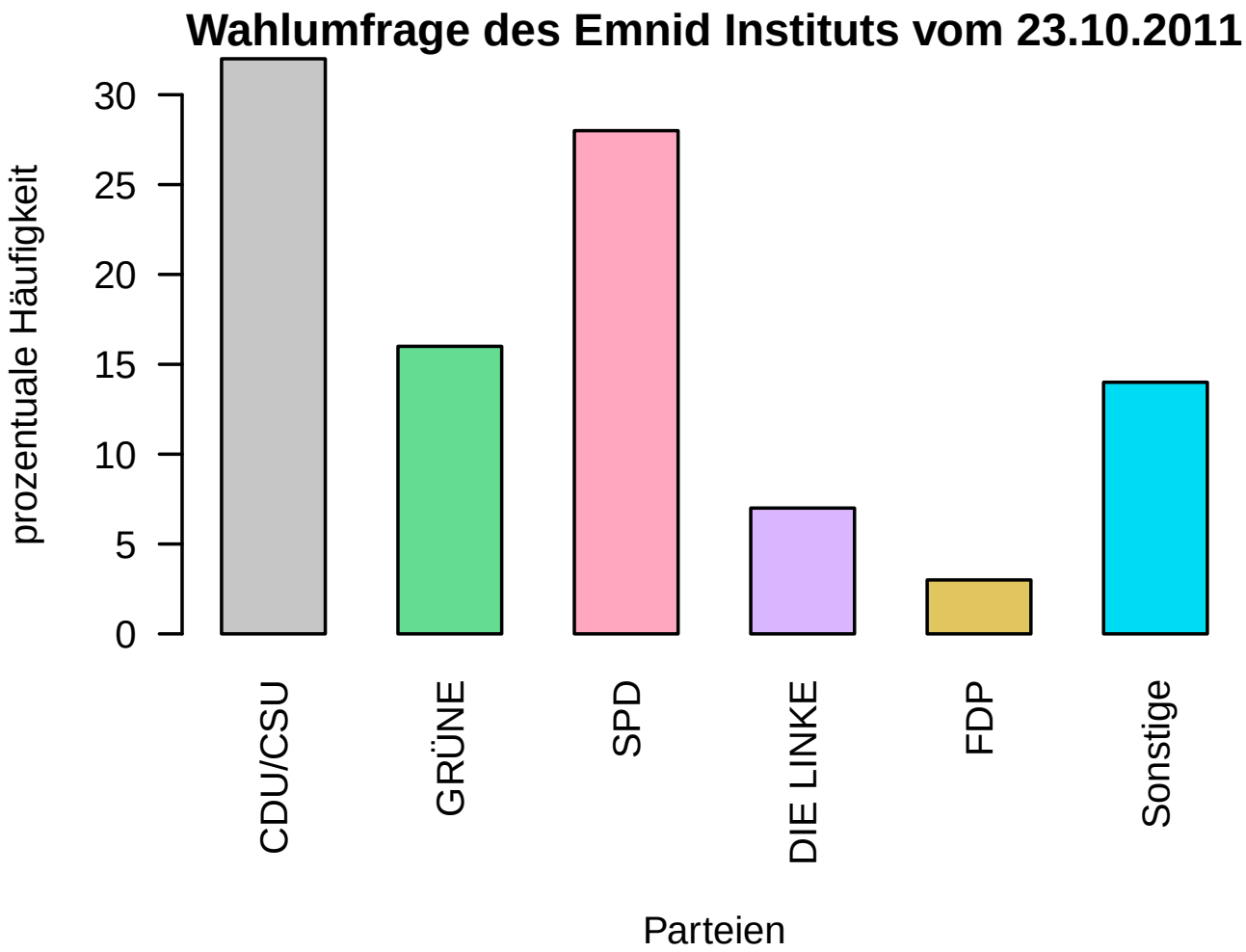


Abb. 12: Wahlumfrage des Emnid Instituts vom 23.10.2011 – Versuch 5

3.4 Lage- und Streuungs-Maße

Idee

Nach „geschickten“ grafischen Darstellungen der Werte eines Merkmals auf einer Gesamtheit interessieren jetzt „geschickte“ algebraische Charakterisierungen der Verteilung solcher Werte.

Ziel ist es, die Verteilung durch möglichst wenige sog. **Maßzahlen** zu beschreiben.

3.4 Lage- und Streuungs-Maße

Dabei stehen zwei Fragen im Vordergrund:

1. Wo liegt die „Mitte“ der Werte?
2. Wie streuen die Werte um die Mitte?

- Die erste Frage basiert auf der Hoffnung, dass sich die Verteilung einigermaßen repräsentativ durch eine einzige Zahl, ein sogenanntes **Lagemaß**, charakterisieren lässt („Einer für Alle“).
- Durch Hinzunahme eines sogenannten Streuungsmaßes soll die Größe der Unsicherheit (= Streuung) der Merkmalswerte charakterisiert werden.

Bemerkung

Maßzahlen werden z.B. dazu verwendet, um verschiedene Gesamtheiten miteinander zu vergleichen. Dabei hängt ihre Interpretation typischerweise von der Problemstellung ab.

3.4 Lage- und Streuungs-Maße

Beispiel: Welcher Schütze schießt besser?

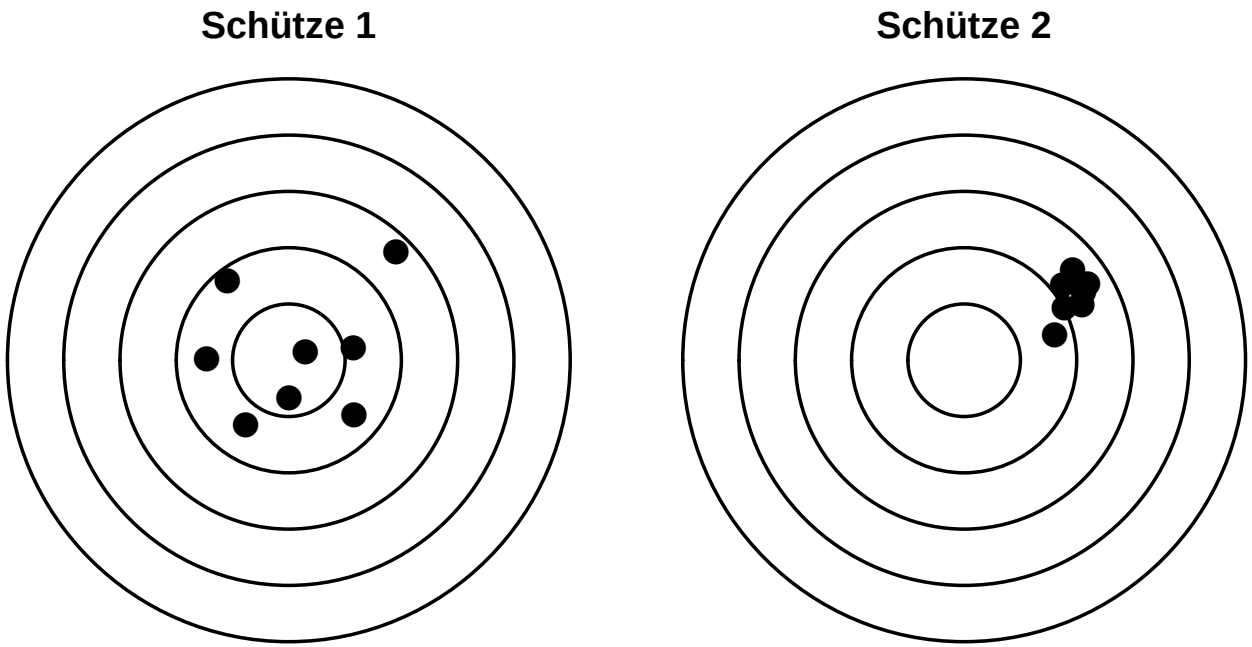


Abb. 13: Welcher Schütze schießt besser?

Schütze 1: Lage gut, Streuung schlecht
Schütze 2: Lage schlecht, Streuung gut

3.4 Lage- und Streuungs-Maße

Beispiel: Schokoriegelproduktion

- In einer Fabrik werden Schokoriegel produziert.
- Verpackungsangabe: Ein Schokoriegel wiegt 60 g.
- Es gibt regelmäßig Kontrollen, dass Verpackungsangaben eingehalten werden. Auch muss sich der Hersteller bei Unterschreitung der Verpackungsangabe schlechter Presse stellen.
- Der Hersteller will also keinesfalls viele zu leichte Schokoriegel produzieren. Andererseits kosten zu schwere Schokoriegel mehr Material und schmälern den Gewinn.
- Idee: Erlauben, ganz wenige ($< 1\%$) zu leichte Schokoriegel zu produzieren.

Iteratives Vorgehen:

1. Einstellung der Lage auf Basis der Streuung.
2. Reduzieren der Streuung.

3.4 Lage- und Streuungs-Maße

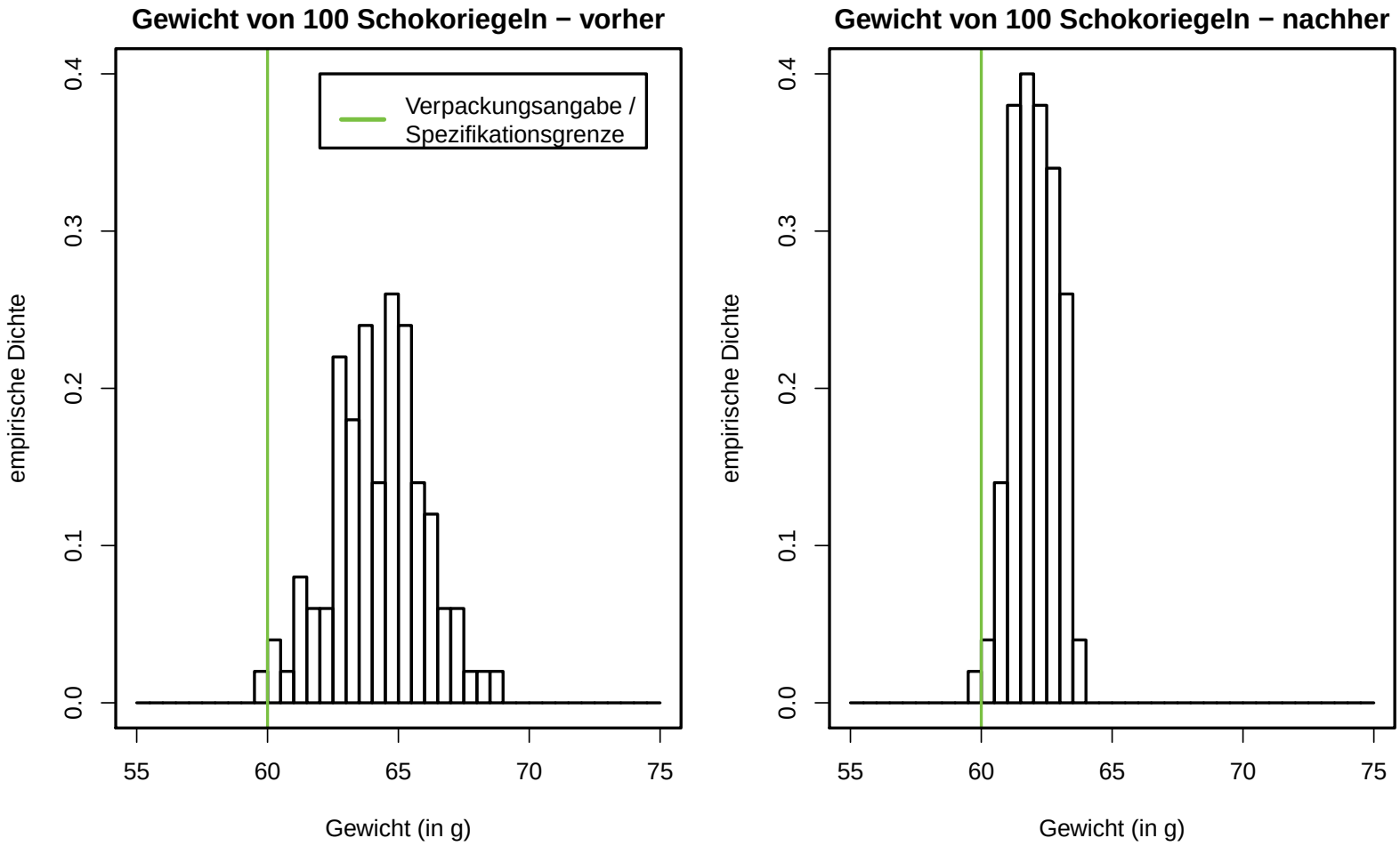


Abb. 14: Ersparnis bei Schokoriegelproduktion

3.4.1 Lagemaße

Idee

Zur Charakterisierung der Merkmalswerte auf einer Gesamtheit durch eine einzige Zahl werden sog. Lagemaße verwendet:

Lagemaß = „Mitte“ der Merkmalswerte $x_1, \ldots, x_n$.

Die wichtigsten Beispiele sind:

Lagemaße 1

- **Arithmetisches Mittel** = Mittelwert (mean)

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

3.4.1 Lagemaße

Lagemaße 2

- **Median** = „Zentralwert“ = 50%-Wert: med_x
  Der Median ist derjenige Wert, für den 50% der Merkmalswerte größer oder gleich und 50% kleiner oder gleich sind.
  Der Median ist der mittlere Wert der Rangliste:

$$\text{med}_x := \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ ungerade} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & n \text{ gerade} \end{cases} \tag{2}$$

- **Modalwert / Modus** = häufigster Wert: mod_x
  Der Modalwert ist derjenige Merkmalswert, der am häufigsten vorkommt.
  Er liegt zwar nicht unbedingt in der Mitte der Merkmalswerte, bietet sich aber natürlich trotzdem als Repräsentant der Merkmalswerte an.

3.4.1 Lagemaße

Bemerkungen

- Je unterschiedlicher diese Werte, desto weniger lässt sich die Verteilung durch einen einzigen Wert charakterisieren.
- Das arithmetische Mittel reagiert am empfindlichsten auf „Ausreißer“, d.h. auf (für die Verteilung) ungewöhnlich große oder kleine Werte, und auf „Schiefe der Verteilung“.
- Der *Modalwert* ist u.U. nicht eindeutig.
- Bei wirklich stetigen Merkmalen eignet sich der Modalwert erst nach einer Klassierung, da evtl. gar keine Merkmalsausprägungen mehrfach beobachtet werden.

3.4.2 Das „richtige“ Lagemaß

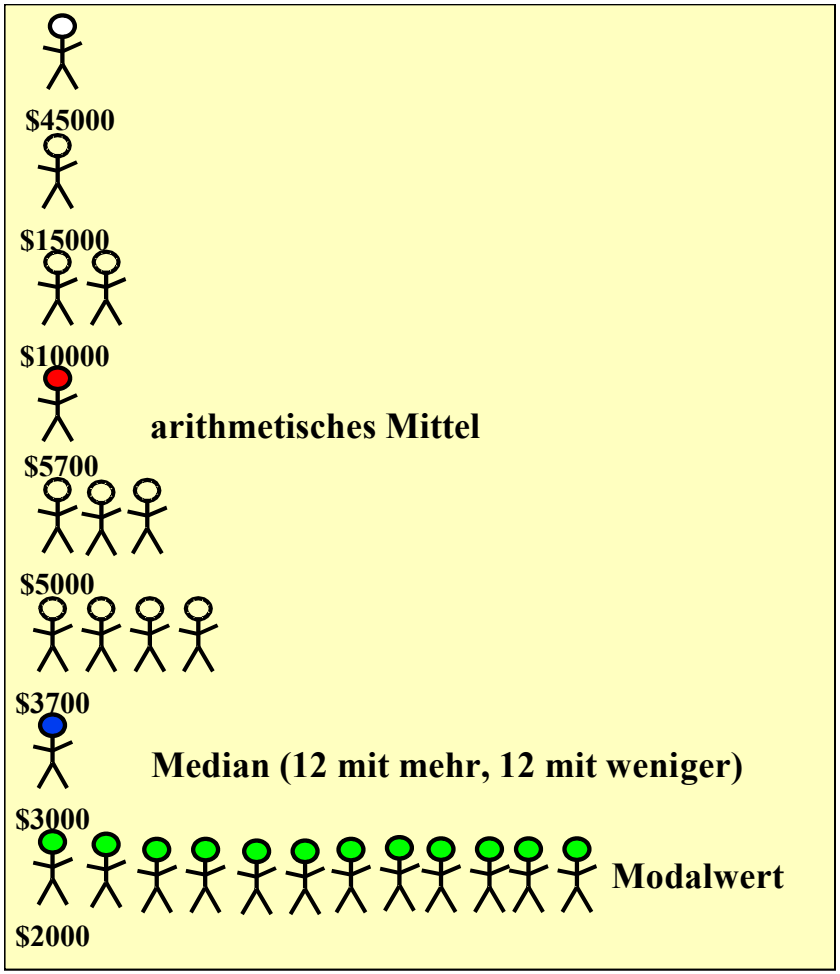


Abb. 15: Beispiel: Mittleres Einkommen von 25 Personen

3.4.2 Das „richtige“ Lagemaß

Bemerkung

- Die *Wahl des Lagemaßes* kann die Aussage entscheidend beeinflussen.
- Aus einem Beispiel mit Anmeldezeiten von Studierenden (Histogramm siehe nächste Folie), folgt:
  - Arithmetisches Mittel bei schiefer Verteilung – wieder nicht gut geeignet.
  - Modus bei stetigen Merkmalen: Macht erst Sinn nach Klassierung. Aber welche Klassierung verwenden?
  - Median repräsentiert die Anmeldezeiten noch am besten, sofern das mit einer einzigen Zahl überhaupt möglich ist.

3.4.2 Das „richtige“ Lagemaß

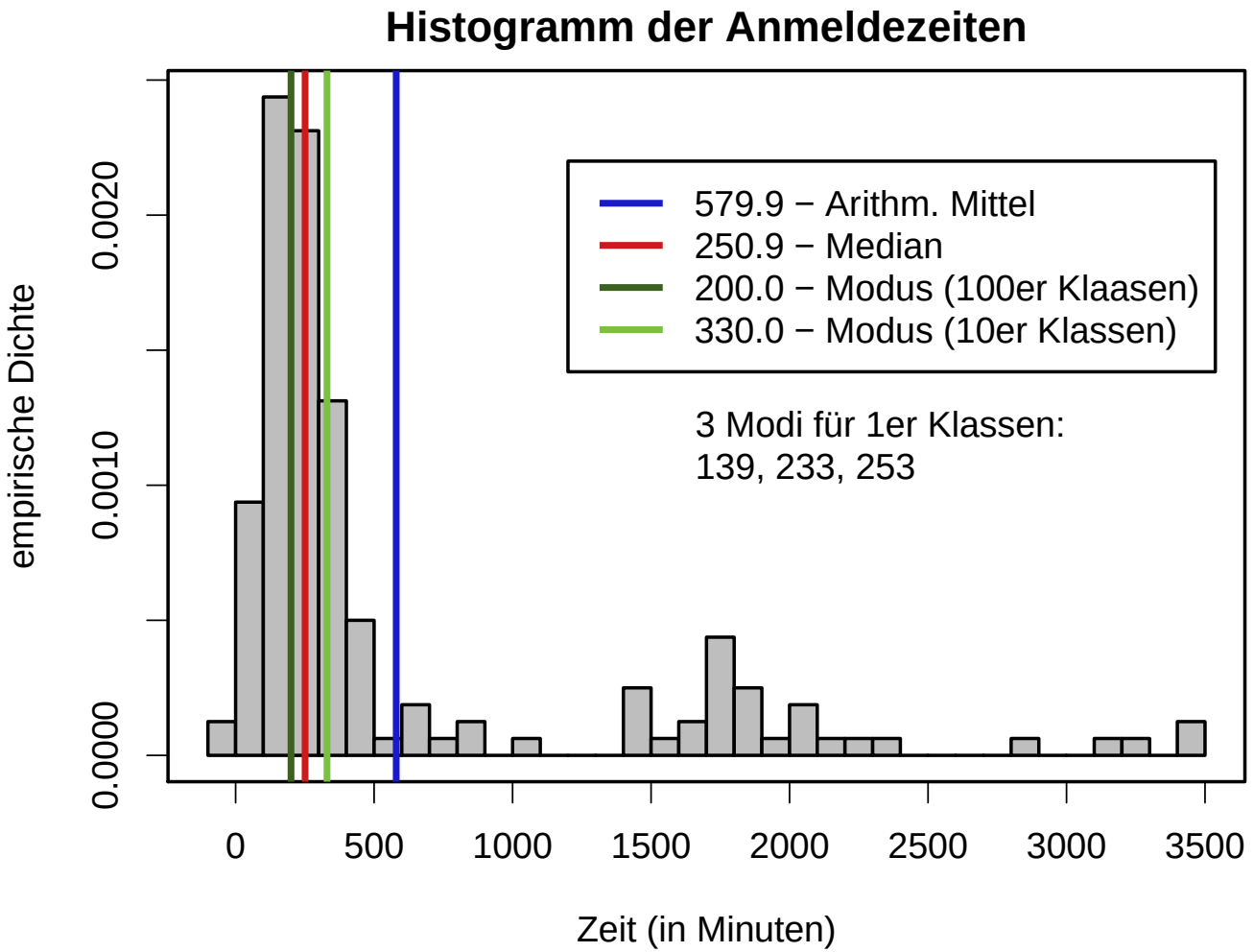


Abb. 16: Beispiel: Arithmetisches Mittel, Median und Modalwerte der Anmeldezeiten von Studierenden

3.4.2 Das „richtige“ Lagemaß

Fortsetzung von Beispiel 1, hier sind sich die Lagemaße recht einig:

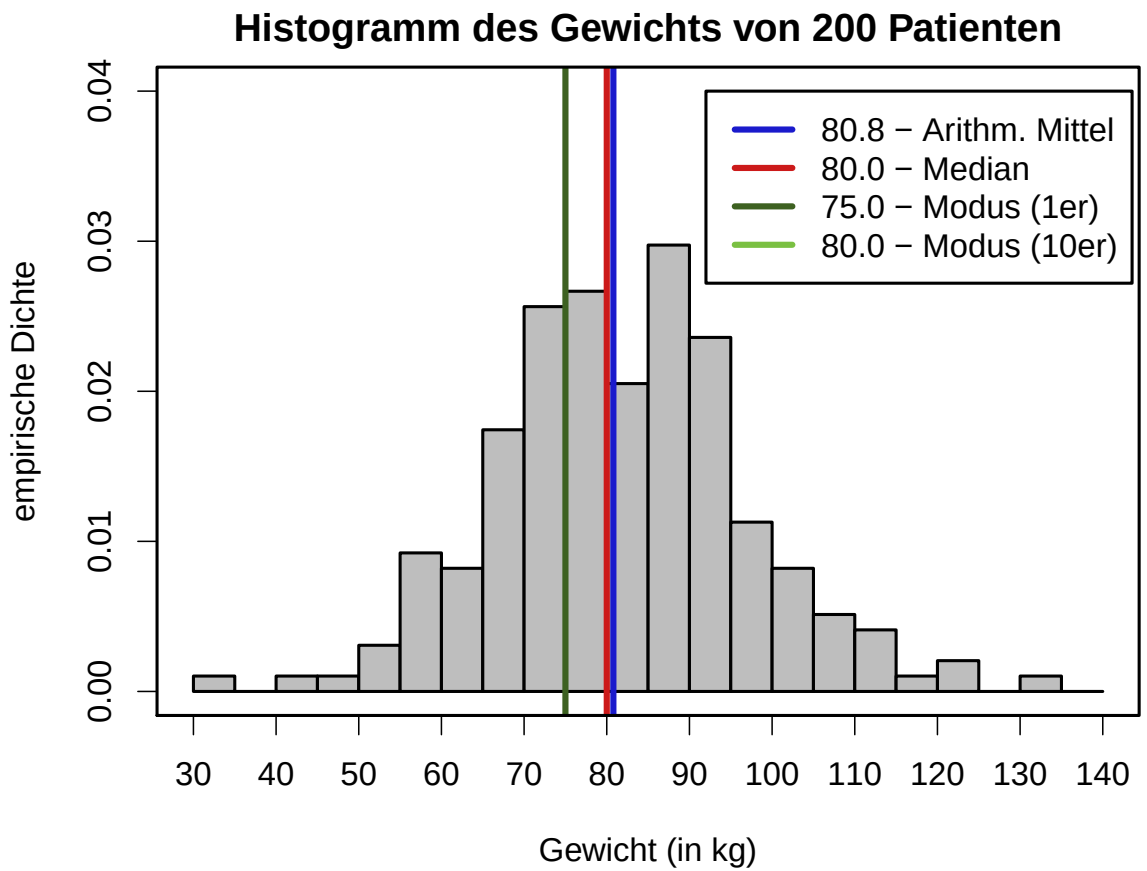


Abb. 17: Beispiel: Arithmetisches Mittel, Median und Modalwerte des Gewichts von 200 Patienten

3.4.3 Weitere Lagemaße / Mittel

Idee

Neben den 3 bisher vorgestellten Lagemaßen gibt es noch viele weitere Lagemaße, die für spezielle Problemstellungen entwickelt wurden. Es folgen einige wichtige Beispiele.

3.4.3 Weitere Lagemaße / Mittel

Lagemaße 3

- **Gewichtetes** (bzw. gewogenes) **arithmetisches Mittel**

$$\bar{x}_w := \sum_{i=1}^{n} w_i x_i, \quad \text{mit } w_i \geq 0, \quad \sum_{i=1}^{n} w_i = 1 \tag{3}$$

- **Geometrisches Mittel**

$$\bar{x}_g := \sqrt[n]{\prod_{i=1}^{n} x_i}, \quad \text{mit } x_i \geq 0 \tag{4}$$

- **Gewichtetes geometrisches Mittel**

$$\bar{x}_{gw} := \prod_{i=1}^{n} x_i^{w_i}, \quad \text{mit } x_i, w_i \geq 0, \quad \sum_{i=1}^{n} w_i = 1 \tag{5}$$

3.4.3 Weitere Lagemaße / Mittel

Beispiele:

- *Ausgaben für Benzin und Motoröl*
  Der für einen Autofahrer relevante mittlere Preisanstieg von Benzin und Motoröl hängt natürlich davon ab, wieviel Benzin und Motoröl ein Auto verbraucht. Es gelte:
  - Preisanstieg: Benzin um 50%, Motoröl um 10%.
  - Ausgabenanteil: Benzin 90%, Motoröl 10%.

  Dann ist $\bar{x}_w := 0.9 \cdot 50 + 0.1 \cdot 10 = 46\%$ der mittlere Preisanstieg.

3.4.3 Weitere Lagemaße / Mittel

- *Mittlere Wachstumsrate*
  - Der Kurs einer Aktie habe sich über 4 Zeitpunkte wie folgt entwickelt:
    1000 EUR $\rightarrow$ 1200 EUR $\rightarrow$ 1500 EUR $\rightarrow$ 1000 EUR.
  - Sie interessieren sich für die mittlere Rendite, wobei Zinsen und Dividenden vernachlässigt werden sollen.
  - Was für ein Mittel würden Sie verwenden?

a) Das **arithmetische Mittel** $\bar{r}$ der Wachstumsraten r_i:
$\bar{r} = 1/3 \sum_{i=1}^{n} r_i = \frac{(20\% + 25\% - 33.33\%)}{3} = 3.89\%,$

b) den **Median** der Wachstumsraten: $r_{\text{med}} = 20\%$ oder

c) die **Differenz** des geometrischen Mittels der Wachstumsfaktoren $1 + r_i$ und 1, also
$\bar{r}_g - 1 = \sqrt[3]{1.2 \cdot 1.25 \cdot 0.66} - 1 = 0$?

- Natürlich muss die mittlere Rendite $= 0$ sein, da der Kurs wieder zu seinem Ausgangswert zurückgekehrt ist.

3.4.3 Weitere Lagemaße / Mittel

Ganz allgemein verwendet man als **Durchschnitt von Wachstumsraten:**

$$\bar{r} := \sqrt[n]{\prod_{i=1}^{n}(1 + r_i)} - 1. \tag{6}$$

3.4.3 Weitere Lagemaße / Mittel

Lagemaße 4

- **Harmonisches Mittel**

$$\bar{x}_h := \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}, \quad \text{falls alle } x_i < 0 \text{ oder alle } x_i > 0 \tag{7}$$

- **Gewichtetes harmonisches Mittel**

$$\bar{x}_{hw} := \frac{1}{\sum_{i=1}^{n} \frac{w_i}{x_i}}, \tag{8}$$

mit $x_i < 0$ oder $x_i > 0$ für alle i, $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$.

3.4.4 Rangmaßzahlen

Idee

Die Idee, eine Verteilung durch den Zentralwert (= 50%-Wert, Median) zu beschreiben, lässt sich einfach auf beliebige $100 \cdot p$%-Werte verallgemeinern.

Solche sogenannten **p-Quantile** sind ebenfalls nützliche Hilfsmittel zur Beschreibung einer Rangliste

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

3.4.4 Rangmaßzahlen

Definition 6

Ein **p-Quantil** Q_p, $p \in [0,1]$, *ist eine Zahl, für die* $100 \cdot p\%$ *der Merkmalswerte einer Gesamtheit kleiner oder gleich sind und* $100 \cdot (1-p)\%$ *größer oder gleich.*

Genauer könnte man für Q_p *z.B. Folgendes fordern:*

$Q_p \geq$ *größtem Merkmalswert einer Gesamtheit, der* $\leq 100 \cdot p\%$ *der Merkmalswerte ist und*
$Q_p \leq$ *nächstgrößerem Merkmalswert der Gesamtheit, also*

$$x_{(\lfloor np \rfloor)} \leq Q_p \leq x_{(\lfloor np \rfloor + 1)}. \tag{9}$$

Wenn eine von beiden Forderungen nicht erfüllbar ist, wird nur die andere verwendet.

3.4.4 Rangmaßzahlen

Die folgende Berechnungsmethode für Quantile entspricht der obigen Berechnung des Medians.

p-Quantil Berechnung: „Standard“ (Nicht in R, dort `type = 2` wählen.)

$$Q_p := \begin{cases} x_{(j)}, & j := \lceil np \rceil, \quad np \text{ nicht ganzzahlig} \\ \dfrac{x_{(j)} + x_{(j+1)}}{2}, & j := np, \quad np \text{ ganzzahlig} \end{cases}$$

3.4.4 Rangmaßzahlen

Diese Art der Berechnung wird im Folgenden immer verwendet, wenn nichts anderes vermerkt ist.

Bezeichnung

- Anstelle von **p-Quantil** sagt man auch $100 \cdot p(\%)$-**Perzentil** oder **(1-p)-Fraktil**.
- 0.25- bzw. 0.75-Quantile heißen auch unteres bzw. oberes **Quartil**: unteres Quartil $q_4 = 0.25$-Quantil; oberes Quartil $q^4 = 0.75$-Quantil.

3.4.4 Rangmaßzahlen

Beispiel 1, Gewicht von 200 Patienten, Erinnerung:

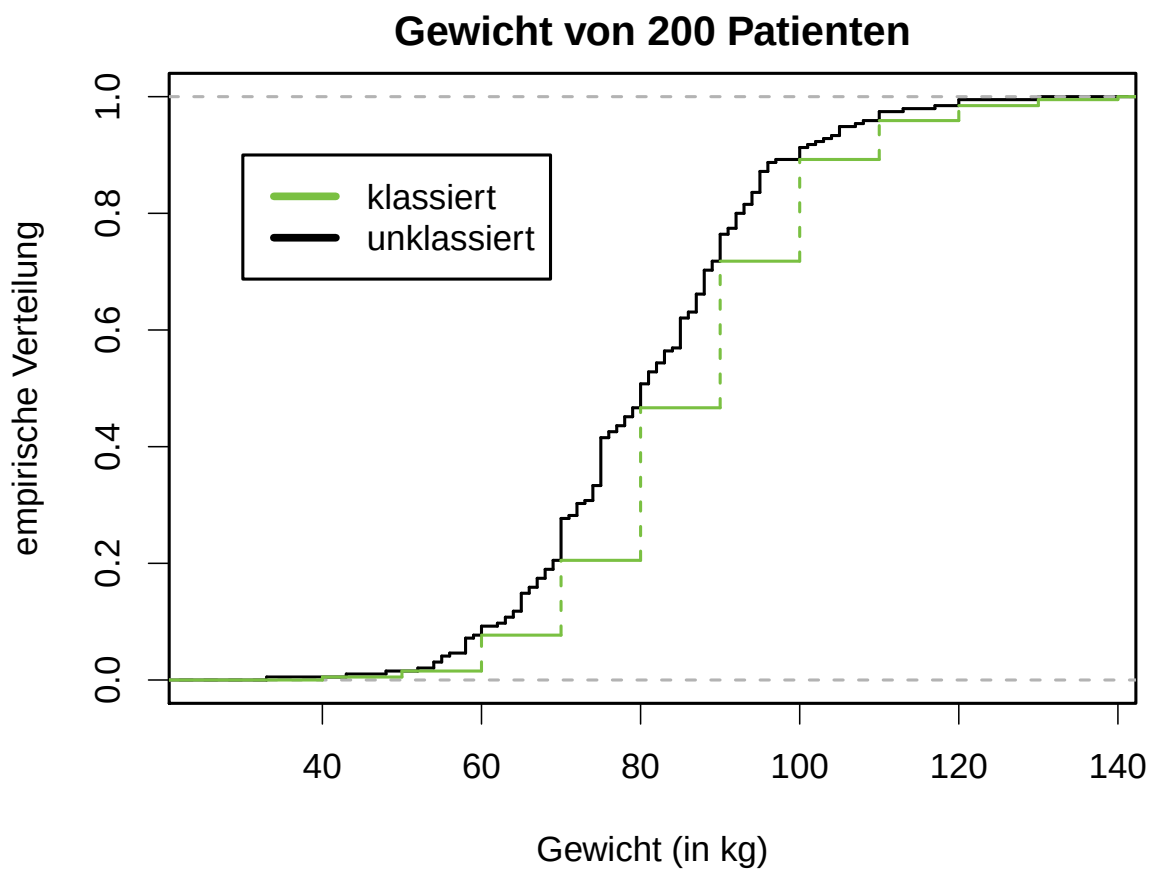


Abb. 18: Empirische Verteilungsfunktion zu Beispiel 1: Gewicht von 200 Patienten

3.4.4 Rangmaßzahlen

Beispiel 1, Gewicht von 200 Patienten:

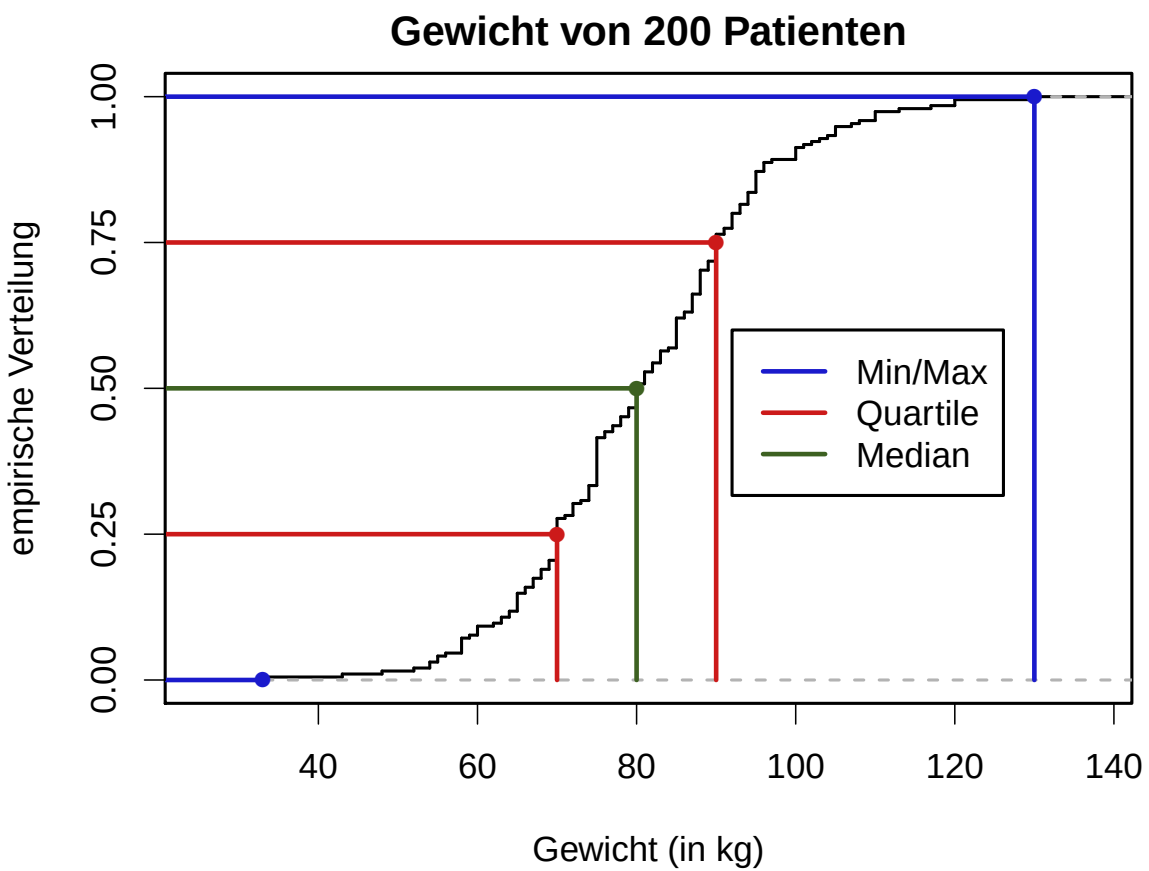


Abb. 19: Rangmaßzahlen in der empirischen Verteilungsfunktion zu Beispiel 1: Gewicht von 200 Patienten

3.4.5 Streuungsmaße

Streuungsmaße 1

- empirische **Varianz**: „Durchschnitt“ der quadrierten Abweichungen vom arithmetischen Mittel

$$\text{var}_x = s_x^2 := \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)} = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{(n-1)} \tag{10}$$

- **Standardabweichung**: Wurzel aus der **Varianz**

$$s_x := \sqrt{\text{var}_x} \tag{11}$$

- **Quartilsdifferenz** (interquartile range)

$$qd_x := q^4 - q_4 \tag{12}$$

- **Spannweite** (range)

$$R_x := \max(x) - \min(x) = x_{(n)} - x_{(1)} \tag{13}$$

3.4.5 Streuungsmaße

Beispiel:

geordnete Liste: -15, 1, 3, 4, 4, 6, 6, 7

$\text{var}_x = 50.86, s_x = 7.13, qd_x = 4, R_x = 22$

Bemerkungen

- Die Spannweite ist am empfindlichsten gegen „Ausreißer“, die Quartilsdifferenz am wenigsten.
- Bei „normalverteilten“ Merkmalen gilt die folgende Beziehung zwischen den Streuungsmaßzahlen: $qd \approx 1.35 s_x$ und $R \approx 6 s_x$.
- Wiederum stellt sich die Frage nach der „richtigen“ Streuung.
- Wiederum hängt die Interpretation einer Verteilung häufig entscheidend von der Wahl der Streuungsmaßzahl ab.

3.4.5 Streuungsmaße

- Beispiel 1, Gewicht von 200 Patienten
  $s_x = 15.14$ kg, $qd_x = 20$ kg, $R_x = 97$ kg

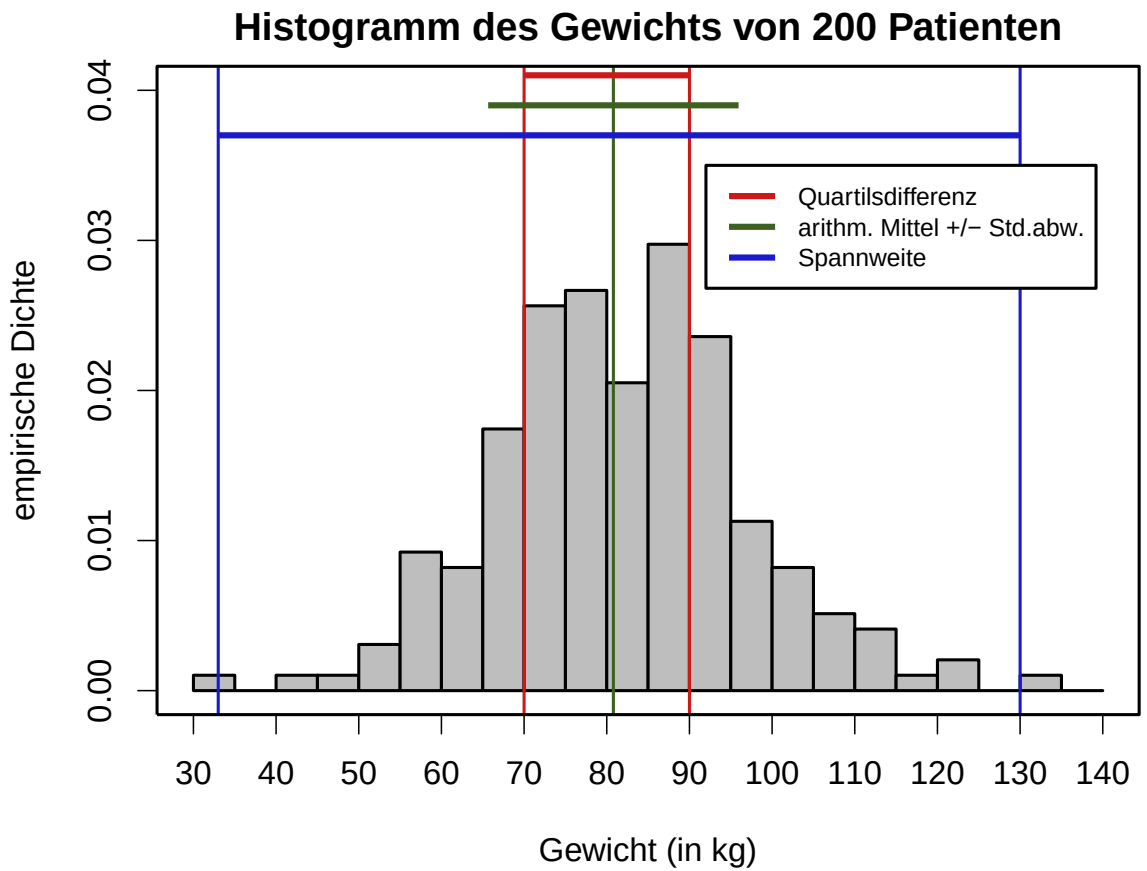


Abb. 20: Beispiel: Streuungsmaße des Gewichts von 200 Patienten

3.4.5 Streuungsmaße

Streuungsmaße 2

- **Variationskoeffizient** (relative Standardabweichung)

$$v_x := \frac{s_x}{\bar{x}} \tag{14}$$

- **Mittlere absolute Medianabweichung**, **MD** (von „Mean Deviation from the median“)

$$\mathrm{md}_x := \frac{1}{n} \sum_{i=1}^{n} |x_i - \mathrm{med}_x| \tag{15}$$

- **Mediane absolute Medianabweichung**, **MAD** (von „Median Absolute Deviation“)

$$\mathrm{mad}_x := \mathrm{med}(|x_i - \mathrm{med}_x|) \tag{16}$$

3.4.5 Streuungsmaße

Bemerkungen

- Die Streuung der Verteilungen (mit positiven Werten) unterschiedlicher Merkmale wird häufig durch den **Variationskoeffizienten** verglichen, weil er „dimensionslos“ ist, d.h. nicht von den Einheiten der verglichenen Merkmale abhängt.
- Ein wesentlicher Vorteil des Medians gegenüber dem arithmetischen Mittel ist die **Robustheit** gegenüber Ausreißern. Siehe auch „robuste Statistik“.

3.4.6 Eigenschaften der Lagemaße

Satz 1 (Minimaleigenschaft des arithmetischen Mittels)

Das arithmetische Mittel minimiert die mittlere quadratische Abweichung von einer Konstanten c:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - c)^2 \text{ für alle } c \in \mathbb{R} \tag{17}$$

Beweis.

$$\frac{d}{dc}\sum_{i=1}^{n}(x_i - c)^2 = -2\sum_{i=1}^{n}(x_i - c_{opt}) = 0 \quad \Rightarrow \quad c_{opt} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

□

Der Mittelwert ist damit derjenige Wert, der minimalen summierten quadratischen Abstand zu allen anderen Werten x_i hat.

3.4.6 Eigenschaften der Lagemaße

Satz 2 (Minimaleigenschaft des Medians)

Der Median minimiert die mittlere absolute Abweichung von einer Konstanten c:

$$\sum_{i=1}^{n} |x_i - \text{med}_x| \leq \sum_{i=1}^{n} |x_i - c| \text{ für alle } c \in \mathbb{R}. \tag{18}$$

Motivation (!)

$$\begin{aligned}
\sum_{i=1}^{n} |x_i - c| &= \sum_{x_i > c} (x_i - c) + \sum_{x_i < c} (c - x_i) + \sum_{x_i = c} (c - x_i) \\
\frac{d}{dc} \sum_{i=1}^{n} |x_i - c| &= \frac{d}{dc} \left(\sum_{x_i > c} (x_i - c) + \sum_{x_i < c} (c - x_i) \right) \\
&= \sum_{x_i > c} (-1) + \sum_{x_i < c} (+1) = 0 \quad \Leftrightarrow c_{opt} = \text{med}_x
\end{aligned}$$

3.4.7 Eigenschaften der Varianz

Satz 3 (Steiner'scher Verschiebungssatz)

Es seien $x_1, \ldots, x_n \in \mathbb{R}, a \in \mathbb{R}$. *Dann gilt:*

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2. \tag{19}$$

Beweis.

siehe Literatur; nicht Teil der Vorlesung □

Bemerkungen zu Steiner'schem Verschiebungssatz:

- Wichtiger Satz zur Zerlegung von Quadratsummen
- Anwendung im folgenden Satz

3.4.7 Eigenschaften der Varianz

Satz 4 (Spezieller Steiner'scher Verschiebungssatz)

Die empirische Varianz lässt sich berechnen als:

$$\text{var}_x = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2. \tag{20}$$

Beweis.

Anwendung von Satz 3 mit $a = 0$ und Division durch $n - 1$. □

3.4.8 Lage- und Streuungsmaße: Skalentypen

Idee

- Bisher wurde stillschweigend angenommen, dass das untersuchte Merkmal metrisch ist.
- Nur dann ist es möglich, mit den Merkmalswerten zu rechnen und also arithmetisches Mittel und Standardabweichung zu berechnen. Diese beiden Maßzahlen sind tatsächlich bei ordinalen und nominalen Merkmalen nicht verwendbar.
- Anders sieht es aus bei Quantilen, wie z.B. dem Median.
- Die obige Standard-Berechnung der Quantile ist sozusagen reparierbar, so dass diese *Maßzahlen* auch bei *ordinalen Merkmalen* verwendbar sind, wenn auch nicht bei nominalen Merkmalen, deren Werte ja gar nicht anzuordnen sind.

3.4.8 Lage- und Streuungsmaße: Skalentypen

Zur Reparatur der p-Quantile wird folgende Berechnung verwendet, die das arithmetische Mittel vermeidet.

p-Quantile für ordinale Merkmale:

$$Q_p := x_{(j)}, \quad j := \lceil np \rceil.$$

- Mit dem „reparierten“ Median haben wir ein angemessenes Lagemaß für ordinale Merkmale.
- Schwieriger sieht es mit Streuungsmaßen aus, denn sowohl bei der Quartilsdifferenz als auch bei der Spannweite muss mit den Merkmalswerten gerechnet werden.
- *Streuungsaussagen bei ordinalen Merkmalen* sind deshalb von folgendem Typ:
  - 50% der Werte liegen zwischen q_4 und q^4 oder
  - 100% der Werte liegen zwischen min und max.
  - Für *nominale Merkmale* steht mit dem Modalwert bisher lediglich ein Lagemaß zur Verfügung. Ein Streuungsmaß wird ab Seite 86 vorgestellt.

3.4.8 Lage- und Streuungsmaße: Skalentypen

Tab. 1: Lagemaße bei den verschiedenen Skalentypen

Skalentyp:	Mittelwert	Median	Modalwert
metrisch	ja	ja	(ja)
ordinal	-	ja	ja
nominal	-	-	ja

Tab. 2: Streuungsmaße bei den verschiedenen Skalentypen

Skalentyp:	Std.Abw., Varianz	Quartile, Quartilsdiff.	Min., Max. Spannweite
metrisch	ja	ja	ja
ordinal	-	ja	ja
nominal	-	-	-

3.4.9 Empirische Entropie

Idee

Ausgehend von der Idee, dass seltene Ereignisse mehr „Information“ beinhalten als häufige, schlug der österreichische Physiker Boltzmann (1844 - 1906) vor, den *Informationsgehalt* eines Ereignisses durch $ln(1/p)$ zu bestimmen, wobei p die Wahrscheinlichkeit des Ereignisses ist.

Die „mittlere“ Information von K möglichen Ereignissen heißt dann die **Entropie** $\sum_{k=1}^{K} p_k ln(1/p_k)$, d.h. die mit den Wahrscheinlichkeiten p_k gewichteten Informationsgehalte $ln(1/p_k)$ der möglichen Ereignisse.

- Diese Idee wird nun zur Definition eines Streuungsmaßes für nominale Merkmale verwendet.
- Dabei werden die Ereignisse betrachtet, dass bei einem Merkmalsträger ein bestimmter Merkmalswert angenommen wird.
- Die Wahrscheinlichkeiten werden durch die relativen Häufigkeiten angenähert.
- Man spricht deshalb auch von empirischer Entropie.

3.4.9 Empirische Entropie

Definition 7 (Empirische Entropie als Streuungsmaß)

Gegeben seien n Merkmalsträger in K Klassen mit absoluten Häufigkeiten H_k und relativen Häufigkeiten h_k, $k = 1, \ldots, K$.
Dann ist die **(normierte) empirische Entropie** *B (nach Boltzmann) definiert durch:*

$$B := \frac{1}{ln(K)} \sum_{\substack{k=1, \\ h_k \neq 0}}^{K} h_k \cdot \ln\left(\frac{1}{h_k}\right) \tag{21}$$

3.4.9 Empirische Entropie

Bemerkung

Das Teilen durch $ln(K)$ bewirkt eine „Normierung“:

- B ist maximal $= 1$, nämlich wenn $h_k = 1/K$, $k = 1, \ldots, K$,
- B ist minimal $= 0$, nämlich wenn $h_j = 1$ und $h_k = 0$ für alle $k \neq j$.

Diese Eigenschaften qualifizieren die empirische Entropie B zu einem Streuungsmaß, denn B ist maximal, wenn die Merkmalswerte maximal streuen, und minimal bei überhaupt keiner Streuung.

Idee

Natürlich hat man bei den meisten Verteilungen noch keine Vorstellung über die Form eines dazugehörigen Histogramms, wenn man nur Lage- und Streuungsmaße kennt.

Hier werden deshalb noch 3 weitere Charakteristiken von Verteilungen angesprochen, nämlich **Anzahl Gipfel**, **Schiefe** und **Wölbung**.

3.5.1 Verteilungscharakterisierungen: Gipfel

Gipfel

Bei mehrgipfligen Häufigkeitsverteilungen, insbesondere bei U-förmigen, sind, im Gegensatz zu eingipfligen Verteilungen, die Lagemaße oft nicht charakteristisch für die Verteilung.

Man spricht von ein-, zwei- und mehrgipfligen Verteilungen.

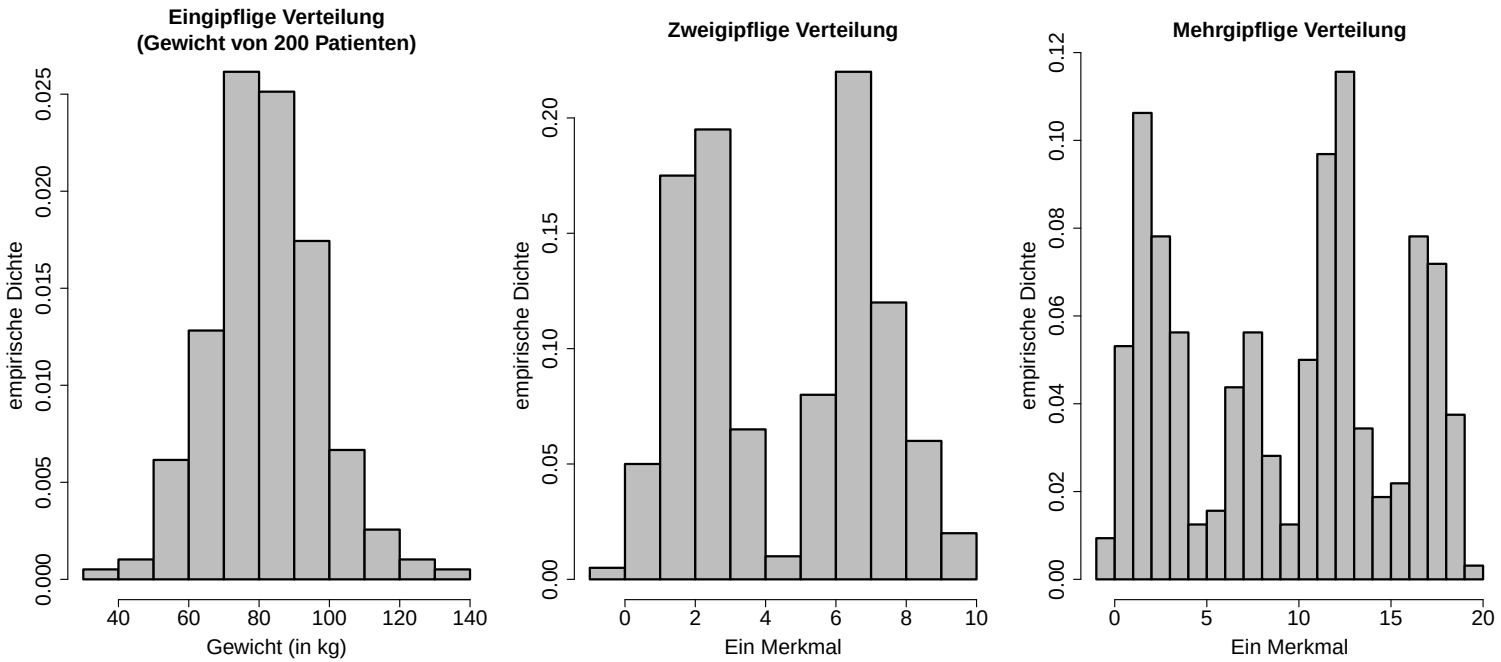


Abb. 21: a) ein-, b) zwei-, c) mehrgipflige Verteilung

3.5.2 Verteilungscharakterisierungen: Schiefe

Schiefe (skewness)

Eingipflige Verteilungen können auf 2 Arten schief sein:

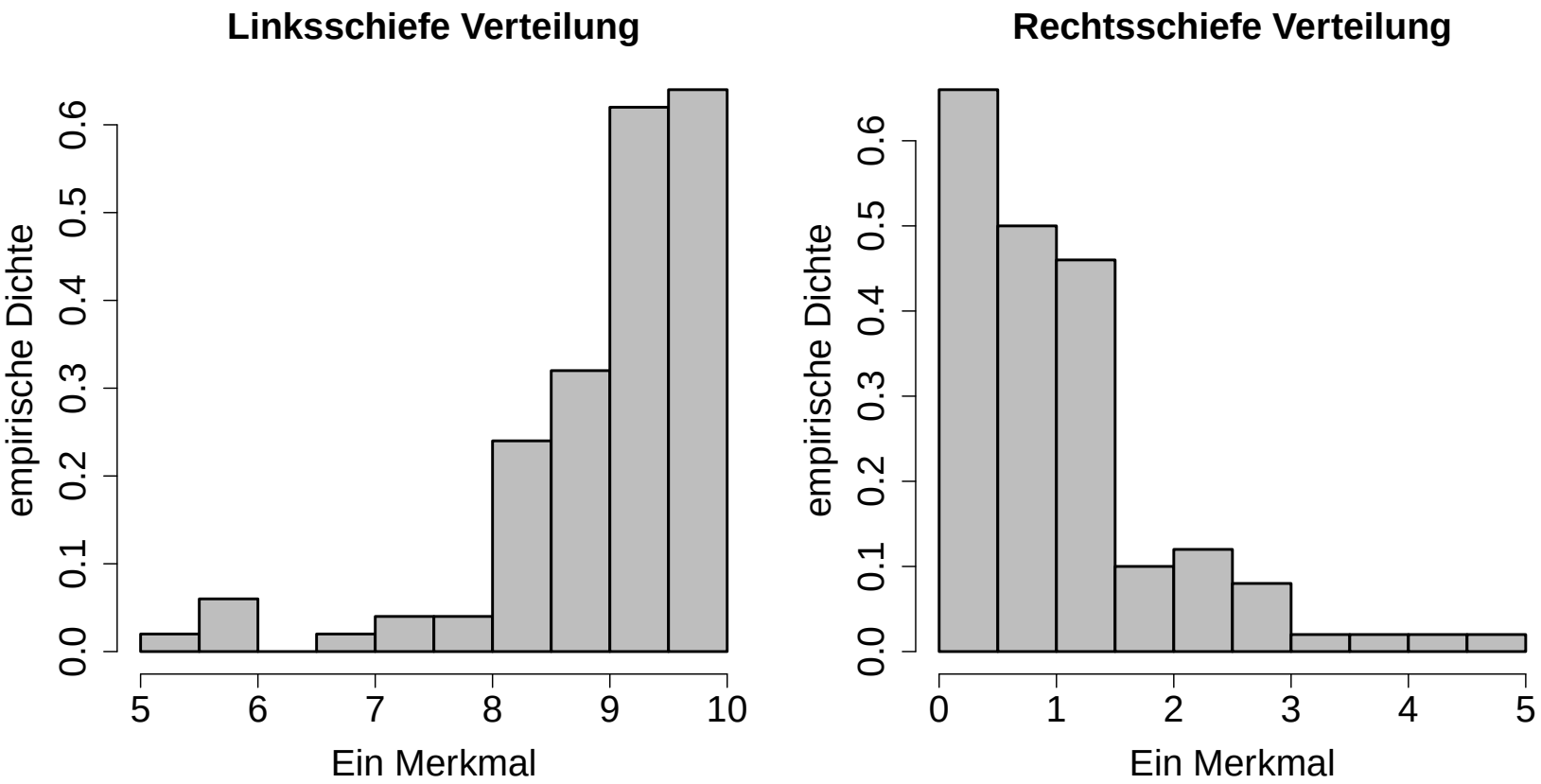


Abb. 22: a) link-, b) rechtsschiefe Verteilung

3.5.2 Verteilungscharakterisierungen: Schiefe

Bei eingipfligen symmetrischen Verteilungen stimmen Modalwert, Median und Mittelwert überein.

Linksschiefe und Rechtsschiefe lassen sich als systematische Verschiebung des Mittelwerts gegenüber Median und Modalwert nach links bzw. rechts charakterisieren.

Definition 8 (Schiefe, Mittelwert-Variante)

Eine eingipflige Verteilung heißt

- **rechtsschief oder linkssteil**, *falls* $\text{mod}_x < \text{med}_x < \bar{x}$
- **linksschief oder rechtssteil**, *falls* $\text{mod}_x > \text{med}_x > \bar{x}$
- **symmetrisch**, *falls* $\text{mod}_x = \text{med}_x = \bar{x}$

3.5.2 Verteilungscharakterisierungen: Schiefe

Beispiel 1: Gewicht von 200 Patienten.

Diese Verteilung ist eingipflig und symmetrisch, denn $\text{mod}_x = 80 = \text{med}_x = 80 \approx \bar{x} = 80.8$.

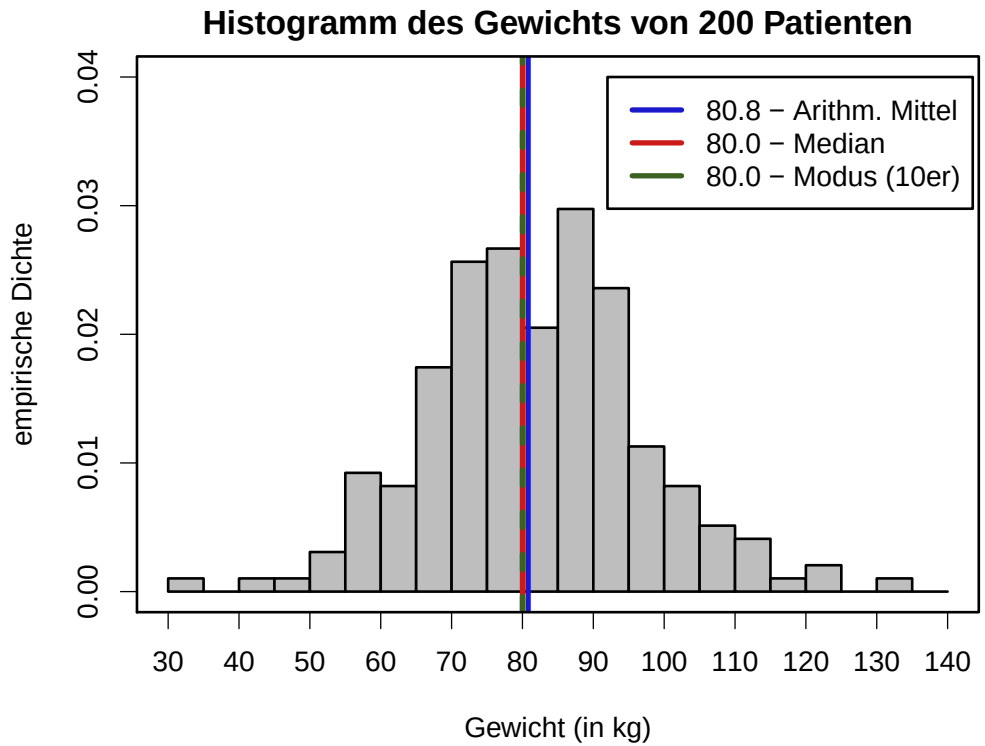


Abb. 23: Lagemaße des Gewichts von 200 Patienten

3.5.3 Verteilungscharakterisierungen: Wölbung

Wölbung (Kurtosis, Exzess)

Idee

Verteilungen können also mehrgipflig sein oder eingipflig und symmetrisch oder schief.

Bei symmetrischen eingipfligen Verteilungen ist außerdem von Interesse, ob die Verteilung spitz oder flach ist.

Maße dafür sind die sogenannten Wölbungskoeffizienten.

3.5.3 Verteilungscharakterisierungen: Wölbung

Formen der Wölbung bei **symmetrischen** Verteilungen

Die Merkmalsausprägungen sind bei

i) **abgeplatteten (platykurtischen)** Verteilungen
gleichmäßig verteilt über begrenzten Bereich;

ii) **spitzen (leptokurtischen)** Verteilungen
stark um Mittelwert konzentriert mit wenigen weit abseits liegenden Werten;

iii) **mesokurtischen** Verteilungen
„ausgewogen“ um den Mittelwert verteilt.

Um die Stärke der Wölbung zu bestimmen, werden unterschiedliche sog. Wölbungskoeffizienten vorgeschlagen (hier nicht diskutiert).

3.5.4 Verteilungscharakterisierungen: Fazit

Fazit: Verteilungscharakterisierungen

- **Lagemaße** charakterisieren die *Mitte der Verteilung*.
- **Streuungsmaße** charakterisieren die *Streuung um die Mitte*.
- **Schiefekoeffizienten** charakterisieren die *Unsymmetrie*.
- **Wölbungskoeffizienten** charakterisieren die *Steilheit*.
- Unnötige **Gruppierung** der Daten führt zu *Informationsverlust*.

3.6 Box-Plots

Idee

- Grafische Darstellungen sind viel anschaulicher als algebraische Charakteristiken.
- Deshalb erfreut sich eine grafische Darstellung von 5 Kennzahlen der Häufigkeitsverteilung von Merkmalswerten großer Beliebtheit.

3.6 Box-Plots

Idee

- Grafische Darstellungen sind viel anschaulicher als algebraische Charakteristiken.
- Deshalb erfreut sich eine grafische Darstellung von 5 Kennzahlen der Häufigkeitsverteilung von Merkmalswerten großer Beliebtheit.
- Dargestellt wird das Pentagramm, bestehend aus den 5 Kennzahlen Minimum, Maximum, unteres und oberes Quartil und Median.
- Die Bezeichnung **Pentagramm** basiert auf der angedeuteten 5-eckigen Darstellung dieser Charakteristiken.

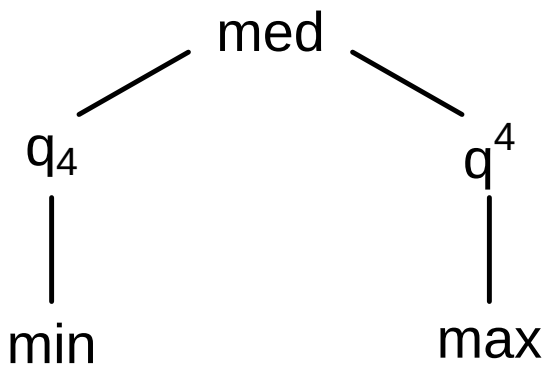


Abb. 24: Pentagramm

3.6 Box-Plots

Definition 9 (Box-Plot)

Der **Box- (und whisker-) Plot** *besteht aus einem Kasten (Box), mit*

- *unterem Quartil q_4 und oberem Quartil q^4 als Begrenzungslinien,*
- *Median med_x als innere Linie,*
- *Verbindungslinien (whiskers, Schnurrbarthaare), die bis zum äußersten Wert gezogen, der*
  - *nicht größer ist als $q^4 + 1.5 \cdot (q^4 - q_4)$ bzw.*
  - *nicht kleiner ist als $q_4 - 1.5 \cdot (q^4 - q_4)$.*

  Die Grenzen $q^4 + 1.5 \cdot (q^4 - q_4)$ und $q_4 - 1.5 \cdot (q^4 - q_4)$ heißen **innere Zäune** *des Box-Plots, die Punkte zwischen Box und inneren Zäunen* **Anrainer**.

Alle Punkte, die jenseits der Verbindungslinien liegen, heißen **Außenpunkte** *und werden mit $\circ$ gekennzeichnet.*

3.6 Box-Plots

Bemerkungen

- In dem Kasten liegen die *mittleren 50% der Rangliste*.
- Der Box-Plot kann sowohl *vertikal* als auch *horizontal* gezeichnet werden.
- Bei Normalverteilung sind nur 0.7% der Werte Außenpunkte.
- Das bedeutet, man muss je nach Anzahl an Beobachtungen eine entsprechende Anzahl an Außenpunkten erwarten.

3.6 Box-Plots

Beispiel: geordnete Liste: -15, 1, 3, 4, 4, 6, 6, 7

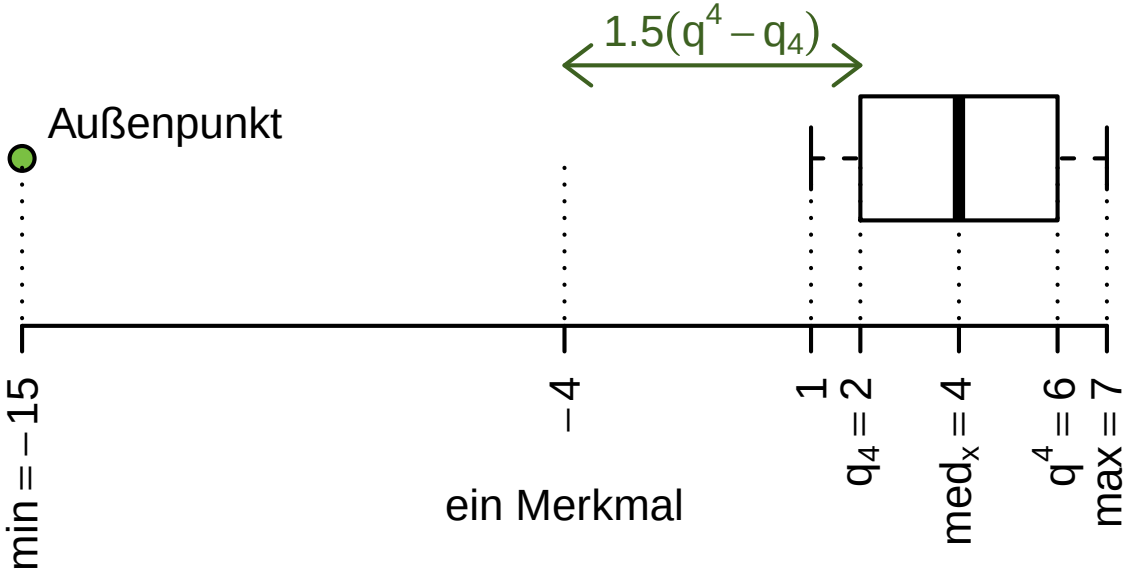


Abb. 25: Boxplot

3.6 Box-Plots

Beispiel 1: Gewicht von 200 Patienten

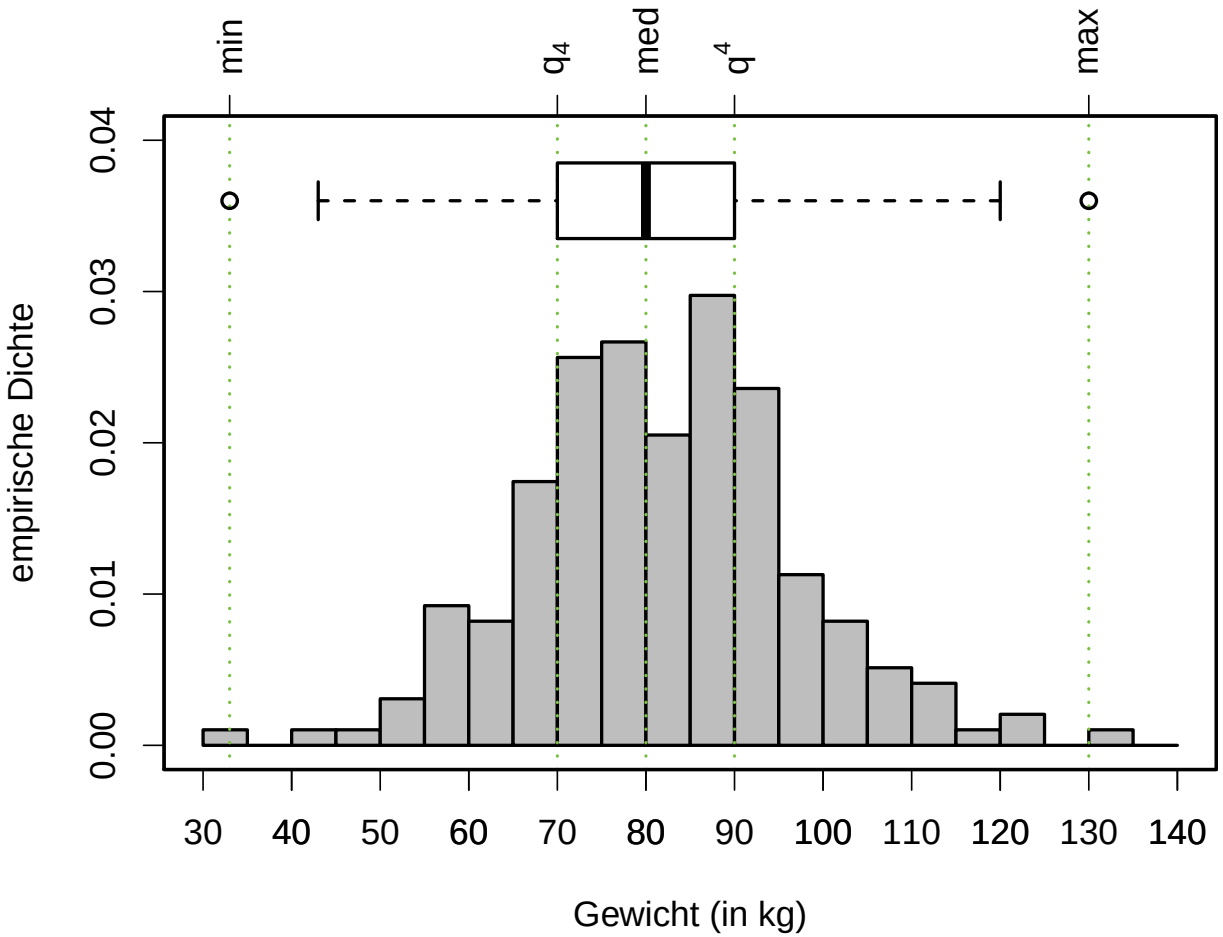


Abb. 26: Histogramm und Boxplot des Gewichts von 200 Patienten

3.6.1 Parallele Box-Plots

Idee: Parallele Box-Plots

- Oft ist es wichtig, verschiedene Verteilungen zu vergleichen, z.B. bei der Unterteilung der Merkmalsträger in logische Gruppen (etwa Männer und Frauen).
- In solchen Fällen wird ein Box-Plot pro Gruppe gezeichnet, und zwar parallel bzgl. derselben Skala (*parallele Box-Plots*).
- Dann werden Lage und Streuung der Gruppen miteinander verglichen.
- Anstatt mehrere Histogramme zu vergleichen, kann man schneller und auf weniger Platz mehrere Box-Plots vergleichen.

3.6.1 Parallele Box-Plots

Beispiel 1: Gewicht von 200 Patienten

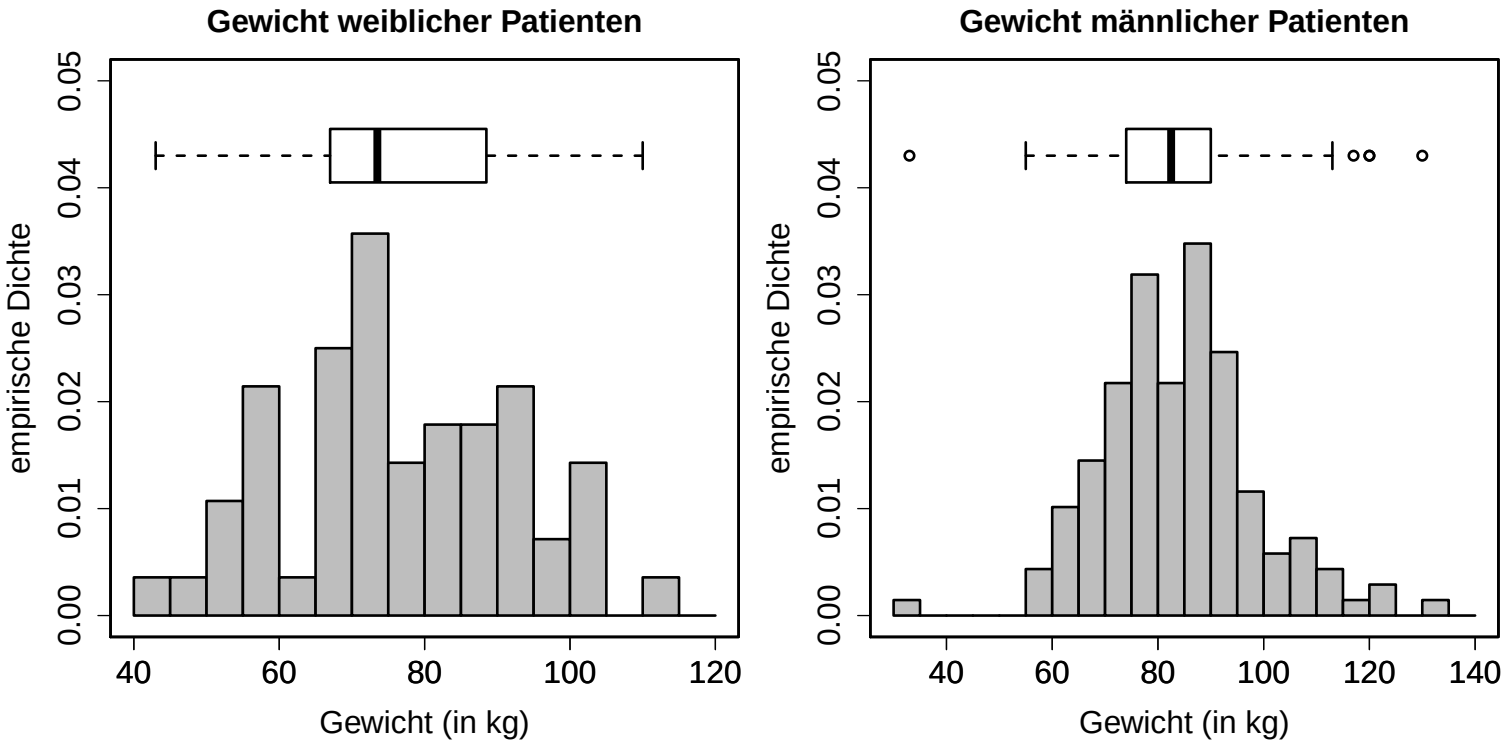


Abb. 27: Der Vergleich des Gewichts von weiblichen und männlichen Patienten fällt so sehr schwer!

3.6.1 Parallele Box-Plots

Beispiel 1: Gewicht von 200 Patienten

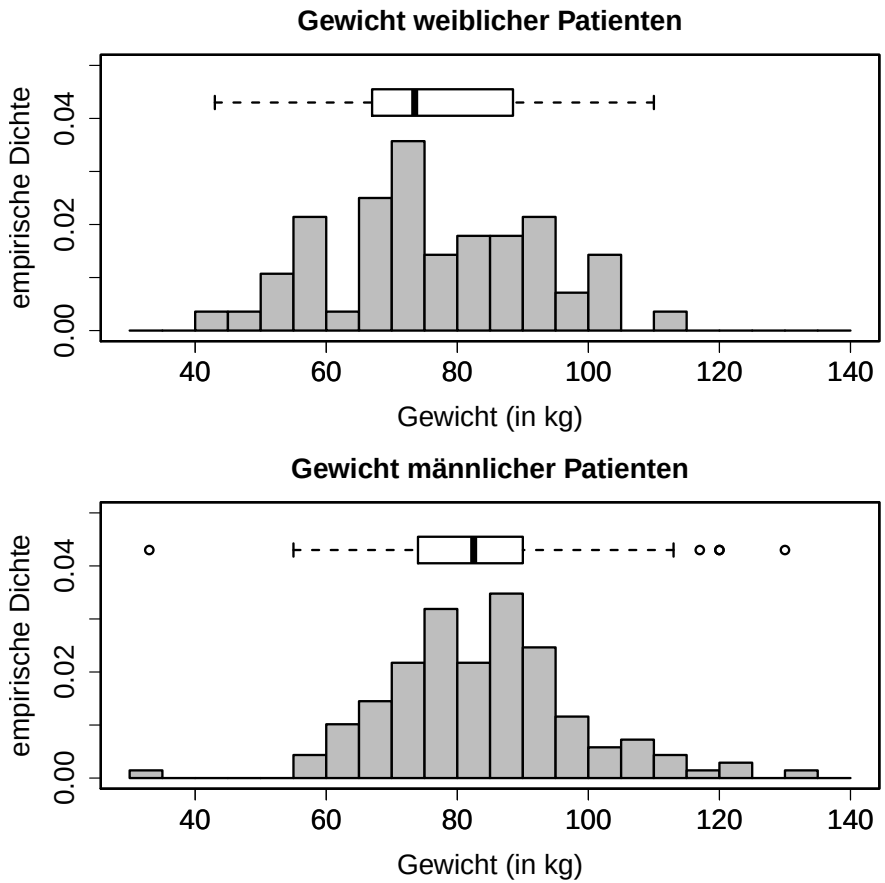


Abb. 28: Der Vergleich des Gewichts von weiblichen und männlichen Patienten fällt so schon leichter.

3.6.1 Parallele Box-Plots

Beispiel 1: Gewicht von 200 Patienten

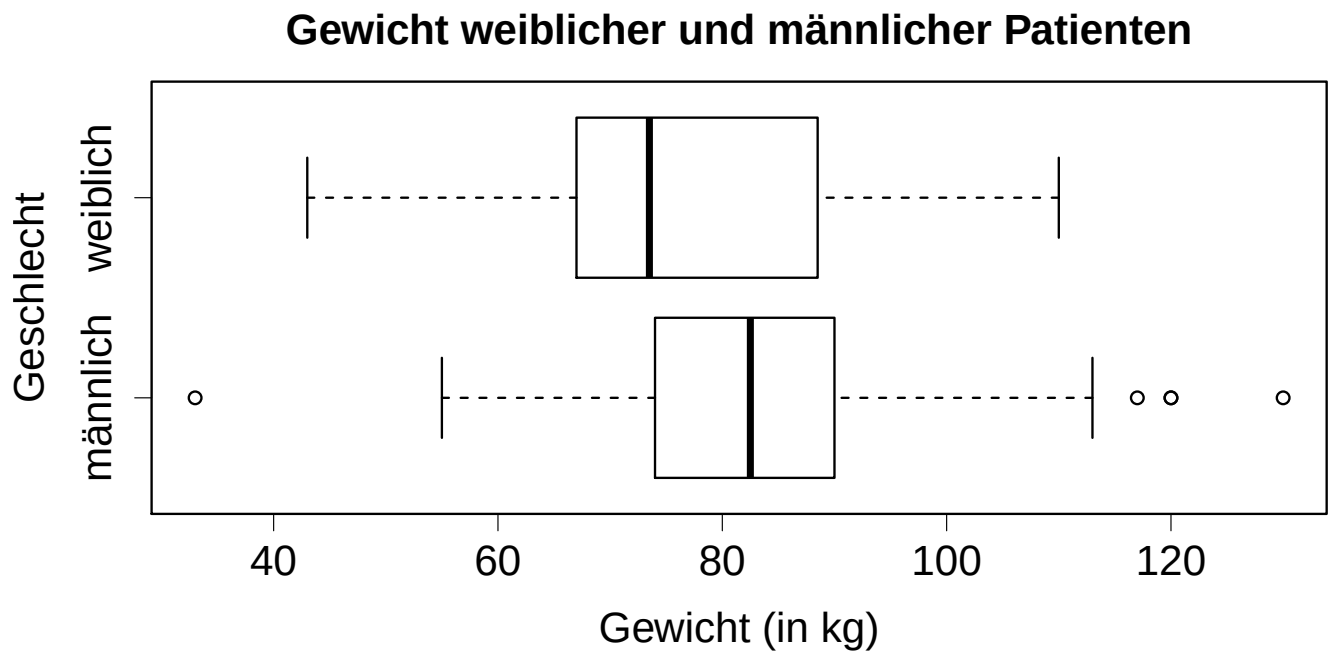


Abb. 29: Vergleich des Gewichts weiblicher und männlicher Patienten

3.6.1 Parallele Box-Plots

Fragen:

- Wo ist das Zentrum der Daten?
- Wie streuen die Daten?
- Wie ist die Schiefe ist die Verteilung?

Antworten:

- Frauen wiegen weniger als Männer (alle Maße der Frauen im Boxplot kleiner als entsprechende Maße für die Männer).
- Die mittleren 50% streuen bei den Frauen stärker als bei den Männern.
- Beide Verteilungen sind recht symmetrisch.

3.7 Übungsaufgaben

- Gegeben seien die folgenden Daten: -2, 0, 0, 2, 5.

  Berechnen Sie alle Maßzahlen zur Lage und Streuung, die auf Folie 84 angegeben sind. Nehmen Sie nacheinander an, dass die Daten auf Kardinalskalen-, Ordinalskalen- und Nominalskalen-Niveau vorliegen und geben Sie jeweils eine geeignete Maßzahl für die Lage und für die Streuung an.

- Gegeben sei ein Datensatz (x_1, x_2, x_3) mit den Werten $x_1 = 1$ und $x_2 = 2$ und $x_3 \in \mathbb{R}$.

  Für welche Werte von x_3 stimmen jeweils zwei der Maßzahlen Modus, Median und Mittelwert überein? Für welche Werte von x_3 stimmen Standardabweichung und Varianz überein?

3.7 Übungsaufgaben

- Gegeben seien die folgenden Daten: -2, 0, 0, 2, 5.

  Mittelwert: $\frac{-2+0+0+2+5}{5} = 1$,
  Median: $x_{(3)} = 0$,
  Modalwert: 0.
  Varianz: $\frac{(-2-1)^2+(0-1)^2+(0-1)^2+(2-1)^2+(5-1)^2}{5-1} = \frac{9+1+1+1+16}{4} = 7$,
  Standardabweichung: $\sqrt{7}$,
  Quartilsdifferenz: 2-0=2, da $Q_{0.25} = x_{(2)} = 0$ und $Q_{0.75} = x(4) = 2$,
  Spannweite: $5 - (-2) = 7$.

- Gegeben sei ein Datensatz mit den Werten $x_1 = 1$, $x_2 = 2$, $x_3 \in \mathbb{R}$.

  Modus=Median: alle $x_3 \in \mathbb{R}$ (aber Modus nicht eindeutig)
  Median=Mittelwert: $x_3 = 0$ oder $x_3 = 1.5$ $x_3 = 3$.
  Modus=Mittelwert: $x_3 = 0$ oder $x_3 = 1.5$ $x_3 = 3$ (aber Modus nicht eindeutig)
  Standardabweichung=Varianz $\Rightarrow$ Varianz=1 $\Rightarrow$ $x_3 = 0$ oder $x_3 = 3$
  (Für Varianz=0 gilt auch Stand.abw.=Varianz, aber wegen $x_1 \neq x_2$ nicht möglich)

4.1 Bivariate Merkmale

Idee

- Bei der Untersuchung eines bivariaten Merkmals geht es um die Darstellung von *bivariaten Häufigkeitsverteilungen*.
- Wir beschäftigen uns dabei mit Situationen, in denen 2 Merkmale an demselben Merkmalsträger gleichzeitig auftreten.
- Wir untersuchen also Paare von Beobachtungen.
- Dabei interessiert insbesondere die *Art des Zusammenhangs der beiden Merkmale*.
- Bei der *Zusammenhangsanalyse* von 2 Merkmalen werden einerseits grafische Darstellungen verwendet, andererseits aber auch statistische Maßzahlen.

4.2 Kontingenztafeln

Bivariate Merkmale

4 Bivariate Merkmale
4.2 Kontingenztafeln

Definition 10 (Kontingenztafel)

Die Merkmalsausprägungen x des univariaten Merkmals X seien in Klassen $A_1, \ldots, A_K$ eingeteilt, die Merkmalsausprägungen y des univariaten Merkmals Y in Klassen $B_1, \ldots, B_L$.
Eine **Kontingenztafel** *ist ein 2-dimensionales Schema, in dem für jede Kombination einer der Klassen $A_1, \ldots, A_K$ mit einer der Klassen $B_1, \ldots, B_L$ die* **absolute Häufigkeit** *eingetragen wird, dass ein Paar (x, y) diese Klassenkombination aufweist.*

4.2 Kontingenztafeln

Tab. 3: Kontingenztafel

Klassen von X	Klassen von Y: B_1	B_2	...	B_L	Zeilen-summen
A_1	H_{11}	H_{12}	...	H_{1L}	Z_1
A_2	H_{21}	H_{22}	...	H_{2L}	Z_2
A_3	H_{31}	H_{32}	...	H_{3L}	Z_3
...			...		...
A_K	H_{K1}	H_{K2}	...	H_{KL}	Z_K
Spalten-summen	S_1	S_2	...	S_L	n

4.2 Kontingenztafeln

Definition 11

- *H_{kl} ist die gemeinsame (absolute) Häufigkeit der Klassen A_k und B_l bzw. die (absolute) Häufigkeit der* **Zelle** *k, l.*
- *Die $Z_k := H_{k1} + H_{k2} + \ldots + H_{kL}$, $k = 1, \ldots, K$ heißen* **Zeilensummen** *und die $S_l := H_{1l} + H_{2l} + \ldots + H_{Kl}$, $l = 1, \ldots, L$ heißen* **Spaltensummen**.
- *Die Spalten- bzw. Zeilensummen repräsentieren die Häufigkeitsverteilung von X bzw. Y. Diese heißen auch* **Randverteilungen**.

Bemerkung

Bei *nominalen, ordinalen* oder *diskreten metrischen* Daten werden die Klassen häufig nur mit einer Merkmalsausprägung besetzt.

4.2 Kontingenztafeln

Erweiterung von **Beispiel 1:**

Patientendaten erhoben bei Untersuchungen und Behandlung nach kardiologischem Notfall; `NA`: fehlender Wert (Not Available)

a) Gewicht (in kg)
b) Größe (in cm)
c) Herzfrequenz (in Hertz)
d) systolischer Blutdruck (in mmHg)
e) Geburtsdatum
f) Untersuchungsdatum
g) Geschlecht (m / w)
h) Diagnose (ACS = Akutes Koronarsyndrom, av = AV-Knoten-Störung, chf = Chronischer Herzfehler)
i) Rhythmus (KF = Kammerflimmern, SM = Schrittmacher, SR = Sinusrhythmus, VHF = Vorhofflimmern)

4.2 Kontingenztafeln

Tab. 4: Kontingenztafeln zu je zwei Merkmalen von 200 Patienten

Diagnose

Geschl.	ACS	av	chf	$\sum$
m	92	19	29	140
w	42	5	12	59
$\sum$	134	24	41	199

Rhythmus

Geschl.	KF	SM	SR	VHF	$\sum$
m	0	1	115	20	136
w	0	0	56	3	59
$\sum$	0	1	171	23	195

Rhythmus

Diagnose	KF	SM	SR	VHF	$\sum$
ACS	0	0	123	8	131
av	0	1	22	2	25
chf	0	0	27	13	40
$\sum$	0	1	172	23	196

4.2 Kontingenztafeln

Interpretation

- Der Anteil von Männern mit Rhythmus „VHF“ ist höher als der der Frauen.
- Der Anteil von Männern mit Diagnose „av“ ist höher als der der Frauen.
- Besonders bei Diagnose „chf“ konnte als Rhythmus „VHF“ beobachtet werden.

Bemerkungen

- Es ist schwierig, schon bei halbwegs großen Zahlen alle Zusammenhänge zu finden.
- Das führt zur Suche nach geeigneten Visualisierungsmöglichkeiten.
- Beispielsweise könnte man für Tabelle 4.4 (a) versuchen, ein Stabdiagramm zu zeichnen, das für jede Diagnose einen nach Geschlecht eingefärbten Anteil enthält.

4.2 Kontingenztafeln

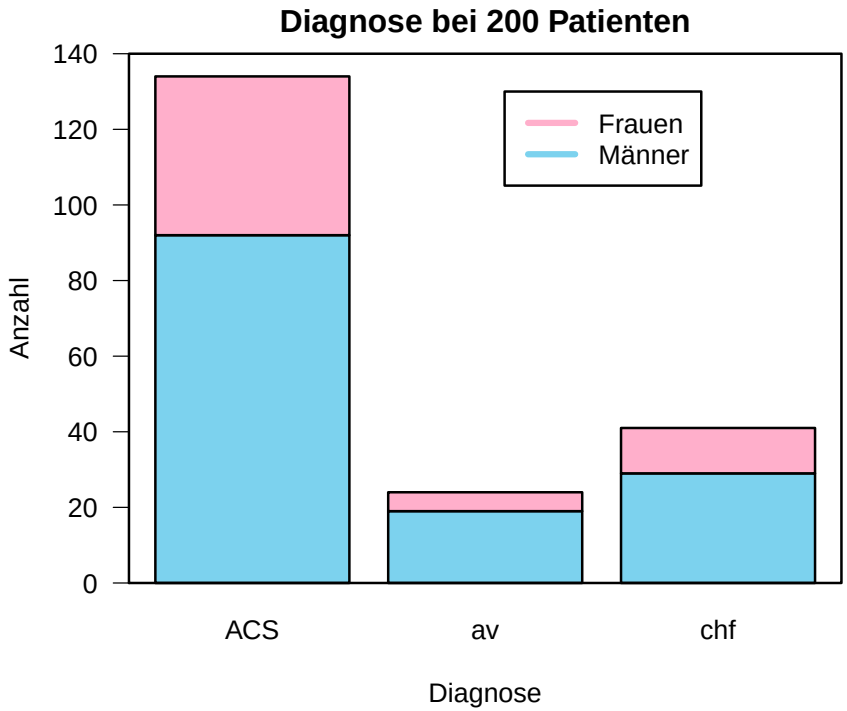


Abb. 30: Visualisierungsversuch des Zusammenhangs von Diagnose und Geschlecht

Die Anteile von Männern und Frauen pro Diagnose sind nur schwer vergleichbar.

4.2.1 Mosaikplots

Definition 12 (Mosaikplot)

Ein **Mosaikplot** *ist ein Verfahren zur Visualisierung zweier oder mehrerer (aber weniger) qualitativer (nominalen, ordinalen oder seltener klassiert metrischen) Merkmale mit folgenden Eigenschaften:*

- *Die Fläche der einzelnen Zellen ist proportional zur Anzahl der Fälle in dieser Zelle (analog zum Stabdiagramm und zum Histogramm)*
- *Die Gesamtfläche ist 1 und wird vollständig ausgenutzt.*
- *Zwischenräume dienen der Übersichtlichkeit und tragen nicht zu den Anteilen bei.*
- *Bei Visualisierung von mehr als 2 Merkmalen wird rekursiv vorgegangen.*

4.2.1 Mosaikplots

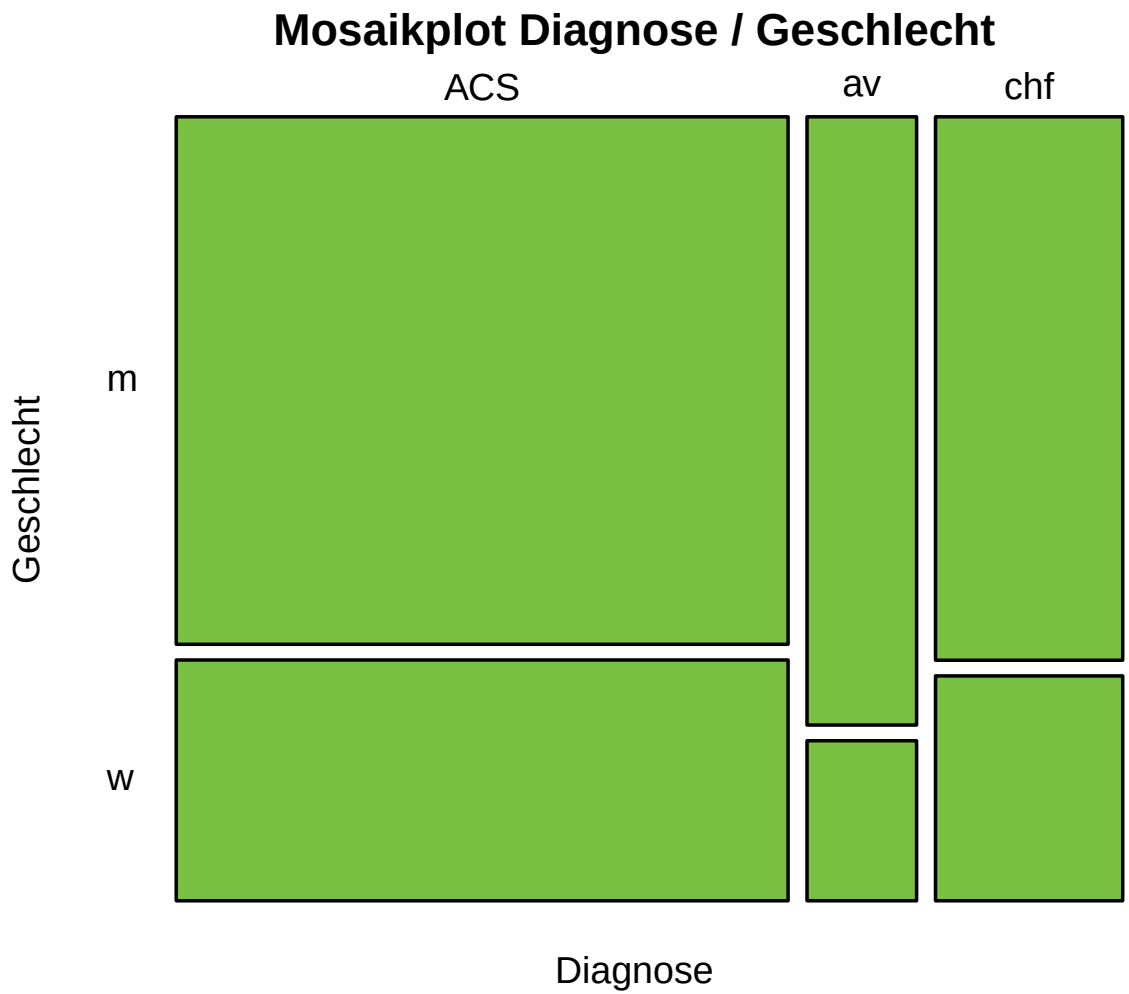


Abb. 31: Mosaikplot des Zusammenhangs von Diagnose und Geschlecht

4.2.1 Mosaikplots

Bemerkungen:

- Durch die pro Faktorstufe des einen Merkmals gleich großen Flächen lassen sich Unterschiede der Verhältnisse der Faktorstufen des anderen Merkmals sehr gut vergleichen.
- Das Umsortieren von Merkmalen und Faktorstufen kann wichtig sein, da es die Darstellung verändert und zu anderer Einsicht führen kann (siehe folgende Abbildung).

4.2.1 Mosaikplots

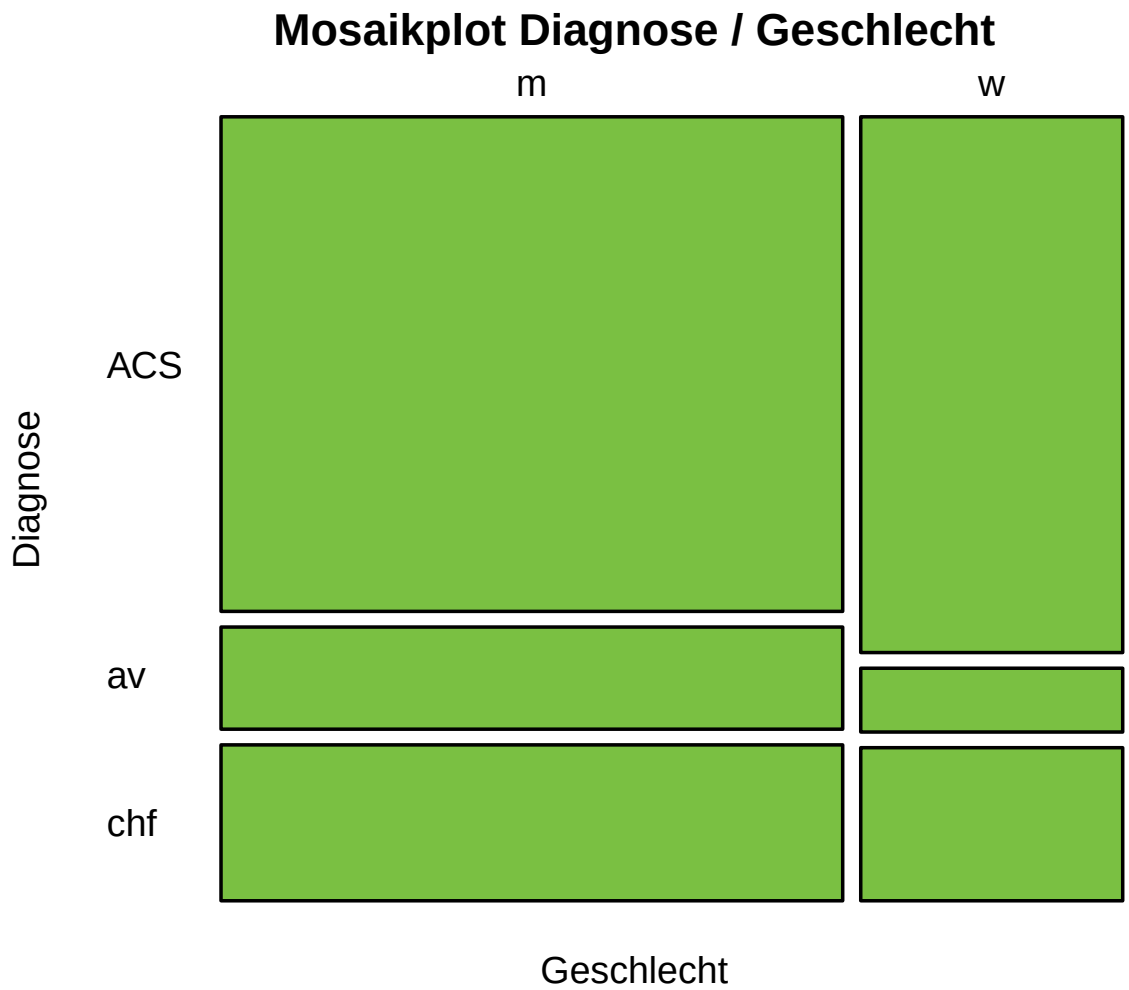


Abb. 32: Mosaikplot des Zusammenhangs von Diagnose und Geschlecht

4.2.1 Mosaikplots

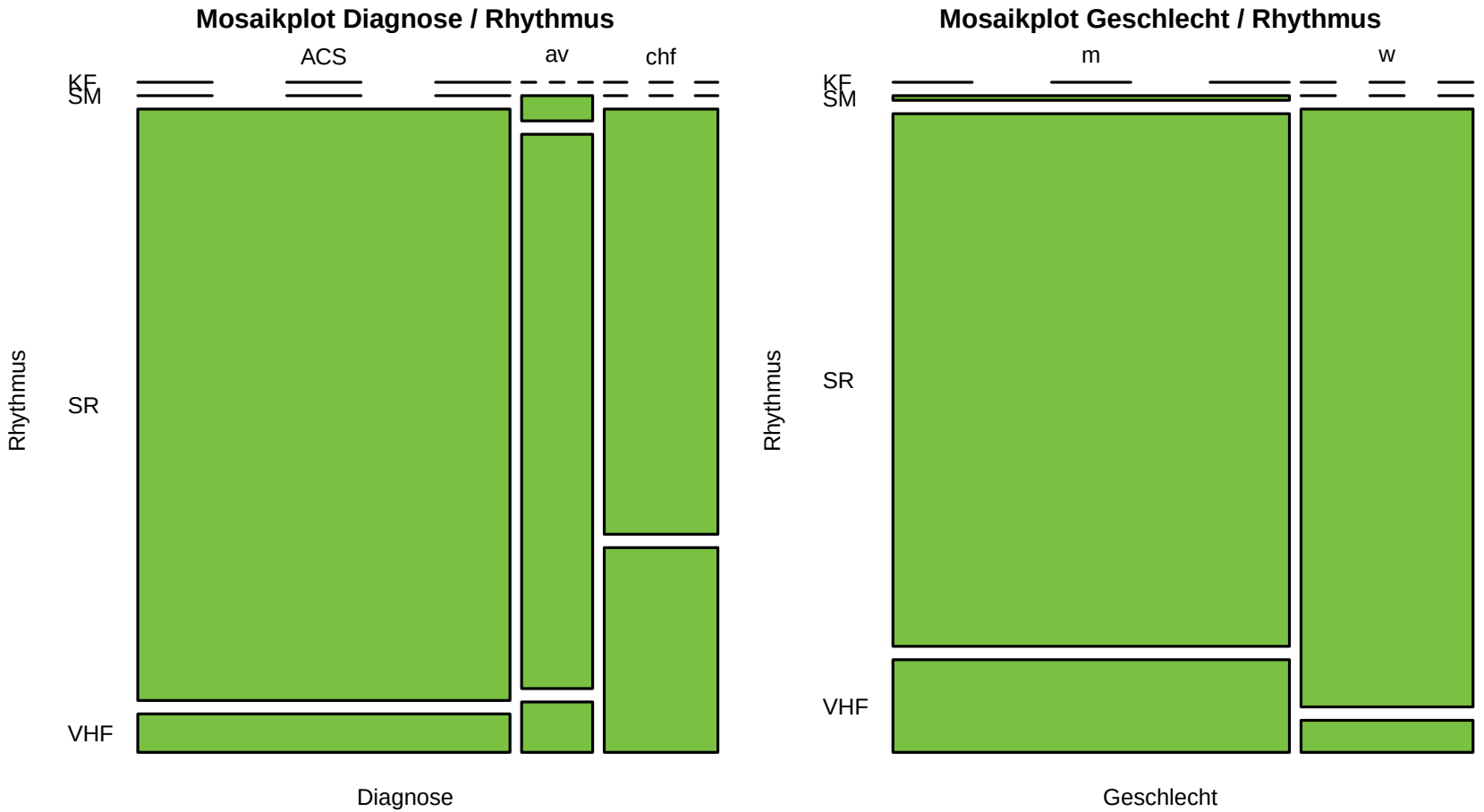


Abb. 33: Mosaikplots der weiteren Zusammenhänge der Kontingenztafeln aus Tabelle 4.4

4.2.1 Mosaikplots

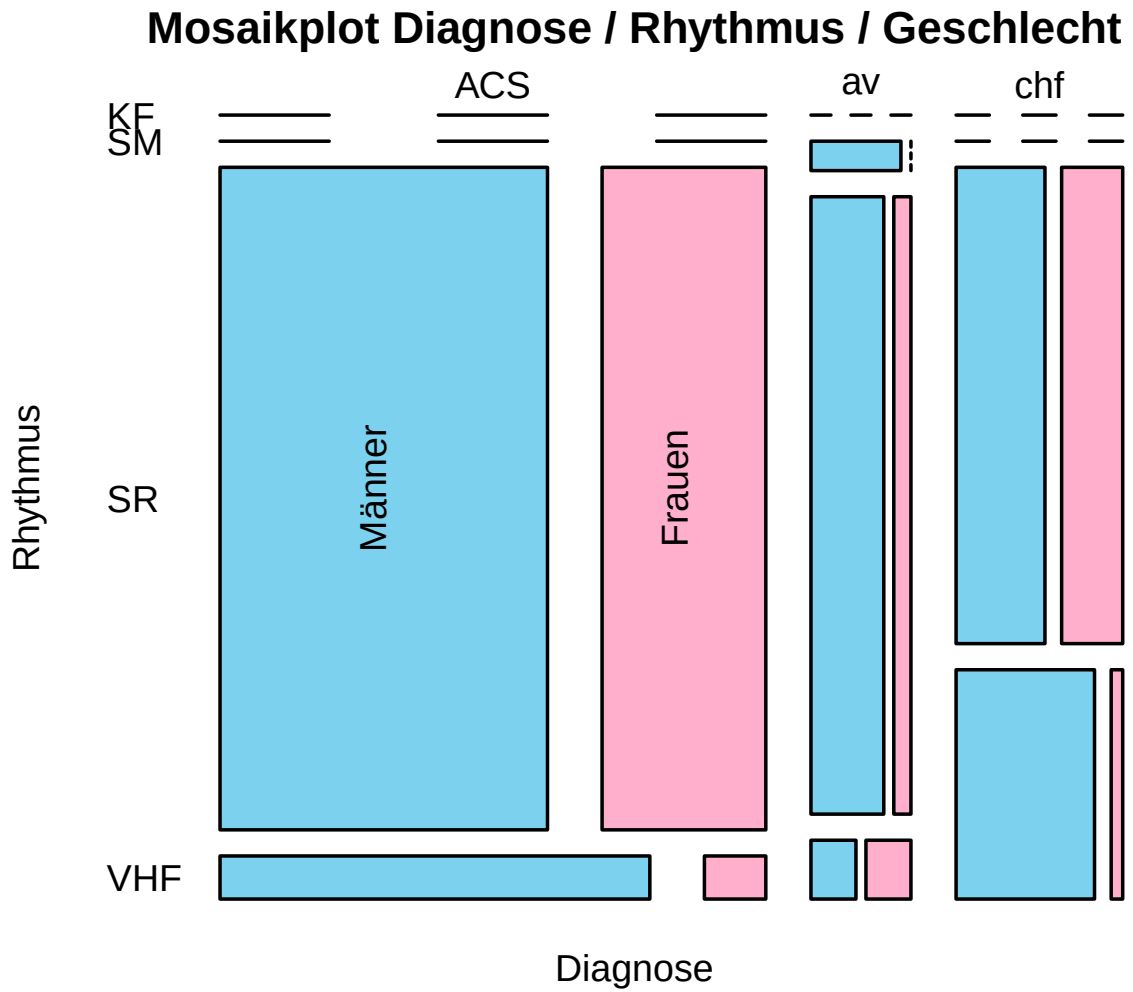


Abb. 34: Mosaikplot: gemeinsame Darstellung von 3 Merkmalen aus beiden Mosaikplots aus Abbildung 4.33

4.2.2 Streudiagramme

Idee

- Bei metrischen bivariaten Merkmalen bilden die Merkmalswertepaare der verschiedenen Merkmalsträger eine *Punktwolke* in der Ebene.
- Diese lässt sich anschaulich in einem x-y-Koordinatensystem darstellen.

Definition 13 (Streudiagramm)

Ein **Streudiagramm (scatterplot)** *ist eine grafische Darstellung von 2 Merkmalen, wobei das eine Merkmal auf der x-Achse, das andere auf der y-Achse eines x-y-Koordinatensystems abgetragen wird.*

Jedem Merkmalsträger $i = 1, \ldots, n$ *entspricht dabei ein Symbol an der Stelle* (x_i, y_i)*, wobei* x_i *bzw.* y_i *der Wert des Merkmals* X *bzw.* Y *ist.*

4.2.2 Streudiagramme

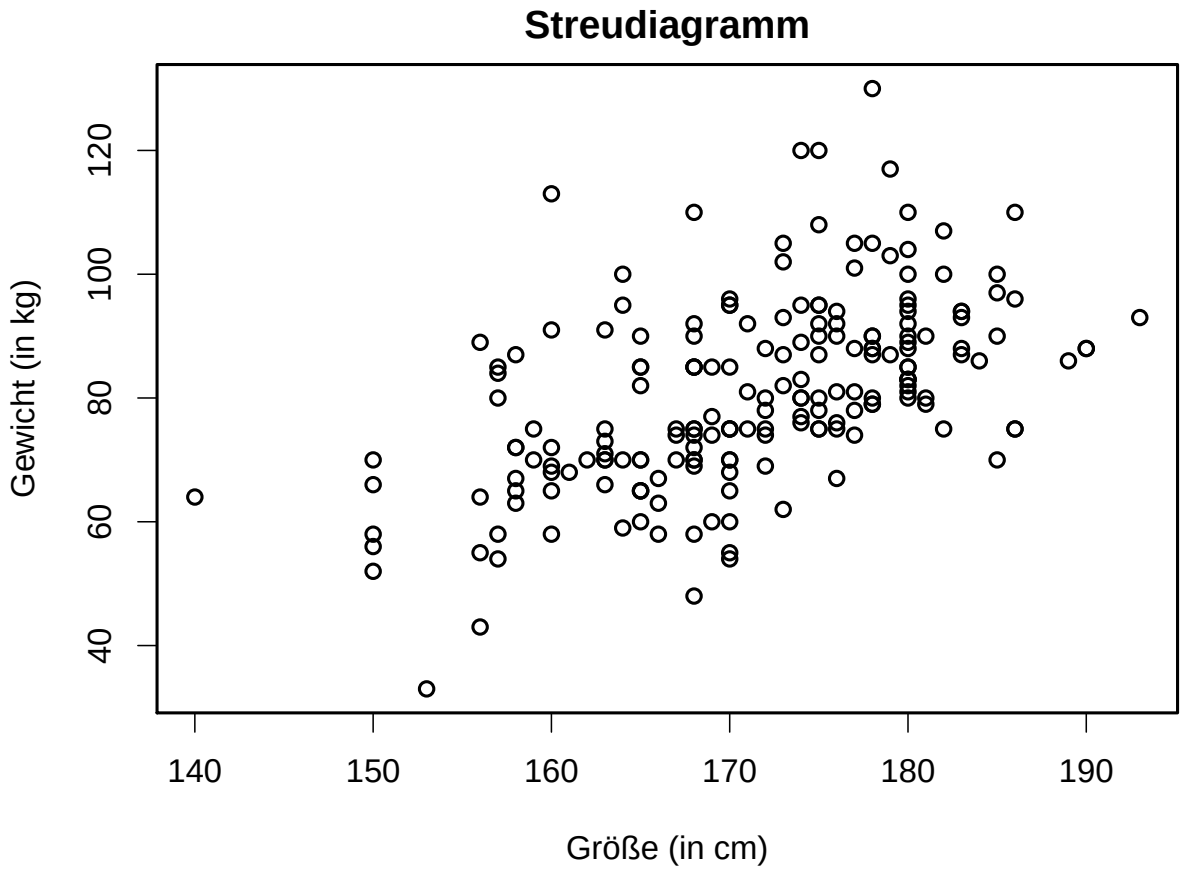


Abb. 35: Streudiagramm (Scatterplot) von Größe und Gewicht von 200 Patienten aus Beispiel 1

4.2.2 Streudiagramme

Interpretation / Idee:

- Offensichtlich steigt „im Mittel“ das Gewicht mit der Körpergröße.
- Sind Sie auch auf die Idee gekommen, dass man zusätzlich das Geschlecht mit einbeziehen sollte?
- Um das zu überprüfen, bietet sich die sogenannte Stratifikation an.

Definition 14 (Stratifikation)

In einem Streudiagramm wird **Stratifikation** *durch Verwendung verschiedener Symbole für die Punkte unterschiedlicher Strata (= Gruppen) realisiert.*

4.2.2 Streudiagramme

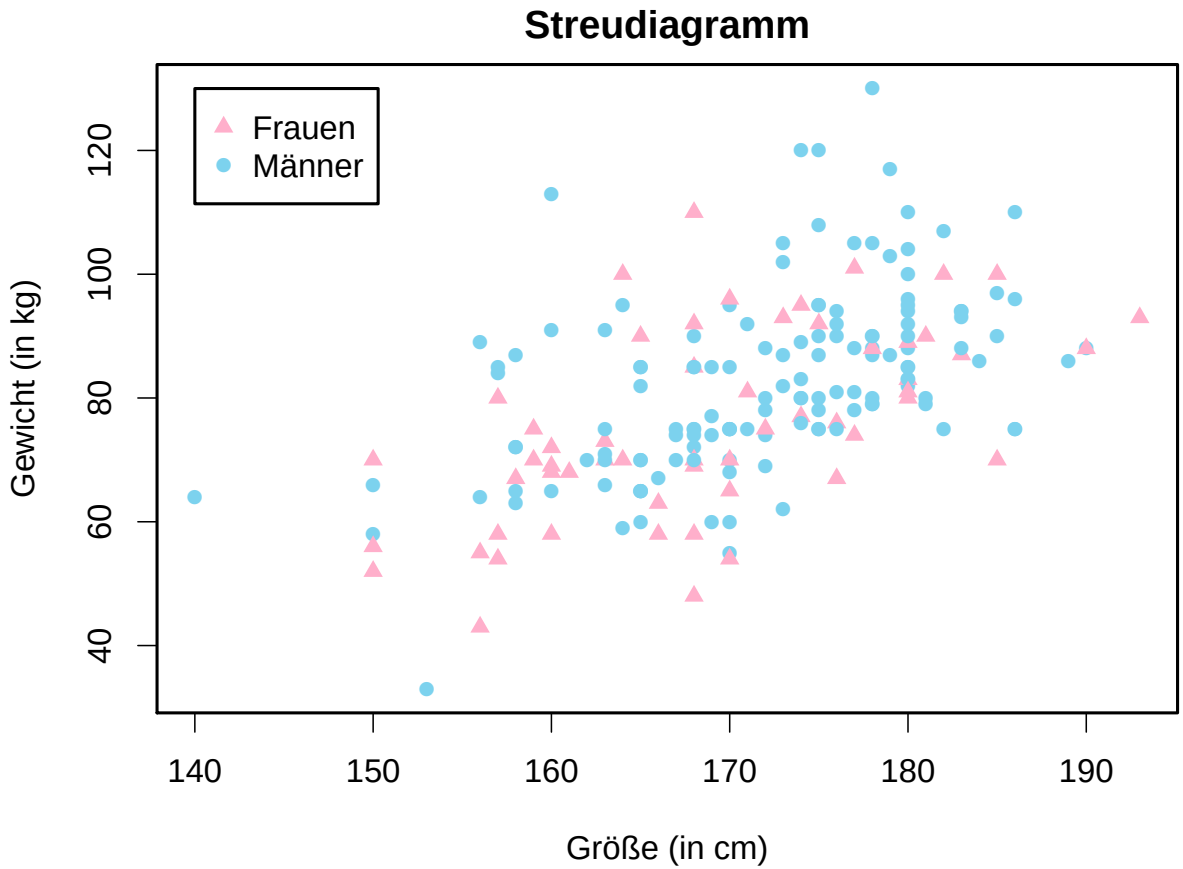


Abb. 36: Streudiagramm (Scatterplot) von Größe und Gewicht von 200 Patienten aus Beispiel 1, stratifiziert nach Geschlecht

4.2.2 Streudiagramme

Interpretation

- Ganz so einfach ist es also doch nicht:
  Zwar liegen Frauen tendenziell etwas unterhalb der Männer, es würde aber wohl niemand ein Vermögen darauf verwetten wollen, dass von den nächsten jeweils 5 Frauen und Männern die Männer größer und schwerer als die Frauen sind.

Bemerkung

- Nachdem wir grafische Darstellungen für bivariate Zusammenhänge kennengelernt haben, wird es im Folgenden darum gehen, die Stärke des Zusammenhangs mit Hilfe von statistischen Maßzahlen zu charakterisieren.

4.3 Kontingenzkoeffizienten – Geschichte

Geschichte

An einem Nachmittag im Jahr 1920 in der Rothamsted Versuchsstation machte der berühmte Statistiker Ronald A. Fisher (1890 – 1962) eine Tasse Tee für Muriel Bristol. Sie protestierte, als er den Tee in die Tasse goss, bevor er die Milch dazu gab und behauptete, dass sie unterscheiden könnte, ob die Milch zuerst oder als zweites dazu gegeben worden sei, und sie würde ersteres vorziehen, ...

4.3 Kontingenzkoeffizienten – Geschichte

- Fisher entwirft Experiment, um Behauptung zu überprüfen
- Muriel muss acht Tassen Tee beurteilen
- Jeweils vier in jeder Reihenfolge (Milch – Tee, Tee – Milch)
- Dargereicht in randomisierter Reihenfolge

Tab. 5: Ergebnis des Experiments von Fisher

Tatsächlich	Beurteilung Muriel	
	Milch-Tee	Tee-Milch
Milch-Tee	3	1
Tee-Milch	1	3

4.3 Kontingenzkoeffizienten – Geschichte

Induktive Statistik: Fisher's Exakter Test

- **Hypothese H_0:** Muriel kann Reihenfolge nicht unterscheiden
- **Teststatistik:** Häufigkeit in der ersten Zelle der Tafel
- **Testentscheidung:** Lehne H_0 ab, falls Teststatistikwert größer als kritischer Wert c. Wähle c so, dass Wahrscheinlichkeit für den Fehler 1. Art (H_0 ablehnen, obwohl wahr) kleiner ist als das vorgegebene Signifikanzniveau α.

Doch bevor Sie in einem späteren Beispiel das Testen erlernen, wollen wir zunächst versuchen, den Zusammenhang zu beschreiben.

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

Idee

- Bei nominalen Merkmalen soll ein Zusammenhangsmaß darüber Auskunft geben, in welchem Maße die Kenntnis des Werts eines Merkmals Information über den Wert des anderen Merkmals beinhaltet.
- Als Grundlage für solche Zusammenhangsmaße bieten sich Kontingenztafeln an.

Tab. 6: Kontingenztafel

Klassen von X	Klassen von Y: y_1	y_2	...	y_L	Zeilen-summen
x_1	H_{11}	H_{12}	...	H_{1L}	Z_1
x_2	H_{21}	H_{22}	...	H_{2L}	Z_2
x_3	H_{31}	H_{32}	...	H_{3L}	Z_3
...			...		...
x_K	H_{K1}	H_{K2}	...	H_{KL}	Z_K
Spalten-summen	S_1	S_2	...	S_L	n

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

Definition 15 (Kontingenzkoeffizient)

Ein Zusammenhangsmaß für 2 nominale Merkmale heißt **Kontingenzkoeffizient**, *wenn es auf einer Kontingenztafel der beiden Merkmale beruht.*
Im Allg. werden dabei die Klassen nur mit einer Merkmalsausprägung besetzt.

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

Beispiel 3 (Abiturart und Bewerbungsergebnis)

Besteht ein Zusammenhang zwischen der Art des Abiturs und der Chance, einen Lehrvertrag als Bankkauffrau/-mann in einem speziellen Unternehmen zu bekommen?
Die Ergebnisse einer Umfrage sind in der folgenden Kontingenztafel zusammengefasst:

Tab. 7: Ergebnisse

		Bewerbungsergebnis			
		angenommen	*warten*	*abgelehnt*	*Summe*
Abitur	*naturwiss.*	*2*	*6*	*4*	*12*
	sprachlich	*4*	*12*	*8*	*24*
	anderes	*1*	*3*	*2*	*6*
	Summe	*7*	*21*	*14*	*42*

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

In Beispiel 3 gibt es keinen Zusammenhang zwischen Abiturart und Bewerbungsergebnis, denn relativ zur Zeilensumme (Anzahl Abschlüsse einer Art) sind alle Zeilen gleich.

Diese Beobachtung führt zu:

Definition 16 ((empirische) Unabhängigkeit)

Zwei Merkmale heißen **(empirisch) unabhängig**, *wenn alle beobachteten Häufigkeiten mit den dazugehörigen Erwartungshäufigkeiten übereinstimmen, d.h.* $H_{ij} = E_{ij}$ *für alle* $i = 1, \ldots, K$, $j = 1, \ldots, L$, *wobei gilt:*

Erwartungshäufigkeit *in Zelle* $(i, j) := E_{ij} := n \frac{S_j}{n} \frac{Z_i}{n}$.

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

Bemerkungen

- Wenn 2 Merkmale **empirisch unabhängig** sind, dann liefert die Kenntnis eines Merkmals keine Information über das Häufigkeitsverhalten des anderen Merkmals, denn

$$
\begin{aligned}
H_{ij} &= E_{ij} = n s_j z_i = s_j Z_i = S_j z_i, \text{ wobei} \\
s_j &:= \frac{S_j}{n} = \text{ Anteil der Werte in Klasse } j \text{ des Merkmals } Y \\
z_i &:= \frac{Z_i}{n} = \text{ Anteil der Werte in Klasse } i \text{ des Merkmals } X,
\end{aligned}
$$

d.h. (z.B.) die Information, dass der Wert von Merkmal X in Klasse i fällt, liefert insofern keine Information über den Wert von Merkmal Y, dass Klasse j von Merkmal Y genau so häufig vorkommen wird, wie diese Klasse insgesamt, d.h. summiert über alle Klassen von Merkmal X.

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

Bei **empirischer Unabhängigkeit**

- sind sämtliche Zeilen gleich, wenn man pro Zeile durch die jeweilige Zeilensumme Z_i teilt, denn es gilt:

$$\frac{E_{ij}}{Z_i} = s_j.$$

  Analoges gilt für Spalten:

$$\frac{E_{ij}}{S_j} = z_i.$$

- hängen die Werte in den Zellen der Kontingenztafel nur von den Rändern der Tafel ab, werden also allein durch die Randverteilungen festgelegt.
- ist die relative Häufigkeit eines Paares (x_i, y_j) gleich dem Produkt der relativen Häufigkeiten der Komponenten:

$$h_{ij} = \frac{H_{ij}}{n} = z_i s_j.$$

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

Beispiel 3: Abiturart und Bewerbungsergebnis

Nach Division durch die jeweiligen Zeilensummen sind alle Zeilen gleich und haben die Werte:

$$s_1 = \frac{7}{42} = \frac{1}{6}, \quad s_2 = \frac{21}{42} = \frac{1}{2}, \quad s_3 = \frac{14}{42} = \frac{1}{3}.$$

4.3.1 Kontingenzkoeffizienten – Unabhängigkeit

Idee

- Die Summe der Abweichungen der Zellenhäufigkeiten von den Erwartungshäufigkeiten kann als Maß für Abhängigkeit und damit für Zusammenhang verwendet werden.
- Alle folgenden Kontingenzkoeffizienten beruhen auf dieser Idee.
- Man verwendet allerdings nicht die Summe der Abweichungen, sondern die *Summe der quadrierten Abweichungen der Zellenhäufigkeiten von den Erwartungshäufigkeiten relativ zu den Erwartungshäufigkeiten*.
- Wegen der Quadrierung werden Abweichungen unterschiedlichen Vorzeichens gleich behandelt.
- Wegen der Division durch die Erwartungshäufigkeit werden Abweichungen bei kleinen Erwartungshäufigkeiten stärker gewichtet.

4.3.1 Kontingenzkoeffizienten – χ^2-Koeffizient

Definition 17 (χ^2-Koeffizient)

$$\chi^2\text{–}Koeffizient := \chi^2 := \sum_{i=1}^{K}\sum_{j=1}^{L}\frac{(H_{ij} - E_{ij})^2}{E_{ij}} \tag{22}$$

4.3.1 Kontingenzkoeffizienten – χ^2-Koeffizient

Bemerkung

Obige Formel für den χ^2-Koeffizienten ist anschaulich, aber unhandlich, da zunächst die Erwartungshäufigkeiten berechnet werden müssen, danach quadrierte Differenzen zu den beobachteten Häufigkeiten, die noch durch die Erwartungshäufigkeiten geteilt werden müssen.

Zur Berechnung verwendet man deshalb die folgende Beziehung:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{K}\sum_{j=1}^{L}\frac{(H_{ij}-\frac{Z_i S_j}{n})^2}{\frac{Z_i S_j}{n}} = n\sum_{i=1}^{K}\sum_{j=1}^{L}\left(\frac{H_{ij}^2}{Z_i S_j}-\frac{2H_{ij}}{n}+\frac{Z_i S_j}{n^2}\right)\\ &= n\left(\sum_{i=1}^{K}\sum_{j=1}^{L}\frac{H_{ij}^2}{Z_i S_j}-1\right), \text{ denn:}\end{aligned}$$

$$\sum_{i=1}^{K}\sum_{j=1}^{L}H_{ij}=\sum_{i=1}^{K}\sum_{j=1}^{L}E_{ij}=n.$$

4.3.1 Kontingenzkoeffizienten – χ^2-Koeffizient

Idee: Anforderungen an einen Kontingenzindex

- Um den χ^2-Koeffizienten als *Kontingenzkoeffizienten* verwenden zu können, normiert man ihn so, dass er nicht nur *nach unten durch 0*, sondern auch *nach oben durch 1* beschränkt ist.
- Kontingenzkoeffizienten sollten
  - bei **empirischer Unabhängigkeit = 0** sein,
  - bei **vollständiger Abhängigkeit = 1**.

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Definition 18 (dichotom)

Ein Merkmal, das nur 2 Merkmalsausprägungen annehmen kann, heißt **dichotom** *(zweiwertig).*

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Idee

Wenn die beiden Merkmale X und Y dichotom sind (alternative Daten), wird die Kontingenztafel zu einer sogenannten **Vierfeldertafel:**

Tab. 8: Vierfeldertafel

	y_1	y_2	Zeilensumme
x_1	a	b	$a+b$
x_2	c	d	$c+d$
Spaltensumme	$a+c$	$b+d$	$n=a+b+c+d$

Für solche Vierfeldertafeln lässt sich der χ^2-Koeffizient relativ einfach nach oben abschätzen. Daraus ergibt sich dann der einfachste Kontingenzkoeffizient.

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Lemma 1

Es gilt:

$$\chi^2 = n \frac{(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \tag{23}$$

Beweis.

$$\begin{aligned}
\chi^2 &= n\left(\frac{a^2}{Z_1 S_1} + \frac{b^2}{Z_1 S_2} + \frac{c^2}{Z_2 S_1} + \frac{d^2}{Z_2 S_2} - 1\right) \\
&= \frac{n(a^2 Z_2 S_2 + b^2 Z_2 S_1 + c^2 Z_1 S_2 + d^2 Z_1 S_1 - Z_1 Z_2 S_1 S_2)}{Z_1 Z_2 S_1 S_2} \\
&= \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}, \text{ wegen} \\
&\quad Z_1 = a + b,\ Z_2 = c + d,\ S_1 = a + c,\ S_2 = b + d.
\end{aligned}$$

□

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Satz 5

Es gilt:

$$0 \leq \chi^2 \leq n. \tag{24}$$

Beweis.

$$\chi^2 = n\frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \max! = n, \text{ wenn}$$

$b = c = 0$ (und $a, d \neq 0$) oder $a = d = 0$ (und $b, c \neq 0$). □

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Motiviert durch die Idee, das ein geeigneter Koeffizient die Form

$$\phi = \pm\sqrt{\frac{\chi^2}{n}}$$

haben könnte, kommen wir zu folgender Definition, die auch dem Vorzeichen noch eine Bedeutung beimisst:

Definition 19 (Phi-Koeffizient)

Der **Phi-Koeffizient** *ist definiert durch:*

$$\phi := \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{25}$$

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Bemerkungen

- Offensichtlich wird in der Definition das negative Vorzeichen verwendet, wenn $ad < bc$ ist, und sonst das positive.
- $|\phi|$ ist in obigem Sinne ein Kontingenzkoeffizient.
- Im Fall $|\phi| = 1$ wird ein Merkmal vollständig durch das andere bestimmt.
- Für ϕ selber gilt: $-1 \leq \phi \leq 1$, wobei
  - $\phi = +1 \Leftrightarrow b = c = 0$ (und $a, d \neq 0$) und
  - $\phi = -1 \Leftrightarrow a = d = 0$ (und $b, c \neq 0$).
- Generell teht ein negatives Vorzeichen von ϕ für einen negativen Zusammenhang, d.h. für die Tendenz, dass Merkmal Y den 2. Wert annimmt, wenn Merkmal X den 1. Wert annimmt.
- Da die Merkmalswerte bei nominalen Merkmalen keine natürliche Reihenfolge haben, macht eine solche Aussage aber wenig Sinn.
- Bei nominalen Merkmalen wird deshalb im Allg. $|\phi|$ als Kontingenzkoeffizient verwendet.

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Beispiel 4 (Lieblingsspielzeug bei Mädchen und Jungen)

Es sollte die Frage untersucht werden, welches Spielzeug für Kinder unterschiedlichen Geschlechts typisch ist. In einem Raum stand eine Kiste mit Puppen und eine Kiste mit Autos. 50 Mädchen und 50 Jungen entnahmen sich „ihr" Spielzeug und es entstand folgendes Ergebnis:

Tab. 9: Ergebnis der Spielzeugentnahme

	Auto	*Puppe*	*Summe*
Jungen	*35*	*15*	*50*
Mädchen	*25*	*25*	*50*
Summe	*60*	*40*	*100*

Damit gilt: $\phi = \dfrac{35 \cdot 25 - 15 \cdot 25}{\sqrt{50 \cdot 50 \cdot 60 \cdot 40}} \approx 0.204.$

4.3.1 Kontingenzkoeffizienten – Phi-Koeffizient

Der Phi-Koeffizient deutet also auf eine „schwache“ Abhängigkeit von Geschlecht und Lieblingsspielzeug hin, was mit der Anschauung der Zahlen übereinstimmt.

4.3.1 Übungsaufgaben

Gegeben sei die folgende Kontigenztafel:

	Erkrankt	Nicht erkrankt	Summe
mit Impfung	6	54	60
ohne Impfung	12	18	30
Summe	18	72	90

Berechnen Sie den Phi-Koeffizienten und interpretieren Sie das Ergebnis.

Welchen Wert erhält man, wenn alle Einträge in der Tabelle mit 10 multipliziert werden?

Welchen Wert erhält man annähernd, wenn auf alle Werte dieselbe sehr große Zahl addiert wird und wie sieht dann der Mosaikplot aus?

4.3.1 Übungsaufgaben

Für den Phi-Koeffizienten erhält man: $\phi = \frac{6 \cdot 18 - 12 \cdot 54}{\sqrt{60 \cdot 30 \cdot 18 \cdot 72}} \approx -0.354$.

Es besteht somit ein schwacher (negativer) Zusammenhang zwischen *Erkrankt* und *mit Impfung* oder gleichwertig dazu ein schwacher (positiver) Zusammenhang zwischen *Erkrankt* und *ohne Impfung*.

Man kann also sagen, dass der Anteil der Erkrankten ohne Impfung größer ist, bei der Interpretation der Kausalität muss man aber vorsichtig sein, hierzu fehlen Hintergrundinformationen.

Wenn man alle Einträge mit einer Konstanten multipliziert, ändert sich der Wert des Phi-Koeffizienten nicht.

Wenn man auf alle Werte dieselbe große Zahl addiert, sind alle vier Einträge annähernd gleich, der Phi-Koeffizient ist nahe bei 0 und der Mosaikplot besteht aus vier etwa gleich großen Quadraten.

4.4 Maßkorrelationskoeffizienten

Idee

- Bei metrischen Merkmalen soll ein Zusammenhangsmaß darüber Auskunft geben, auf welche Weise die Größe des Werts des einen Merkmals die Größe des Werts des anderen Merkmals beeinflusst.
- Als Grundlage für solche Zusammenhangsmaße bieten sich Streudiagramme an.

4.4 Regression – Korrelation

Wie groß ist der Korrelationskoeffizient bei folgenden Streudiagrammen?

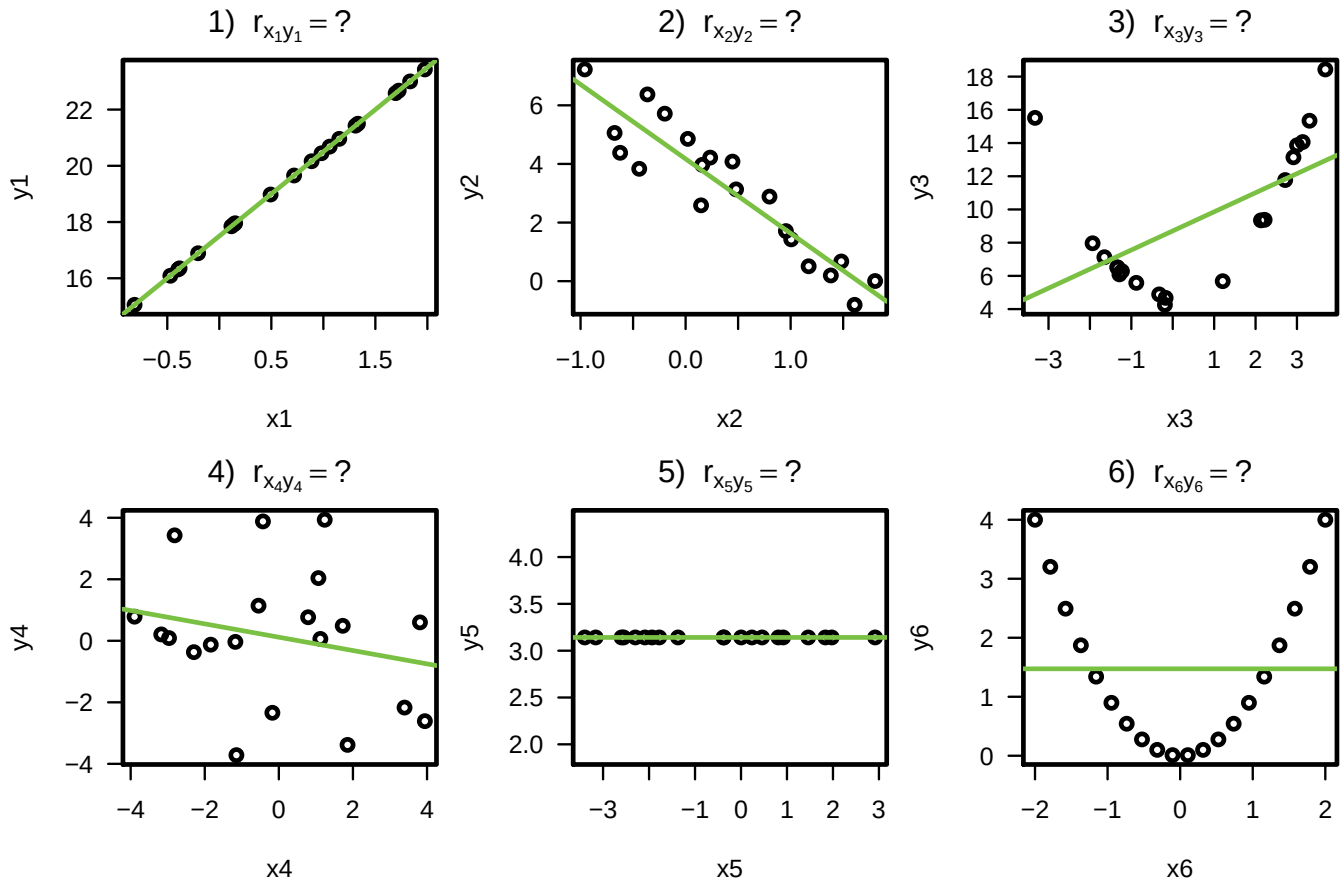


Abb. 37: Quiz: Wie groß ist die jeweilige Korrelation?

4.4 Korrelationskoeffizienten – Bravais-Pearson

(Bravais-Pearson) Korrelationskoeffizient

Idee (W. Krämer (1994): Statistik verstehen, Campus)

- Das verbreitetste Zusammenhangsmaß für metrische Merkmale ist der Korrelationskoeffizient.
- Die zugrundeliegenden Ideen stammen von *Galton* (1888).
- Heute ist der Koeffizient nach seinem Schüler Pearson benannt. Galton argumentierte, dass wir einen Merkmalswert als groß empfinden, wenn er deutlich größer als der Durchschnitt ist, und eine Abweichung vom Mittelwert als umso größer, je weniger die Daten streuen.
- Die Größe eines Merkmalswerts wird also auf den Durchschnittswert bezogen, Abweichungen vom Durchschnitt werden in Anzahl Standardabweichungen gemessen.

4.4 Korrelationskoeffizienten – Bravais-Pearson

Beispiel 5 (Größe und Gewicht von 13 Männern)

Tab. 10: Größe und Gewicht von 13 Männern

	Merkmal	
Träger	*Größe*	*Gewicht*
1	*170*	*60*
2	*172*	*76*
3	*175*	*60*
4	*176*	*75*
5	*177*	*66*
6	*180*	*65*
7	*180*	*78*
8	*183*	*75*
9	*185*	*87*
10	*187*	*72*
11	*188*	*90*
12	*190*	*82*
13	*194*	*92*

4.4 Korrelationskoeffizienten – Bravais-Pearson

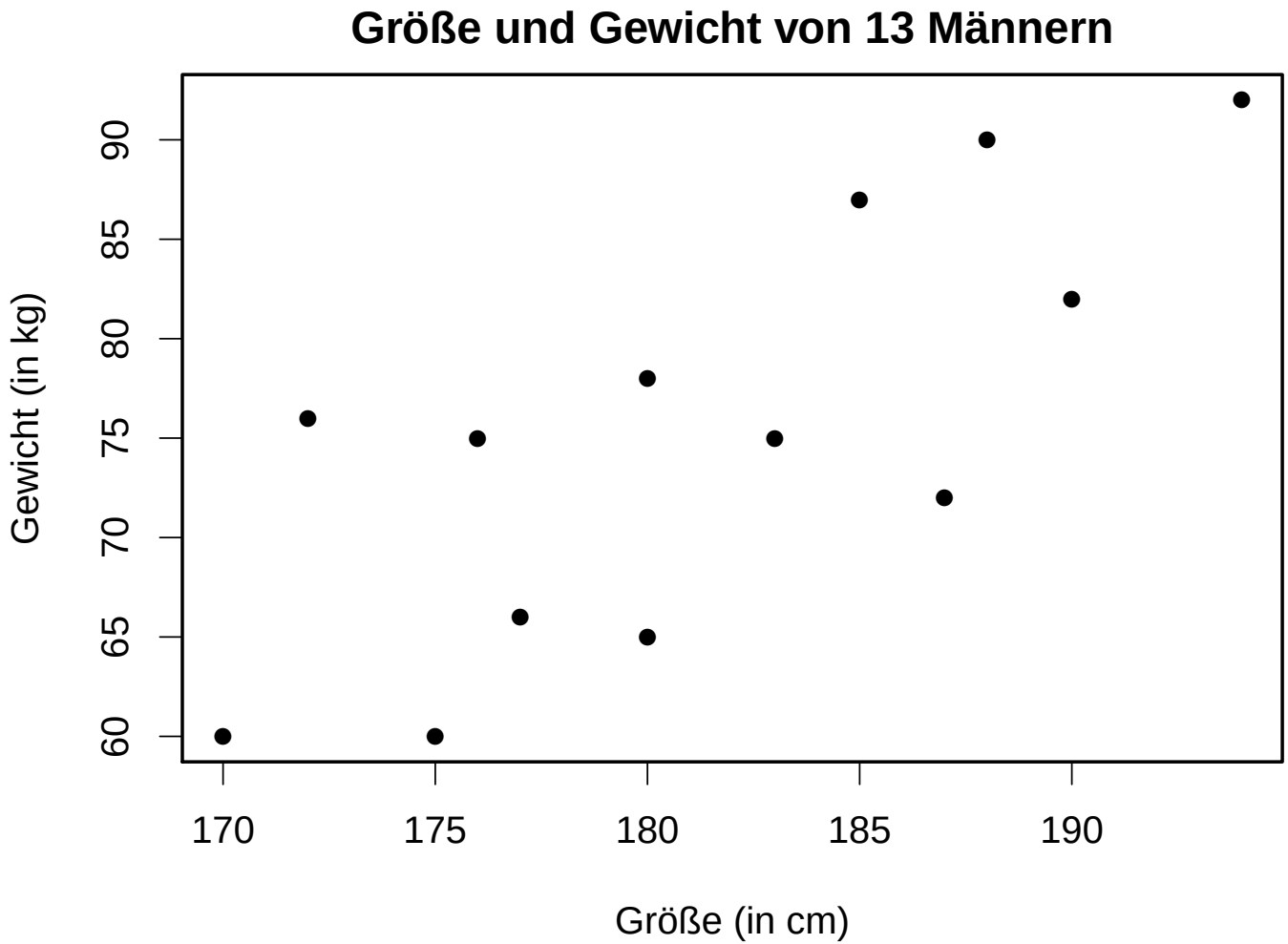


Abb. 38: Größe und Gewicht von 13 Männern

4.4 Korrelationskoeffizienten – Bravais-Pearson

Interpretation

- Wir interessieren uns für den Zusammenhang von Größe und Gewicht von Männern.
- Dazu werden bei 13 Männern mittleren Alters Größe und Gewicht gemessen.
- Nach Galton sprechen wir von einem „großen“ Mann, wenn er „groß in Bezug auf den Durchschnitt (181.3 cm)“ ist, und ein Mann ist „schwer“, wenn er „schwer in Bezug auf den Durchschnitt (75.2 kg)“ ist.
- Abweichungen vom Mittelwert werden gemessen in Anzahl Standardabweichungen, also in den Einheiten $S_{\text{Größe}} = 7.3$ cm und $S_{\text{Gewicht}} = 10.6$ kg.

4.4 Korrelationskoeffizienten – Bravais-Pearson

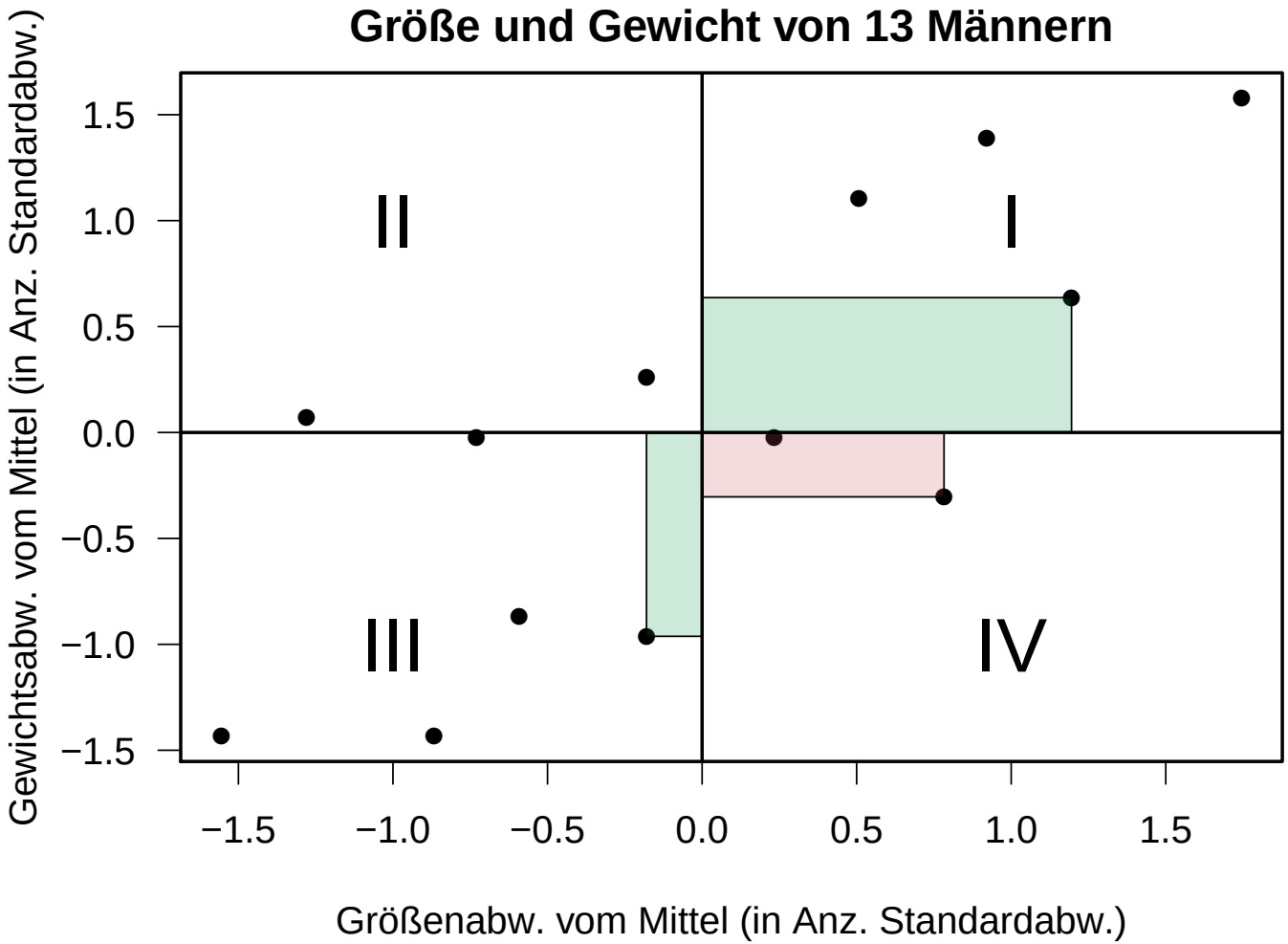


Abb. 39: Größe und Gewicht von 13 Männern – transformiert

4.4 Korrelationskoeffizienten – Bravais-Pearson

Idee

- Wenn die Größe eines Merkmalswerts auf den Durchschnittswert bezogen wird, entspricht das einer *Nullpunktverschiebung in das arithmetische Mittel* der Punkte, wie in Abb. 38 angedeutet.
- Wenn Abweichungen vom Durchschnitt in „Anzahl Standardabweichungen“ gemessen werden, entspricht das der Verwendung neuer Einheiten, wie in Abb. 39 angedeutet.
- Beide Transformationen zusammen entsprechen dem Übergang zu *standardisierten Abweichungen vom Mittelwert:*

$$\tilde{x}_i := \frac{x_i - \bar{x}}{s_x}, \quad i = 1, \ldots, n.$$

4.4 Korrelationskoeffizienten – Bravais-Pearson

- Nummeriert man die Quadranten in dem neuen Koordinatenkreuz wie in Abb. 39, so spricht man von *positiver Korrelation*, wenn sich die Punkte in den Quadranten I und III häufen, und von *negativer Korrelation*, wenn die Punkte sich in den Quadranten II und IV häufen.
- Abweichungen vom Durchschnitt werden proportional zur Fläche des Rechtecks vom Mittelwert zum Streudigramm-Punkt gerechnet.
- Dabei werden die Vorzeichen der beiden Koordinaten eines Punktes insofern berücksichtigt, dass Punkte in den Quadranten I und III als positive Abweichungen gezählt werden (grün in der Skizze) und Punkte in Quadranten II und IV als negative Abweichungen (rot in de Skizze).

4.4 Korrelationskoeffizienten – Bravais-Pearson

Definition 20 ((empirischer) Korrelationskoeffizient)

Der **(empirische) Korrelationskoeffizient** r_{xy} *der Merkmale* X *und* Y *ist definiert als die „mittlere Fläche mit dem Achsenschnittpunkt“:*

$$r_{xy} := \frac{1}{n-1}\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} \tag{26}$$

Bemerkung

Wie bei der Berechnung der Standardabweichungen teilt man nicht durch n, sondern durch $n-1$.

4.4 Korrelationskoeffizienten – Bravais-Pearson

Berechnung des Korrelationskoeffizienten

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

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

Beispiel 5: Größe und Gewicht von 13 Männern

$$r_{xy} = \frac{178026 - 13 \cdot 75.2308 \cdot 181.3077}{\sqrt{(427977 - 13 \cdot 181.3077^2)(74932 - 13 \cdot 75.2308^2)}} = \frac{707.0}{927.8} = 0.76$$

4.4 Korrelationskoeffizienten – Bravais-Pearson

Bemerkung

- Der Korrelationskoeffizient standardisiert eine Größe mit Hilfe der Standardabweichungen der beiden beteiligten Merkmale, die ganz ähnlich wie die Varianzen der Merkmale aufgebaut ist.
- Anstelle der quadrierten Abweichungen $(x_i - \bar{x})^2$ bzw. $(y_i - \bar{y})^2$ werden sogenannte „Kreuzprodukte“ $(x_i - \bar{x})(y_i - \bar{y})$ aufsummiert.
- **Der Korrelationskoeffizient r_{xy} liegt zwischen -1 und +1** (Bew. folgt).
- Z.B. ist $r_{xy} = +1$, wenn $y = +x$ und $r_{xy} = -1$, wenn $y = -x$.
- Allerdings ist r_{xy} **undefiniert**, falls X oder Y nicht variiert (s_x oder $s_y = 0$).

4.4 Korrelationskoeffizienten – Kovarianz

Die statistische Maßzahl im Zähler des Korrelationskoeffizienten hat einen eigenen Namen:

Definition 21 (Kovarianz)

Die (empirische) **Kovarianz** cov_{xy} *der Merkmale X und Y ist definiert als:*

$$\text{cov}_{xy} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{27}$$

Beispiel 5: Größe und Gewicht von 13 Männern

$$\text{cov}_{xy} = \frac{(178026 - 13 \cdot 75.2308 \cdot 181.3077)}{12} = \frac{707.0}{12} = 58.9$$

4.4 Korrelationskoeffizienten – Kovarianz

Bemerkungen

- Die Kovarianz ist im Gegensatz zum Korrelationskoeffizienten nicht dimensionslos.
- Die Größenordnung der Kovarianz hängt von den *Messeinheiten* ab.
- Insofern ist die Kovarianz als Zusammenhangsmaß nicht besonders gut geeignet.

4.4 Korrelation und Unabhängigkeit

Definition 22 ((Empirische) Unkorreliertheit)

Merkmale, deren Korrelationskoeffizient gleich Null ist, heißen **(empirisch) unkorreliert**.

Bemerkung

Aus Unabhängigkeit folgt also Unkorreliertheit, aber nicht umgekehrt!

4.4 Korrelationskoeffizienten für ordinale Daten

Rangkorrelationskoeffizienten

Idee

- Korrelationskoeffizienten für 2 ordinale Merkmale beruhen auf **Rangzahlen**.
- Anforderungen an Rangkorrelationskoeffizienten r:
  - i) r ist normiert, d.h. $-1 \leq r \leq 1$
  - ii) $r = 1$ bei gleicher Rangordnung, d.h. falls $r_{xi} = r_{yi}$, $i = 1, \ldots, n$
  - iii) $r = -1$ bei inverser Rangordnung, d.h. falls $r_{xi} = n + 1 - r_{yi}$, $i = 1, \ldots, n$
  - iv) $r = 0$ bei empirischer Unabhängigkeit.

4.4 Korrelationskoeffizienten für ordinale Daten

Spearman'scher Rangkorrelationskoeffizient

Idee

- Der Bravais-Pearson-Maßkorrelationskoeffizient hat eine direkte Entsprechung für ordinale Merkmale.
- Dabei wird statt mit den Merkmalswerten mit den Rangzahlen gerechnet.
- Wir verwenden bei der Definition gleich die Berechnungsformel.

Definition 23 (Spearman'scher Rangkorrelationskoeffizient)

Der **Spearman'sche Rangkorrelationskoeffizient** *ist definiert durch:*

$$r_s := \frac{\sum_{i=1}^{n} r_{xi} r_{yi} - n\bar{r}_x \bar{r}_y}{\sqrt{\left(\sum_{i=1}^{n} r_{xi}^2 - n\bar{r}_x^2\right)\left(\sum_{i=1}^{n} r_{yi}^2 - n\bar{r}_y^2\right)}} \tag{28}$$

4.4 Korrelationskoeffizienten für ordinale Daten

Satz 6 (Spearman Korrelation ohne Bindungen)

Falls keine Rangzahl mehrfach auftritt, gilt für den Spearman-Korrelationskoeffizienten:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{29}$$

wobei $d_i := r_{xi} - r_{yi}$, $i = 1, \ldots, n$.

Beweis.

ohne Beweis, siehe Literatur □

4.4 Korrelationskoeffizienten für ordinale Daten

Bemerkung

Obige *Anforderungen an Rangkorrelationskoeffizienten* sind zumindest im dem Fall, dass keine Rangzahl mehrfach auftritt, bei Spearman erfüllt:

i) Normierung: $-1 \leq r_S \leq 1$ ist klar wegen Maßkorrelation.

ii) $r_S = 1$ bei gleicher Rangordnung, d.h. falls $r_{xi} = r_{yi}$, $i = 1, \ldots, n$; klar, da alle $d_i = 0$

iii) $r_S = -1$ bei inverser Rangordnung, d.h. falls $r_{xi} = n + 1 - r_{yi}$, $i = 1, \ldots, n$; (hier ohne Beweis, siehe Literatur).

iv) $r_S = 0$ bei empirischer Unabhängigkeit
Falls keine Rangzahl mehrfach auftritt, sind die Merkmale X und Y niemals (empirisch) unabhängig (ohne Bew.), sie können aber empirisch unkorreliert sein (vgl. Abschnitt 4).

4.4 Korrelationskoeffizienten für ordinale Daten

Bemerkung

Der Spearman'sche Rangkorrelationskoeffizient nimmt immer dann den *Wert 1* an, wenn zwischen den Merkmalen X und Y eine monoton wachsende Beziehung besteht, d.h. wenn für alle (x_i, y_i), (x_j, y_j) gilt: wenn $x_i < x_j$ ist, dann ist auch $y_i < y_j$.

4.4 Korrelationskoeffizienten für ordinale Daten

Beispiel 6 (Leistung und soziale Position in der Gruppe)

- *Es soll untersucht werden, ob ein Zusammenhang zwischen der schulischen Leistung und der Position in einer Gruppe besteht.*
- *Dazu wurden die 6 Mitglieder der Gruppe sowohl bzgl. ihrer Leistung, als auch bzgl. ihrer Stellung in der Gruppe rangiert:*

Tab. 11: Leistung und soziale Position in der Gruppe

Name	$r_{Leistung}$	$r_{Sympathie}$	$\|d\|$	d^2
Rainer	*1*	*2*	*1*	*1*
Horst	*2*	*3*	*1*	*1*
Klaus	*3*	*1*	*2*	*4*
Mario	*4*	*4*	*0*	*0*
Peter	*5*	*6*	*1*	*1*
Tilo	*6*	*5*	*1*	*1*

4.4 Korrelationskoeffizienten für ordinale Daten

Daraus ergibt sich der Spearman'sche Rangkorrelationskoeffizient als

$$r_S = 1 - \frac{6 \cdot 8}{6 \cdot 35} = 0.77.$$

Der interessierende Zusammenhang ist also nicht abzustreiten.

4.5 Regression

Korrelation und Linearität

Idee

- Der Korrelationskoeffizient ist auch deshalb so beliebt, weil er ein *Maß für die Linearität eines Zusammenhangs* darstellt.
- Es gilt $r_{xy} = \pm 1$, genau wenn die Punkte (x_i, y_i) auf einer Geraden liegen, und es gilt $r_{xy} = 0$, wenn keine lineare Beziehung besteht.
- Um den Grad der Linearität eines Zusammenhangs quantifizieren zu können, ist es zunächst notwendig, sich auf ein Optimalitätskriterium zu einigen, nach dem man eine „optimal an die Punkte angepasste Gerade“ bestimmt.
- Das beliebteste Kriterium ist das Prinzip der Kleinsten Quadrate, nach dem die Gerade so bestimmt wird, dass die Quadratsumme derjenigen Abstände der Punkte von der Geraden minimal werden, die senkrecht zu der x-Achse gemessen werden.

4.5 Regression

Beispiel 5: Größe und Gewicht von 13 Männern

Tab. 12: Größe und Gewicht von 13 Männern

	Merkmal		Gerade
Träger	x=Größe	y=Gewicht	y_a
1	170	60	62.7
2	172	76	64.9
3	175	60	68.2
4	176	75	69.3
5	177	66	70.4
6	180	65	73.7
7	180	78	73.7
8	183	75	77.0
9	185	87	79.2
10	187	72	81.4
11	188	90	82.6
12	190	82	84.8
13	194	92	89.2

4.5 Regression

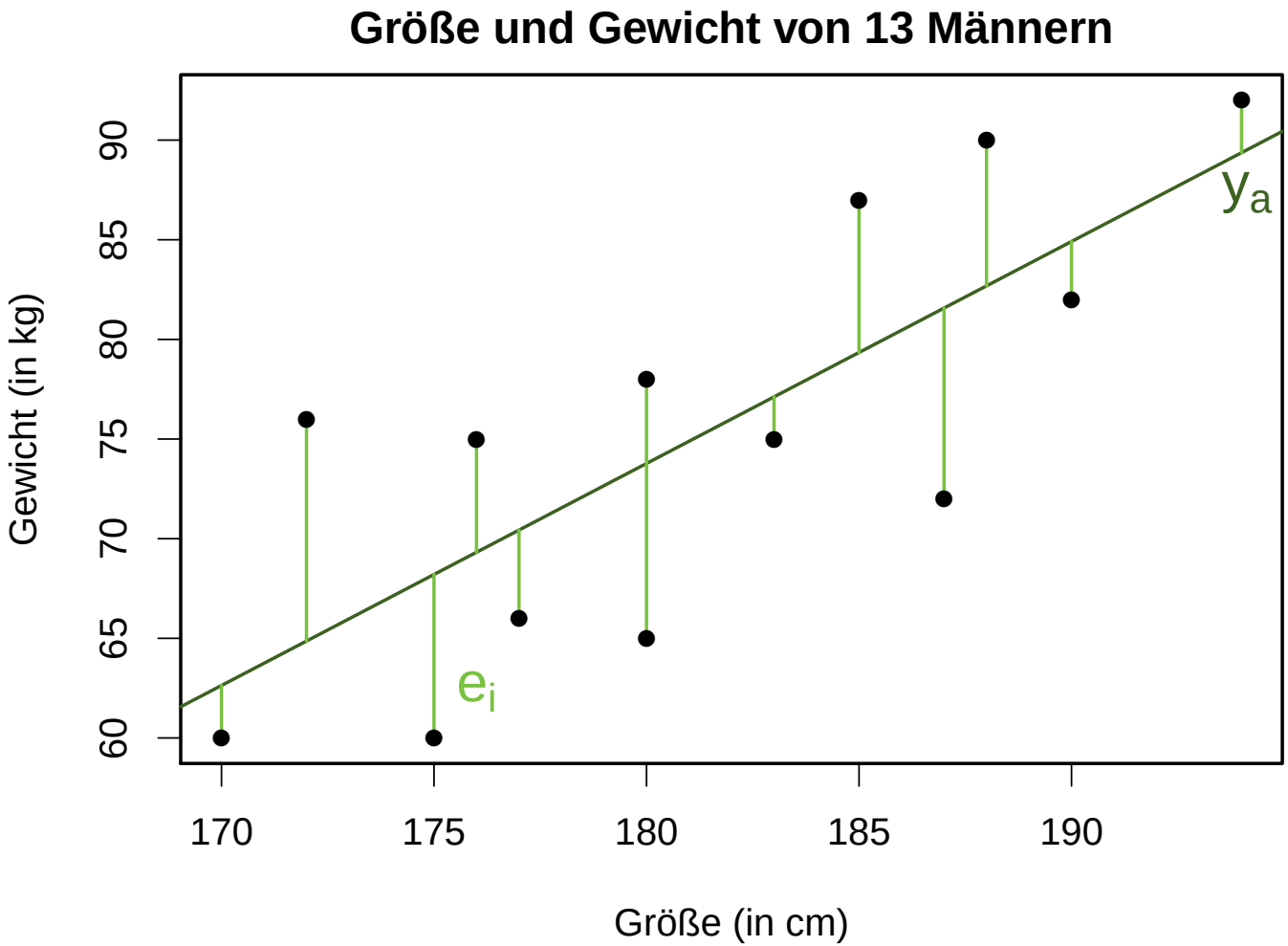


Abb. 40: Größe und Gewicht von 13 Männern

4.5 Regression

Interpretation

- Zu jedem Punkt (x_i, y_i) und jeder Gerade mit Achsenabschnitt b_0 und Steigung b_1 wird der Abstand $e_i := y_i - b_0 - b_1 x_i$ senkrecht zur x-Achse berechnet (e = error).
- Bei der optimalen Gerade ist $S = e_1^2 + \ldots + e_n^2$ minimal.
- Die Kleinste-Quadrate-Gerade hat hier die Form: $y_a = -126.73 + 1.114 \cdot x$ (y_a steht für „angepasster“ y-Wert).
- Wie wird eine solche Gerade konstruiert?

4.5 Regression

Definition 24 (Definitionen zur Regression)

- *Die Bestimmung einer optimal angepassten Gerade nach dem Prinzip der Kleinsten Quadrate heißt* **Regression** *oder* **Methode der kleinsten Quadrate**.
- *Dabei werden Achsenabschnitt b_0 und Steigung b_1 berechnet, so dass die* **Fehlerquadratsumme**

$$S(b_0, b_1) = \sum_{i=1}^{n} e_i^2(b_0, b_1) := \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{30}$$

  minimal wird.
- *Die so bestimmte Gerade heißt* **Regressionsgerade**.
- $y_{ai} := b_0 + b_1 x_i$ *heißt* **Anpassung (Fit) von** y_i
- $e_i := y_i - y_{ai}$ *heißt* **Residuum**, *jeweils für die Beobachtung* $i = 1, \ldots, n$.

4.5 Regression

Satz 7

Für Achsenabschnitt b_0, Steigung b_1 und minimale Fehlerquadratsumme $S_{\min}$ der Regressionsgerade gilt:

$$\begin{aligned} b_1 &= \frac{\text{cov}_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}, \\ b_0 &= \bar{y} - b_1 \bar{x}, \\ S_{\min} &:= (n-1)(1-r_{xy}^2)s_y^2. \end{aligned}$$

Beweis zu Satz 7

Wir bestimmen b_0 und b_1 durch Nullsetzen der 1. Ableitungen

$$\frac{\partial S(b_0,b_1)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

Daraus folgt: $\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0$, also: $b_0 = \bar{y} - b_1\bar{x}$.

$$\frac{\partial S(b_0,b_1)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i = 0$$

4.5 Regression

Daher: $\sum_{i=1}^{n} y_i x_i - n(\bar{y} - b_1\bar{x})\bar{x} - b_1 \sum_{i=1}^{n} x_i^2 = 0$, also:

$$b_1 = \frac{\left(\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}\right)}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$= \frac{cov_{xy}}{s_x^2} = \frac{r_{xy} s_x s_y}{s_x^2} = r_{xy}\frac{s_y}{s_x}.$$

Für die minimale Fehlerquadratsumme gilt:

$$S_{min} = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n}((y_i - \bar{y}) - b_1(x_i - \bar{x}))^2$$
$$= (n-1)(s_y^2 - 2b_1 \,\mathrm{cov}_{xy} + b_1^2 s_x^2) = (n-1)(s_y^2 - b_1^2 s_x^2), \text{ denn } \mathrm{cov}_{xy} = b_1 s_x^2$$
$$= (n-1)(1 - r_{xy}^2)s_y^2$$

4.5 Regression – Korrelation

Idee

- Als Nächstes ist eine statistische Maßzahl gesucht, die die Güte der Anpassung der optimalen Gerade an die Daten beschreibt.
- Tatsächlich kennen wir mit dem Korrelationskoeffizienten schon eine solche Maßzahl.
- Es ist jedoch Konvention, die *Anpassungsgüte* mit dem Quadrat des Korrelationskoeffizienten zu messen.

Definition 25 (Bestimmtheitsmaß)

Das **Bestimmtheitsmaß** *(der Daten durch die Regressionsgerade) ist definiert durch* $R^2 := r_{xy}^2$

4.5 Regression – Korrelation

Beispiel 5: Größe und Gewicht von 13 Männern

$$
\begin{aligned}
b_1 &= \frac{\text{cov}_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x} = 0.76\frac{10.6}{7.4} = 1.104, \\
b_0 &= \bar{y} - b_1\bar{x} = 75.2 - 1.104 \cdot 181.3 = -125.0, \\
y_a &= -125.0 + 1.104 \cdot x, \\
R^2 &= r_{xy}^2 = 0.76^2 = 0.58.
\end{aligned}
$$

y_a wurde in der letzten Tabelle mit aufgeführt und in der dazugehörigen Skizze eingezeichnet.

4.5 Regression – Korrelation

Folgerung

Der Korrelationskoeffizient ist genau dann $+1$ bzw. -1, wenn die Merkmalsausprägungen auf einer Gerade mit positiver bzw. negativer Steigung liegen.

4.5 Regression – Korrelation

Wie groß ist der Korrelationskoeffizient bei folgenden Streudiagrammen?

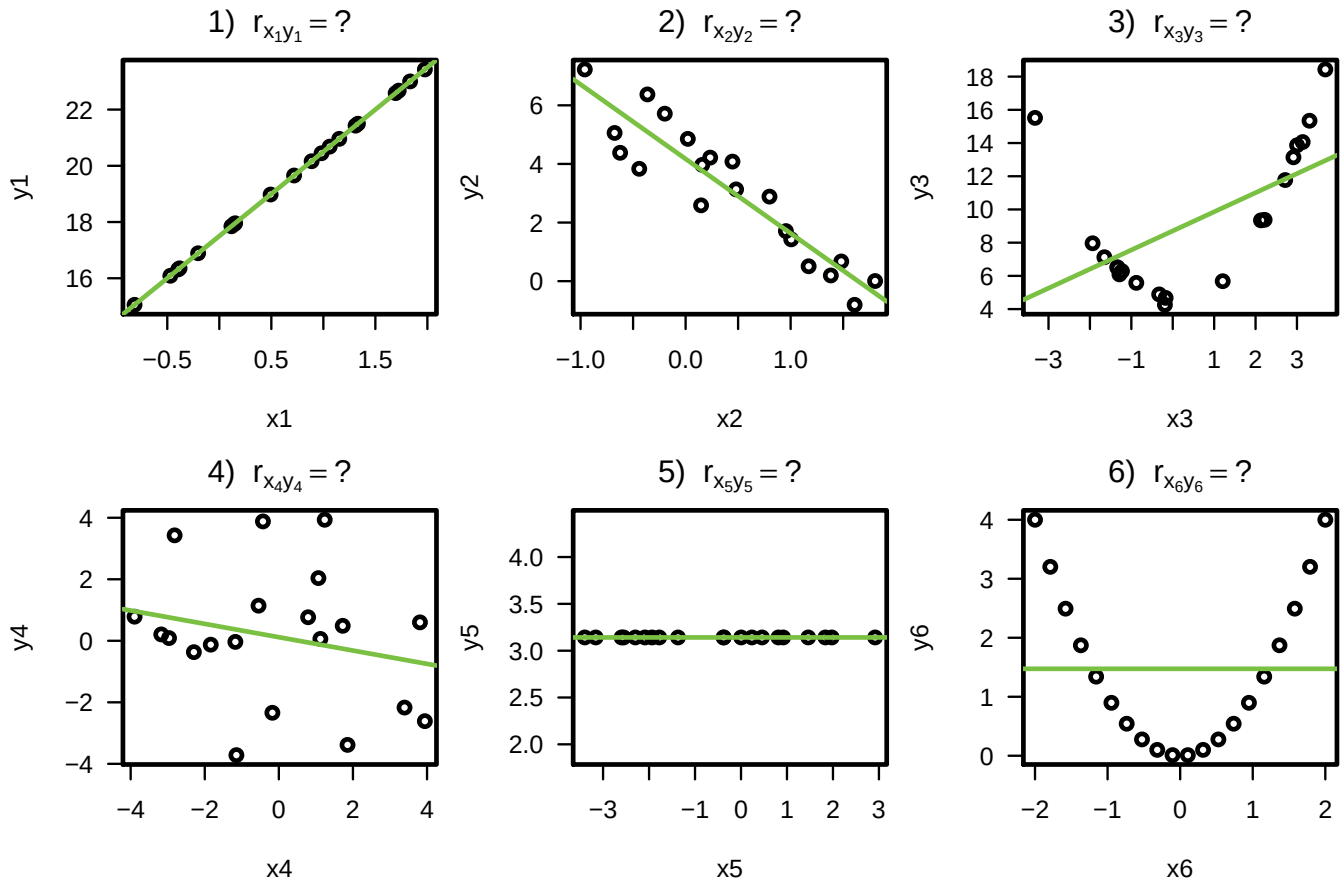


Abb. 41: Quiz: Wie groß ist die jeweilige Korrelation?

4.5 Regression – Korrelation

Bemerkungen

- Die „wahren“ Korrelationskoeffizienten werden in der Vorlesung angegeben.
- Der Korrelationskoeffizient gibt den Grad des linearen Zusammenhangs an.
- Ganz und gar *nicht-lineare Zusammenhänge* können denselben Korrelationskoeffizienten haben wie ein nahezu linearer.
- Die folgenden 4 Streudiagramme haben denselben Korrelationskoeffizienten von $r_{xy} = 0.82$.

4.5 Regression – Korrelation

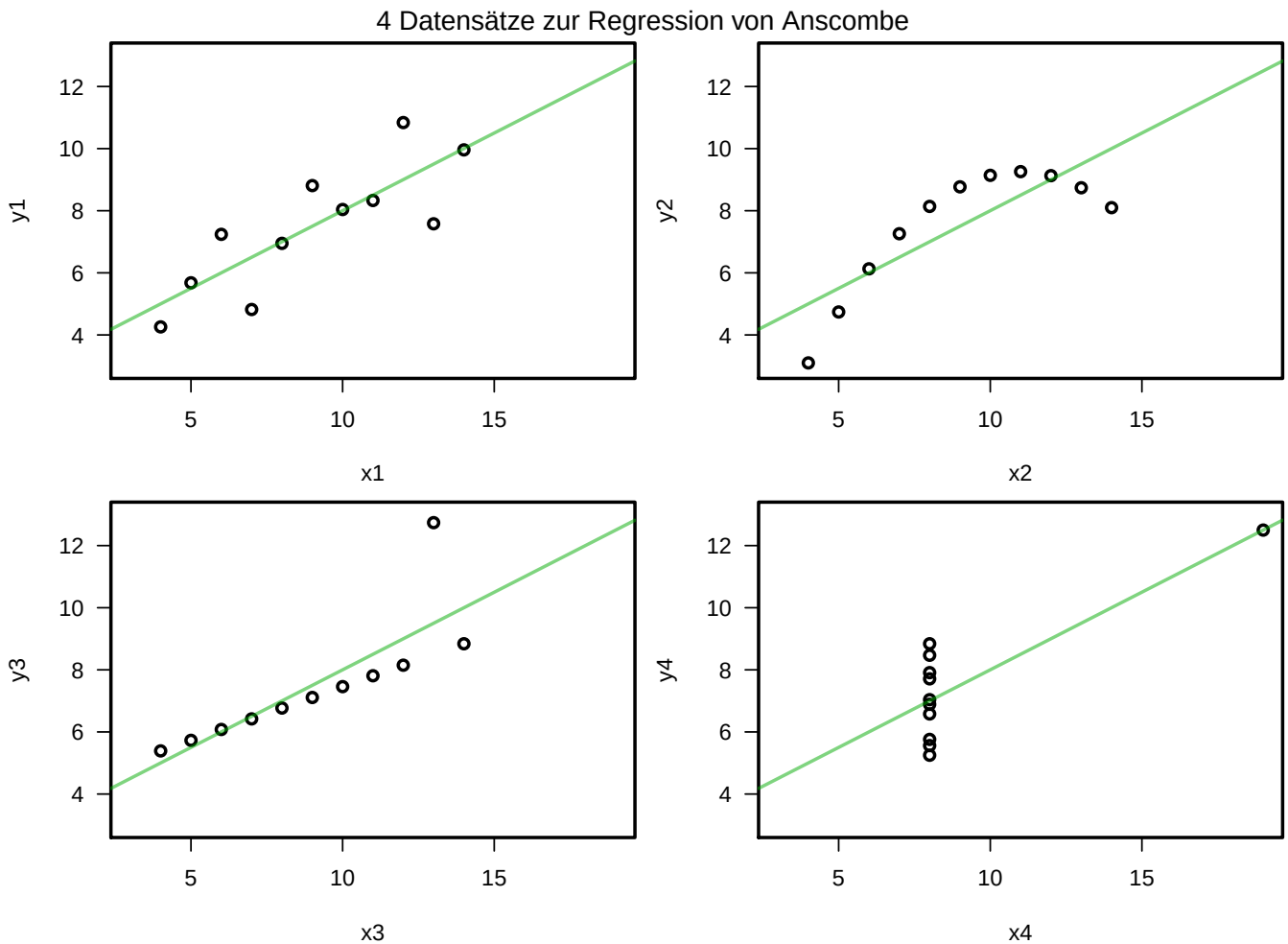


Abb. 42: Beispiel: 4 Mal $r_{xy} = 0.82$; aus: Anscombe (1973): „Graphs in statistical analysis“, American Statistician, 27, p. 17–21

4.5 Regression – Korrelation

Vorteile von Streudiagrammen

- Bei der Darstellung des Zusammenhangs zwischen 2 Merkmalen durch ein Streudiagramm gibt es, im Gegensatz zur Verwendung des Korrelationskoeffizienten, *keinen Informationsverlust* – im Prinzip (s.u.)!
- Sämtliche Information ist ablesbar, die Art des Zusammenhangs, Ausreißer, auch nicht-lineare Zusammenhänge.
- Dagegen misst der Korrelationskoeffizient nur den Grad des *linearen* Zusammenhangs!

4.5 Regression – Korrelation

Probleme von Streudiagrammen

- Mehrere Punkte können an derselben Stelle liegen und verdecken sich gegenseitig.
- Extrem viele Punkte verursachen einen schwarzen Punktehaufen, in dem man wichtige zugrundeliegende Strukturen nicht mehr sieht.

Als Lösung empfiehlt sich in beiden Fällen der Einsatz von

- Transparenz (jeder Punkt mit einigen Prozent Deckung: „je mehr Punkte desto schwärzer“),
- zufällige Auswahl und Darstellung einer kleineren Stichprobe,
- Jitter (absichtliches Hinzufügen einer kleinen Streuung oder damit nicht alle Datenpunkte an exakt einer Stelle liegen)
- Dichteschätzung und zeichnen der Höhenlinien.

4.5 Regression – Korrelation

Wir stellen diastolischen und systolischen gemessenen Blutdruck dar:

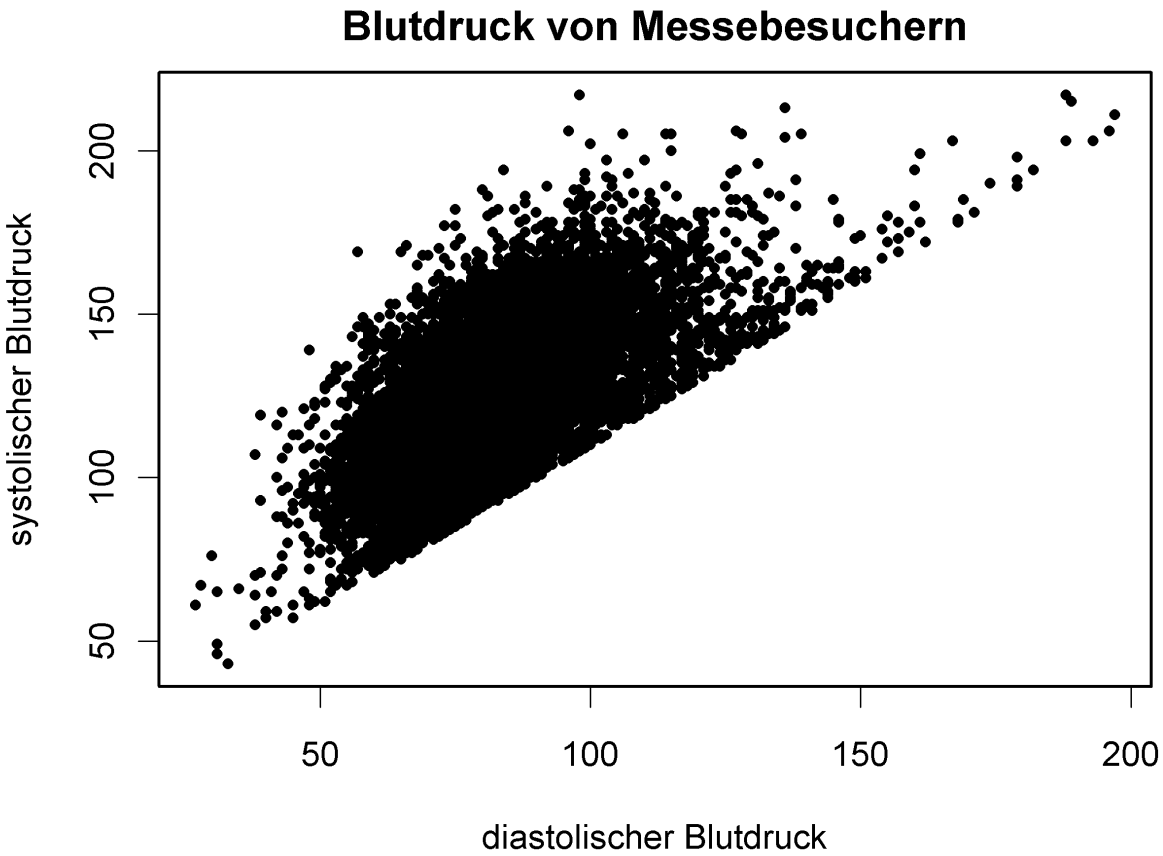


Abb. 43: Diastolischer und systolischer Blutdruck

4.5 Regression – Korrelation

... und jetzt nochmal mit Transparenz::

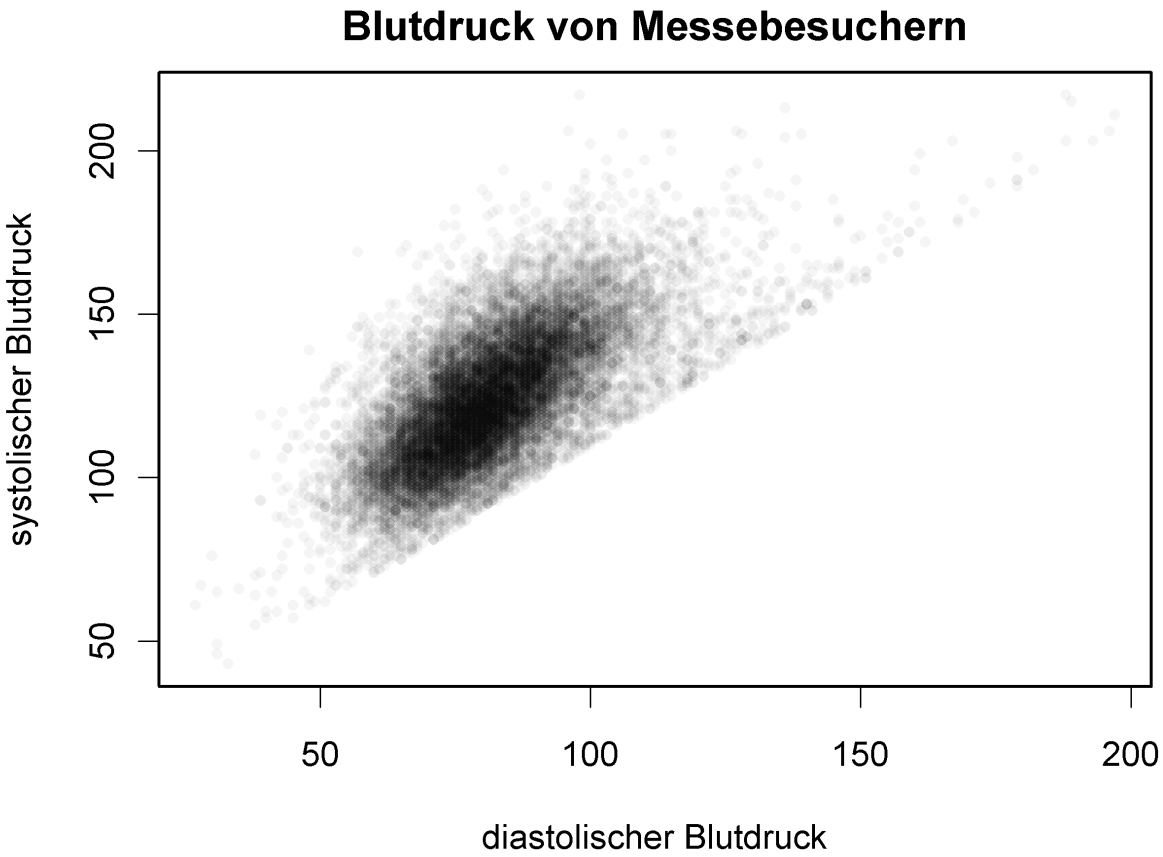


Abb. 44: Diastolischer und systolischer Blutdruck

4.5 Regression – Korrelation

Ein weiteres Extrembeispiel zur Transparenz mit simulierten Daten (es versteckt sich eine Gruppe von Beobachtungen mit perfektem Zusammenhang in allgemeinen Chaos):

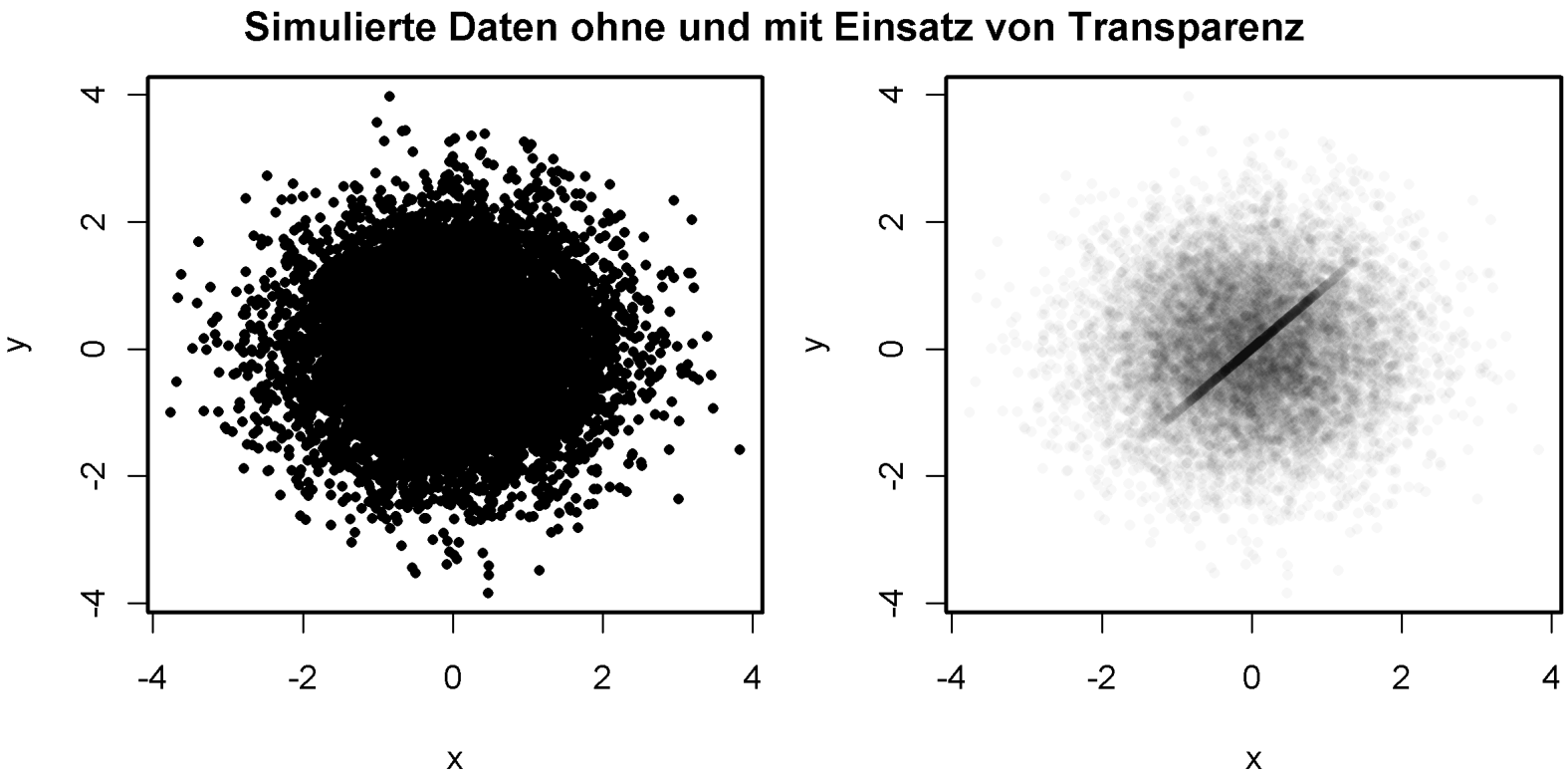


Abb. 45: Vergleich der Darstellungen ohne und mit Einsatz von Transparenz

4.5 Regression – Korrelation

In diesem Beispiel zu Motorendaten (Ethanolgemisch und Kompression) sieht man die Nützlichkeit von zusätzlicher Streuung:

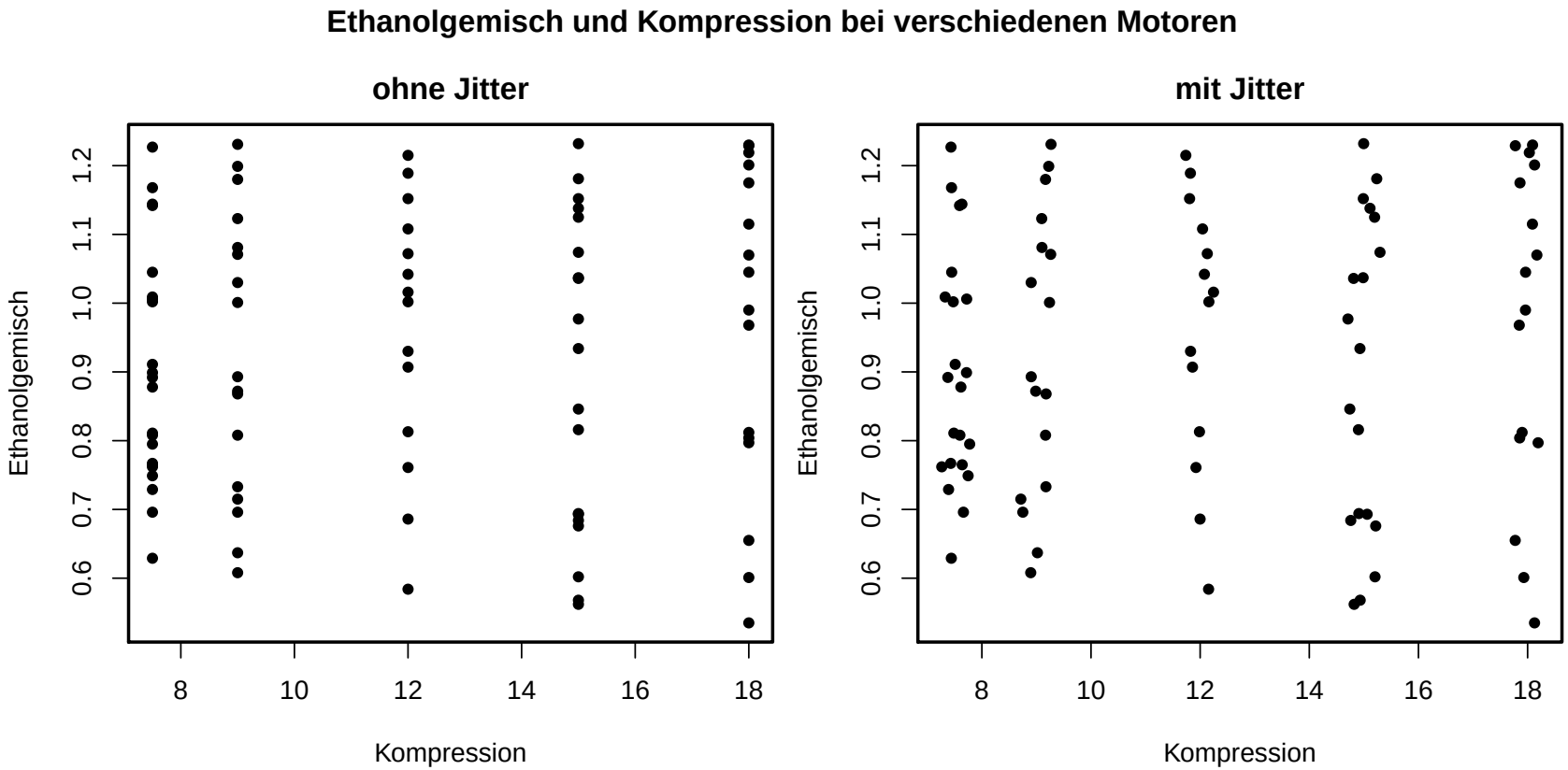


Abb. 46: Vergleich der Darstellungen ohne und mit Einsatz von Jitter

4.5 Regression – Korrelation

Bemerkungen

- Einsatz von Jitter macht Sinn, wenn es nicht allzuviele Beobachtungen gibt, die allerdings gehäuft in einzelnen Punkten auftreten.
  Man beachte, dass künstlich eine Streuung hinzugefügt wird, die eigentlich nicht vorhanden ist. So wird leicht der Eindruck in Richtung der Jitterstreuung verfälscht.
- Transparenz empfiehlt sich bei sehr vielen Beobachtungen, um Strukturen aufzudecken.

4.5 Regression – Korrelation

Achtung: Interpretation von Streudiagrammen

- Ein Streudiagramm oder eine Korrelation sagt nichts über die Interpretation des Zusammenhangs!
- Nicht selten sind sogenannte Scheinkorrelationen.
- Das sind Korrelationen, die entweder durch Zufall oder deswegen zustande kommen, weil eine sogenannte Hintergrundvariable für beide beobachteten Merkmale verantwortlich ist.

4.5 Regression – Korrelation

Beispiele für zufällige Korrelationen:

- in den 60er/70er Jahren die negative Korrelation zwischen Rocklänge und Dow-Jones Index,
- nach dem 1. Weltkrieg die positive Korrelation zwischen der Anzahl Störche und der Anzahl Geburten.

Beispiele für das Wirken einer Hintergrundvariable:

- Größe von Geschwistern (gleiche Eltern!),
- Wasserstand von Rhein und Donau (Regen, Schneeschmelze !) und evtl. auch
- die gleichzeitige Abnahme der Anzahl Störche und der Anzahl Geburten (Industrialisierung nach dem 1. Weltkrieg).

4.5 Regression – Korrelation

Achtung: Kausalitätsrichtung

Der Korrelationskoeffizient sagt über Kausalität oder Kausalitätsrichtung nichts aus! Insbesondere bei der Kausalitätsrichtung neigt man häufig zu Trugschlüssen.

Beispiel:

- Bei einem Naturvolk wurde eine negative Korrelation zwischen Kopfläusen und Fieber festgestellt.
- Natürlich wäre es falsch, daraus abzuleiten, dass Läuse gut sind für die Gesundheit sind. Tatsächlich vertreibt Fieber die Läuse!

4.5 Zusammenfassung

Zusammenfassung für bivariate Merkmale: Wann kann welche Methode verwendet werden?

Tab. 13: Welche Methode für welche Daten?

Methode	metrisch	ordinal	nominal
grafische Darstellungen			
Kontingenztafel	(✓)	✓	✓
Mosaikplot		✓	✓
Streudiagramm	✓		
Parallele Boxplots	✓	✓	
Zusammenhangsmaße			
Kontingenzkoeffizient		✓	✓
Phi-Koeffizient		✓	✓
Rangkorrelationskoeffizient	✓	✓	
Korrelationskoeffizient	✓		
Zusammenhangsbeschreibungen			
Regression	✓		

4.6 Übungsaufgaben

Gegeben seien drei Beobachtungen eines Datensatzes mit zwei Variablen X und Y:

$$x_1 = 4,\ x_2 = 1,\ x_3 = 1, \qquad y_1 = -1,\ y_2 = 0,\ y_3 = 1.$$

- Berechnen Sie für die beiden Variablen die Varianz und die Standardabweichung.
- Berechnen Sie für die beiden Variablen den Korrelationskoeffizienten nach Bravais-Pearson.
- Berechnen Sie die Regressionsparameter des linearen Modells $y = c + d\,x$, bei dem also Y durch X vorhergesagt wird.

4.6 Übungsaufgaben

$$
\begin{aligned}
x_1 = 4, \quad x_2 = 1, \quad x_3 = 1 \quad &\Rightarrow \quad \bar{x} = 2 \\
y_1 = -1, \quad y_2 = 0, \quad y_3 = 1 \quad &\Rightarrow \quad \bar{y} = 0
\end{aligned}
$$

$$
\begin{aligned}
s_x^2 &= \frac{(4-2)^2 + (1-2)^2 + (1-2)^2}{2} = \frac{4+1+1}{2} = 3 \quad \Rightarrow \quad s_x = \sqrt{3} \\
s_y^2 &= \frac{(-1-0)^2 + (0-0)^2 + (1-0)^2}{2} = \frac{1+1}{2} = 1 \quad \Rightarrow \quad s_y = 1
\end{aligned}
$$

$$
\begin{aligned}
s_{xy} &= \frac{(4-2)(-1-0) + (1-2)(0-0) + (1-2)(1-0)}{2} = -\frac{3}{2} \\
r_{xy} &= \frac{s_{xy}}{s_x \, s_y} = \frac{-3}{2\sqrt{3}} = -\frac{\sqrt{3}}{2} \approx -0.866
\end{aligned}
$$

4.6 Übungsaufgaben

$$x_1 = 4, \quad x_2 = 1, \quad x_3 = 1 \quad \Rightarrow \quad \bar{x} = 2$$
$$y_1 = -1, \quad y_2 = 0, \quad y_3 = 1 \quad \Rightarrow \quad \bar{y} = 0$$

$$s_x^2 = 3, \quad s_y^2 = 1, \quad s_{xy} = -\frac{3}{2}$$

$$\begin{aligned} \mathrm{d} &= \frac{s_{xy}}{s_x^2} = -\frac{3}{2 \cdot 3} = -\frac{1}{2} \\ \mathrm{c} &= \bar{y} - \mathrm{d}\,\bar{x} = 0 - \left(-\frac{1}{2}\right) \cdot 2 = 1 \\ \Rightarrow \quad y &= c + d\,x = 1 - \frac{1}{2}x \end{aligned}$$

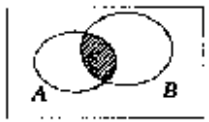
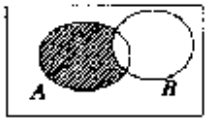
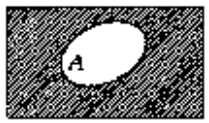
5.1 Wahrscheinlichkeitstheorie

Nach der deskriptiven Statistik werden wir uns nun der induktiven Statistik zuwenden, um mit Hilfe von Wahrscheinlichkeiten und deren Verteilungen Schlussfolgerungen ziehen zu können.

Beim Aufbau des Modells für die Wahrscheinlichkeitsrechnung wird die Mengenlehre verwendet. Die folgenden Folien dienen der Erinnerung an Schulmathematik!

5.1 Exkurs: Mengenlehre

Gegeben 2 Ereignisse A, B, dann werden die folgenden Mengenoperationen benötigt:

Bezeichnung	Sprechweise	Interpretation	Symbolische Darstellung
$A \cap B$	A und B	A und B treten gleichzeitig ein.	
$A \cup B$	A oder B	Es tritt A oder B ein.	
$A \setminus B$ **A-B**	A minus B	Es tritt A, aber nicht B ein.	
$\overline{A} = \Omega \setminus A$	nicht A	Es tritt das Gegenteil von A ein.	
$A \triangle B$	A Diskrepanz B ***symmetrische Differenz***	Es tritt entweder A oder B ein, aber nicht beide zugleich.	

Venn-Diagramme

Abb. 47: Venn-Diagramme der Mengenoperationen

5.1 Exkurs: Mengenlehre

Beispiel: Werfen von 2 verschiedenfarbigen Würfeln

A: Die Augenzahlen beider Würfel sind gleich (Pasch)

B: Die Augensumme beider Würfel ist gleich 6.

Operation zwischen A und B	Interpretation	Angabe der zugehörigen Versuchsergebnisse	Anzahl
$A \cap B$	Man erhält Pasch und gleichzeitig Augensumme 6.	(3,3)	1
$A \cup B$	Pasch oder Augensumme 6	(1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (1,5), (2,4), (4,2), (5,1)	10
$A \setminus B$	Pasch, aber nicht Augensumme 6	(1,1), (2,2), (4,4), (5,5), (6,6)	5
$\overline{A}$	kein Pasch	30 Versuchsergebnisse	
$A \triangle B$	Entweder Pasch oder Augensumme 6	(1,1), (2,2), (4,4), (5,5), (6,6), (1,5), (2,4), (4,2), (5,1)	9

Abb. 48: Mögliche Operationen bei Würfelbeispiel

5.1 Exkurs: Mengenlehre

Satz 8 (Eigenschaften von Mengensystemen)

(i) **Kommutativgesetze**:
$A \cup B = B \cup A$ *und* $A \cap B = B \cap A$

(ii) **Assoziativgesetze**:
$A \cup (B \cup C) = (A \cup B) \cup C$ *und* $(A \cap B) \cap C = A \cap (B \cap C)$

(iii) **Distributivgesetze**:
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ *und*
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(iv) **Doppeltes Komplement**:
$\overline{(\bar{A})} = A$

(v) **Sonderstellung von $\emptyset$, Ω**:
$A \cap \Omega = A$, $A \cup \Omega = \Omega$, $A \cap \emptyset = \emptyset$, $A \cup \emptyset = A$

(vi) **Mengen und ihr Komplement**:
$A \cap \bar{A} = \emptyset$, $A \cup \bar{A} = \Omega$, $A \cap A = A$, $A \cup A = A$

5.1 Exkurs: Mengenlehre

Satz 8 (Eigenschaften von Mengensystemen)

(vii) **De Morgan's Gesetze**:
$\overline{(A \cup B)} = \bar{A} \cap \bar{B}$ und $\overline{(A \cap B)} = \bar{A} \cup \bar{B}$

(viii) **Mengendifferenz**:
$A - B = A \cap \bar{B}$

(ix) **Mengendiskrepanz**:
$A \,\Delta\, B = (A - B) \cup (B - A)$

(x) **Komplementarität**:
$A = (A \cap B) \cup \left(A \cap \bar{B}\right)$ und $(A \cap B) \cap \left(A \cap \bar{B}\right) = \emptyset$

(xi) **Teilmengen**:
Sei $A \subset B$, dann gilt: $A \cap B = A$ und $A \cup B = B$

Beweis: Illustration mit **Venn Diagrammen**.

5.1 Exkurs: Mengenlehre

Beispiele für formale Beweise.

(vii) zz. $\overline{(A \cup B)} = \bar{A} \cap \bar{B}$, d.h.
zz. a) $\overline{(A \cup B)} \subset \bar{A} \cup \bar{B}$ und b) $\overline{(A \cup B)} \supset \bar{A} \cap \bar{B}$

a) $\omega \in \overline{(A \cup B)} \Rightarrow \omega \notin (A \cup B) \Rightarrow \omega \notin A \wedge \omega \notin B \Rightarrow \omega \in \bar{A} \wedge \omega \in \bar{B}$
$\Rightarrow \omega \in (\bar{A} \cap \bar{B})$

b) $\omega \in (\bar{A} \cap \bar{B}) \Rightarrow \omega \notin A \wedge \omega \notin B \Rightarrow \omega \notin (A \cup B)$

(ix) $A = A \cap \Omega = A \cap (B \cup \bar{B}) = (A \cap B) \cup (A \cap \bar{B})$
$(A \cap B) \cap (A \cap \bar{B}) = (A \cap A) \cap (B \cap \bar{B}) = A \cap \emptyset = \emptyset$

□

5.2 Grundgesamtheit: Ereignisse

Definition 26 (Grundgesamtheit)

Eine **Grundgesamtheit** *(***Stichprobenraum***)* Ω *ist die Menge aller (prinzipiell) denkbaren* **Versuchsergebnisse** ω*.*

5.2 Grundgesamtheit: Ereignisse

Beispiele

- Werfen einer **Münze**: $\Omega = \{\omega_1, \omega_2\} = \{\text{Kopf, Zahl}\}$
- Werfen eines **Würfels**:
  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\} = \{1, 2, 3, 4, 5, 6\}$
- Werfen von **3 verschiedenen Münzen**: pro Münze 2 Möglichkeiten, insgesamt $2 \cdot 2 \cdot 2 = 8$ Versuchsergebnisse,
  $\Omega =$ {(Z,Z,Z), (Z,Z,W), (Z,W,Z), (W,Z,Z), (Z,W,W), (W,Z,W), (W,W,Z), (W,W,W)}
- Werfen von **2 verschiedenfarbigen Würfeln**: pro Würfel 6 Möglichkeiten, insgesamt $6 \cdot 6 = 36$ Versuchsergebnisse,
  $\Omega = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (2,6), \ldots, (6,6)\}$
- **Körpergröße und Gewicht**: $\Omega = \{\omega = (\omega_K, \omega_G) \,|\, \omega_K, \omega_G > 0\}$

5.2 Grundgesamtheit: Ereignisse

Definition 27 (Ereignis)

*Ein (**zufälliges**) **Ereignis** A ist eine Teilmenge des Stichprobenraums Ω.*

Beispiele

- Werfen eines **Würfels**: Ereignis $A = \{2, 4, 6\}$:
  Würfeln einer geraden Zahl
- Werfen von **3 verschiedenen Münzen**:
  $A =$\{(Z,W,W), (W,Z,W), (W,W,Z)\}: genau 1mal Zahl
- Werfen von **2 verschiedenfarbigen Würfeln**:
  $A = \{(1, 4), (2, 3), (3, 2), (4, 1)\}$: Würfelsumme $= 5$
- **Körpergröße und Gewicht**:
  $A = \{\omega = (\omega_K, \omega_G) \,|\, \omega_K > 1.70, \omega_G < 68.5\}$

5.2 Grundgesamtheit: Ereignisse

Definition 28

Ein **Ereignis** A **tritt ein**, *wenn ein Versuchsergebnis in A liegt. Die leere Menge $\emptyset$ heißt* **unmögliches Ereignis**, *die Gesamtmenge Ω heißt* **sicheres Ereignis**, *die einzelnen Versuchsergebnisse ω heißen* **Elementarereignisse**.

Definition 29 (Axiomatische Definition des Ereignisraums)

Eine Menge $\mathcal{A}$ von Teilmengen eines Stichprobenraums Ω heißt **Ereignisalgebra**, *wenn gilt:*

(i) $\Omega \in \mathcal{A}$,

(ii) *Falls* $A \in \mathcal{A}$, *dann* $\bar{A} \in \mathcal{A}$.

(iii) *Falls* $A_1, A_2 \in \mathcal{A}$, *dann* $A_1 \cup A_2 \in \mathcal{A}$.

Andere Bezeichnungen sind: **Boole'sche (Mengen-)Algebra** *oder* **(Mengen-)Körper** *bzw.* **Ereignisraum** *oder* **Ereigniskörper**.

5.2 Grundgesamtheit: Ereignisse

Satz 9 (Eigenschaften von Boole'schen Algebren)

Sei $\mathcal{A}$ eine Ereignisalgebra. Dann gilt:

(a) $\emptyset \in \mathcal{A}$

(b) *Falls* $A_1, A_2 \in \mathcal{A}$, *dann* $A_1 \cap A_2 \in \mathcal{A}$.

(c) *Falls* $A_1, A_2, \ldots, A_n \in \mathcal{A}$, *dann* $\bigcap_{i=1}^{n} A_i \in \mathcal{A}$, $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$.

5.2 Grundgesamtheit: Ereignisse

Definition 30

Ereignisse $A_1, A_2 \in \mathcal{A}$ *heißen* **unvereinbar** *(***disjunkt***), wenn* $A_1 \cap A_2 = \emptyset$. *Ereignisse* $A_1, A_2, \ldots, A_n \in \mathcal{A}$ *heißen ein* **vollständiges Ereignissystem** *(***Partition***), wenn sie paarweise unvereinbar sind und durch sie eine Zerlegung der Grundgesamtheit* Ω *gegeben ist, d.h. wenn* $A_i \cap A_j = \emptyset$, $i \neq j$ *und* $\bigcup_{i=1}^{n} A_i = \Omega$.

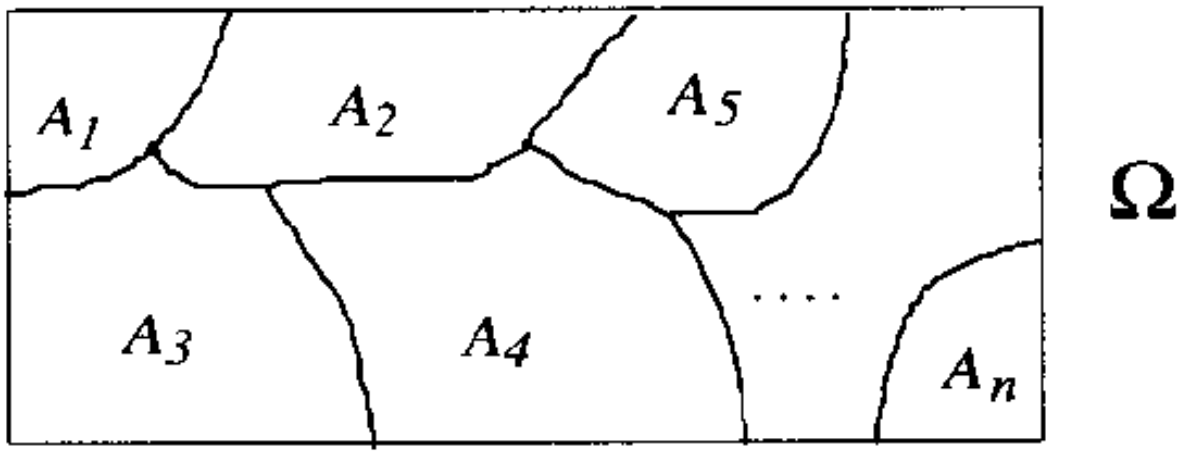


Abb. 49: Skizze (vollständiges Ereignissystem)

5.2 Grundgesamtheit: Ereignisse

Konstruktion einer Ereignisalgebra

Man geht von allen interessierenden Ereignissen (= Teilmengen) einer Grundgesamtheit aus. Wenn notwendig, nimmt man dann

(i) das sichere Ereignis,

(ii) sämtliche Komplementärereignisse und

(iii) alle endlichen Vereinigungen und Durchschnitte von Ereignissen mit hinzu.

Beispiel: Lebensdauer $x \geq 0$ einer Glühbirne

$\Omega = [0, \infty)$, interessante Ereignisse $A_a = \{x | x \geq a\}, \ a > 0$:
Lebensdauer größer oder gleich a (Tage).

5.2 Grundgesamtheit: Ereignisse

Definition 31

Sei $E := \{A_i\}_{i=1,\ldots,n}$ eine Menge von Teilmengen eines Stichprobenraums Ω. Dann heißt die Ereignisalgebra $\mathcal{A}(E)$, die wie oben aus E konstruiert wird, die **von den A_i, $i = 1, \ldots, n$, erzeugte Ereignisalgebra**.

Bemerkung

Bei einem endlichen Stichprobenraum $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ (= endlich viele Elementarereignisse) ist jede Teilmenge ein Ereignis in der von den Elementarereignissen ω_i, $i = 1, \ldots, n$, erzeugten Ereignisalgebra $\mathcal{A}(\{\omega_1, \omega_2, \ldots, \omega_n\}) = \mathcal{P}(\Omega) =$ Potenzmenge von Ω.

5.2 Grundgesamtheit: Ereignisse

Bemerkung

Unendliche Stichprobenräume sind u.a. deswegen problematisch, weil u.U. bei obiger Konstruktion einer Ereignisalgebra, die vorgegebene Teilmengen enthält, „wichtige“ andere Teilmengen nicht erzeugt werden.

Beispiel: Lebensdauer $x \geq 0$ einer Glühbirne

$\Omega = [0, \infty)$, vorgegebene Ereignisse $A_a = \{x | x \geq a\}, \ a > 0$.
Es gilt: $[0, a) \cap [b, \infty) = \emptyset$ oder $[b, a)$. Also sind die Elementarereignisse $\{b\}$ nicht in der erzeugten Ereignisalgebra, aber „wichtig“!

5.2 Grundgesamtheit: Ereignisse

Definition 32 (σ-Algebra, σ-Körper)

Eine Mengenalgebra (oder Mengenkörper) $\mathcal{A}$ heißt σ-**Algebra** *(oder* σ-**Körper***), wenn gilt:*

$$\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}, \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{A} \text{ für alle Folgen } A_1, A_2, \ldots \in \mathcal{A}.$$

„Konstruktion“ einer σ-Algebra

Ausgehend von allen interessierenden Ereignissen einer Grundgesamtheit nimmt man (wenn notwendig) (i) das sichere Ereignis, (ii) sämtliche Komplementärereignisse und (iii) **alle abzählbaren** Vereinigungen und Durchschnitte von Ereignissen mit hinzu.

5.2 Grundgesamtheit: Ereignisse

Definition 33

Sei $E := \{A_i\}_{i \in I}$, I beliebige Indexmenge, eine Menge von Teilmengen eines Stichprobenraums Ω. Dann heißt die σ-Algebra $\mathcal{A}(E)$, die wie oben aus E „konstruiert“ wird, die **von den $\{A_i\}_{i \in I}$ erzeugte σ-Algebra**.

Beispiele

- Lebensdauer $x \geq 0$ einer Glühbirne:
  $[0, a_i) \cap [b, \infty) = [b, a_i) \to \{b\}$ für $b < a_i \to b$
  $\Rightarrow$ Elementarereignisse $\{b\}$ sind in der erzeugten σ-Algebra !
- **Körpergröße und Gewicht**:
  $\Omega = \{\omega = (\omega_K, \omega_G) \,|\, \omega_K, \omega_G > 0\}$, interessante Ereignisse:
  $A_{mk} := \{\omega = (\omega_K, \omega_G) \,|\, \omega_K > m, 0 < \omega_G < k\}$, $m, k > 0$.
  Interessant sind also insbesondere schlanke (große und leichte) Menschen.

5.3 Relative Häufigkeit

Ziel:

Bewertung eines jeden Ereignisses A aus einer σ-Algebra $\mathcal{A}$ mit der Chance seines Auftretens, also einer Zahl zwischen 0 und 1.

Wir betrachten die **Grundgesamtheit** Ω, die alle möglichen Versuchsergebnisse ω_i, $i \in I$, eines Versuchs enthält, und die von den ω_i, $i \in I$, erzeugte σ-**Algebra** $\mathcal{A}$. Wir wiederholen diesen Versuch n Mal. Dabei interessieren wir uns dafür, welches Ereignis jeweils eintritt.

5.3 Relative Häufigkeit

Definition 34 (Erinnerung: Häufigkeit)

Die **absolute Häufigkeit** $H_n(A)$ *des Eintretens von* A *in den ersten* $n > 0$ *Versuchen ist definiert als die Anzahl des Auftretens von* A *in diesen* n *Versuchen.*
Die **relative Häufigkeit** $h_n(A)$ *ist die absolute Häufigkeit geteilt durch die Anzahl Versuche:* $h_n(A) := H_n(A)/n$.

5.3 Relative Häufigkeit

Satz 10 (Eigenschaften der relativen Häufigkeit)

(i) $h_n(A)$ *kann nur die Werte* $0, 1/n, 2/n, \ldots, n/n = 1$ *annehmen.*

(ii) $0 \leq h_n(A) \leq 1$ *für alle* $A \in \mathcal{A}$.

(iii) $h_n(\emptyset) = 0/n = 0$, $h_n(\Omega) = n/n = 1$.

(iv) *relative Häufigkeiten sind (sub-)additiv (!):*
$h_n(A \cup B) = h_n(A) + h_n(B) - h_n(A \cap B)$

(v) $h_n(A \cup B) = h_n(A) + h_n(B)$, *falls* $A \cap B = \emptyset$

(vi) $h_n(\bar{A}) = 1 - h_n(A)$

Beweis.

(iv) Venn-Diagramm

(vi) $1 = h_n(A \cup \bar{A}) = h_n(A) + h_n(\bar{A})$ wegen (v).

□

5.4 Wahrscheinlichkeit

Ziel:

Wir betrachten wieder die **Grundgesamtheit** Ω und eine dazugehörige σ-**Algebra** $\mathcal{A}$. Jedem Ereignis $A \in \mathcal{A}$ soll die (theoretische) Wahrscheinlichkeit seines Eintretens zugeordnet werden. Jeder Teilmenge A von Ω, die zu der σ-Algebra gehört, soll also eine reelle Zahl $\in [0, 1]$ zugeordnet werden.

Man benötigt also eine Funktion, die einer Menge eine Zahl zuordnet (Mengenfunktion) und die mit den Eigenschaften von Häufigkeit „verträglich“ ist.

5.4 Wahrscheinlichkeit

Definition 35

Eine Mengenfunktion M auf einer σ-Algebra $\mathcal{A}$ von Ω ordnet jedem Ereignis $A \in \mathcal{A}$ eine reelle Zahl zu ($-\infty$ und ∞ evtl. eingeschlossen).

Definition 36 (Wahrscheinlichkeitsfunktion)

Eine **Wahrscheinlichkeitsfunktion** *P ist eine Mengenfunktion auf einer σ-Algebra $\mathcal{A}$ von Ω mit Wertebereich $[0, 1]$ und folgenden Eigenschaften:*

(i) $P(A) \geq 0$ *für alle* $A \in \mathcal{A}$

(ii) $P(\Omega) = 1$

(iii) *Für alle Folgen von paarweise unvereinbaren Ereignissen $A_1, A_2, \ldots (A_i \cap A_j = \emptyset,\ i \neq j)$ gilt:*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

5.4 Wahrscheinlichkeit

Sprechweise

$P(A)$ heißt auch „**Wahrscheinlichkeit des Ereignisses** A“ oder „**Wahrscheinlichkeit, dass das Ereignis** A **eintritt**“, d.h. die Wahrscheinlichkeit, dass ein Versuchsergebnis in A liegt.

Beispiele

- Werfen einer **Münze**: $\Omega = \{\omega_1, \omega_2\} = \{\text{Kopf, Zahl}\}$: $P(K) = P(Z) = 0.5$ erfüllt offenbar die Axiome.
- Werfen eines **Würfels**: $P(1) = P(2) = \ldots = P(6) = 1/6$
- Werfen von **3 verschiedenen Münzen**: $P(\text{Z,Z,Z}) = P(\text{Z,Z,W}) = P(\text{Z,W,Z}) = P(\text{W,Z,Z}) = P(\text{Z,W,W}) = P(\text{W,Z,W}) = P(\text{W,W,Z}) = P(\text{W,W,W}) = 1/8$

5.4 Wahrscheinlichkeit

Satz 11 (Rechnen mit Wahrscheinlichkeiten)

(i) $P(\emptyset) = 0$

(ii) $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ *für paarweise unvereinbare* $A_1, A_2, \ldots, A_n$

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

(iv) $P(A) = P(A \cap B) + P(A \cap \bar{B})$

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

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

5.4 Wahrscheinlichkeit

Satz 11 (Rechnen mit Wahrscheinlichkeiten)

(vii) Einschluss- / Ausschluss-Formel:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum\sum_{i<j} P(A_i \cap A_j)$$

$$+ \sum\sum\sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$$

Siehe Skizze an Tafel...

(viii) Sei $A \subset B$, dann gilt: $P(A) \leq P(B)$

(ix) Boole's Ungleichung: $P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$

5.4 Wahrscheinlichkeit

Beweis.

(i) $A_1 = A_2 = \ldots = \emptyset$ in Axiom (iii)

(ii) $A \cup \bar{A} = \Omega$, $A \cap \bar{A} = \emptyset$ in Eigenschaft (ii)

(viii) $B = (B \cap A) \cup (B \cap \bar{A}) = A \cup (B \cap \bar{A})$, $A \cap (B \cap \bar{A}) = \emptyset$
also: $P(B) = P(A) + P(B \cap \bar{A}) \geq P(A)$

□

Definition 37 (Wahrscheinlichkeitsraum)

Ein **Wahrscheinlichkeitsraum** *ist ein Tripel* $(\Omega, \mathcal{A}, P)$*, wobei* Ω *eine Grundgesamtheit,* $\mathcal{A}$ *eine* σ*-Algebra auf* Ω *und* P *eine Wahrscheinlichkeitsfunktion auf* $\mathcal{A}$ *ist.*

5.5 Übungsaufgaben

Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$.

- Wann gilt $P(A \cup B) = P(A) + P(B)$ und wann gilt $P(A \cup B) > P(A) + P(B)$?
- Welche Wahrscheinlichkeit ist größer, $P(A \cap B)$ oder $P(A) \cdot P(B)$?
- Warum gilt für Wahrscheinlichkeiten stets $P(A) \geq 0$ und $P(A) \leq 1$?
- Folgt aus $P(A - B) \leq P(B - A)$ stets $P(A) \leq p(B)$?

5.5 Übungsaufgaben

Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$.

- Wann gilt $P(A \cup B) = P(A) + P(B)$ und wann gilt $P(A \cup B) > P(A) + P(B)$?

  Wegen Satz 10 (vi) ist der erste Ausdruck gleichbedeutend mit $P(A \cap B) = 0$, d.h. A und B sind unvereinbar. Der zweite Ausdruck ist gleichbedeutend mit $P(A \cap B) < 0$, was nie gilt.

- Welche Wahrscheinlichkeit ist größer, $P(A \cap B)$ oder $P(A) \cdot P(B)$?

  Beides ist möglich, siehe Kapitel zur Unabhängigkeit.

- Warum gilt für Wahrscheinlichkeiten stets $P(A) \geq 0$ und $P(A) \leq 1$?

  Das erste folgt aus Definition 36 (i), das zweite aus Definition 36 (ii) und Satz 11 (viii) mit $B = \Omega$.

- Folgt aus $P(A - B) \leq P(B - A)$ stets $P(A) \leq P(B)$?

  Ja, direkt aus dem zweimaligen Anwenden von Satz 11 (v).

5.6 Endliche Grundgesamtheiten

Ziel:

Formeln zur Berechnung von Wahrscheinlichkeiten

5.6.1 Klassische Wahrscheinlichkeit

Definition 38

Eine Mengenfunktion P mit den Eigenschaften (i), (ii) heißt **Wahrscheinlichkeitsfunktion mit gleichwahrscheinlichen Elementarereignissen**:
Seien $\omega_1, \omega_2, \ldots, \omega_N$ die N Elementarereignisse der endlichen Grundgesamtheit Ω, dann gilt:

(i) $P(\omega_1) = P(\omega_2) = \ldots = P(\omega_N) = 1/N$

(ii) $P(A) = N(A)/N$, *wobei* $N(A)$:= *Anzahl Elementarereignisse in Ereignis* $A \in \mathcal{A}$.

Bemerkung:

P = Zahl der günstigen Ereignisse / Zahl der möglichen Ereignisse

5.6.1 Klassische Wahrscheinlichkeit

Beispiele

- Werfen eines **Würfels**:
  $A = \{2, 4, 6\}$: Würfeln einer geraden Zahl: $P(A) = 3/6$
- Werfen von **3 verschiedenen Münzen**: $A = \{$(Z,W,W), (W,Z,W), (W,W,Z)$\}$: genau 1mal Zahl: $P(A) = 3/8$

5.6.2 Kombinatorik

Annahme

Jedes Ergebnis eines Experiments ist ein n-Tupel.

Beispiele sind 3,4,6,7, aber auch Experimente, deren Ergebnisse aus einer Stichprobe der Größe n einer Grundgesamtheit bestehen.

Typisch sind sogenannte **Urnenexperimente**, bei denen aus einer Urne mit M Kugeln eine Stichprobe der Größe n gezogen wird. Dabei soll jede Kugel in der Urne (!) die gleiche Chance haben, gezogen zu werden (**Zufallsauswahl**). Man unterscheidet **2 Arten des Stichprobenziehens**:

Ziehen ohne Zurücklegen bzw. **Ziehen mit Zurücklegen**.

5.6.2 Kombinatorik

Wahrscheinlichkeit einer Menge von n-Tupeln

- Sei A eine Menge von n-Tupeln, die auf eine spezielle Weise aus einer Urne mit M Kugeln konstruiert wird.
- Bestimme $N(A)$ auf die folgende Weise:
  - Bestimme die Anzahl N_1 der Objekte, die als erstes Element eines n-Tupels in A gewählt werden können.
  - Bestimme die Anzahl N_2 der Objekte, die als zweites Element gewählt werden können, unter der Annahme, dass die Anzahl N_2 nicht von der Wahl des ersten Elements abhängt.
  - ...
  - Bestimme die Anzahl N_n der Objekte, die als n-tes Element gewählt werden können, unter der Annahme, dass die Anzahl N_n nicht von der Wahl der ersten $(n-1)$ Elemente abhängt.

  Dann gilt: $\mathbf{N(A) = N_1 \cdot N_2 \cdot \ldots \cdot N_n}$.
- Sei nun $A \subset \Omega := \{\text{alle relevanten } n\text{-Tupel aus } M \text{ Objekten}\}$. Dann gilt: $P(A) = N(A)/N(\Omega)$ ist die Wahrscheinlichkeit von A in der Grundgesamtheit Ω.

5.6.2 Kombinatorik: Permutationen ohne Wiederholungen

Definition 39

Sei eine Gruppe von n Objekten fest vorgegeben. Eine **Permutation** *dieser Objekte ist eine Umordnung der Objekte.*
Sind alle gegebenen Objekte voneinander verschieden, so spricht man von **Permutationen ohne Wiederholungen**. *Treten gewisse Objekte mehrfach auf, so spricht man von* **Permutationen mit Wiederholungen**.

5.6.2 Kombinatorik: Permutationen ohne Wiederholungen

Satz 12

Sei $A :=$ Permutationen von n Objekten ohne Wiederholung (n-maliges Ziehen aus n Objekten ohne Zurücklegen)
Dann gilt: $N(A) = n!$.

Beweis.

Anwendung des allgemeinen Prinzips in diesem Abschnitt (Kombinatorik): Um eine Permutation eines n-Tupels zu erhalten, kann man an der ersten Stelle zwischen $N_1 = n$ Objekten wählen, an der zweiten zwischen $N_2 = n - 1, \ldots,$ und an der letzten zwischen $N_n = n - (n-1) = 1$ Objekten.
Also: $N(A) = n \cdot (n-1) \cdot \ldots \cdot 1 = n!$. □

5.6.2 Kombinatorik: Permutationen ohne Wiederholungen

Beispiele

- **12 Personen in Kinoreihe mit 12 Plätzen**
  Wieviele Sitzanordnungen gibt es?
  $N(A) = 12! = 479\,001\,600$
- **10 Paare aus 10 Herren und 10 Damen**
  Wieviele Konstellationen gibt es?
  Damen nehmen feste Plätze $1 - 10$ ein. Herren werden auf diese Plätze „permutiert“.
  $N(A) = 10! = 3\,628\,800$

5.6.2 Kombinatorik: Permutationen mit Wiederholungen

Sei $A :=$ Permutationen von n nicht unbedingt unterschiedlichen Objekten (n-maliges Ziehen aus n Objekten ohne Zurücklegen!)

Beispiel: 4-stellige Zahlen aus den 4 Ziffern 1,1,3,3

Wieviele verschiedene Zahlen gibt es?

1133, 1313, 3113, 1331, 3131, 3311: $N(A) = 6$

Zurückführung auf Permutationen ohne Wiederholung „Farbtrick“:
Färbe identische Elemente verschieden ein.
Permutiere ohne Wiederholungen.
Entfärbe! Dabei fallen jeweils (!) soviele Möglichkeiten zusammen, wie sich durch Permutation innerhalb der Gruppen gleicher Elemente ergeben.

5.6.2 Kombinatorik: Permutationen mit Wiederholungen

Gegebene Elemente	Hilfskonstruktion mittels Fettdruck	Permutationen mit Wiederholung	Anzahl
1,1,3	**1**13 1**1**3	113	
	131 13**1**	131	$3!/2! = 3$
	3**1**1 31**1**	311	
a,a,b,b	a**ab**b a**ab**b	aabb	
	aab**b** **a**ab**b**		
	ab**ab** a**ba**b	abab	
	aba**b** **ab**ab		$4!/(2! \cdot 2!) = 6$
	ab**ba** a**b**b**a**	abba	
	abb**a** **ab**ba		
	...	baab, baba, bbaa	

Tab. 14: Beispiele für Permutationen mit Wiederholungen

5.6.2 Kombinatorik: Permutationen mit Wiederholungen

Satz 13

Die n Objekte zerfallen in k Gruppen identischer Objekte mit den Anzahlen $n_1, n_2, \ldots, n_k$, so dass $n_1 + n_2 + \ldots + n_k = n$.
Dann gilt: $N(A) = n!/(n_1! \cdot n_2! \cdot \ldots \cdot n_k!)$

5.6.2 Kombinatorik: Variationen / Kombinationen

Bei Permutationen sind sämtliche Elemente der n-Tupel vorher bekannt, nur die Reihenfolge ist variabel.

Bei Variationen und Kombinationen werden die Elemente der n-Tupel in A aus einem Grundvorrat von Elementen ausgewählt. Falls die Reihenfolge des Ziehens wesentlich ist, spricht man von **Variationen**, sonst von **Kombinationen**. Man unterscheidet Variationen und Kombinationen **mit und ohne Zurücklegen**.

Sprechweise

Anstelle von Variationen und Kombinationen mit und ohne Zurücklegen spricht man auch von **mit und ohne Wiederholungen** (s. Beispiel).

5.6.2 Kombinatorik: Variationen / Kombinationen

Beispiel: Eisdiele

Aus 3 Eissorten sollen Portionen mit 2 Kugeln zusammengestellt werden. Spielt die Reihenfolge eine Rolle (spitze Eisbecher), so erhält man Variationen. Kommt es nur auf die Kombination an, so erhält man Kombinationen (runde Eisschalen).

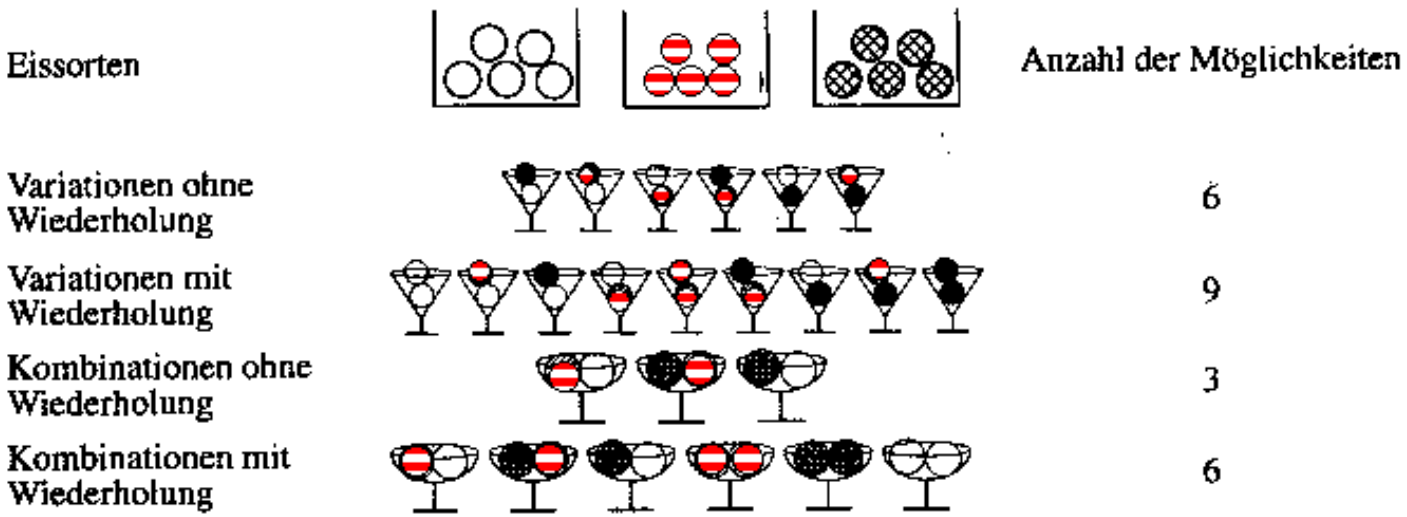


Abb. 50: Veranschaulichung des Beispiels

5.6.2 Kombinatorik: Variationen ohne Zurücklegen

Definition 40

*n-**Variationen ohne Zurücklegen** sind n-Tupel aus M unterschiedlichen Objekten ($n \leq M$) unter Beachtung der Reihenfolge, wobei kein Objekt mehr als einmal ausgewählt werden darf.*

Bemerkung

Für das erste Element des n-Tupels stehen also M Elemente zur Auswahl,
für das zweite Element nur noch $(M-1), \ldots,$
für das n-te Element nur noch $(M-(n-1))$.

5.6.2 Kombinatorik: Variationen ohne Zurücklegen

Satz 14

Sei $A :=$ n-Variationen von M Objekten ohne Zurücklegen
(n-maliges Ziehen aus M Objekten ohne Zurücklegen)
Dann gilt: $N(A) = M \cdot (M-1) \cdot \ldots \cdot (M-n+1)$.

Beispiele

- **Eisdiele**
  2 Kugeln aus 3 Eissorten ohne Wiederholungen unter Beachtung der Reihenfolge: $N(A) = 3 \cdot 2 = 6$
- **Medaillen**
  Wieviele Möglichkeiten der Medaillenverteilung Gold/Silber/Bronze gibt es bei einem Wettkampf von 10 Teilnehmern?
  $N(A) = 10 \cdot 9 \cdot 8 = 720$

5.6.2 Kombinatorik: Variationen mit Zurücklegen

Definition 41

n-**Variationen mit Zurücklegen** *sind n-Tupel aus M unterschiedlichen Objekten unter Beachtung der Reihenfolge, wobei die Objekte mehr als einmal ausgewählt werden dürfen.*

Satz 15

Sei A := n-Variationen von M Objekten mit Zurücklegen
(n-maliges Ziehen aus M Objekten mit Zurücklegen).
Dann gilt: $N(A) = M^n$.

Bemerkung

Offenbar muss hier nicht mehr $n \leq M$ gelten!

5.6.2 Kombinatorik: Variationen mit Zurücklegen

Beispiele

- Werfen von **3 verschiedenen Münzen**
  $N(\Omega) = 2^3 = 8$ mögliche Versuchsergebnisse
- **Eisdiele**
  2 Kugeln aus 3 Eissorten mit Wiederholungen unter Beachtung der Reihenfolge: $N(A) = 3^2 = 9$

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Definition 42

n-**Kombinationen ohne Zurücklegen** *sind n-Tupel aus M unterschiedlichen Objekten* ($n \leq M$) *ohne Beachtung der Reihenfolge, wobei kein Objekt mehr als einmal ausgewählt werden darf.*

Bemerkung

Zwei *n*-Tupel, die sich nur durch die Reihenfolge ihrer Elemente unterscheiden, entsprechen also derselben Kombination.

Ziel

Auswahl einer *n*-elementigen Teilmenge aus einer *M*-elementigen Menge.

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Zurückführung auf Variationen ohne Zurücklegen

Ziehe zunächst unter Beachtung der Reihenfolge:

$$\text{„}N(A) = M \cdot (M-1) \cdot \ldots \cdot (M-n+1)\text{“}.$$

Danach eliminiere alle n-Tupel, die in anderer Reihenfolge schon einmal vorkamen. Von jedem n-Tupel kommen alle Permutationen vor!

$$N(A) = M \cdot (M-1) \cdot \ldots \cdot (M-n+1)/n!$$

Beachte: Anzahl Terme im Zähler und Nenner gleich!

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Satz 16

Sei A := n-Kombinationen von M Objekten ohne Zurücklegen
(n-maliges Ziehen aus M Objekten ohne Zurücklegen).
Dann gilt:

$$N(A) = \binom{M}{n} := \frac{M!}{n!(M-n)!} : \textbf{Binomialkoeffizienten}$$

Beweis.

$$N(A) = \frac{M \cdot (M-1) \cdot \ldots \cdot (M-n+1)}{n!} = \frac{M!}{n!(M-n)!}$$

□

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Satz 17 (Eigenschaften von Binomialkoeffizienten)

(i) $\binom{M}{0} = \binom{M}{M}$

(ii) $\binom{M}{n} = \binom{M}{M-n}$: *Symmetrie*

(iii) $\binom{M}{n} + \binom{M}{n+1} = \binom{M+1}{n+1}$

Es gilt (ohne Beweis):

$$(a+b)^M = \binom{M}{0} a^0 b^M + \binom{M}{1} a^1 b^{M-1} + \ldots + \binom{M}{M-1} a^{M-1} b^1 + \binom{M}{M} a^M b^0$$

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Beispiele

- **Eisdiele**
  2 Kugeln aus 3 Eissorten ohne Wiederholungen ohne Beachtung der Reihenfolge:

$$N(A) = \binom{3}{2} = \frac{3!}{2! \cdot 1!} = 3$$

- **Potenzmenge**
  Wieviele verschiedene Teilmengen einer Menge Ω mit M Elementen gibt es?

$$N(\Omega) = \binom{M}{0} + \binom{M}{1} + \ldots + \binom{M}{M-1} + \binom{M}{M} = (1+1)^M = 2^M$$

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Beispiele (Fortsetzung)

- **Lotto**
  Wieviele verschiedene mögliche Tippscheine gibt es?

$$N(A) = \binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6!} = 13\,983\,816$$

Wie groß ist die Wahrscheinlichkeit, 6 Richtige zu haben?
$A =$ gezogene Zahlenkombination

$$\Rightarrow P(A) = \frac{1}{13\,983\,816}.$$

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Beispiele (Fortsetzung)

- **Urne mit K roten und $(M-K)$ weißen Kugeln**
  Wie groß ist die Chance, beim n-fachen Ziehen ohne Zurücklegen genau k rote Kugeln zu bekommen?

  $A := \{k \text{ rote Kugeln und } (n-k) \text{ weiße Kugeln gezogen}\}$
  $\Omega := \{\text{alle } n\text{-elementigen Teilmengen aus } M \text{ Kugeln}\}$

$$P(A) = \frac{\binom{K}{k}\binom{M-K}{n-k}}{\binom{M}{n}}$$

5.6.2 Kombinatorik: Kombinationen ohne Zurücklegen

Beispiele (Fortsetzung)

- **Kartenspiel**
  Wie groß ist die Chance bei einem Kartenspiel 6 Herz auf der Hand zu haben?
  Es gibt $K = 13$ Herz-Karten und $M - K = 52 - 13$ andere! Jeder Spieler bekommt $n = 13$ Karten. Also:

$$P(A) = \frac{\binom{13}{6}\binom{52-13}{13-6}}{\binom{52}{13}} = ?$$

5.6.2 Kombinatorik: Kombinationen mit Zurücklegen

Definition 43 (Kombinationen mit Zurücklegen)

n-**Kombinationen mit Zurücklegen** *sind n-Tupel aus M unterschiedlichen Objekten ohne Beachtung der Reihenfolge, wobei die Objekte mehrmals ausgewählt werden dürfen.*

Satz 18

Sei A := n-Kombinationen von M Objekten mit Zurücklegen (n-maliges Ziehen aus M Objekten mit Zurücklegen).
Dann gilt:

$$N(A) = \binom{M+n-1}{n} = \frac{(M+n-1)\cdot(M+n-2)\cdot\ldots\cdot M}{n!}$$

5.6.2 Kombinatorik: Kombinationen mit Zurücklegen

Beweis (wird nicht in der Vorlesung behandelt).

Sei (o.B.d.A.) $X := \{1, 2, \ldots, M\}$ die Menge der Objekte, aus der n Elemente mit Zurücklegen gezogen werden. Es genügt, eine bijektive Abbildung von A auf die Menge der n-elementigen Teilmengen von $\{1, 2, \ldots, M + n - 1\}$ anzugeben.

Seien $a_1 \leq a_2 \leq \ldots \leq a_n$ die Elemente eines n-Tupels in A. Dann ordne jedem a_{i+1} mit $a_i = a_{i+1}$ den Wert $M + i$ zu und allen anderen a_i den Wert i. Diese Abbildung ist offenbar injektiv.

Bsp. $M = 8,\ n = 6 : 1, 2, 2, 2, 5, 5 \rightarrow 1, 2, 10, 11, 5, 13$

Sei umgekehrt $b_1 < b_2 < \ldots < b_n$ eine n-elementige Teilmenge aus $\{1, 2, \ldots, M + n - 1\}$. Dann lässt sich ein Urbild in A wie folgt konstruieren:

Für alle $b_i > M$ wiederhole die Zahl an der Stelle $b_i - M$, ansonsten übernehme b_i. Damit ist die Abbildung auch surjektiv.

Bsp. $1, 2, 2, 2, 5, 5 \leftarrow 1, 2, 10, 11, 5, 13$. □

5.6.2 Kombinatorik: Kombinationen mit Zurücklegen

Beispiele

- **Eisdiele**
  2 Kugeln aus 3 Eissorten mit Wiederholungen ohne Beachtung der Reihenfolge:

$$N(A) = \binom{4}{2} = \frac{4 \cdot 3}{2} = 6$$

- **Würfelbilder**
  Wieviele Würfelbilder gibt es bei 2 gleichfarbigen Würfeln?

$$N(A) = \binom{6+2-1}{2} = \frac{7 \cdot 6}{2} = 21$$

5.6.3 Elementarereignisse mit ungleichen Wahrscheinlichkeiten

Konstruktion einer Wahrscheinlichkeitsfunktion

Falls die Elementarereignisse nicht alle die gleiche Wahrscheinlichkeit haben, kann man zur Definition einer Wahrscheinlichkeitsfunktion auf den Teilmengen einer endlichen Grundgesamtheit Ω folgendermaßen vorgehen:

Sei $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ und $\mathcal{A}$ eine σ-Algebra auf Ω, dann wähle

$$p_j := P(\{\omega_j\}) \text{ mit } \sum_{j=1}^{N} p_j = 1$$

und setze

$$P(A) := \sum_{\omega_j \in A} p_j$$

für ein beliebiges Ereignis $A \in \mathcal{A}$.

5.7 Wahrscheinlichkeit in unendlichen Grundgesamtheiten

Bei der klassischen Wahrscheinlichkeit wird vorausgesetzt, dass die Anzahl der Elementarereignisse endlich ist. Der Wahrscheinlichkeitsbegriff soll jetzt auf den Fall von „unendlich vielen gleichwahrscheinlichen Elementarereignissen“ verallgemeinert werden.

Annahme:

Die Grundgesamtheit Ω aller möglichen Versuchsergebnisse fällt ein endlich begrenztes „Gebiet“ vollständig und gleichmäßig aus.

5.7 Wahrscheinlichkeit in unendlichen Grundgesamtheiten

Definition 44

Die **geometrische Wahrscheinlichkeit** *eines Ereignisses A aus einer σ-Algebra $\mathcal{A}$ auf Ω ist definiert durch:*
$P(A)$:= „Ausdehnung“ von A / „Ausdehnung“ von Ω.

Beispiele für „Ausdehnung“ sind Länge (1D), Fläche (2D), Volumen (3D) usw.

Bemerkung

Die Gleichwahrscheinlichkeit des Eintretens aller Versuchsergebnisse wird dadurch zum Ausdruck gebracht, dass die Wahrscheinlichkeit eines Ereignisses A nicht von seiner Lage in Ω abhängt, sondern nur von seiner „Ausdehnung“.

5.7 Wahrscheinlichkeit in unendlichen Grundgesamtheiten

Beispiel: Verabredung

2 Personen wollen garantiert zwischen 8.00 Uhr und 9.00 Uhr an einem vereinbarten Ort sein. Jeder Zeitpunkt innerhalb dieser Grenzen wird als „gleichwahrscheinlich“ angesehen. Jede Person wartet maximal 15 Minuten, dann geht sie wieder.

Mit welcher Wahrscheinlichkeit treffen sich beide (Ereignis A)?

$$P(A) = \frac{\text{Fläche von } A}{\text{Fläche von } \Omega} = \frac{60^2 - 45^2}{60^2} = 0.4375$$

(siehe Skizze)

5.7 Wahrscheinlichkeit in unendlichen Grundgesamtheiten

ACHTUNG

So wie die „klassische“ Wahrscheinlichkeit einer speziellen Wahrscheinlichkeitsfunktion auf einer endlichen Grundgesamtheit entspricht (vgl. 5.6.3), so entspricht die „geometrische“ Wahrscheinlichkeit einer speziellen Wahrscheinlichkeitsfunktion auf unendlichen Grundgesamtheiten, nämlich einer „Gleichverteilung“ über das gesamte „Gebiet“ von Ω.

Natürlich hängt im allgemeinen Fall die Wahrscheinlichkeit eines Ereignisses nicht nur von seiner Ausdehnung, sondern auch von seiner Lage ab. Die „Verteilung“ der „Wahrscheinlichkeitsmasse“ ist also ungleichmäßig (vgl. folgende Skizze). Lediglich die „Gesamtmasse“ ist immer gleich 1!

5.7 Wahrscheinlichkeit in unendlichen Grundgesamtheiten

Massenverteilung: Beispiel in zwei Dimensionen

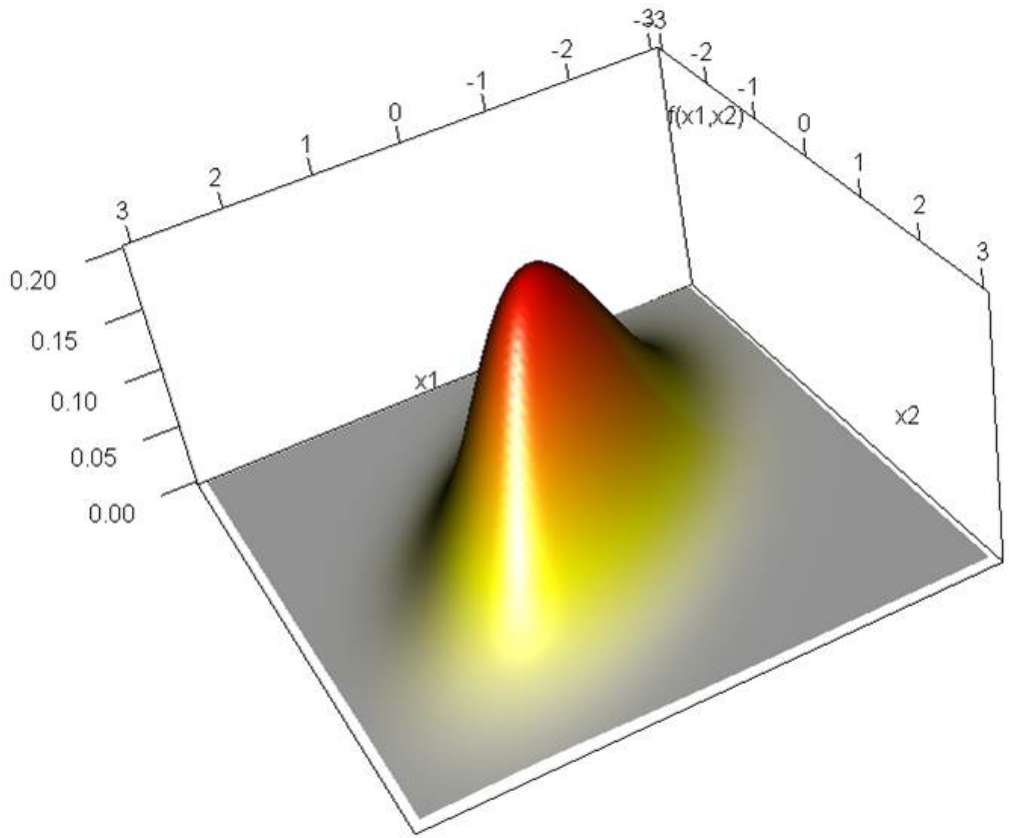


Abb. 51: Massenverteilung

5.8 Übungsaufgaben

An einem Fußballturnier nehmen 8 Mannschaften teil.

- Es sollen alle Mannschaften gegeneinander spielen. Wieviele Spiele gibt es?
- Es werden zwei Gruppen zu je vier Mannschaften gebildet. Innerhalb einer Gruppe sollen alle Mannschaften gegeneinander spielen. Dann spielen die Gruppenersten, Gruppenzweiten usw. gegeneinander. Wieviele Spiele gibt es insgesamt?

5.8 Übungsaufgaben

An einem Fußballturnier nehmen 8 Mannschaften teil.

- Es sollen alle Mannschaften gegeneinander spielen. Wieviele Spiele gibt es?

  Man bestimmt die Anzahl an Möglichkeiten, von 8 Mannschaften 2 auszuwählen (Kombination ohne Zurücklegen). Es gibt $\binom{8}{2} = \frac{8 \cdot 7}{2 \cdot 1} = 28$ Spiele.

- Es werden zwei Gruppen zu je vier Mannschaften gebildet. Innerhalb einer Gruppe sollen alle Mannschaften gegeneinander spielen. Dann spielen die Gruppenersten, Gruppenzweiten usw. gegeneinander. Wieviele Spiele gibt es insgesamt?

  In jeder der beiden Gruppen gibt es zunächst $\binom{4}{2} = \frac{4 \cdot 3}{2 \cdot 1} = 6$ Spiele. Mit den 4 Platzierungsspielen gibt es insgesamt $6 + 6 + 4 = 16$ Spiele.

5.8 Übungsaufgaben

Ein Lampengeschäft möchte sein Schaufenster mit 4 roten, 3 blauen und 2 gelben Glühlampen dekorieren. Wieviele Möglichkeiten gibt es, wenn

- es keine weiteren Einschränkungen gibt?
- die Glühlampen gleicher Farbe jeweils nebeneinander angeordnet werden sollen?
- die Reihe mit einer blauen Glühlampe beginnen und aufhören soll?

5.8 Übungsaufgaben

Ein Lampengeschäft möchte sein Schaufenster mit 4 roten, 3 blauen und 2 gelben Glühlampen dekorieren. Wieviele Möglichkeiten gibt es, wenn

- es keine weiteren Einschränkungen gibt?

  Es gibt $\frac{9!}{4!3!2!} = 1260$ Möglichkeiten.

- die Glühlampen gleicher Farbe jeweils nebeneinander angeordnet werden sollen?

  Es muss nur die Reihenfolge der Farben festgelegt werden, also $3! = 6$ Möglichkeiten.

- die Reihe mit einer blauen Glühlampe beginnen und aufhören soll?

  Es müssen nur noch 7 Glühlampen verteilt werden, davon eine blaue. Es gibt $\frac{7!}{4!1!2!} = \frac{7 \cdot 6 \cdot 5}{2} = 105$ Möglichkeiten.

6.1 Bedingte Wahrscheinlichkeit

Zusatzinformation

Nicht selten soll die Wahrscheinlichkeit eines Ereignisses A berechnet werden unter der Voraussetzung, dass ein Ereignis B schon eingetreten ist (Zusatzinformation). Dadurch wird die Menge der möglichen Versuchsergebnisse „a-priori“ auf die Menge B eingeschränkt. Das Eintreten von B ist nicht mehr zufällig, es ist sicher. Man spricht von „bedingter Wahrscheinlichkeit“ des Ereignisses A, gegeben das Ereignis B.

6.1 Bedingte Wahrscheinlichkeit

Definition 45

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, wobei Ω eine Grundgesamtheit, $\mathcal{A}$ eine σ-Algebra auf Ω und P eine Wahrscheinlichkeitsfunktion auf $\mathcal{A}$ ist. Seien A, B zwei Ereignisse in $\mathcal{A}$. Dann ist die **bedingte Wahrscheinlichkeit von** A **unter der Bedingung** B *definiert durch*

$$P_B(A) = P(A|B) := P(A \cap B)/P(B), \text{ falls } P(B) > 0.$$

6.1 Bedingte Wahrscheinlichkeit

Satz 19 (Eigenschaften der bedingten Wahrscheinlichkeit)

(i) $P(\emptyset|B) = 0$

(ii) $P(\bigcup_{i=1}^{n} A_i|B) = \sum_{i=1}^{n} P(A_i|B)$ *für paarweise unvereinbare* A_i

(iii) $P(\bar{A}|B) = 1 - P(A|B)$

(iv) $P(A_1|B) = P(A_1 \cap A_2|B) + P(A_1 \cap \bar{A}_2|B)$

(v) $P(A_1 - A_2|B) = P(A_1|B) - P(A_1 \cap A_2|B)$

(vi) $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$

(vii) $P(\bigcup_{i=1}^{n} A_i|B) \leq \sum_{i=1}^{n} P(A_i|B)$

(viii) *Sei* $A_1 \subset A_2$, *dann gilt:* $P(A_1|B) \leq P(A_2|B)$

Beweis.

Literatur. □

6.1 Bedingte Wahrscheinlichkeit

Bemerkung

Bedingte Wahrscheinlichkeit bei **endlichen Grundgesamtheiten mit gleichwahrscheinlichen Elementarereignissen**:
$P(A|B) = N(A \cap B)/N(B)$, denn $P(A) = N(A)/N(\Omega)$.

Beispiele

- Werfen von **3 verschiedenen Münzen**:
  $A = \{$(Z,W,W), (W,Z,W), (W,W,Z)$\}$: genau 1mal Zahl
  $B = \{$(W,Z,Z), (W,Z,W), (W,W,Z), (W,W,W)$\}$:
  Wappen auf erster Münze
  $P(A|B) = P(A \cap B)/P(B) = N(A \cap B)/N(B) = 2/4 = 0.5$

6.1 Bedingte Wahrscheinlichkeit

Satz 20 (Multiplikationsregel)

Sei $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum. Seien* A_i, $i = 1, 2, \ldots, n$, n *Ereignisse in* $\mathcal{A}$ *mit* $P(A_1 \cap \ldots \cap A_{n-1}) > 0$, *dann gilt:*

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|(A_1 \cap A_2) \cdots P(A_n|(A_1 \cap \ldots \cap A_{n-1}))$$

Beweis.

Nach der Definition der bedingten Wahrscheinlichkeit gilt:
$P(A \cap B) = P(A)P(B|A)$, falls $P(A) > 0$. Genauso gilt:
$P(A \cap B \cap C) = P(A \cap B)P(C|A \cap B) = P(A)P(B|A)P(C|A \cap B)$,
falls $P(A \cap B) > 0$.
Die Behauptung folgt durch Induktion. □

6.1 Bedingte Wahrscheinlichkeit

Beispiele

- **Urne mit 10 roten und 15 weißen Kugeln**
  Wie groß ist die Wahrscheinlichkeit, dass bei 3 Ziehungen ohne Zurücklegen zuerst eine rote, dann eine weiße und dann wieder eine rote Kugel gewählt wird?
  $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) = (10/25)(15/24)(9/23)$
- **Urne mit 3 roten und 7 weißen Kugeln**
  Betrachte folgendes Spiel: Bei jeder Ziehung wird eine Kugel zufällig gezogen, die Farbe wird notiert, und die Kugel wird zurückgelegt zusammen mit 2 zusätzlichen Kugeln derselben Farbe. Wie groß ist die Wahrscheinlichkeit, dass bei den ersten 3 Ziehungen immer eine rote Kugel gewählt wird?
  $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) = (3/10)(5/12)(7/14) = 1/16$

6.1 Unabhängigkeit

Definition 46

*Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Seien A, B zwei Ereignisse in $\mathcal{A}$. Dann heißen A und B (***statistisch bzw. stochastisch***) **unabhängige Ereignisse**, wenn eine der folgenden Bedingungen gilt:*

(i) $P(A \cap B) = P(A)P(B)$,

(ii) $P(A|B) = P(A)$, *falls* $P(B) > 0$,

(iii) $P(B|A) = P(B)$, *falls* $P(A) > 0$.

Äquivalenz der Bedingungen

(i) $\Rightarrow$ (ii): $P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A)$,
falls $P(B) > 0$

(ii) $\Rightarrow$ (iii): $P(B|A) = P(A \cap B)/P(A) = P(A|B)P(B)/P(A) =$
$P(A)P(B)/P(A) = P(B)$, falls $P(A), P(B) > 0$

(iii) $\Rightarrow$ (i): $P(A \cap B) = P(B|A)P(A) = P(B)P(A)$, falls $P(A) > 0$
$P(A \cap B) = P(B)P(A)$, falls $P(A) = 0$ oder $P(B) = 0$

6.1 Unabhängigkeit

Beispiel: Werfen von 2 verschiedenen Würfeln

Sei A das Ereignis einer ungeraden Augensumme,
B das Ereignis einer 1 auf dem ersten Würfel,
C das Ereignis „Augensumme $= 7$“.
Sind A und B, A und C bzw. B und C unabhängig?

$P(A|B) = N(A \cap B)/N(B) = 3/6 = 0.5 = P(A)$,

$P(A|C) = 1 \neq 0.5 = P(A) \Rightarrow A$ und C sind nicht unabhängig

$P(C|B) = N(C \cap B)/N(B) = 1/6 = 6/36 = P(C)$

6.1 Unabhängigkeit

Satz 21 (Eigenschaften von unabhängigen Ereignissen)

(i) *Seien* $P(A) \neq 0$ *und* $P(B) \neq 0$. *Dann gilt:*
A *und* B *unabhängig* $\Rightarrow A \cap B \neq \emptyset$

(ii) A *und* B *unabhängig* $\Rightarrow A$ *und* $\bar{B}$ *unabhängig,* $\bar{A}$ *und* B *unabhängig,* $\bar{A}$ *und* $\bar{B}$ *unabhängig*

Beweis.

(i) A und B unabhängig $\Rightarrow P(A \cap B) = P(A)P(B) \neq 0 = P(\emptyset)$

(ii) $P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(\bar{B})$ (z. B.)

□

6.1 Unabhängigkeit

Definition 47

*Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Seien A_i, $i = 1, \ldots, n$, n Ereignisse in $\mathcal{A}$. Dann heißen die A_i (***statistisch bzw. stochastisch***) **unabhängige Ereignisse**, wenn sämtliche (!) der folgenden Bedingungen gelten:*

$$
\begin{aligned}
P(A_i \cap A_j) &= P(A_i)P(A_j) \quad \text{für } i \neq j, \\
P(A_i \cap A_j \cap A_k) &= P(A_i)P(A_j)P(A_k) \text{ für } i \neq j,\ j \neq k,\ i \neq k, \\
&\ldots, \\
P(\bigcap_{i=1}^{n} A_i) &= \prod_{i=1}^{n} P(A_i)
\end{aligned}
$$

ACHTUNG

Es werden alle Bedingungen benötigt! Z. B. impliziert paarweise Unabhängigkeit NICHT „globale“ Unabhängigkeit!

6.1 Unabhängigkeit

Beispiel: Werfen von 2 verschiedenen Würfeln

$A_1 :=$ ungerade Zahl auf dem ersten Würfel
$A_2 :=$ ungerade Zahl auf dem zweiten Würfel
$A_3 :=$ ungerade Augensumme

$$P(A_1)P(A_2) = 0.5 \cdot 0.5 = P(A_1 \cap A_2),$$

$$P(A_1)P(A_3) = 0.5 \cdot 0.5 = P(A_3|A_1)P(A_1) = P(A_1 \cap A_3),$$

$$P(A_2)P(A_3) = 0.5 \cdot 0.5 = P(A_3|A_2)P(A_2) = P(A_2 \cap A_3), \text{ aber}$$

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq 1/8 = P(A_1)P(A_2)P(A_3)$$

6.1 Unabhängigkeit

Intuition

Beim **Ziehen ohne Zurücklegen** liegt sicher eine Abhängigkeit zwischen den Ergebnissen von zwei Ziehungen vor!

Bei **Stichprobennahme mit Zurücklegen** sind die einzelnen Ziehungen voneinander (stochastisch) unabhängig!

Bei **Versuchen mit denselben Testpersonen** (Längsschnittansatz) liegt i.a. eine Abhängigkeit der Versuchsergebnisse vor, bei einem **Wechsel der Versuchspersonen** (Querschnittsansatz) kann hingegen Unabhängigkeit erreicht werden!

Bemerkung

Ziel von Unabhängigkeitsuntersuchungen ist (häufig) Bestimmung von $P(A \cap B)$! Es gilt immer:

$P(A \cap B) = P(A|B)P(B)$, falls $P(B) > 0$

$P(A \cap B) = P(B|A)P(A)$, falls $P(A) > 0$

6.2 Totale Wahrscheinlichkeit

Ziel:

Bei einem Versuch, bei dem das Versuchsergebnis durch verschiedene Quellen gleichzeitig beeinflusst wird, soll die Wahrscheinlichkeit eines Ereignisses B bestimmt werden. Durch „Fallunterscheidung“ bei einem der Einflussfaktoren sei es möglich, die Gesamt- (**totale**) **Wahrscheinlichkeit** von B in leicht berechenbare Teile aufzuspalten. Genauer hat man ein vollständiges System (Partition) von Ereignissen $A_1, A_2, \ldots, A_n$, für das die Wahrscheinlichkeiten $P(B|A_1), P(B|A_2), \ldots, P(B|A_n)$ bekannt sind. Daraus lässt sich dann $P(B)$ zusammensetzen.

6.2 Totale Wahrscheinlichkeit

Ziel (Fortsetzung):

Diese Vorgehensweise ist besonders nützlich für **Experimente mit mehreren Stufen**. Dabei stehen die Partitionsindizes für die verschiedenen Stufen. Z. B. Stufe 1: Auswahl einer Urne, Stufe 2: Auswahl einer Kugel aus der gewählten Urne (vgl. folgendes Urnen-Beispiel).

Skizze (Ereignis B und vollständiges Ereignissystem):

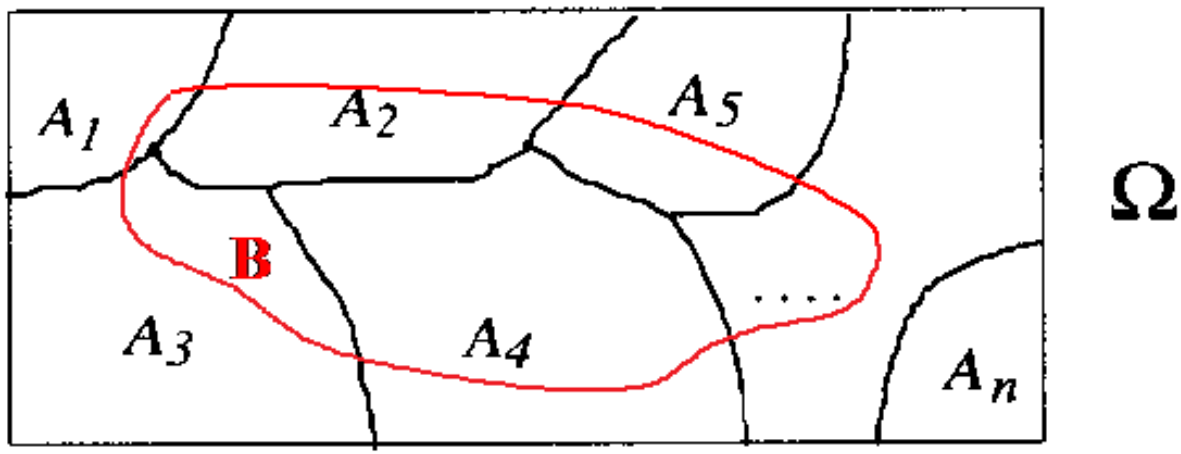


Abb. 52: Ereignis B und vollständiges Ereignissystem

6.2 Totale Wahrscheinlichkeit

Satz 22 (von der totalen Wahrscheinlichkeit)

Sei $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum. Seien* A_i, $i = 1, 2, \ldots, n$, *eine Partition von* Ω, *d.h.* $A_i \cap A_j = \emptyset$, $i \neq j$, *und* $\bigcup_{i=1}^{n} A_i = \Omega$, *mit* $P(A_i) > 0$.

Dann gilt für jedes $B \in \mathcal{A}$:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Beweis.

$B = \bigcup_{i=1}^{n} (B \cap A_i)$ und die $B \cap A_i$ sind alle disjunkt. Also:

$$P(B) = P(\bigcup_{i=1}^{n} (B \cap A_i)) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$ □

6.2 Totale Wahrscheinlichkeit

Spezialfall

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}), \text{ falls } 0 < P(A) < 1$$

Bemerkung

Der Satz bleibt richtig für $n = \infty$.

6.2 Totale Wahrscheinlichkeit

Beispiele

- **Urnen $i = 1, \ldots, 5$ mit i roten und $10 - i$ weißen Kugeln**
  Wähle zunächst zufällig eine Urne und dann zufällig eine Kugel aus der gewählten Urne. Wie groß ist die Wahrscheinlichkeit, dass eine rote Kugel gewählt wird?

$$P(B) = \sum_{i=1}^{5} P(B|A_i)P(A_i) = \sum_{i=1}^{5} \frac{i}{10} \cdot \frac{1}{5} = \frac{1}{50} \sum_{i=1}^{5} i = \frac{1}{50} \cdot \frac{5 \cdot 6}{2} = \frac{3}{10}$$

6.2 Totale Wahrscheinlichkeit

Beispiele (Fortsetzung)

- **Gesamtproduktionsausschuss**
  In einem Betrieb werde die Gesamtproduktion eines bestimmten Teiles zu 20 % durch Maschine 1, zu 55 % durch Maschine 2 und zu 25 % durch Maschine 3 hergestellt. Die gefertigten Teile liegen willkürlich gemischt in einem Behälter. Sei B das Ereignis, dass ein zufällig entnommenes Teil Ausschuss ist. Sei A_i das Ereignis, dass dieses Teil von Maschine i stammt. Die Qualität der einzelnen Maschinen sei bekannt: Maschine 1, 2, 3 produziert 2 %, 1 %, 4 % Ausschuss.
  Wie groß ist die Wahrscheinlichkeit von B?
  $P(B) = 0.02 \cdot 0.20 + 0.01 \cdot 0.55 + 0.04 \cdot 0.25 = 0.0195$

6.3 Bayes'sche Formel

Ziel:

Im Zusammenhang mit der totalen Wahrscheinlichkeit ist die folgende Frage von Interesse: Wie groß ist der Anteil der einzelnen Ereignisse $A_1, A_2, \ldots, A_n$ an der Gesamtwahrscheinlichkeit $P(B)$? Es interessieren also für die bedingten Wahrscheinlichkeiten $P(A_i|B) := P(A_i \cap B)/P(B)$.

Idee:

Schließen von den bekannten bedingten Wahrscheinlichkeiten $P(B|A_i)$ mit Hilfe des Satzes von der totalen Wahrscheinlichkeit auf die unbekannten (in gewisser Weise „umgekehrten“) bedingten Wahrscheinlichkeiten $P(A_i|B)$

6.3 Bayes'sche Formel

Definition 48 (Formel von Bayes (1702-1761))

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Seien A_i, $i = 1, 2, \ldots, n$, eine Partition von Ω mit $P(A_i) > 0$. Dann gilt für jedes $B \in \mathcal{A}$ mit $P(B) > 0$:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Beweis.

$P(A_i|B) := P(A_i \cap B)/P(B) = P(B|A_i)P(A_i)/P(B)$ □

Bemerkung

Der Satz bleibt richtig für $n = \infty$.

6.3 Bayes'sche Formel

Beispiel

- **Gesamtproduktionsausschuss**
  Aus welcher Maschine stammt das defekte Teil?
  $P(A_1|B) = 0.02 \cdot 0.20/0.0195 = 0.2051$
  $P(A_2|B) = 0.01 \cdot 0.55/0.0195 = 0.2821$
  $P(A_3|B) = 0.04 \cdot 0.25/0.0195 = 0.5128$

6.3 Bayes'sche Formel

Maximale Mutmaßlichkeit (maximum-likelihood)

Bei einem Zufallsexperiment kann genau eines der n Ereignisse $A_1, A_2, \ldots, A_n$ eintreten. Die direkte Beobachtung, welches Ereignis wirklich eingetreten ist, sei aber unmöglich!
Es wird angenommen, dass die Eintrittswahrscheinlichkeiten der A_i gleich $P(A_i)$ sind (**a-priori-Verteilung**).
Bei der Versuchsdurchführung beobachtet man das Eintreten des Ereignisses B (**Versuchsergebnis**).
Daraus berechnet man neue Wahrscheinlichkeiten der Ereignisse A_i (**a-posteriori-Verteilung**), nämlich die bedingten Wahrscheinlichkeiten $P(A_1|B), P(A_2|B), \ldots, P(A_n|B)$ (mit Hilfe der bedingten Wahrscheinlichkeiten $P(B|A_1), P(B|A_2), \ldots, P(B|A_n)$).
Es soll nun eine Vermutung darüber abgegeben werden, welches der Ereignisse wirklich eingetreten ist. Nach dem **Prinzip der maximalen Mutmaßlichkeit** ist das wahrscheinlichste Ereignis das plausibelste. Man entscheidet sich deshalb für das Ereignis mit der maximalen **a-posteriori Wahrscheinlichkeit** $P(A_i|B)$!

6.3 Bayes'sche Formel

Beispiel

- **Gesamtproduktionsausschuss**
  Von welcher Maschine stammt ein Ausschussteil?
  Von Maschine 3 ($P(A_3|B) = 0.5128$)! Diese würde man also (zunächst) für jeglichen Ausschuss verantwortlich machen!

6.3 Zusammenfassung

Sätze (Fortsetzung)

Totale Wahrscheinlichkeit

Seien A_i, $i = 1, 2, \ldots, n$, eine Partition von Ω mit $P(A_i) > 0$. Dann gilt für jedes $B \in \mathcal{A}$:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Formel von Bayes

Für jedes $B \in \mathcal{A}$ mit $P(B) > 0$ gilt:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

6.4 Übungsaufgaben

Nehmen Sie an, dass zwei Personen viele Twitter-Nachrichten schreiben. Person T1 schreibt 5 Nachrichten pro Tag und Person T2 schreibt 45 Nachrichten pro Tag. Dabei sei die Wahrscheinlichkeit, dass es sich bei einer Nachricht um Fake-News F handelt, bei T1 80% und bei T2 nur 20%.

Wie hoch ist die Wahrscheinlichkeit, dass es sich bei einer zufällig ausgewählten Bachricht um Fake-News handelt?

Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Fake-News von Person T1 stammt?

6.4 Übungsaufgaben

Wie hoch ist die Wahrscheinlichkeit, dass es sich bei einer zufällig ausgewählten Bachricht um Fake-News handelt?

Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Fake-News von Person T1 stammt?

Berechnung mit Formel von der totalen Wahrscheinlichkeit und Formel von Bayes. Sei T_i, $i = 1, 2$ das Ereignis *Die Nachricht stammt von Person Ti* und F das Ereignis *Es handelt sich um Fake-News*. Dann gilt:

$$\begin{aligned} P(F) &= P(F|T_1) \cdot P(T_1) + P(F|T_2) \cdot P(T_2) = 0.8 \cdot 0.1 + 0.2 \cdot 0.9 \\ &= 0.26 \end{aligned}$$

$$P(T_1|F) = \frac{P(F|T_1) \cdot P(T_1 F)}{P(F)} = \frac{0.8 \cdot 0.1}{0.26} = \frac{0.08}{0.26} = \frac{4}{13} \approx 0.31$$

Das heißt, dass T1 nur 10% der Nachrichten schreibt, aber viele falsche, und wenn eine falsche Nachricht aufkommt, steigt die Wahrscheinlichkeit, dass die Nachricht von T1 kommt, von 10% auf 31%.

6.4 Übungsaufgaben

Zwei Taxigesellschaften sind in einer Stadt tätig. Die Taxis der Gesellschaft A sind grün, die der Gesellschaft B blau. Die Gesellschaft A stellt 15% der Taxis, die Gesellschaft B die verbleibenden 85%. Eines Nachts kommt es zu einem Unfall mit Fahrerflucht. Das fliehende Auto war ein Taxi. Ein Zeuge sagt aus, es habe sich um ein grünes Taxi gehandelt.

Das Gericht lässt den Zeugen auf seine Fähigkeit untersuchen, grüne und blaue Taxis unter nächtlichen Sichtbedingungen zu unterscheiden. Das Untersuchungsergebnis ist: In 80% der Fälle identifiziert der Zeuge die Farbe zutreffend, in 20% der Fälle irrt er sich.

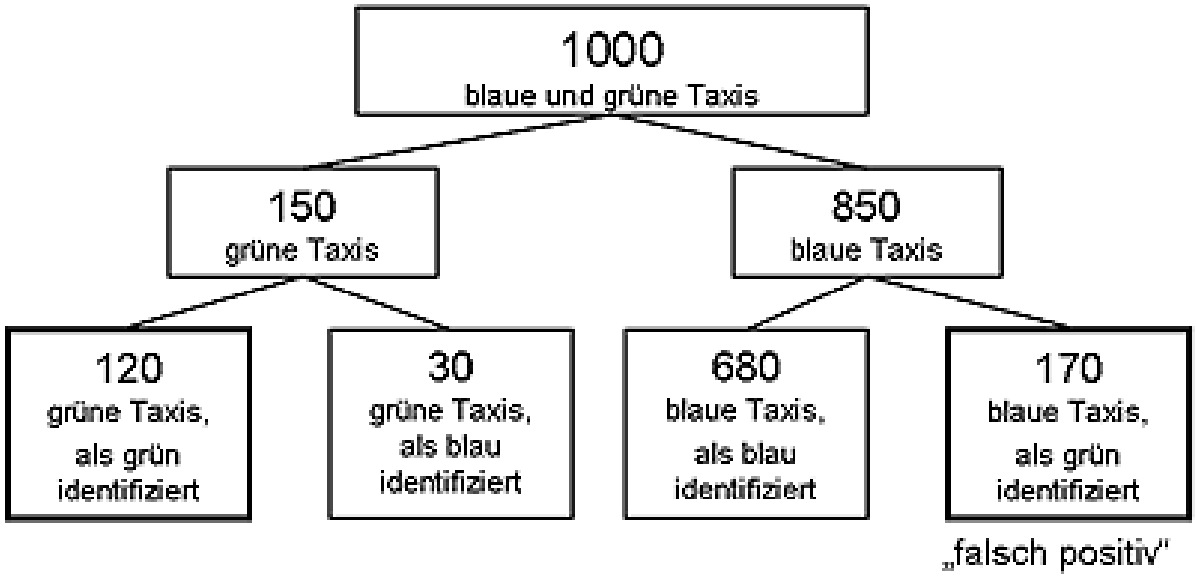
Wie hoch ist die Wahrscheinlichkeit, dass es sich bei dem fliehenden Taxi um ein Taxi der Gesellschaft A gehandelt hat?

6.4 Übungsaufgaben

Erste Möglichkeit:

Berechnung von Anzahlen über Baum, dann Wahrscheinlichkeit berechnen als Anzahl *günstiger* Fälle geteilt durch Anzahl *möglicher* Fälle. Der Baum ist unten abgebildet.

Die Wahrscheinlichkeit ergibt sich zu $\frac{120}{120+170} \approx 0.41$, also etwa 41%.



6.4 Übungsaufgaben

Zweite Möglichkeit:

Berechnung mit Hilfe der Formel von Bayes. Sei dazu A das Ereignis *Das fliehende Taxi ist ein Taxi der Gesellschaft A* und Z das Ereignis *Der Zeuge sagt aus, dass es sich um ein Taxi der Gesellschaft A handelt*. Dann gilt:

$$\begin{aligned} P(A|Z) &= \frac{P(Z|A) \cdot P(A)}{P(Z)} = \frac{P(Z|A) \cdot P(A)}{P(Z|A) \cdot P(A) + P(Z|\bar{A}) \cdot P(\bar{A})} \\ &= \frac{0.8 \cdot 0.15}{0.8 \cdot 0.15 + 0.2 \cdot 0.85} = \frac{0.12}{0.12 + 0.17} = \frac{12}{29} \approx 0.41 \end{aligned}$$

Bemerkung: Ohne Zeugenaussage beträgt die Wahrscheinlichkeit $P(A) = 0.15$.

7.1 Begriffsbildung

Idee:

Als Ergebnis von Zufallsexperimenten treten i.a. Zahlen auf, die bei der Wiederholung des Versuchs nicht gleich bleiben, sondern aufgrund der Wirkung des Zufalls „streuen“.

Diese sogenannten **Zufallsgrößen** kommen entweder direkt als Messergebnis des Zufallsexperiments vor (metrische Daten) oder ergeben sich indirekt, indem aus dem Ergebnis des Experiments ein interessierender Wert mittels einer genau definierten Bildungsvorschrift bestimmt wird.

Eine **Zufallsvariable** ist also eine Funktion, die einem Ergebnis eines Zufallsexperiments eine (reelle) Zahl zuordnet.

7.1 Begriffsbildung

Idee (Fortsetzung):

Anstatt mit den Wahrscheinlichkeiten selber wird i.a. mit Zufallsvariablen und deren „Verteilung“ gearbeitet. **Verteilungsfunktionen** beschreiben die Wahrscheinlichkeiten von gewissen Ereignissen, die mit Hilfe der Werte von Zufallsvariablen definiert wurden.

Sie haben den großen Vorteil, dass sie Funktionen von den reellen Zahlen (dem Wertebereich von Zufallsvariablen) in das Intervall $[0, 1]$ sind. Solche Funktionen kann man mit den Methoden der Analysis bearbeiten, man kann sie grafisch darstellen und mit ihnen rechnen.

Deshalb werden Zufallsexperimente i.a. durch **Annahmen über die Verteilung** einer passenden Zufallsvariablen modelliert, und die Beschreibung des Wahrscheinlichkeitsraums wird vollständig umgangen.

7.1 Begriffsbildung

Definition 49 (Zufallsvariable)

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine **Zufallsvariable** *ist eine Funktion von der Grundgesamtheit Ω in $\mathbb{R}$, die die* **Messbarkeitseigenschaft** *besitzt, dass nämlich jedes $A_r := \{\omega \mid X(\omega) \leq r\}$ für jedes $r \in \mathbb{R}$ in der σ-Algebra $\mathcal{A}$ liegt.*

Die **Messbarkeitseigenschaft** wird zur Definition der Verteilungsfunktion benötigt. Sie stellt selten eine Einschränkung dar, muss aber natürlich überprüft werden, wenn man sich für die Zufallsvariable selber interessiert und nicht nur für eine bestimmte Verteilung.

Zufallsvariablen werden im folgenden mit großen Buchstaben bezeichnet, ihre Werte, die sogenannten **Realisierungen**, mit entsprechenden kleinen Buchstaben: $x = X(\omega)$.

7.1 Begriffsbildung

Beispiele

- Werfen einer **Münze**: $\Omega = \{\omega_1, \omega_2\} = \{\text{Kopf, Zahl}\}$,
  $X(\omega) := 1$, wenn $\omega = \text{Kopf}$
  $X(\omega) := 0$, wenn $\omega = \text{Zahl}$.
- Werfen eines **Würfels**:
  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\} = \{1, 2, 3, 4, 5, 6\}$, $X(\omega) := \omega$
- Werfen von **2 verschiedenfarbigen Würfeln**:
  $\Omega = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (2,6), \ldots, (6,6)\}$
  $X(\omega_1, \omega_2) := \omega_1 + \omega_2$, $Y(\omega_1, \omega_2) := |\omega_1 - \omega_2|$
- **Körpergröße und Gewicht**:
  $\Omega = \{\omega = (\omega_K, \omega_G) \,|\, \omega_K, \omega_G > 0\}$, $X(\omega_K, \omega_G) := \omega_G$

7.1 Verteilungsfunktion

Definition 50

*Die (***kumulative***)* **Verteilungsfunktion** F_X *einer Zufallsvariable* X *ist definiert als:*
$F_X(x) := P(X \leq x) := P(\{\omega \mid X(\omega) \leq x\}) = P(A_x)$ *für jedes* $x \in \mathbb{R}$.

Bemerkungen

Die Verteilungsfunktion einer Zufallsvariablen beschreibt die Wahrscheinlichkeiten spezieller Mengen einer „zu der Variable gehörenden σ-Algebra“, die die Mengen $\{\omega \mid X(\omega) \leq x\}$ enthält (**Messbarkeitseigenschaft**!).

Offenbar werden mit wachsendem x die „Wahrscheinlichkeiten der darunter liegenden Werte **kumuliert**“.

7.1 Verteilungsfunktion

Achtung:

Verschiedene Zufallsvariablen können dieselbe Verteilungsfunktion haben (vgl. Beispiel mit 2 verschiedenfarbigen Würfeln auf folgender Folie).

Beispiele

- Werfen einer **Münze**:
  $\Omega = \{\omega_1, \omega_2\} = \{\text{Kopf, Zahl}\}$,
  Sei $X(\omega) :=$ Anzahl Köpfe (0 oder 1), dann gilt:
  $$F_X(x) = \begin{cases} 0 & \text{für } x < 0, \\ 0.5 & \text{für } 0 \leq x < 1, \\ 1 & \text{für } 1 \leq x \end{cases}$$
- Werfen eines **Würfels**:
  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\} = \{1, 2, 3, 4, 5, 6\}$, $X(\omega) := \omega$
  Wie sieht die Verteilungsfunktion aus? Übung!

7.1 Verteilungsfunktion

Beispiele (Fortsetzung)

- Werfen von **2 verschiedenfarbigen Würfeln**:
  $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(2,6),\ldots,(6,6)\}$
  Die zu $Y(\omega_1,\omega_2) := |\omega_1 - \omega_2|$ gehörende Verteilungsfunktion hat eine Form wie in der folgenden Skizze; denn es gibt
  6 „Würfelbilder“ mit absoluter Differenz $= 0$,
  10 mit abs. Diff. $= 1$ $((1,2),(2,3),(3,4),(4,5),(5,6)$ + Umkehrungen), ...,
  4 „Würfelbilder“ mit abs. Diff. $= 4$ $((1,5),(2,6)$ + Umkehr.)
  und 2 „Würfelbilder“ mit abs. Differenz $= 5$ $((1,6),(6,1))$.
  Außerdem gilt: Sei $X_k :=$ Wert auf dem k-ten Würfel, $k = 1,2$.
  Dann haben die X_k offenbar dieselbe Verteilungsfunktion.

7.1 Verteilungsfunktion

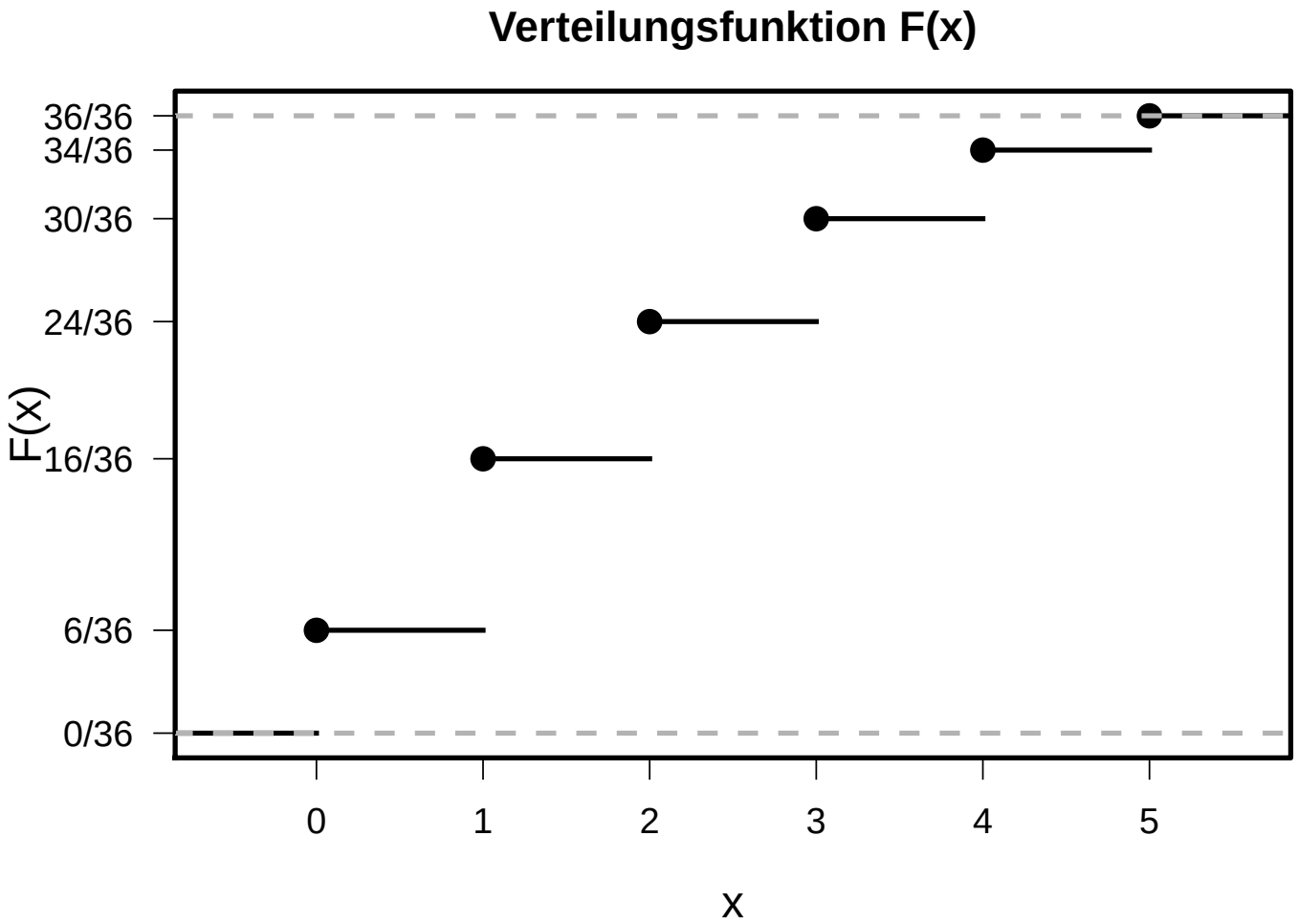


Abb. 54: Verteilungsfunktion

7.1 Verteilungsfunktion

Satz 23 (Eigenschaften der Verteilungsfunktion)

Sei F_X die Verteilungsfunktion einer Zufallsvariablen X. Dann gilt:

(i) $F_X(-\infty) := \lim_{x \to -\infty} F_X(x) = 0$ *und* $F_X(+\infty) := \lim_{x \to +\infty} F_X(x) = 1$

(ii) F_X *ist monoton wachsend:* $F_X(a) \leq F_X(b)$ *für* $a < b$

(iii) F_X *ist rechtsstetig:* $\lim_{0 < h \to 0} F_X(x + h) = F_X(x)$

Definition 51

*Jede Funktion von $\mathbb{R}$ in das Intervall $[0, 1]$ mit den Eigenschaften (i), (ii), (iii) aus obigem Satz definiert eine (***kumulative***) **Verteilungsfunktion**.*

7.2 Dichtefunktionen

Es gibt zwei Klassen von Zufallsvariablen, deren Verteilung sich besonders einfach durch sogenannte **Dichtefunktionen** beschreiben lässt:
Diskrete und stetige Zufallsvariablen.

7.2.1 Diskrete Zufallsvariablen

Definition 52

Eine **Zufallsvariable** X *heißt* **diskret**, *wenn sie höchstens abzählbar viele Werte* $\{x_1, x_2, x_3, \ldots\}$ *annimmt.*
Wenn eine Zufallsvariable diskret ist, dann heißt auch die dazugehörige **Verteilungsfunktion diskret**.

Für eine diskrete Verteilungsfunktion gilt:

$$F_X(x) = \begin{cases} 0, & x < x_i \text{ für alle } i, \\ \sum_{i=1}^{n} P(X = x_i), & x_n \leq x < x_{n+1}, \\ 1, & x \geq x_i \text{ für alle } i. \end{cases}$$

7.2.1 Diskrete Zufallsvariablen

Definition 53

Die Funktion

$$f_X(x) := P(X = x_i), \text{ falls } x = x_i, \ f_X(x) := 0, \text{ sonst},$$

heißt **diskrete Dichtefunktion** *(oder* **Zähldichte***) von* X*.*
Die Menge der x_i *mit* $f(x_i) > 0$ *heißt* **Träger** *von* f*. Diese* x_i *heißen auch* **Massenpunkte** *der Verteilung.*
für eine diskrete Verteilungsfunktion gilt offenbar:

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i).$$

7.2.1 Diskrete Zufallsvariablen

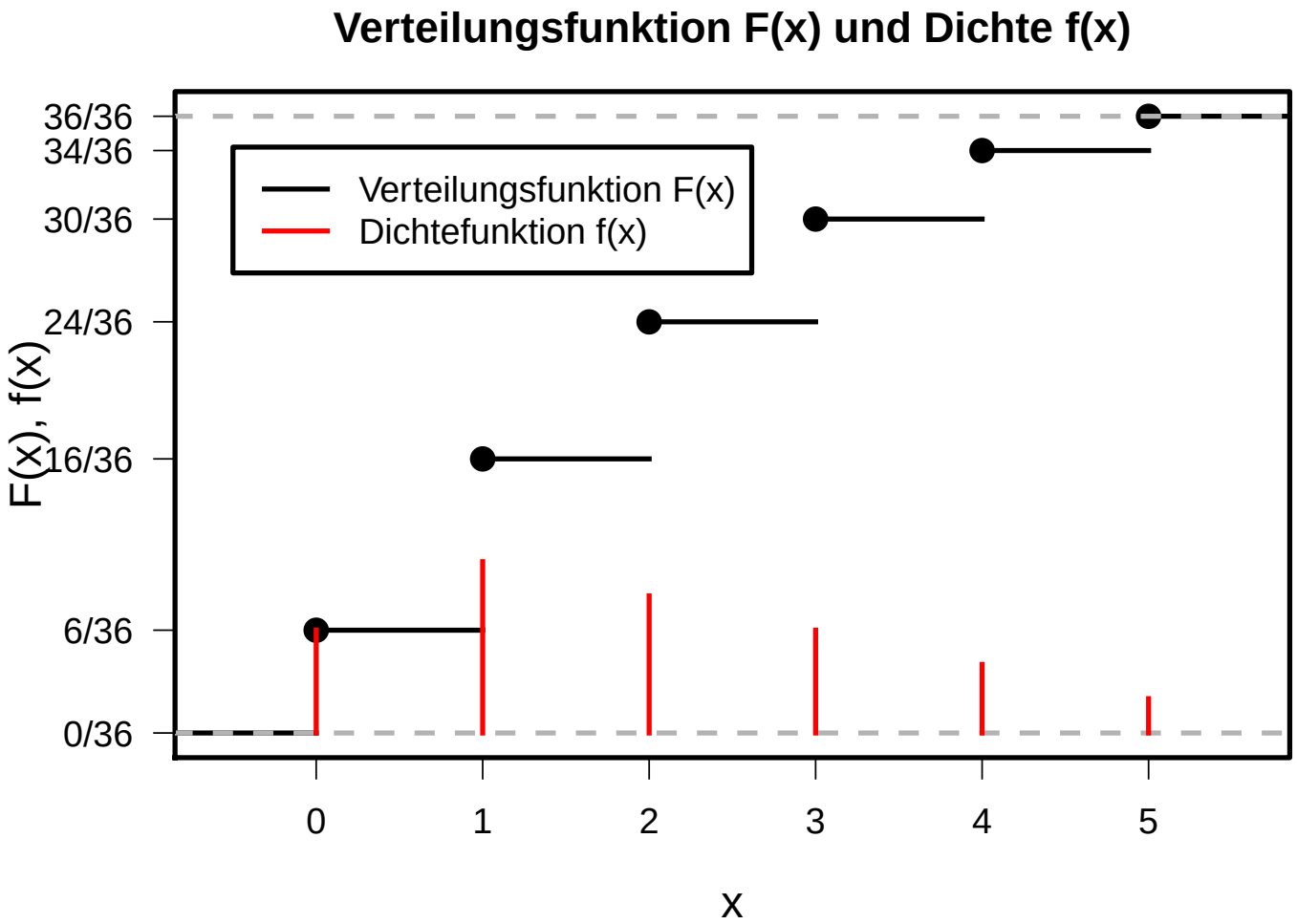


Abb. 55: Skizze: Diskrete Dichte

7.2.1 Diskrete Zufallsvariablen

Satz 24

Sei X eine diskrete Zufallsvariable, F_X und f_X die dazugehörige Verteilungsfunktion bzw. Dichtefunktion. Dann kann F_X aus f_X berechnet werden und umgekehrt.

Beweis.

Es bleibt, f_X aus F_X zu berechnen:
$f_X(x_i) = F_X(x_i) - \lim_{0<h\to 0} F_X(x_i - h)$ und
$f_X(x) = 0$ für $x \neq$ Massenpunkt. □

7.2.1 Diskrete Zufallsvariablen

Definition 54

Jede Funktion von den reellen Zahlen in das Intervall $[0, 1]$ *definiert eine* **diskrete Dichtefunktion**, *wenn für eine höchstens abzählbare Menge* $x_1, x_2, x_3, \ldots$ *gilt:*

(i) $f(x_i) > 0$ *für* $i = 1, 2, 3, \ldots$

(ii) $f(x) = 0$ *für* $x \neq x_i$, $i = 1, 2, 3, \ldots$

(iii) $\sum_i f(x_i) = 1$

7.2.2 Stetige Zufallsvariablen

Definition 55

Eine **Zufallsvariable** X*, die als Wertebereich (einen Teilbereich der) reelle(n) Zahlen hat, heißt* **stetig**, *wenn sich ihre Verteilungsfunktion wie folgt durch eine sogenannte* **Dichtefunktion** $f_X(x)$ *darstellen lässt:*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t.$$

7.2.2 Stetige Zufallsvariablen

Satz 25

Sei X eine stetige Zufallsvariable, F_X und f_X die dazugehörige Verteilungsfunktion bzw. Dichtefunktion. Dann kann F_X aus f_X berechnet werden und umgekehrt.

Beweis.

Es bleibt, f_X aus F_X zu berechnen:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

für die Punkte x, wo F_X differenzierbar ist. □

7.2.2 Stetige Zufallsvariablen

Bemerkungen

Intervalle von reellen Zahlen haben die Wahrscheinlichkeit

$$P((a,b]) = F(b) - F(a) = \int_a^b f_X(x)\,\mathrm{d}x, \text{ wenn } a < b.$$

Einzelne reelle Zahlen haben die Wahrscheinlichkeit 0:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \lim_{\Delta x \to 0} \frac{F_X(x+\Delta x) - F_X(x-\Delta x)}{2\Delta x}, \text{ also}$$

$$f_X(x)\,2\,\Delta x \approx F_X(x+\Delta x) - F_X(x-\Delta x) = P(x-\Delta x < X \leq x+\Delta x)$$

Wegen $\Delta x \to 0$ erhält man also $P(X = x) = 0$ für alle x, wo F_X differenzierbar ist.

Die **Fläche unterhalb der Dichtefunktion** ist

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = F_X(\infty) = 1$$

7.2.2 Stetige Zufallsvariablen

Definition 56

Jede Funktion $f : \mathbb{R} \to [0, \infty)$ *definiert eine* **Dichtefunktion**, *wenn gilt:*

(i) $f(x) \geq 0$ *für alle* x *und*

(ii) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$

7.2.3 Andere Zufallsvariablen

Zufallsvariablen müssen weder diskret noch stetig sein!

Diskrete Zufallsvariablen haben stückweise konstante Verteilungsfunktionen,
stetige Zufallsvariablen stetige Verteilungsfunktionen.

Es gibt aber auch Verteilungsfunktionen, die sowohl Sprünge aufweisen, als auch stetige Abschnitte (**gemischte Verteilungen**).

7.3.1 Erwartungswert

Definition 57

Der **Erwartungswert** μ_X *oder* $\mathrm{E}[X]$ *einer Zufallsvariable* X *ist definiert durch*

$$\mathrm{E}[X] = \sum_i x_i \, P(X = x_i) = \sum_i x_i \, f_X(x_i)$$

für diskretes X *mit Massenpunkten* $x_1, x_2, \ldots,$ *(vorausgesetzt die Reihe ist absolut konvergent)*

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x \, f_X(x) \, \mathrm{d}x$$

für stetiges X *mit Dichte* f_X *(vorausgesetzt das Integral existiert).*

7.3.1 Erwartungswert

Bemerkungen

Für diskrete Zufallsvariablen mit unendlich vielen Massenpunkten ist der Erwartungswert nur definiert, wenn obige Reihe **absolut konvergiert**. Sonst sagt man, dass der **Erwartungswert nicht existiert**.

Für stetige Zufallsvariablen wird ganz analog vorgegangen. Man sagt, dass der Erwartungswert nur existiert, wenn $\int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x$ existiert.

Falls nichts anderes erwähnt, wird im Folgenden angenommen, dass sämtliche Erwartungswerte existieren.

Für diskrete Zufallsvariablen mit endlich vielen Massenpunkten ist der Erwartungswert ein **gewichtetes arithmetisches Mittel** der Massenpunkte der Zufallsvariablen, wobei die Gewichte den Wahrscheinlichkeiten entsprechen, mit denen die Massenpunkte angenommen werden.

Der Erwartungswert ist **Schwerpunkt** der Massenverteilung, die durch die Dichtefunktion gegeben ist.

7.3.1 Erwartungswert

Beachte:

Bei der Definition des Erwartungswerts werden nur die Werte von Zufallsvariablen und ihre Wahrscheinlichkeit verwendet, nicht der Wahrscheinlichkeitsraum oder die Zuordnungsvorschrift der Zufallsvariable. Der Erwartungswert ist also eine **Eigenschaft einer Verteilung**, nicht einer Zufallsvariable!

Beispiel: Werfen von 2 verschiedenfarbigen Würfeln

$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(2,6),\ldots,(6,6)\}$
$X(\omega_1,\omega_2) := \omega_1 + \omega_2,\ \ Y(\omega_1,\omega_2) := |\omega_1 - \omega_2|$:

$$\mathrm{E}[X] = \sum_{i=2}^{12} i\, P(X=i) =$$

$$(2+12)\cdot\tfrac{1}{36} + (3+11)\cdot\tfrac{2}{36} + (4+10)\cdot\tfrac{3}{36} + (5+9)\cdot\tfrac{4}{36} + (6+8)\cdot\tfrac{5}{36} + 7\cdot\tfrac{6}{36} = 7$$

$$\mathrm{E}[Y] = \sum_{i=0}^{5} i\, P(Y=i) = 0\cdot\tfrac{6}{36} + 1\cdot\tfrac{10}{36} + 2\cdot\tfrac{8}{36} + 3\cdot\tfrac{6}{36} + 4\cdot\tfrac{4}{36} + 5\cdot\tfrac{2}{36} = \tfrac{70}{36}$$

7.3.1 Erwartungswert

Beachte:

Die Bezeichnung **Erwartungswert** ist insofern unglücklich, dass der berechnete Wert u.U. gar nicht realisiert werden kann (z.B. 70/36) und deshalb auch **nicht** „**erwartet**“ wird.

7.3.1 Erwartungswert

Beispiele (Fortsetzung)

- **Länge einer Telefonverbindung**
  Zufallsvariable $X :=$ Länge einer Telefonverbindung ≥ 0,
  $F_X(x) := (1 - e^{-\lambda x}),\ x \geq 0,\ \lambda > 0 \Rightarrow f_X(x) = \lambda\, e^{-\lambda x},\ x \geq 0.$

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, \mathrm{d}x = \int_{-\infty}^{+\infty} x\, \lambda e^{-\lambda x}\, \mathrm{d}x = \frac{1}{\lambda}$$

- **Erwartungswert existiert nicht**
  $f_X(x) := 1/x^2,\ x \in [1, \infty),\ f_X(x) := 0$, sonst, definiert eine Dichte „auf $[1, \infty)$“. Aber:

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, \mathrm{d}x = \int_{1}^{+\infty} x\, \frac{1}{x^2}\, \mathrm{d}x = \lim_{b \to \infty} \log b = \infty$$

7.3.1 Erwartungswert

Bemerkung

Der Erwartungswert einer Verteilung ist insofern ein **Lagemaß** der Verteilung, dass er einen „mittleren Wert“ der Verteilung darstellt. **Streuungsmaße** messen die Variabilität einer Verteilung „um ein Lagemaß herum“. Ein Beispiel für ein Streuungsmaß ist die Varianz.

7.3.2 Varianz

Definition 58

Sei X eine Zufallsvariable mit Erwartungswert μ_X. Die **Varianz** σ_X^2 *oder* $\mathrm{var}(X)$ *von X ist definiert durch*

$$\mathrm{var}(X) = \sum_i (x_i - \mu_X)^2 \, f_X(x_i)$$

für diskretes X mit Massenpunkten $x_1, x_2, \ldots,$

$$\mathrm{var}(X) = \int_{-\infty}^{+\infty} (x - \mu_X)^2 \, f_X(x) \, \mathrm{d}x$$

für stetiges X mit Dichte f_X.

7.3.2 Varianz

Bemerkungen

Die Varianz einer Zufallsvariablen ist der **Erwartungswert des quadrierten Abstandes von ihrem Erwartungswert**.

Für diskrete Zufallsvariablen mit endlich vielen Massenpunkten ist die Varianz ein **gewichtetes arithmetisches Mittel der quadrierten Abweichungen der Massenpunkte der Zufallsvariablen von ihrem Erwartungswert**, wobei die Gewichte den Wahrscheinlichkeiten entsprechen, mit denen die Werte angenommen werden.

Die Varianz ist das **Trägheitsmoment** der Massenverteilung, die durch die Dichtefunktion gegeben ist, bezogen auf eine Achse durch den Schwerpunkt.

7.3.2 Varianz

Definition 59

Sei X eine Zufallsvariable. Die **Standardabweichung** σ_X *von X ist definiert als*

$$\sigma_X = \sqrt{\text{var}(X)}.$$

Bemerkung

Die Standardabweichung hat, im Gegensatz zur Varianz, die **gleichen Einheiten** wie die Zufallsvariable X.

7.3.2 Varianz

Beispiele

- Werfen von **2 verschiedenfarbigen Würfeln**:
  $X(\omega_1, \omega_2) := \omega_1 + \omega_2$:
  $\mathrm{var}(X) = \sum_{i=2}^{12} (i-7)^2 \, P(X = i) =$
  $(25+25) \cdot \frac{1}{36} + (16+16) \cdot \frac{2}{36} + (9+9) \cdot \frac{3}{36} + (4+4) \cdot \frac{4}{36} + (1+1) \cdot \frac{5}{36} = \frac{210}{36}$
- **Länge einer Telefonverbindung**

$$\mathrm{var}(X) = \int_{-\infty}^{+\infty} (x - \mu_X)^2 \, f_X(x) \, \mathrm{d}x = \int_{-\infty}^{+\infty} \left(x - \frac{1}{\lambda} \right)^2 \lambda e^{-\lambda x} \, \mathrm{d}x = \frac{1}{\lambda^2}$$

$$\sigma_X = 1/\lambda = \mathrm{E}[X]$$

Bemerkung

Die Varianz ist insofern ein „sinnvolles“ **Streuungsmaß**, dass bei gleichem Erwartungswert diejenige Verteilung die kleinere Varianz hat, die „enger um den Erwartungswert herum“ liegt.

7.3.2 Funktionen von Zufallsvariablen

Satz 26 (Varianzen von linearen Tranformationen)

Sei X eine Zufallsvariable und a, b Konstanten, dann gilt:

$$\operatorname{var}(a + bX) = b^2 \operatorname{var}(X).$$

Beweis.

$$\begin{aligned}\operatorname{var}(a + bX) &= \mathrm{E}[(a + bX - \mathrm{E}[a - bX])^2] = \mathrm{E}[(a + bX - a - b\,\mathrm{E}(X))^2] \\ &= b^2\,\mathrm{E}[(X - \mathrm{E}(X))^2]\end{aligned}$$

□

7.3.2 Funktionen von Zufallsvariablen

Bemerkung: Standardisierung

Allgemein gilt:

$\mathrm{E}[a + bX] = a + b\mu_X,\ \mathrm{var}(a + bX) = b^2\,\mathrm{var}(X)$. Also gilt:

- $\mathrm{E}[X - \mu_X] = 0,\ \mathrm{var}(X - \mu_X) = \mathrm{var}(X)$: **Zentrieren**
- $\mathrm{E}[X/\sigma_X] = \mu_X/\sigma_X,\ \mathrm{var}(X/\sigma_X) = 1$: **Normieren**
- $\mathrm{E}[(X - \mu_X)/\sigma_X] = 0,\ \mathrm{var}((X - \mu_X)/\sigma_X) = 1$: **Standardisieren**

Die Zufallsvariable $(X - \mu_X)/\sigma_X$ hat also immer Erwartungswert 0 und Varianz 1.

7.3.3 Momente

Definition 60

*Sei X eine Zufallsvariable. Das r-**te Moment** μ'_r von X ist definiert als*

$$\mu'_r = \mathrm{E}[X^r].$$

Definition 61

*Sei X eine Zufallsvariable. Das r-**te zentrale Moment** μ_r von X ist definiert als*

$$\mu_r = \mathrm{E}[(X - \mu_X)^r].$$

7.3.3 Momente

Definition 62

Sei X eine Zufallsvariable. Das 3. zentrale Moment μ_3 von X heißt Asymmetriemaß bzw. **Schiefe** *von X.*
Der Quotient μ_3/σ_X^3 heißt **Schiefekoeffizient**.
Eine Verteilung heißt **linksschief** *(= rechtssteil) bzw.* **rechtsschief** *(=linkssteil), wenn $\mu_3 < 0$ bzw. > 0.*

Definition 63

Eine diskrete oder stetige Zufallsvariable X heißt **symmetrisch** *verteilt* **um ihren Erwartungswert**, *wenn für die dazugehörige Dichte gilt:*

$$f_X(\mu_X - x) = f_X(\mu_X + x) \text{ für alle } x \in \mathbb{R}.$$

7.3.3 Momente

Satz 27

für symmetrische diskrete oder stetige Zufallsvariablen gilt $\mu_3 = 0$.

Beweis.

$$\mathrm{E}[(X - \mu_X)^3] = \int_{-\infty}^{\mu_X} (x - \mu_X)^3 \, f_X(x) \, \mathrm{d}x + \int_{\mu_X}^{+\infty} (x - \mu_X)^3 \, f_X(x) \, \mathrm{d}x = 0$$

□

Bemerkung

Die Umkehrung gilt nicht!

7.3.3 Momente

Definition 64

Sei X eine Zufallsvariable. Das 4. zentrale Moment μ_4 von X bzgl. μ_X heißt **Wölbung** *oder* **Kurtosis** *von X.*
$(\mu_4/\sigma_X^4 - 3)$ *heißt* **Wölbungskoeffizient**.

Bemerkung

Der Wölbungskoeffizient ist nur interpretierbar bei symmetrischen Verteilungen. Es gilt:
Symmetrische Verteilungen mit negativem Wölbungskoeffizienten sind flacher in der Nähe des Erwartungswerts als die „Standard-Normalverteilung“, symmetrische Verteilungen mit positivem Wölbungskoeffizienten sind spitzer.

7.3.4 Quantile

Definition 65

*Sei X eine Zufallsvariable. Das q-**Quantil** ξ_q von X ist definiert als die kleinste Zahl ξ mit $F_X(\xi) \geq q$.*

Bemerkung

Falls X eine stetige Zufallsvariable ist, dann ist das q-Quantil die kleinste Zahl ξ mit $F_X(\xi) = q$.

7.3.4 Quantile

Definition 66

Sei X eine Zufallsvariable. Der **Median** med_X, $\mathrm{med}(X)$ *oder* $\xi_{0.5}$ *von X ist das* 0.5*-Quantil.*

Bemerkung (Bezeichnungsweise)

Falls X eine stetige Zufallsvariable ist, dann gilt:

$$\int_{-\infty}^{\mathrm{med}(X)} f_X(x)\,\mathrm{d}x = \frac{1}{2} = \int_{\mathrm{med}(X)}^{+\infty} f_X(x)\,\mathrm{d}x,$$

d.h. der Median (= Zentralwert) ist eine Zahl, so dass die **Hälfte der „Wahrscheinlichkeitsmasse“ links und die andere Hälfte rechts** davon liegt!

7.4.1 Diskrete Gleichverteilung

Definition 67

Jede diskrete Dichtefunktion der Art

$$f(x) = f(x; N) = \frac{1}{N}, \; x = 1, 2, \ldots, N \text{ und } f(x) = 0, \; \text{sonst},$$

wobei N eine natürliche Zahl ist, definiert die Dichte einer **diskreten Gleichverteilung**. *Eine Zufallsvariable mit einer solchen Dichte heißt* **diskret gleichverteilt**.

7.4.1 Diskrete Gleichverteilung

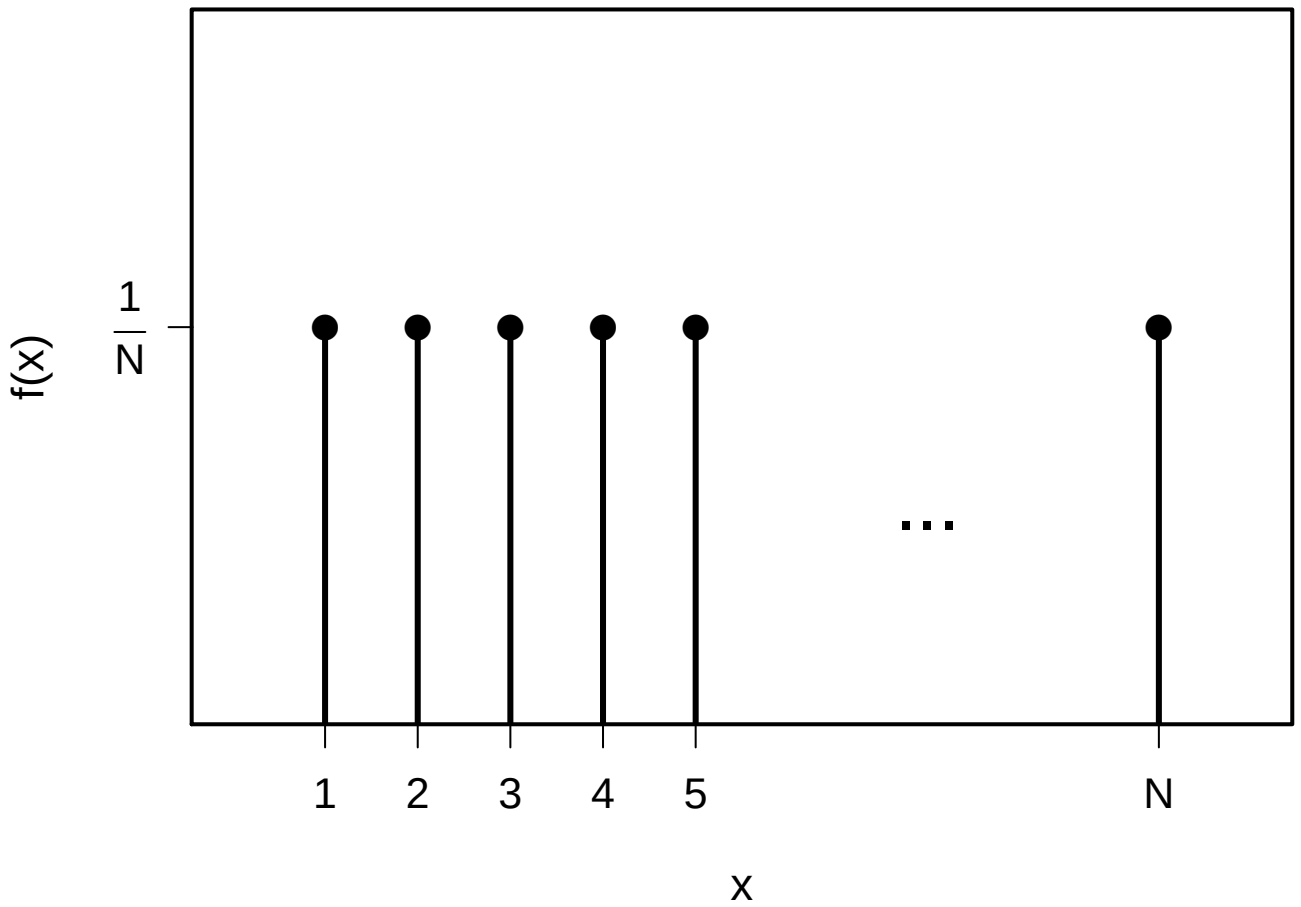


Abb. 56: Dichte einer diskreten Gleichverteilung

7.4.1 Diskrete Gleichverteilung

Satz 28

Sei X diskret gleichverteilt, dann gilt:

$$\mathrm{E}[X] = \frac{(N+1)}{2}, \ \mathrm{var}(X) = \frac{(N^2-1)}{12}$$

Beweis.

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{i=1}^{N} i \frac{1}{N} = \frac{(N+1)}{2}, \\
\mathrm{var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \sum_{j=1}^{N} \frac{j^2}{N} - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{N(N+1)(2N+1)}{6N} - \frac{(N+1)^2}{4} = \frac{(N+1)(N-1)}{12}
\end{aligned}$$

□

7.4.2 Bernoulli Verteilung

Definition 68

Jede diskrete Dichtefunktion der Art

$$f_x = f(x; p) = p^x(1-p)^{1-x} \text{ für } x = 0, 1 \text{ und } f(x) = 0, \text{ sonst},$$

wobei $0 \leq p \leq 1$, *definiert die Dichte einer* **Bernoulliverteilung** *(***mit Parameter** p*).*
Eine Zufallsvariable mit einer solcher Dichte heißt **bernoulliverteilt**.

Bezeichnung: $q := 1 - p$

7.4.2 Bernoulli Verteilung

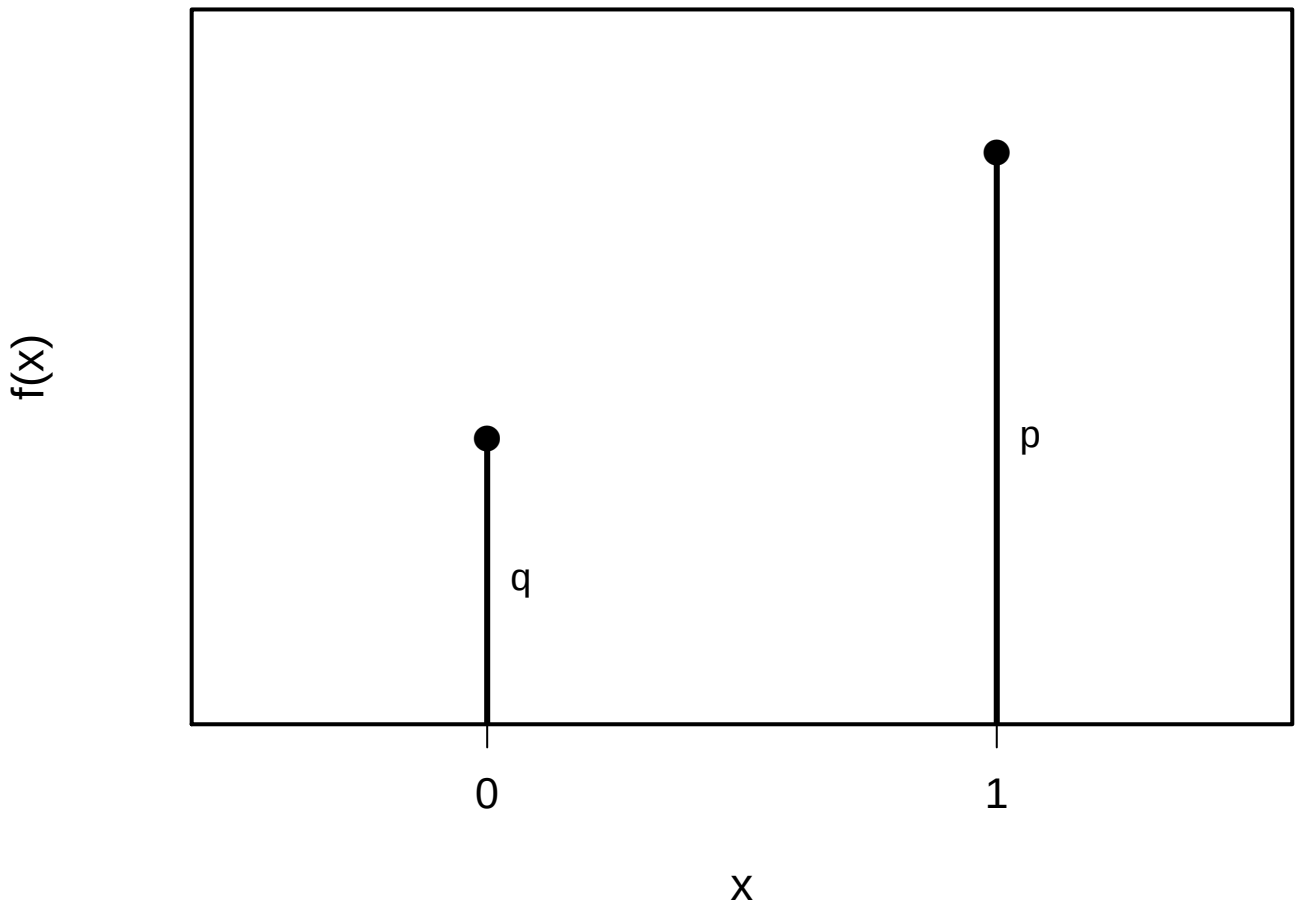


Abb. 57: Dichte einer Bernoulliverteilung (mit $q := 1 - p$)

7.4.2 Bernoulli Verteilung

Satz 29

Sei X bernoulliverteilt, dann gilt:

$$\begin{aligned} \mathrm{E}[X] &= p, \\ \mathrm{var}(X) &= pq \end{aligned}$$

Beweis.

$\mathrm{E}[X] = 0 \cdot q + 1 \cdot p = p,$

$\mathrm{var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = 0^2 q + 1^2 p - p^2 = pq$ □

7.4.2 Bernoulli Verteilung

Beispiele

- **Bernoulliexperiment**
  Das Versuchsergebnis kann ERFOLG oder MISSERFOLG sein. Eine Zufallsvariable mit dem Wert 1 bei Erfolg und dem Wert 0 bei Misserfolg hat eine Bernoulli Verteilung mit dem Parameter $p = P(\text{Erfolg})$.
- **Indikatorfunktion**
  Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und A in der σ-Algebra $\mathcal{A}$. Sei die Zufallsvariable X die Indikatorfunktion für A, d. h. $X(\omega) = 1$, falls $\omega \in A$, $X(\omega) = 0$, sonst. X ist bernoulliverteilt mit dem Parameter $p = P(X = 1) = P(A)$.

7.4.3 Binomial Verteilung

Definition 69

Jede diskrete Dichtefunktion der Art

$$f(x) = f(x; n, p) = \binom{n}{x} p^x q^{n-x} \text{ für } x = 0, 1, \ldots, n \text{ und } f(x) = 0, \text{ sonst},$$

wobei $n \in \mathbb{N}$, $0 \leq p \leq 1$ *und* $q := 1 - p$, *definiert die Dichte einer* **Binomialverteilung** *(***mit Parametern** n, p*). Eine Zufallsvariable mit einer solchen Dichte heißt* **binomialverteilt**.

7.4.3 Binomial Verteilung

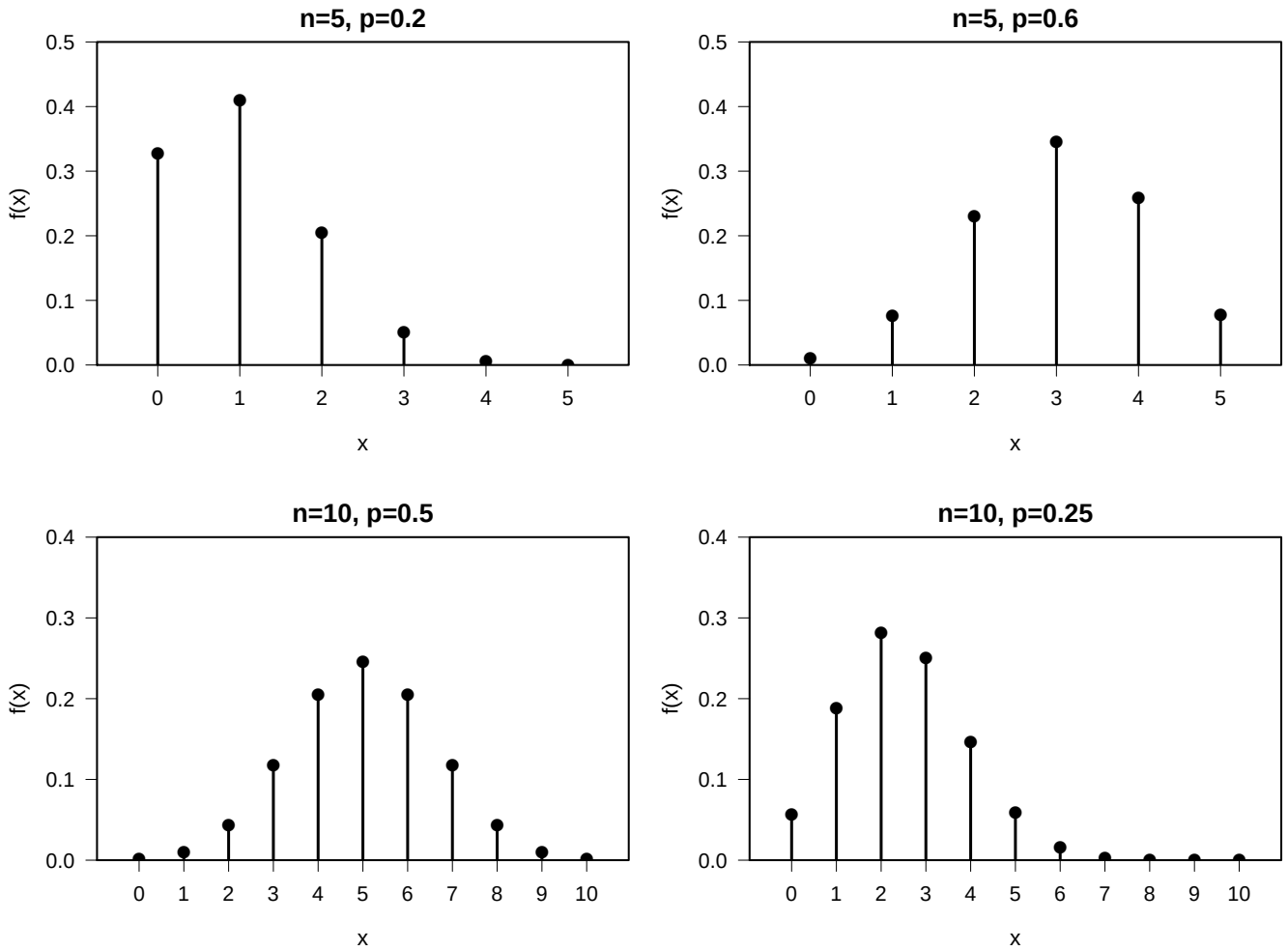


Abb. 58: Dichten von Binomialverteilungen

7.4.3 Binomial Verteilung

Satz 30

Sei X binomialverteilt, dann gilt:

$$\begin{aligned} \mathrm{E}[X] &= np, \\ \mathrm{var}(X) &= npq \end{aligned}$$

Beweis.

Beweisidee:
Eine binomialverteilte Zufallsvariable kann als Summe von n unabhängigen bernoulliverteilten Zufallsvariablen geschrieben werden. Wegen der Unabhängigkeit kann man sowohl für den Erwartungswert als auch für die Varianz dann mit n multiplizieren. □

7.4.3 Binomial Verteilung

Beispiele

- **Binomialexperiment**
  Betrachte das Zufallsexperiment, das aus n „unabhängigen“ Wiederholungen desselben Bernoulliexperiments besteht. Der Stichprobenraum hat also die Form:
  $\Omega = \{\omega = (\omega_1, \omega_2, \ldots, \omega_n) \,|\, \omega_i = \text{Erfolg oder } \omega_i = \text{Misserfolg}\}$.
  Da die Versuche unabhängig voneinander sind, errechnet sich die Wahrscheinlichkeit eines Ergebnisses des Gesamtexperiments durch die Multiplikation der Wahrscheinlichkeiten der Ergebnisse der Einzelexperimente.
  Sei die Zufallsvariable $X :=$ Anzahl Erfolge in n unabhängigen Bernoulliexperimenten. Dann gilt:
  $P(X = x) = \binom{n}{x} p^x q^{n-x}$ für $x = 0, 1, \ldots, n$,
  da die Wahrscheinlichkeit von x Erfolgen und $(n - x)$ Misserfolgen gleich $p^x q^{n-x}$ ist und $\binom{n}{x}$ verschiedene Kombinationen von Bernoulliexperimenten „erfolgreich“ sein können.

7.4.3 Binomial Verteilung

Beispiele (Fortsetzung)

- **Anzahl Defekte mit Zurücklegen**
  Beim Ziehen mit Zurücklegen aus einem Behälter mit K defekten und $(M - K)$ nicht defekten Teilen interessiert man sich für die Wahrscheinlichkeit von x defekten Teilen bei n Versuchen. Sei
  $p := P(\text{Erfolg}) := P(\text{Teil defekt bei einer Ziehung}) = K/M$.
  Sei $X :=$ Anzahl defekte Teile. Dann gilt:
  $P(X = x) = \binom{n}{x} \left(\frac{K}{M}\right)^x (1 - \frac{K}{M})^{n-x}$ für $x = 0, 1, \ldots, n$.

Bemerkung

Bernoulliverteilung $\hat{=}$ Binomialverteilung mit $n = 1$

7.5.1 Gleichverteilung (uniform distribution)

Definition 70

Eine stetige Dichtefunktion der Art

$$f(x) = f(x; a, b) = \frac{1}{b-a}, \ x \in [a, b], \ \textit{und } f(x) = 0, \ \textit{sonst},$$

wobei a, b reelle Zahlen sind, definiert die Dichte der **stetigen Gleichverteilung auf dem Intervall** $[a, b]$. *Eine Zufallsvariable mit einer solcher Dichte heißt (***stetig***)* **gleichverteilt**.

7.5.1 Gleichverteilung (uniform distribution)

Beispiel: Rundungsfehler

X = Fehler beim Runden von Messwerten auf 1 Stelle nach dem Komma:
Wertebereich: -0.05 bis $+0.05$

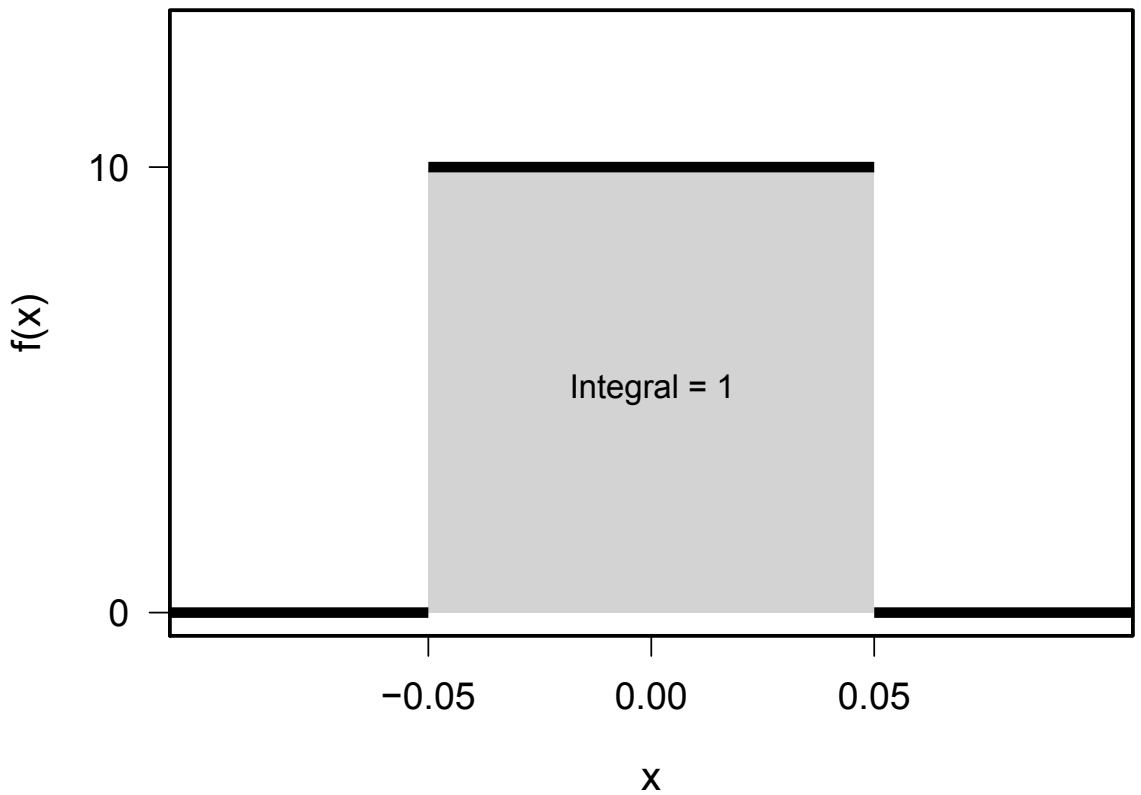


Abb. 59: Dichtefunktion

7.5.1 Gleichverteilung (uniform distribution)

Bemerkungen

- Wegen der Form der Dichte nennt man diese Verteilung auch oft **Rechteckverteilung**.
- Die Verteilung ist **wohldefiniert**, da die Fläche des Rechtecks offensichtlich $= 1$ ist.
- **Verteilungsfunktion einer stetigen Gleichverteilung**:

$$F(x) = \begin{cases} 0, & x < a, \\ \int_a^x \frac{1}{b-a} \, \mathrm{d}y = \frac{x-a}{b-a}, & x \in [a, b], \\ 1, & x > b. \end{cases}$$

7.5.1 Gleichverteilung (uniform distribution)

Satz 31

Sei X (stetig) gleichverteilt, dann gilt:

$$\mathrm{E}[X] = \frac{a+b}{2}, \ \mathrm{var}(X) = \frac{(b-a)^2}{12}$$

Beweis.

$$\begin{aligned}
\mathrm{E}[X] &= \int_a^b x \frac{1}{b-a} \, \mathrm{d}x = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2} \\
\mathrm{var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \int_a^b \frac{x^2}{b-a} \, \mathrm{d}x - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}
\end{aligned}$$

7.5.2 Normalverteilung

Definition 71

Eine stetige Dichtefunktion der Art

$$f(x) = f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

wobei $\sigma > 0$ *und* $\mu \in \mathbb{R}$*, definiert die Dichte der* **Normalverteilung** *oder auch* **Gauss-Verteilung** *mit den Parametern* μ, σ^2*. Eine Zufallsvariable* X *mit einer solchen Dichte heißt* **normalverteilt**.

Kurzschreibweise: $X \sim \mathcal{N}(\mu, \sigma^2)$

7.5.2 Normalverteilung

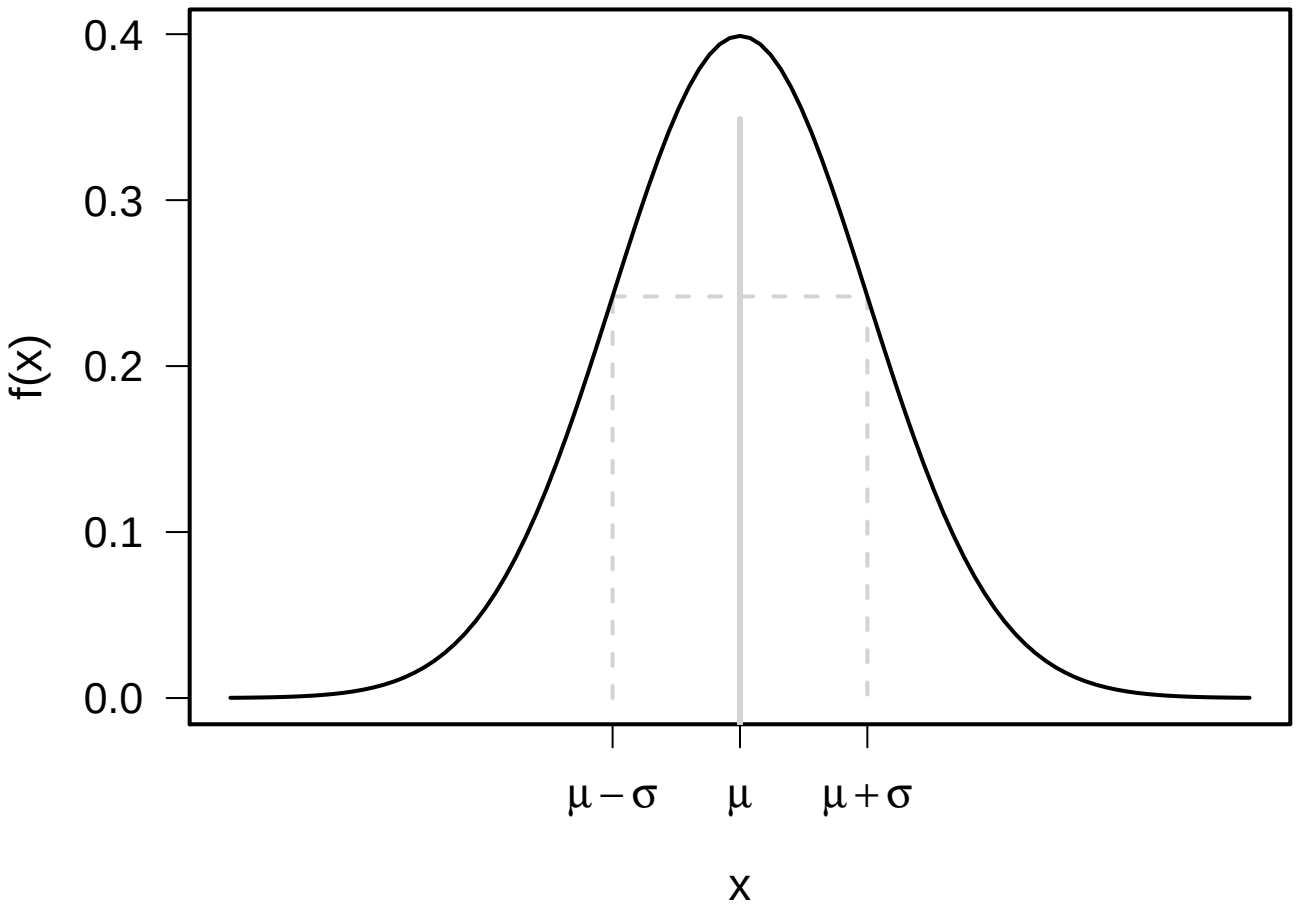


Abb. 60: Dichte der Normalverteilung

7.5.2 Normalverteilung

Bemerkung

für die **Verteilungsfunktion** $\Phi_{\mu,\sigma^2}(x)$ **einer Normalverteilung** gilt:

$$\Phi_{\mu,\sigma^2}(x) := \int_{-\infty}^{x_0} f(z)\,\mathrm{d}z = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}\mathrm{d}z = \int_{-\infty}^{\mu+\sigma y} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\mathrm{d}y$$

$$= \Phi_{(0,1)}\left(\frac{x-\mu}{\sigma}\right),\ y := \frac{z-\mu}{\sigma}$$

7.5.2 Normalverteilung

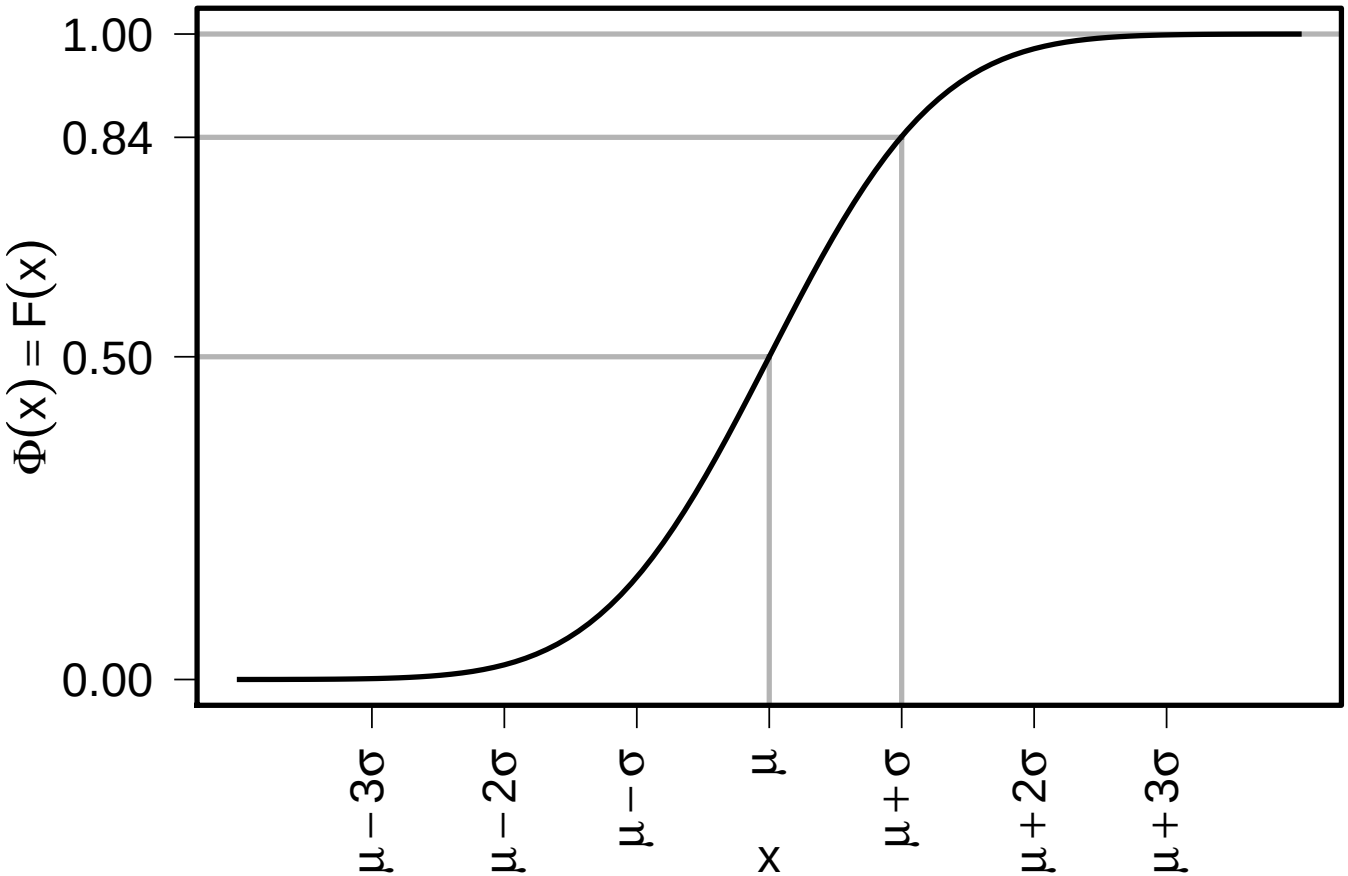


Abb. 61: Verteilungsfunktion der Normalverteilung

7.5.2 Normalverteilung

Bemerkungen

- Bei der Standardnormalverteilung $\mathcal{N}(0,1)$ gilt:
  Schiefe: $\mu_3 = 0$, **Wölbung:** $\mu_4 = 3$
- μ ist der **Erwartungswert** der Normalverteilung und bestimmt die **Lage** der Dichte $f(x)$.
- σ ist die **Standardabweichung** und bestimmt die **Streuung** der Verteilung bzw. die Breite der Dichte $f(x)$.
- σ^2 ist die **Varianz** der Normalverteilung.
- $X \sim \mathcal{N}(\mu, \sigma^2)$: normalverteilt mit Erwartungswert μ, Varianz σ^2
- **Standard-Normalverteilung**: $\mu = 0,\ \sigma = \sigma^2 = 1 :\ \mathcal{N}(0,1)$
- **Transformation auf Standard-Normalverteilung**:
  Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, dann ist $Y := \left(\frac{X-\mu}{\sigma}\right) \sim \mathcal{N}(0,1)$.
- **Bezeichnung:** $\Phi := \Phi_{0,1} : \mathcal{N}(0,1) -$ Verteilungsfunktion

7.5.2 Normalverteilung

Satz 32

Sei X normalverteilt, dann gilt:

$$\mathrm{E}[X] = \mu, \ \mathrm{var}(X) = \sigma^2$$

Satz 33

Sei X $\mathcal{N}(\mu, \sigma^2)$-verteilt, dann gilt:

$$P(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Bemerkungen

$\Phi(x) = 1 - \Phi(-x)$ wegen der Symmetrie um Null

$\Phi(x), \ x \geq 0$, ist vertafelt

7.5.2 Normalverteilung

Bemerkung

für Normalverteilungen $\mathcal{N}(\mu, \sigma^2)$ gilt:

- **Lagemaße** $\text{mod}_x = \text{med}_x = \mu$
- **Streuungsmaße** $qd \approx 1.35 \cdot \sigma_x$, also $1.5 \cdot qd \approx 2\sigma_x$

7.5.2 Normalverteilung

Bemerkung

Bei Normalverteilungen $\mathcal{N}(\mu, \sigma^2)$ werden häufig symmetrische Intervalle um den Erwartungswert verwendet, die den Wahrscheinlichkeiten 0.5, 0.68, 0.95, 0.999 entsprechen. Die folgende Tabelle gibt die ungefähre halbe Breite dieser Intervalle in Quartilsabständen und Standardabweichungen an.

Wahrscheinlichkeit	50 %	68 %	95 %	99.9 %
Faktor für *qd*	0.5	0.75	1.5	2.25
Faktor für x	0.68	1	2	3

Das bedeutet z.B., dass gilt:
$P([\mu - 2\sigma, \mu + 2\sigma]) \approx P([\mu - 1.5qd, \mu + 1.5qd]) \approx 0.95$.

Diese Intervalle sind insbesondere wegen ihrer einfachen Darstellung in der Praxis sehr beliebt.

7.6 Übungsaufgaben

Ein Würfel habe vier Seiten mit den Zahlen 0, 1, 2 und 5. Die Wahrscheinlichkeit, eine bestimmte Zahl zu würfeln, sei für alle Zahlen gleich groß.

a) Bestimmen Sie die Zähldichte der Zufallsvariablen X, die das Würfelergebnis bezeichnet.

b) Bestimmen und skizzieren Sie die Verteilungsfunktion von X.

c) Bestimmen Sie den Erwartungswert und die Varianz von X.

7.6 Übungsaufgaben

a) Bestimmen Sie die Zähldichte der Zufallsvariablen X, die das Würfelergebnis bezeichnet.

$$f(x) = \begin{cases} 1/4, & x \in \{0, 1, 2, 5\}, \\ 0, & \text{sonst.} \end{cases}$$

b) Bestimmen und skizzieren Sie die Verteilungsfunktion von X.

$$F(x) = \begin{cases} 0, & x < 0, \\ 1/4, & 0 \leq x < 1, \\ 1/2, & 1 \leq x < 2, \\ 3/4, & 2 \leq x < 5, \\ 1, & 5 \leq x. \end{cases}$$

Für die Skizze siehe Tafelbild.

7.6 Übungsaufgaben

c) Bestimmen Sie den Erwartungswert und die Varianz von X.

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_{i=1}^{4} x_i \, P(X = x_i) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 5 \cdot \frac{1}{4} \\
&= (0 + 1 + 2 + 5) \cdot \frac{1}{4} = \frac{8}{4} = 2 \\
\mathrm{var}(X) &= \sum_{i=1}^{4} (x_i - \mathrm{E}[X])^2 \, P(X = i) \\
&= (0 - 2)^2 \cdot \frac{1}{4} + (1 - 2)^2 \cdot \frac{1}{4} + (2 - 2)^2 \cdot \frac{1}{4} + (5 - 2)^2 \cdot \frac{1}{4} \\
&= (4 + 1 + 0 + 9) \cdot \frac{1}{4} = \frac{14}{4} = \frac{7}{2}
\end{aligned}
$$

7.6 Übungsaufgaben

Es sei X eine bimomialverteilte Zufallsvariable mit $n = 10$ und $p = 0.5$.

a) Bestimmen Sie den Erwartungswert von die Varianz von X und von $Y = 2 \cdot X + 3$.

b) Bestimmen und skizzieren Sie die Verteilungsfunktion von Y.

c) Standardisieren Sie die Zufallsvariablen X und Y.

7.6 Übungsaufgaben

a) Bestimmen Sie den Erwartungswert von die Varianz von X und von $Y = 2 \cdot X + 3$.

$$\begin{aligned} \mathrm{E}(X) &= n \cdot p = 10 \cdot 0.5 = 5 \\ \mathrm{E}(Y) &= \mathrm{E}(2 \cdot X + 3) = 2 \cdot \mathrm{E}(X) + 3 = 2 \cdot 5 + 3 = 13 \\ \mathrm{var}(X) &= n \cdot p \cdot (1 - p) = 10 \cdot 0.5 \cdot 0.5 = 2.5 \\ \mathrm{var}(Y) &= \mathrm{var}(2 \cdot X + 3) = 2^2 \cdot \mathrm{var}(X) = 4 \cdot 2.5 = 10 \end{aligned}$$

c) Standardisieren Sie die Zufallsvariablen X und Y.

$$\begin{aligned} \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{var}(X)}} &= \frac{X - 5}{\sqrt{2.5}} \\ \frac{Y - \mathrm{E}(Y)}{\sqrt{\mathrm{var}(Y)}} &= \frac{Y - 13}{\sqrt{10}} \end{aligned}$$

7.6 Übungsaufgaben

b) Bestimmen und skizzieren Sie die Verteilungsfunktion von Y.

Verteilungsfunktion von X:

$$F_X(x) = P(X \leq x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \text{ für } x \in \mathbb{R}$$

$$= \sum_{k=0}^{\lfloor x \rfloor} \binom{10}{k} 0.5^k 0.5^{10-k} = \sum_{k=0}^{\lfloor x \rfloor} \binom{10}{k} 0.5^{10}$$

Verteilungsfunktion von Y:

$$F_Y(y) = P(Y \leq y) = P(2 \cdot X + 3 \leq y) = P\left(X \leq \frac{y-3}{2}\right)$$

$$= \sum_{k=0}^{\lfloor \frac{y-3}{2} \rfloor} \binom{10}{k} 0.5^{10}$$

8.1 Schätzen

Definition: Sei $X_1, \ldots, X_N$ eine Zufallsstichprobe mit der Dichte $f_X(x, \theta)$. Sei $\tau(\theta)$ eine Funktion der unbekannten Parameter $\theta = (\theta_1, \ldots, \theta_K)$.

- Ein **Punktschätzer** ist eine Statistik $T(X_1, \ldots, X_N)$, deren Realisierung benutzt wird, um die Unbekannte $\tau(\theta)$ möglichst gut zu repräsentieren.
- Ein **Intervallschätzer** ist ein Paar von Statistiken $T_1(X_1, \ldots, X_N)$ und $T_2(X_1, \ldots, X_N)$ mit $T_1(X_1, \ldots, X_N) < T_2(X_1, \ldots, X_N)$, so dass $P_\theta(T_1(X_1, \ldots, X_N) < \tau(\theta) < T_2(X_1, \ldots, X_N)) = \gamma$, wobei $0 < \gamma < 1$ eine vorgegebene Wahrscheinlichkeit ist.
- $\gamma = 1 - \alpha$ heißt **Konfidenzniveau** (Vertrauensniveau), T_1 und T_2 **untere** bzw. **obere Konfidenzgrenzen** (Vertrauensgrenzen) für $\tau(\theta)$.

8.1 Schätzen

- Ein Intervall $(T_1(x_1, \ldots, x_N), T_2(x_1, \ldots, x_N))$ von Realisierungen eines Intervallschätzers heißt **zweiseitiges** $100\gamma\%$**-Konfidenzintervall für** $\tau(\theta)$.
- Ein Punktschätzer $T(X_1, \ldots, X_N)$ heißt **unverzerrter (erwartungstreuer)** Schätzer für $\tau(\theta)$, wenn gilt: $E_\theta[T] = E_\theta[T(X_1, \ldots, X_N)] = \tau(\theta)$.
- Ein unverzerrter Schätzer $T(X_1, \ldots, X_N)$ für $\tau(\theta)$ heißt **bester unverzerrter Schätzer**, wenn für alle θ gilt, dass $\text{var}_\theta(T) = E_\theta[(T - \tau(\theta))^2]$ minimal ist für alle unverzerrten Schätzer.

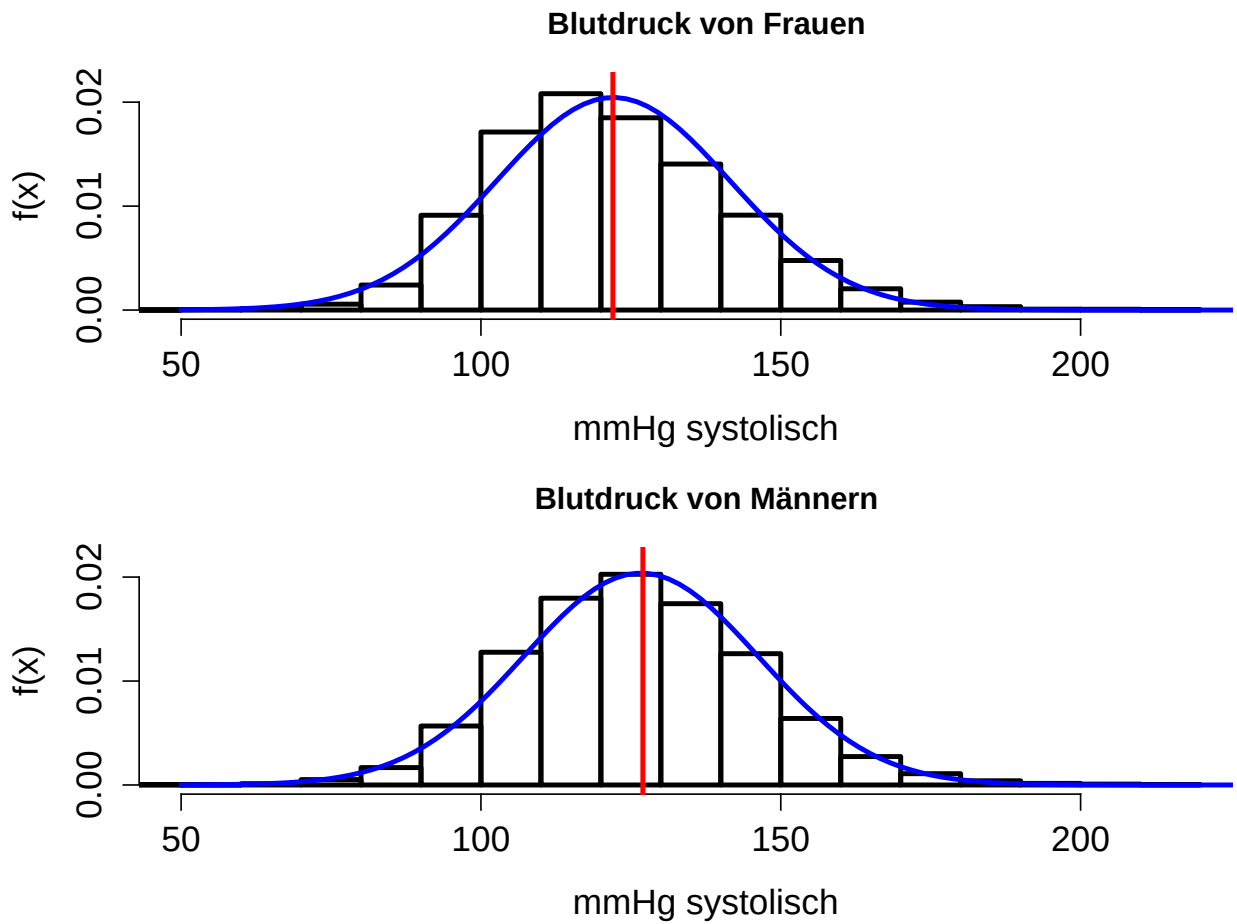
8.1 Schätzen am Beispiel

Beispiel: Schätzer für Blutdruckwerte von Frauen und Männern in einem Medizindatensatz

```
FRAUEN: Schätzer aus der Stichprobe (n=9190)
            Mittel = 122.07; Standardabw. = 19.51

   Quantile 0%  2.5%   25%   50%   75% 97.5%  100%
            43    89   108   120   134   164   217
   Quantile einer $N(122.07, 19.51^2)$ Verteilung
          -Inf    84   109   122   135   160   Inf

MÄNNER: Schätzer aus der Stichprobe (n=7173)
            Mittel = 126.68; Standardabw. = 19.57

   Quantile 0%  2.5%   25%   50%   75% 97.5%  100%
            46    91   113   126   139   167   213
   Quantile einer $N(126.68, 19.57^2)$ Verteilung
          -Inf    88   113   127   140   165   Inf
```

8.1 Schätzen am Beispiel

Beispiel: Schätzer für Blutdruckwerte von Frauen und Männern im Medizindatensatz



8.1 Schätzen am Beispiel

Das $1 - \alpha$ Konfidenzintervall für μ bei unbekanntem σ bei unabhängig identisch $\mathcal{N}(\mu, \sigma)$-verteilten Zufallsvariablen X_i:

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$

wobei s der Schätzer der Standardabweichung und $t_{n-1;1-\alpha/2}$ das $1 - \alpha/2$ Quantil einer t-Verteilung mit $n - 1$ Freiheitsgraden ist.

8.1 Schätzen am Beispiel

Das Konfidenzintervall für den Blutdruck von Frauen bei $\alpha = 0.05$ ist damit konkret:

$$\left[122.07 - t_{9189;0.975}\frac{19.51}{95.864} \quad ; \quad 122.07 + t_{9189;0.975}\frac{19.51}{95.864}\right]$$

$$= \left[122.07 - 1.96\frac{19.51}{95.864} \quad ; \quad 122.07 + 1.96\frac{19.51}{95.864}\right]$$

$$= [121.67 \quad ; \quad 122.47]$$

Das entsprechende Konfidenzintervall für Männer:

$$\left[126.68 - t_{7172;0.975}\frac{19.57}{84.69} \quad ; \quad 126.68 + t_{7172;0.975}\frac{19.57}{84.69}\right]$$

$$= [126.23 \quad ; \quad 127.13]$$

8.2 Testen

Definitionen:

- Eine **statistische Hypothese** H_0 für einen unbekannten Parameter $\theta = (\theta_1, \ldots, \theta_K)$ einer Verteilung ist eine Annahme über die Verteilung einer Zufallsvariablen.
- Ein **Test** einer statistischen Hypothese H_0 ist eine Regel oder Prozedur zur Entscheidung über die **Ablehnung** (bzw. Nicht-Ablehnung) der statistischen Hypothese.
- Man spricht von einem **Typ I Fehler**, wenn die Hypothese H_0 abgelehnt wird, obwohl sie richtig ist.
- Die Größe des Typ I Fehlers wird **Signifikanzniveau** (α) des Tests genannt.
- Ist die Wahrscheinlichkeit, bei der Testentscheidung den Typ I Fehler zu begehen, kleiner oder gleich α, so lehnt man den Test ab und spricht von einem **signifikanten** Ergebnis.
- Man spricht von einem **Typ II Fehler**, wenn die Hypothese H_0 nicht abgelehnt wird, obwohl sie falsch ist.

8.2 t-Test

Beispiel: Unterscheiden sich die Blutdruckwerte von Frauen und Männern?

- ***t*-Test:** (Ein-Stichproben-Fall)
  Falls die X_i unabhängig $\mathcal{N}(\mu, \sigma^2)$ normalverteilt sind mit unbekannter Varianz, dann gilt:

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}, \quad t \sim t_{n-1},$$

  wobei s der übliche Schätzer der Standardabw. σ ist.
  Man spricht: Die **Teststatistik** t ist t-verteilt mit $n-1$ Freiheitsgraden.
- Dies kann zur Überprüfung von Hypothesen genutzt werden, wie z.B.: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

8.2 t-Test

- t-**Test:** (Zwei-Stichproben-Fall)
  Analog zum Ein-Stichproben-Fall kann die Teststatistik

$$t = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{s_X^2/n + s_Y^2/m}}$$

für den Vergleich zweier Erwartungswerte mit unbekannten Varianzen verwendet werden, wobei s_X und s_Y die üblichen Schätzer der Standardabweichung sind und n und m die Stichprobenumfänge. Unter $\mu_X - \mu_Y = \delta_0$ ist t wieder t-verteilt mit k Freiheitsgraden, die gegeben sind durch:

$$k = \left\lfloor \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_Y^2}{m}\right)^2} \right\rfloor .$$

8.2 t-Test

- mögliche Hypothesen, Alternativen und Ablehnungsbereich der t-Tests:

(a) $H_0 : \mu_X - \mu_Y = \delta_0$ (zweiseitig)
$H_1 : \mu_X - \mu_Y \neq \delta_0$
ablehnen, wenn: $|t| > t_{1-\alpha/2}(k)$

(b) $H_0 : \mu_X - \mu_Y \geq \delta_0$ (einseitig)
$H_1 : \mu_X - \mu_Y < \delta_0$
ablehnen, wenn: $t < -t_{1-\alpha}(k)$

(c) $H_0 : \mu_X - \mu_Y \leq \delta_0$ (einseitig)
$H_1 : \mu_X - \mu_Y > \delta_0$
ablehnen, wenn: $t > t_{1-\alpha}(k)$

8.2 t-Test am Beispiel

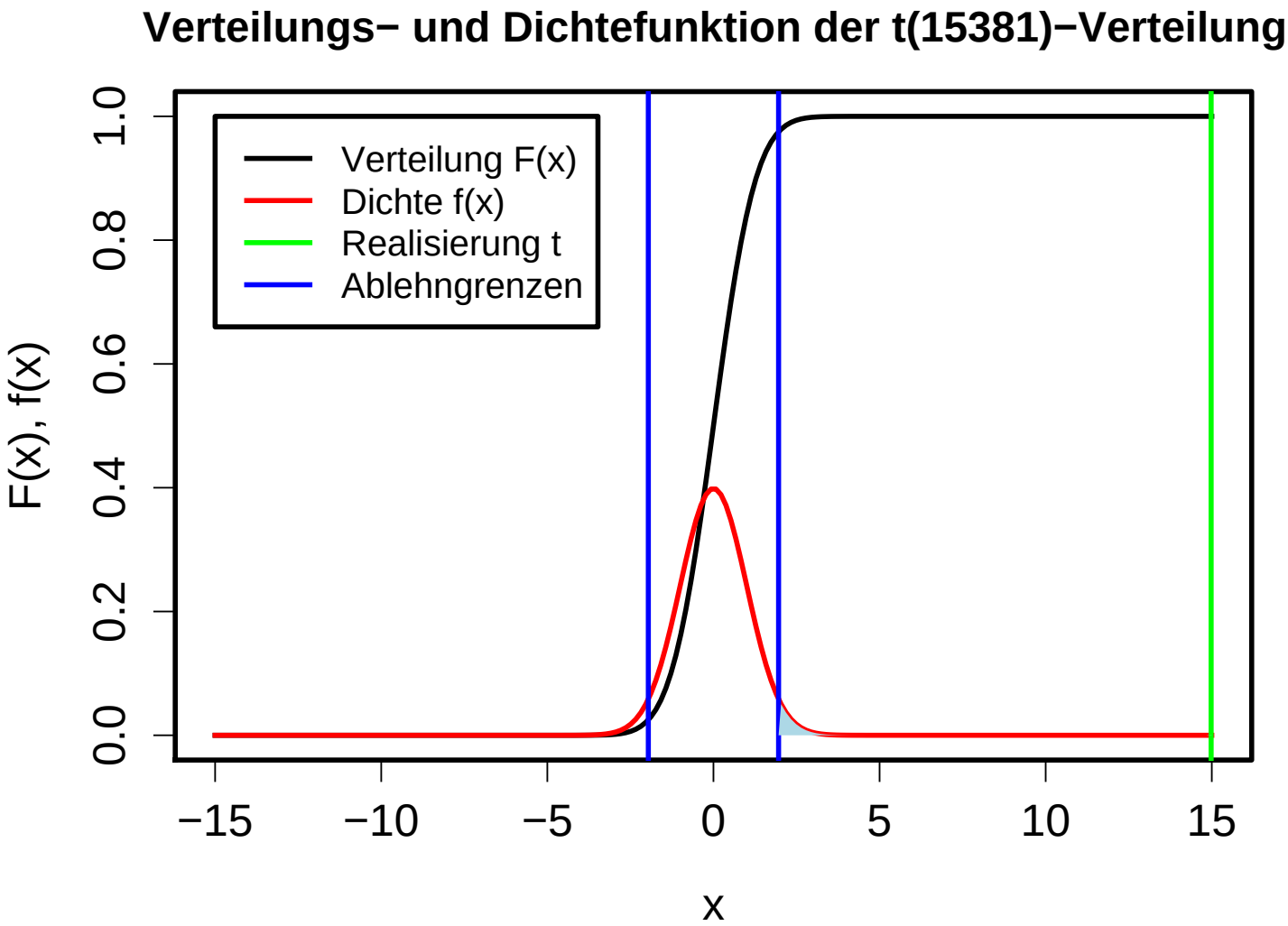
Beispiel: Unterscheiden sich die Blutdruckwerte von Frauen und Männern?
Einsetzen:

$$
\begin{aligned}
t &= \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{s_X^2/n + s_Y^2/m}} \\
&= \frac{(122.07 - 126.68) - 0}{\sqrt{380/9190 + 383/7173}} = \frac{-4.61}{\sqrt{0.095}} = -14.98. \\
k &= \left\lfloor \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_Y^2}{m}\right)^2} \right\rfloor \\
&= \left\lfloor \frac{\left(\frac{380}{9190} + \frac{383}{7173}\right)^2}{\frac{1}{9189}\left(\frac{380}{9190}\right)^2 + \frac{1}{7172}\left(\frac{383}{7173}\right)^2} \right\rfloor = 15381
\end{aligned}
$$

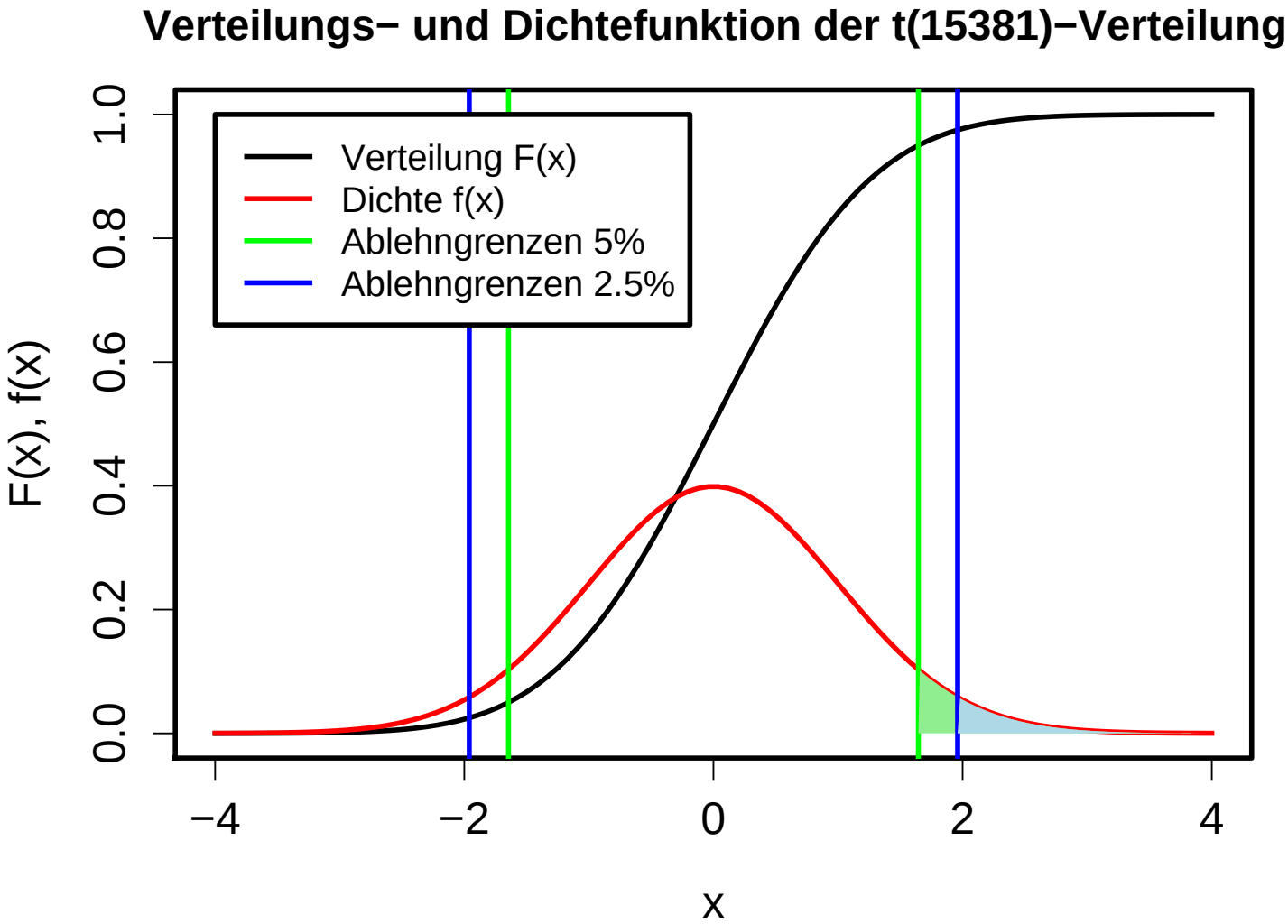
8.2 t-Test am Beispiel

$$
\begin{aligned}
H_0 : \mu_X - \mu_Y = \delta_0 \quad &\rightarrow \quad H_0 : \mu_X = \mu_Y \\
H_1 : \mu_X - \mu_Y \neq \delta_0 \quad &\rightarrow \quad H_1 : \mu_X \neq \mu_Y \\
|t| > t_{1-\alpha/2}(k) \quad &\rightarrow \quad 14.98 > t_{1-0.05/2}(15381) \\
& \qquad\quad = t_{0.975}(15381) = 1.96
\end{aligned}
$$

8.2 t-Test am Beispiel



8.2 t-Test am Beispiel



8.2 Test

Offensichtlich muss man sich Gedanken machen, ob folgende Einstellungen des Tests vernünftig sind:

- gerichtete (?) Nullhypothese
- Relevanz, nicht alleine Signifikanz
- Stichprobengröße
- Verteilungsannahme
- Gibt es bereits bekannte Parameter?

8.2 Wilcoxon / Mann-Whitney-Tests

Eine (zu?) oft verwendete Alternative des t-Tests auf Lageunterschiede ist der Wilcoxon-Rangsummentest oder Mann-Whitney-(U)-Test.

- Die Wilcoxon-Rangsummenstatistik ist

$$W_{m,n} = \sum_{i=1}^{m} R(X_i)$$

  mit $R(X_i)$ der Rang von X_i in der gepoolten, geordneten Stichprobe.
- Der Test ist verteilungsfrei und wird bei Abweichung von der Normalverteilung häufig als Alternative zum t-Test verwendet, er darf allerdings nur bei symmetrischer Verteilung verwendet werden (Voraussetzung!), was dann meist auch nicht erfüllt ist!
- Die exakte Verteilung von $W_{m,n}$ unter der Bedingung der Nullhypothese kann mittels kombinatorischer Überlegungen mathematisch leicht gefunden werden, ist aber für große m und n kaum beherrschbar.
- Für $m > 25$ oder $n > 25$ kann die Teststatistik durch die Normalverteilung $\left(N\left(\frac{m\,(n+m+1)}{2}; \frac{n\,m\,(n+m+1)}{12}\right)\right)$ approximiert werden.

8.2 Fisher Test (Unabhängigkeit i.d. Kontingenztafel)

Bei dem exakten Test von Fisher werden in der Kontingenztafel Zeilen- und Spaltensummen angeschaut: Man berechnet die bedingte Wahrscheinlichkeit für die Zellhäufigkeiten, gegeben die Randsummen.

- Es werden die Wahrscheinlichkeiten für den vorgegebenen Fall und für die extremeren Fälle berechnet und addiert.
- Diese Wahrscheinlichkeiten folgen, wenn die Nullhypothese stimmt, einer hypergeometrischen Verteilung:
- Die berechnete Wahrscheinlichkeit ist der p-Wert.

8.2 Fisher Test (Unabhängigkeit i.d. Kontingenztafel)

Erinnerung:

Tab. 15: Ergebnis des Experiments von Fisher

Tatsächlich	Beurteilung Muriel	
	Milch-Tee	Tee-Milch
Milch-Tee	3	1
Tee-Milch	1	3

p-Wert am Beispiel:

$$p = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} \approx 0.24$$

8.2 χ^2 Test

In Kontingenztafeln mit mehr Spalten oder Zeilen oder sehr großen Besetzungszahlen ist es schwierig oder gar unmöglich, einen exakten Test zu verwenden.

Dann geht man über zum χ^2 Test, der

- den bereits kennengelernten χ^2 Kontingenzkoeffizienten als Teststatistik verwendet,
- welcher χ^2-verteilt ist mit m und n Freiheitsgraden.
- Es muss also nur noch mit dem entsprechenden Quantil der χ^2-Verteilung verglichen werden, um zu einer Testentscheidung zu kommen.

9.1 Statistische Grafik

Statistische Grafik dient dazu

- Informationen in Daten zu finden
- Zusammenhänge in Daten zu erkennen
- Daten schnell und ohne Worte zu beschreiben
- Informationen ohne Worte schnell an andere weiterzugeben

Informationen in Grafiken sollen neutral und objektiv weitergegeben werden. Das ist schwierig, weil

- der Statistiker schon ein Vorurteil hat,
- der Kunde schon einen konkreten Wunsch zum Ergebnis hat,
- die Wahl von Farben, Anordnungen usw. die Wahrnehmung (z.T. unbewusst oder unterbewusst) beeinflusst.

9.1 Statistische Grafik

Anforderungen an statistische Grafik sind

- vollständige, selbsterklärende Beschriftung (siehe Seite 386)
- gute Lesbarkeit
- Vergleichbarkeit
- Objektivität

Viele Grafiken, die man im täglichen Leben sieht, z.B. in Zeitungen, Zeitschriften, Werbematerial, sind leider oft *verfälschend*, weil

- sie ohne besseres Wissen und Reflektion unabsichtlich so erstellt werden,
- sie absichtlich den Blick des Betrachters auf bestimmte Teilgebiete lenken oder von anderen ablenken sollen.

9.1 Statistische Grafik

Die **Beschriftung einer Grafik** ist von zentraler Bedeutung für gute Datenanalyse und Präsentationen.

Zu einer geeigneten Beschriftung zählen:

- Beschriftungen an den Achsen (mit Einheit!)
- Titel und eventuell Untertitel
- Beschriftung muss gut lesbar sein (Größe und Schriftart)
- Inhalt einer Grafik soll verstanden werden können, ohne den Fließtext zu lesen
- unterschiedliche Farben, Symbole und Linienarten
  - sollen so gewählt werden, dass sie gut unterscheidbar sind
  - sollen aber auch nicht von anderen Farben, Symbolen und Linienarten ablenken
  - müssen in einer Legende erklärt werden

9.2 Gute und schlechte statistische Grafik

Die folgenden Beispiele sollen dazu dienen, den Blick für Probleme zu schärfen.

Als abschreckende Beispiele haben wir bereits die schlechte Vergleichbarkeit der Kreissegmente bei Kreisdiagrammen und die wegen der Perspektive stark verfälschende Darstellung bei den 3D Tortendiagrammen gesehen (S. 43 ff.).

9.2 Gute und schlechte statistische Grafik

Beispiel 7 (Umsatzentwicklung zweier Firmen)

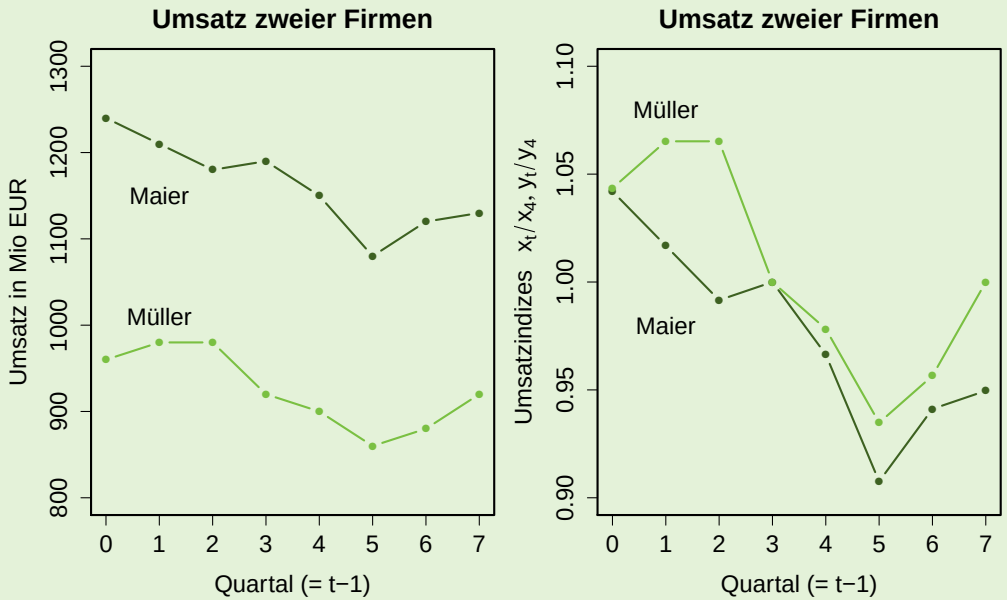


Abb. 62: Umsatzverläufe zweier Firmen, einmal in Mio EUR und einmal skaliert auf den Wert in Quartal 4

Auf der nächsten Folie werden mit zwei Grafiken die Entwicklung des Umsatzes von Firma Maier vor und nach Quartal 5 dargestellt – mit unterschiedlicher Skalierung der vertikalen (y) Achse.

9.2 Gute und schlechte statistische Grafik

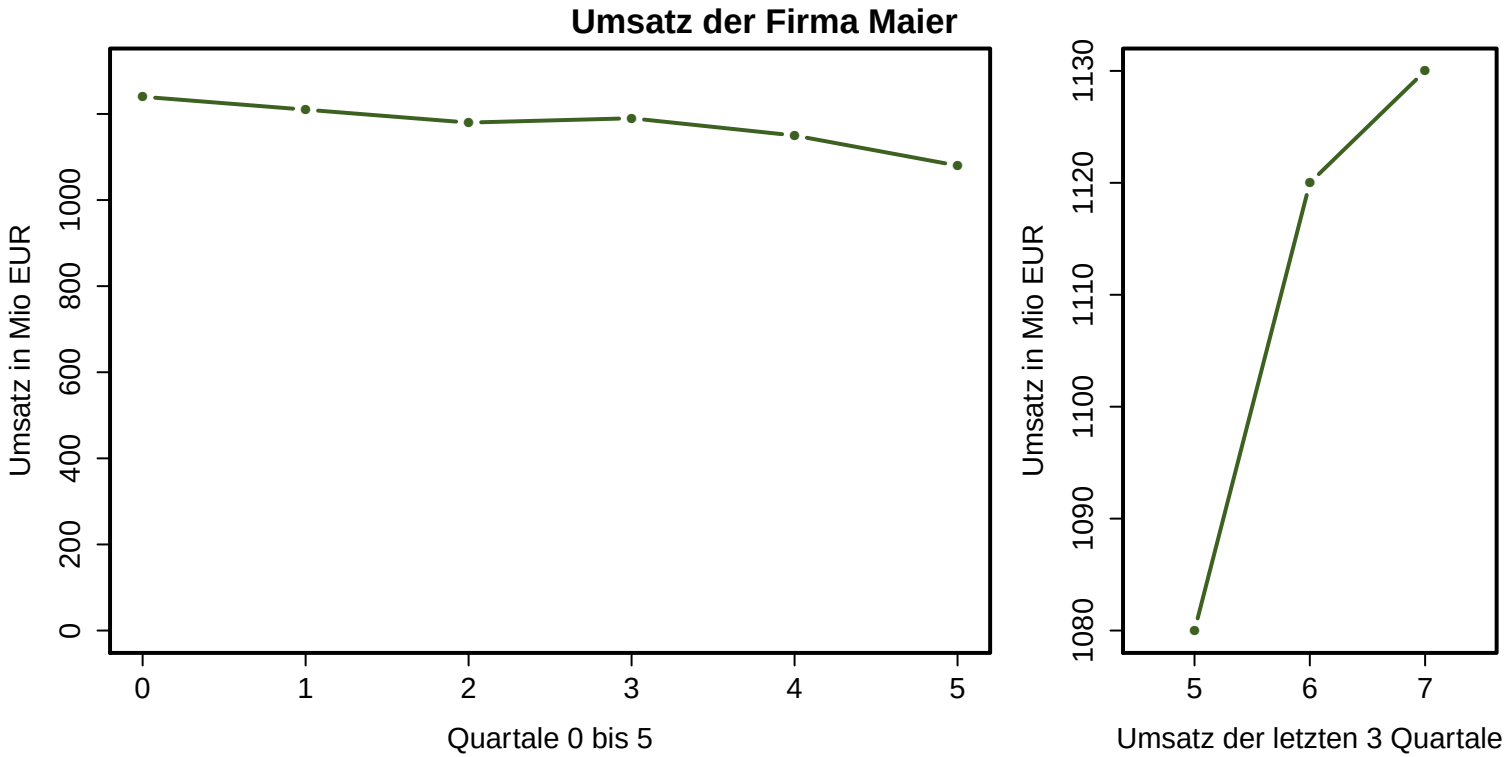


Abb. 63: Umsatzverlauf der Firma Maier – verfälscht dargestellt

Die Gefahr der verfälschenden Darstellung bei unterschiedlich gewählter Skalierung ist direkt offensichtlich.

9.2 Gute und schlechte statistische Grafik

Bemerkungen

- Immer die selbe Skalierung der entsprechenden Achsen verwenden, an denen etwas verglichen werden soll!
- Man kann die Wahl der Skalierung nicht der Voreinstellung des Computerprogramms überlassen.
- Je nachdem, wie weit der Wertebereich von der Null weg liegt und variiert, sollte man überlegen, ob die Null inkludiert werden sollte oder nicht.
- Ist die Null in der Nähe, sollte man sie aufnehmen, da sie eine natürliche Basis für Vergleiche liefert (bzw. die 1 bei Indizes / Verhältniszahlen).
- Ist die Null weit weg vom Geschehen, sieht man nicht mehr gut Unterschiede der dargestellten Beobachtungen.
- Einzelne Punkte, die weit vom Geschehen entfernt liegen, werden oft nicht dargestellt (wie Fernpunkte im Boxplot),

9.2 Gute und schlechte statistische Grafik

Beispiel 8 (Kursverlauf der BVB Aktie)

- *Kursverlauf der BVB Aktie vom 17. November 2000 bis zum 24. November 2010.*
- *Tageswerte, es liegen jeweils vor:*
  Eröffnung, Schluss, Tageshoch, Tagestief, gehandeltes Volumen
- *Wir starten mit einer Grafik der letzten 2 Jahre.*

9.2 Gute und schlechte statistische Grafik

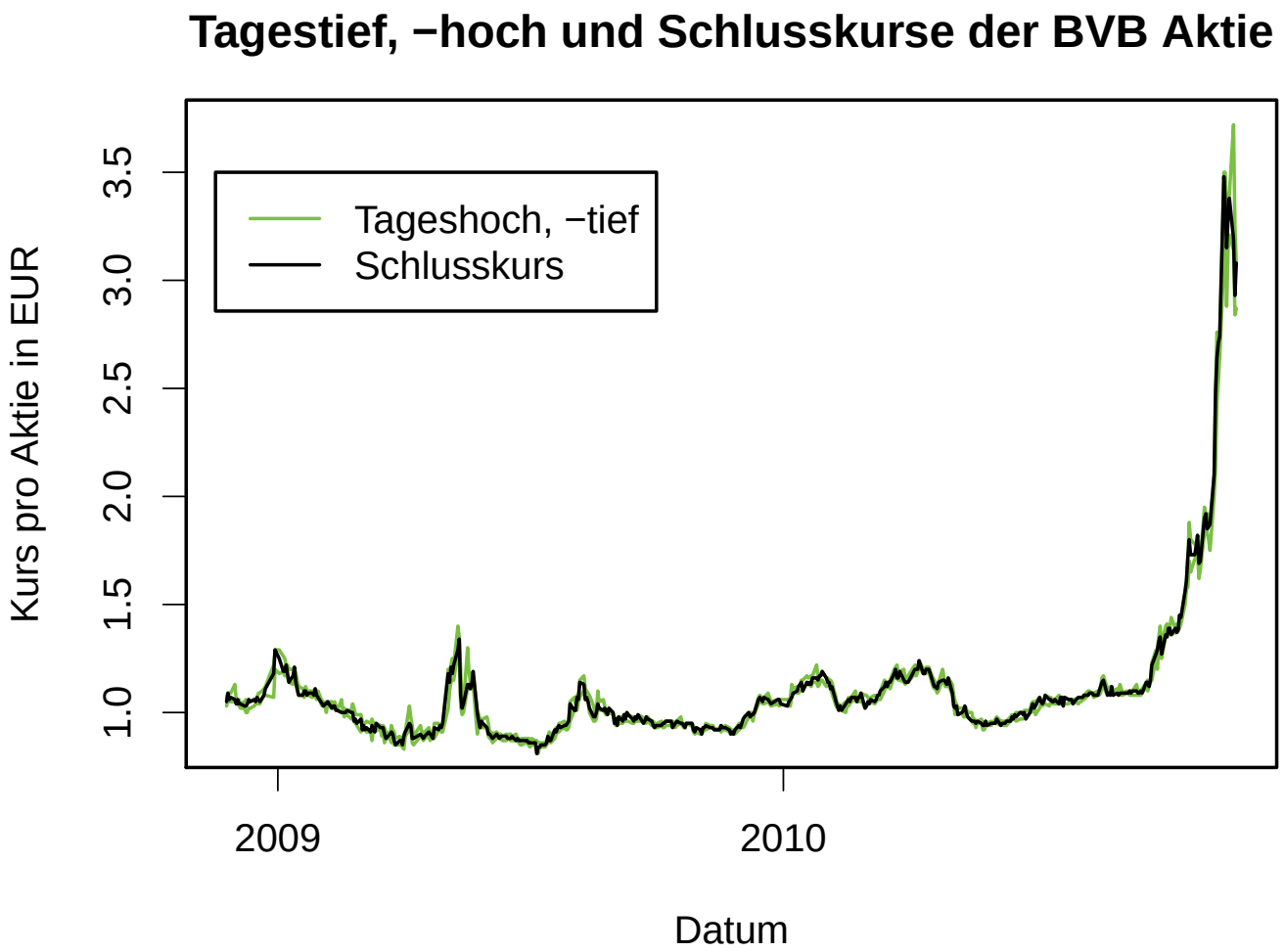


Abb. 64: Kursverlauf der BVB Aktie der letzten 2 Jahre (25.11.2008 – 24.11.2010)

9.2 Gute und schlechte statistische Grafik

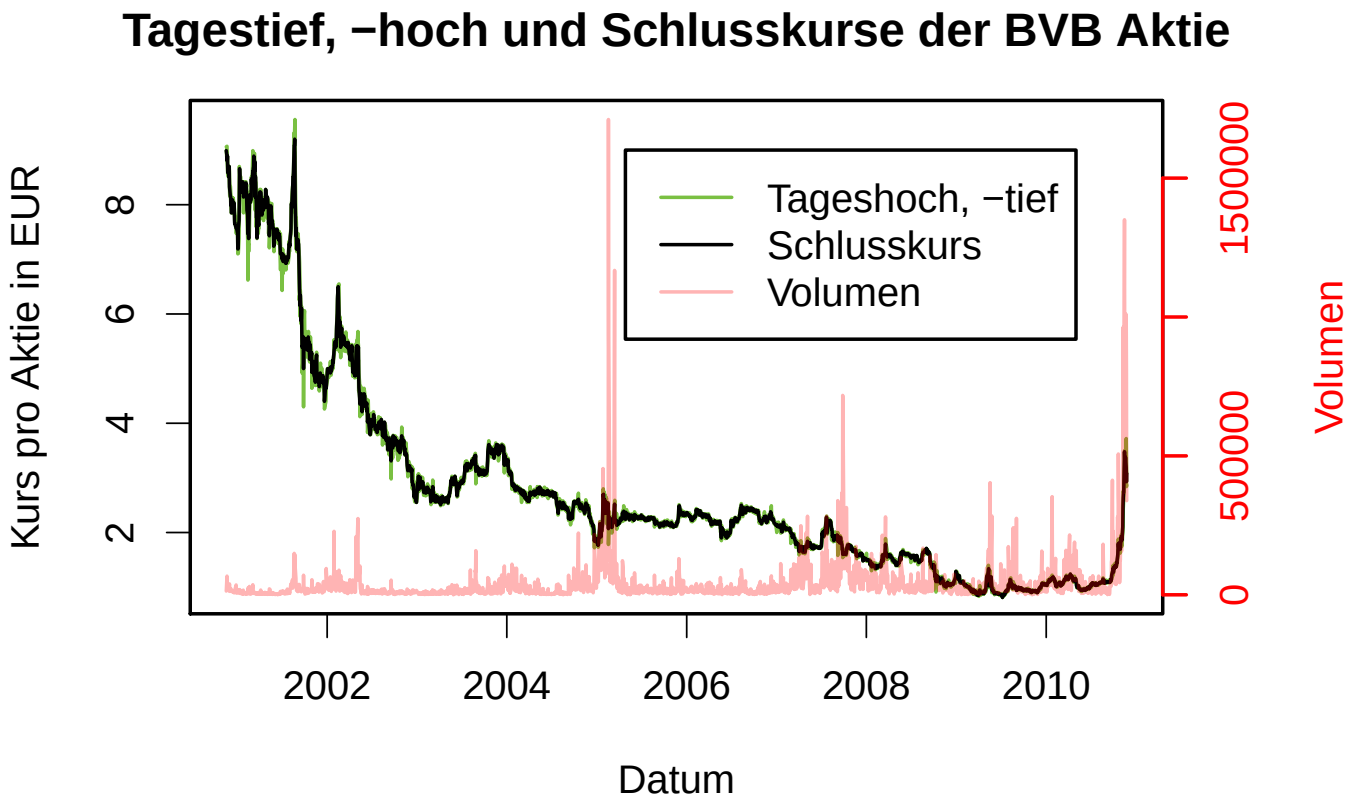


Abb. 65: Kursverlauf der BVB Aktie (17.11.2000 – 24.11.2010)

Hier sieht man wieder, dass die Präsentation eines Teilabschnitts der Daten immer vorsichtig interpretiert werden muss.

9.2 Gute und schlechte statistische Grafik

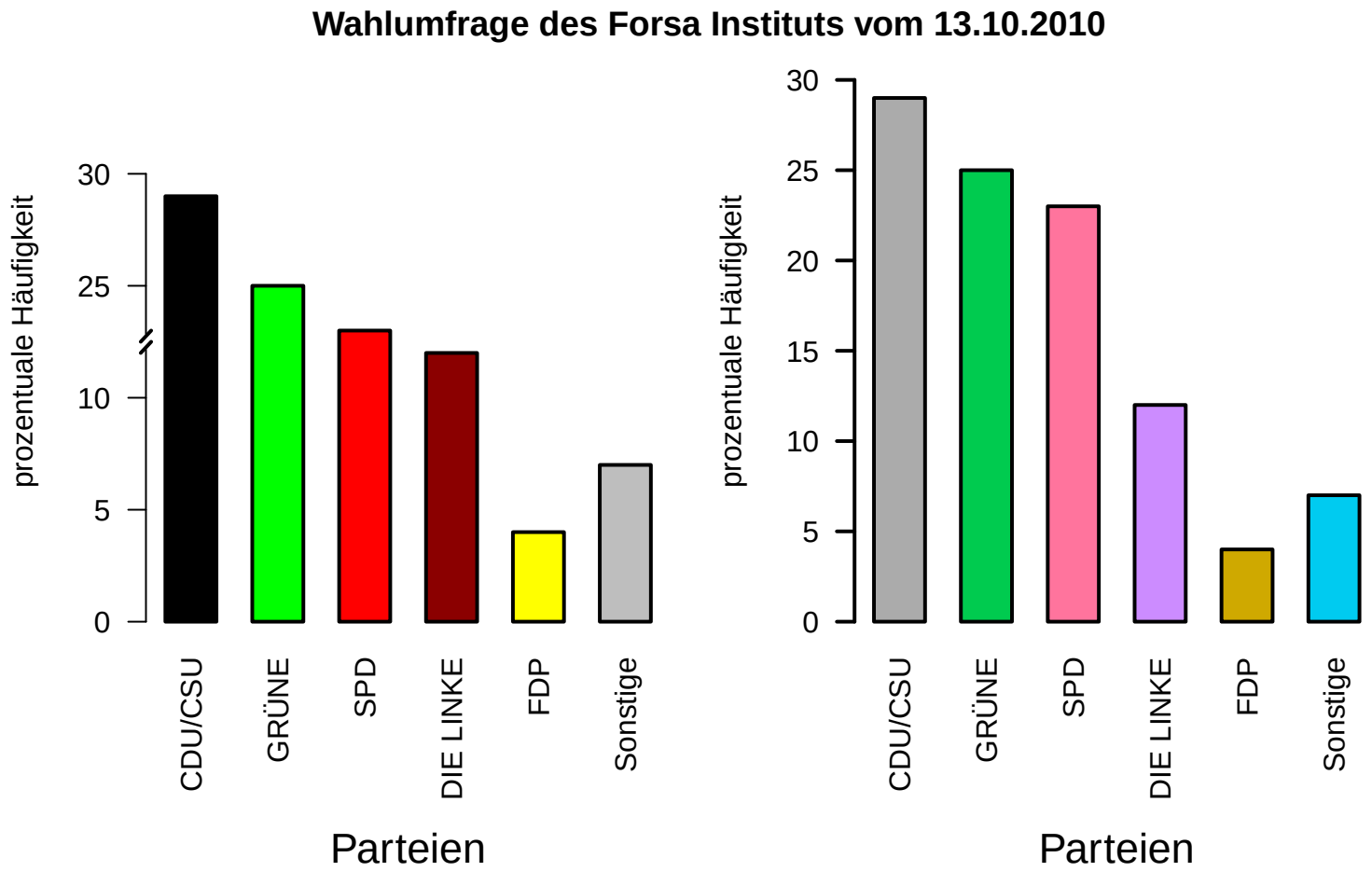


Abb. 66: Wahlumfrage des Forsa Instituts vom 13.10.2010 – linker Teil verfälscht dargestellt

9.2 Gute und schlechte statistische Grafik

Bemerkungen

- Lücken in einer Achse (linker Teil) werden leider immer wieder verwendet. **Lücken in der Achse sind aber fast niemals sinnvoll!**
- Auf der vorherigen Folie sieht man einen Vergleich inklusive einer anderen Farbdarstellung.
- Es bietet sich die Verwendung eines perzeptiven Farbraums ab, der also die menschliche Farbwahrnehmung nachahmt. Ein Beispiel ist der hcl (hue, chroma, luminance), bei dem Farbton, Farbintensität und Helligkeit separat eingestellt werden können.
- Leider wird sehr oft der rgb (red, green, blue) Farbraum verwendet (dieser wird z.B. für die Kommunikation zwischen Rechner und Monitor verwendet). Hier ist es sehr schwierig, vergleichbare Farben darzustellen.

9.2 Anforderungen an „gute“ Grafik (Tufte, 1982)

- **Grafik** ist *„anziehend“*, erweckt Neugier, nicht voll von „Chartjunk“
- **Beschriftung** ist klar, präzise und *unaufdringlich*, nicht „zerronnen“, klobig oder überladen
- **Wörter** werden *ausgeschrieben*, keine mysteriöse und aufwendige Kodierung
- **Wörter** werden *von links nach rechts* geschrieben, nicht vertikal oder in verschiedene Richtungen
- **Beschriftung** benutzt *Groß- und Kleinschreibung*, nicht nur Großschreibung
- **Schatten, Schraffierungen und Farbe** werden *minimal* eingesetzt: 5-10% der Betrachter sind rot-grün-blind!
- **Grafik** ist *selbsterklärend, wenig Text und wenig Legende* reicht zur Erklärung
- Exzellente Grafik vermittelt dem Betrachter **die „Wahrheit“ über die Daten in kurzer Zeit, mit kleinem „Tintenverbrauch“, auf kleinstem Raum**.

9.3 Weitere Bemerkungen zu Grafiken

(Cleveland (1994): „The Elements of Graphing Data“)

Inhalt einer Grafik

Es sind im Wesentlichen zwei Arten von Information, die in einer Grafik kodiert sind:

- **tabellarische Information**
  über die wahren Werte der Beobachtungen und
- **physische Information**,
  d.h. das Bild, das sich durch eine Umsetzung dieser Werte (in cm und Farbe) auf dem Medium ergibt.

Eine Beurteilung einer Grafik richtet sich danach, wie gut die tabellarische Information entnommen werden kann, und inwieweit die physische Information die wesentlichen Aspekte der Grafik transportiert.

9.3 Weitere Bemerkungen zu Grafiken

Nachschlagen

Um die **tabellarische Information** zu erhalten, muss man

- die Grafik abtasten, um die Skalenwerte innerhalb des Datenfensters wieder zu finden,
- Werte zwischen den angegebenen Skalenwerten interpolieren können und
- die Legende (Key) mit der Darstellung der Werte abgleichen können.

9.3 Weitere Bemerkungen zu Grafiken

Mustererkennung

Die **physische Information** liefert uns Aussagen über die Struktur der Daten. Dafür müssen wir

- Beobachtungen erkennen,
- Gruppen von Beobachtungen vereinigen und
- verschiedene Beobachtungen vergleichen, d.h. unterscheiden, anordnen und ins Verhältnis setzen.

Die Qualität einer Darstellung hängt davon ab, ob diese Operationen gut und schnell ausgeführt werden können.

Die Richtigkeit der tabellarischen Information ist eine selbstverständliche Forderung, die Richtigkeit der physischen Information ist sowohl schwerer zu definieren als auch, wenn sie definiert ist, schwerer zu erreichen.

9.3 Weitere Bemerkungen zu Grafiken

Überlagerte Kurven

Sowohl die Differenzen als auch das Verhältnis zweier Größen wird durch eine Überlagerung von Kurven schlecht dargestellt.

Zum einen werden Abstände zwischen Kurven verzerrt wahrgenommen, d.h. die Mustererkennung ist falsch, zum anderen ist das Nachschlagen der Abstände eine vielschrittige Aktion.

Daher ist es besser, die Zielgröße (Abstand, Verhältnis,...) selber abzubilden, oder die Verläufe auf parallelen Panels.

9.3 Weitere Bemerkungen zu Grafiken

Visuelle Referenzgitter

Die Wichtigkeit von Rastern leitet Cleveland vom sogenannten Weber'schen „Gesetz“ über den Vergleich der Länge zweier Linien ab.

Demnach ist die Wahrscheinlichkeit, mit der eine um w Einheiten längere Linie korrekterweise als größer erkannt wird, umgekehrt proportional zur Länge der kürzeren Linie.

Wahrgenommen wird also eher der prozentuale, denn der absolute Unterschied.

10.1 Histogramm

Es sei $x_1 = 1.4$, $x_2 = 1.6$, $x_3 = 2.0$, $x_4 = 2.3$ und $x_5 = 6.5$.

- Welches Histogramm beschreibt die Daten mathematisch korrekt?

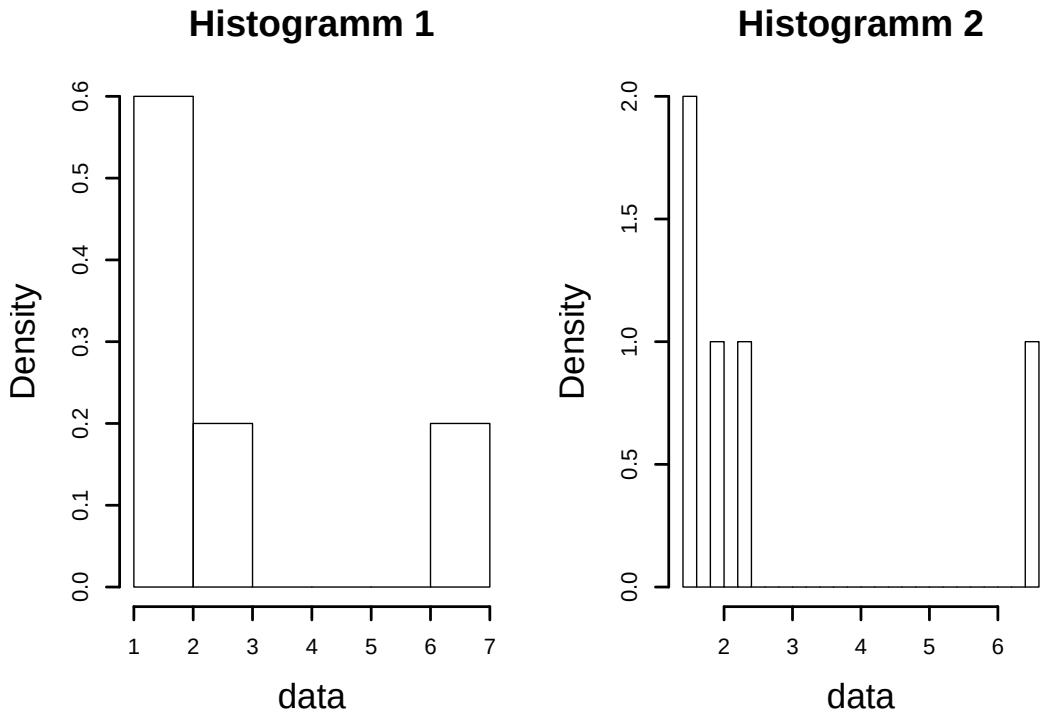


Abb. 67: Potentielle Histogramme

- Welches Histogramm ist prinzipiell besser zur Visualisierung der Daten geeignet?

10.1 Mittelwerte

- Ein Schüler wechselt von einem Gymnasium auf eine Gesamtschule. Darauf steigt an beiden Schulen der mittlere IQ. (arithmetisches Mittel). Ist das möglich? Begründen Sie die Antwort.
- Wie berechnet man einen geeigneten Mittelwert von Wachstumsraten?

10.1 Unabhängigkeit

Wir nehmen an, dass zwei zufällig ausgewählte Menschen in Deutschland jeweils 2000 (in Deutschland zufällig ausgewählte) Bekannte haben und dass Deutschland genau 80 Millionen Einwohner hat.

- Wie groß ist die dann die Wahrscheinlichkeit, dass die beiden Menschen mindestens einen gemeinsamen Bekannten haben?
- Warum ist diese Wahrscheinlichkeit in der Realität deutlich kleiner, d.h. welche Annahme ist unrealistisch?

10.1 Bedingte Wahrscheinlichkeit

Eine Ehefrau ist ermordet worden. Es ist bekannt, dass der Ehemann seine Ehefrau geschlagen hat. Welche Wahrscheinlichkeit muss betrachtet werden, um ohne weitere Zusatzinformationen abzuschätzen, ob der Ehemann der Mörder ist?

- $P(\text{Ehemann ist Mörder} \mid \text{Ehemann hat Frau geschlagen und Frau ist ermordet worden}) = 8/9$
- $P(\text{Ehemann ist Mörder} \mid \text{Ehemann hat Frau geschlagen}) = 1/2500$

Quelle: I. Good: „When batterer becomes murderer,“ Nature 391, 1969, S. 481

10.1 Verteilungen

Jeder vierte Bundesbürger stirbt an Krebs. Sie lesen in der Zeitung von drei unzusammenhängenden Todesfällen.

Mit welcher Wahrscheinlichkeit

- starben alle drei an Krebs?
- starb genau einer an Krebs?
- starb mindestens einer an Krebs?

10.1 Lage und Streuung

Ein Würfel habe sechs Seiten mit den Zahlen 1, 2, 3 und dreimal die 6.

a) Bestimmen Sie die Zähldichte der Zufallsvariablen X, die das Würfelergebnis bezeichnet.

b) Bestimmen Sie den Erwartungswert und die Varianz von X.

c) Was wäre das Ergebnis für (arithmetisches) Mittel und Varianz, wenn es sich um einen Datensatz (mit den Werten 1, 2, 3, 6, 6, 6) und nicht um eine Zufallsvariable handeln würde.

10.1 Wahrscheinlichkeit

Von wegen Pitbull & Co.! Schäferhunde führen die Gefährlichkeits-Statistik an

Beißer oder Schoßhündchen? 17 Hunderassen und ihre Mischlings-Nachkommen stehen „auf dem Index", werden per se in der seit rund einem Jahr in der Stadt Göttingen bestehenden Hundesteuerverordnung als Kampfhunde eingestuft. Resultat für die Besitzer: Sie müssen 1.200 Mark ins Stadtsäckel zahlen. Das stößt auf Protest: 29 Kampfhunde sind gemeldet, und: „In nahezu allen Fällen laufen Klageverfahren vor dem Verwaltungsgericht", weiß Stadtpressesprecher Detlef Johannson.

Johannson weiter: „Wir sehen dem Ausgang dieser Verfahren zuversichtlich entgegen." Das sieht man beim Verband für das deutsche Hundewesen (VDH) anders. Hier verweist man auf ein Urteil des Oberverwaltungsgerichts in Magdeburg, das einem Hundehalter recht gegeben hat, der gegen die hohe Besteuerung seines Kampfhundes geklagt hatte. In der Urteilsbegründung heißt es, die Gemeinde habe den Grundsatz der Gleichmäßigkeit zu beachten.

Der VDH kritisiert den falschen Ansatz der Kampfhundesteuer: Es sei falsch, die Gefährlichkeit eines Hundes an seiner Rasse festzumachen. Belegt wird das durch eine polizeiliche Statistik aus Nordrhein Westfalen. Unter den gefährlichen Hunden, die von der Polizei getötet werden mußten, befinden sich 105 Schäferhunde, 94 Mischlinge und 15 Pitbulls. Alle anderen Rassen unter „ferner liefen". Wird zugebissen, sind es also meist Schäferhunde, was wiederum an der großen Beliebtheit dieser Rasse liegt. Das Innenministerium kam zu dem Schluß: „Aus dieser Untersuchung lassen sich keine Hinweise auf die besondere Gefährlichkeit von Kampfhunden ableiten." Der VDH weiß: „Die Zusammenstellung der Rasselisten ist völlig willkürlich erfolgt." Die Gefährlichkeit eines Hundes sei keine Frage der Rasse, sondern hänge von Erziehung und Haltung ab.

Auch die Stadt Göttingen scheint nicht so ganz glücklich zu sein mit der Kampfhunde-Steuer: „Uns ist klar, daß auf diesem Weg die Problematik nicht gelöst werden kann. Schon gar nicht ist dieser Teil der Hundesteuer ein nennenswerter Einnahmebeitrag für den städtischen Haushalt. Angestrebt wird aber eine eindämmende Wirkung auf die, wie viele schreckliche Beispiele in der Bundesrepublik zeigen, fürwahr nicht immer artgerechte Haltung dieser Hunde."

Der Hundeverband hat da ein paar Verbesserungsvorschläge. Etwa statt einer horrenden Kampfhundesteuer eine Steuererleichterung für die Hundehalter einzuführen, die mit ihrem Hund eine Begleithundeprüfung abgelegt haben. Vorbildlich sei die Gefahren-Hunde-Verordnung des Landes Nordrhein-Westfalen. Sie setzt bei den Hundehaltern an, geht konsequent gegen die Verantwortungslosen vor, hilft denen, die Probleme mit ihrem Hund haben. Erster Erfolg: Die Teilnehmerzahlen an Schulungen für Hund und Halter stiegen deutlich an.

Der EXTRA TIP trug diese Vorschläge der Göttinger Stadtverwaltung vor und hat da vielleicht etwas in Gang gesetzt. Der Vorschlag hört sich interessant an", so Detlef Johannson. Man werde sich wegen der in Nordrhein-Westfalen gemachten Erfahrungen kundig machen und eventuell eine Initiative im Rahmen des Niedersächsischen Städtetages anregen: „Denn auch uns geht es nur nachrangig um die Einnahmen, vor allem aber um die Ausübung einer Ordnungsfunktion. Wenn eine Kampfhundeverordnung in diesem Sinne auf anderem Wege ebenfalls hilfreich sein kann, wird man darüber – am besten im Verbund des Bundeslandes – nachdenken müssen." **Stefanie Arndt**

Hundesteuersatzung:

Wer ist denn nun ein Kampfhund?

Kampfhunde sind solche Hunde, bei denen nach Ihrer besonderen Veranlagung, Erziehung und / oder Charaktereigenschaft die erhöhte Gefahr einer Verletzung von Personen besteht.

Kampfhunde im Sinne der Vorschrift sind insbesondere American Staffordshire Terrier, Bandog, Bulldog, Bullmastiff, Bullterrier, Chinesischer Kampfhund, Dog Argentino, Dog de Bordeaux, Fila Brasileiro, Mastiff, Mastino Espagnol, Mastino Neapolitano, Pit-Bull-Terrier, Römischer Kampfhund, Rhodesian Ridgeback, Staffordshire Bullterrier, Tosa-Inu.

Darüberhinaus gelten als Kampfhunde im Sinne der Satzung alle Kreuzungen, an denen die vorgenannten Hunderassen beteiligt sind sowie alle Hunde, bei denen sich die Eigenschaft als Kampfhund im Einzelfall aus ihrer Ausbildung mit dem Ziel einer gesteigerten Aggressivität und Gefährlichkeit ergibt.

Aus der Hundesteuersatzung der Stadt Göttingen.

Der „liebste Hund der Welt": American Staffortshire Terrier Joey mit Britta und Thomas Bayer. Foto: Schachtschneider

Abb. 68: Wo liegt der Fehler?

10.1 t-Test

Es haben 8 Journalisten und 18 Germanisten eine Klausur geschrieben, die aus vielen kleinen Teilaufgaben bestand (das Klausurergebnis wird damit als metrisches Maerkmal aufgefasst). Die Journalisten haben im Durchschnitt 60.5 Punkte erzielt und die Germanisten 56.5 Punkte. Die (empirische) Standardabweichung der Punktzahlen betrug jeweils 4 und 6. Wir nehmen an, dass die Zufallsvariablen für beide Punktzahlen normalverteilt sind. Die wahren Erwartungswerte seien mit μ_J und μ_G bezeichnet.

- Es soll nun mit einem t-Test bestimmt werden, ob die Journalisten signifikant höhere Punktzahlen erzielen als die Germanisten (gerichtete Hypothese). Wie lauten H_0 und H_1?
- Wie lauten H_0 und H_1, wenn untersucht werden soll, ob sich die Punktzahlen signifikant unterscheiden (ungerichtete Hypothese)?
- Wie lautet der Wert der Teststatistik des t-Tests?
- In welchen der beiden Fälle lehnt der entsprechende Test die Nullhypothese zum Niveau $\alpha = 0.05$ ab (signifkantes Ergebnis)?

10.1 t-Test

- Wie lauten H_0 und H_1?

  $H_0 : \mu_J \leq \mu_G$ und $H_1 : \mu_J > \mu_G$.

- Wie lauten H_0 und H_1, wenn untersucht werden soll, ob sich die Punktzahlen signifikant unterscheiden (ungerichtete Hypothese)?

  $H_0 : \mu_J = \mu_G$ und $H_1 : \mu_J \neq \mu_G$.

- Wie lautet der Wert der Teststatistik des t-Tests?

$$\begin{aligned} t &= \frac{(\bar{X}_J - \bar{X}_G) - 0}{\sqrt{s_J^2/n + s_G^2/m}} \\ &= \frac{(60.5 - 56.5)}{\sqrt{4^2/8 + 6^2/18}} = \frac{4}{\sqrt{2+2}} = 2. \end{aligned}$$

10.1 t-Test

- In welchen der beiden Fälle lehnt der entsprechende Test die Nullhypothese ab (signifkantes Ergebnis)?

  Die Anzahl der Freiheitsgrade berechnet sich zu

$$k = \left\lfloor \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_Y^2}{m}\right)^2} \right\rfloor = \left\lfloor \frac{\left(\frac{16}{8} + \frac{36}{18}\right)^2}{\frac{1}{7}\left(\frac{16}{8}\right)^2 + \frac{1}{17}\left(\frac{36}{18}\right)^2} \right\rfloor = 9.$$

  Für den einseitigen Test ist der kritische Wert c das 95%-Quantil einer t_9-Verteilung, also $c = 1.833$.
  Wegen $t = 2 > c$ ist das Ergebnis zum Niveau 0.05 statistisch signifikant.

  Für den zweiseitigen Test ist der kritische Wert c das 97.5%-Quantil einer t_9-Verteilung, also $c = 2.262$.
  Wegen $|t| = 2 \leq c$ ist das Ergebnis zum Niveau 0.05 nicht statistisch signifikant.

10.1 Korrelation und Kausalität

Zusammenhang zwischen der Anzahl in den USA verkaufter japanischer Autos und der Anzahl der Selbstmorde durch Autounfälle in den USA

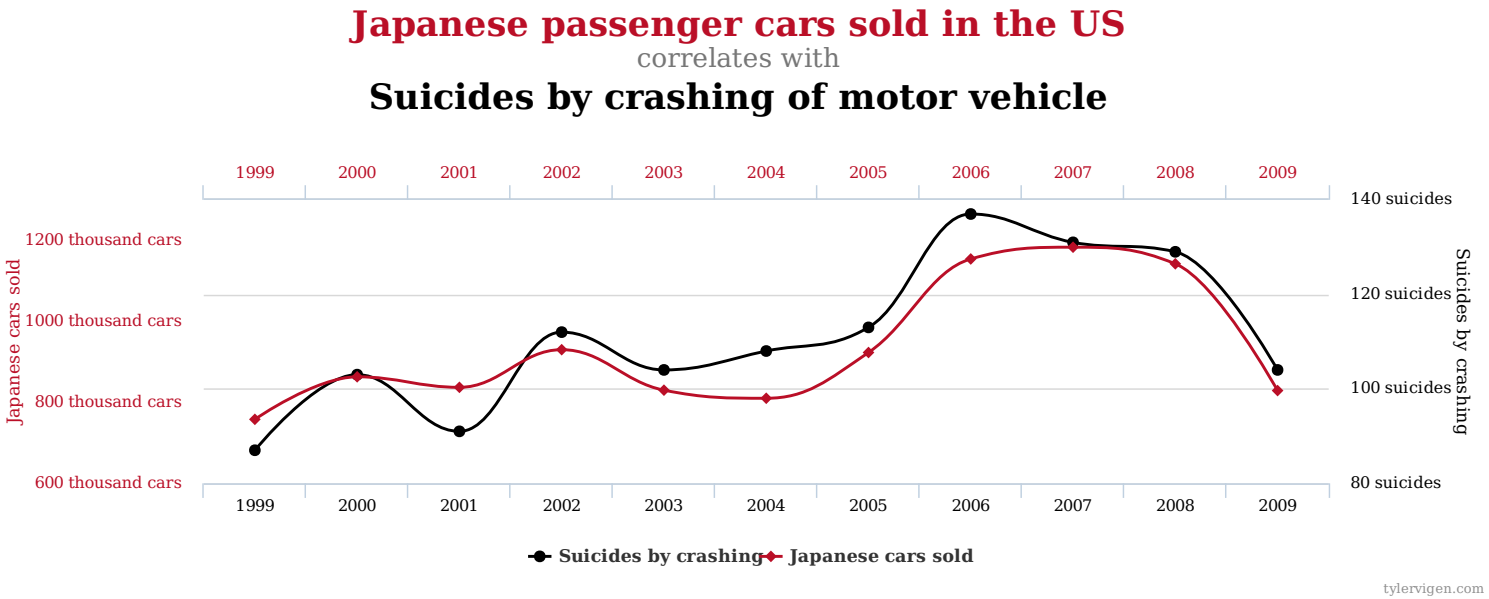
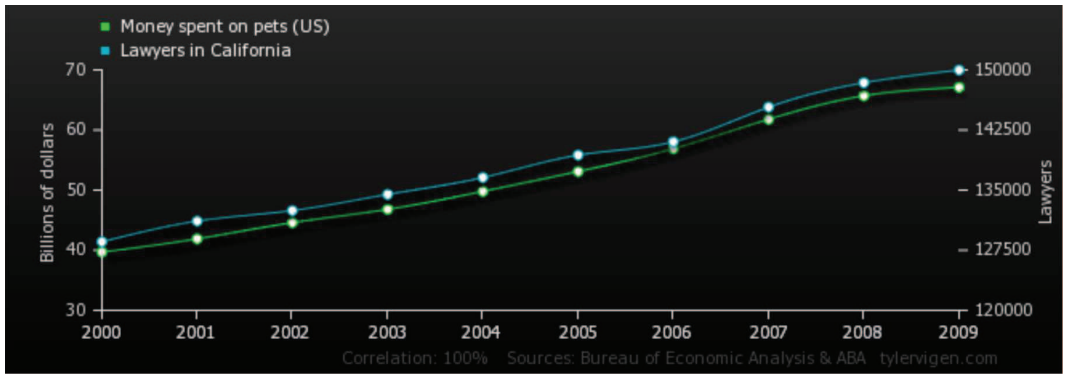


Abb. 69: Japanische Autos provozieren Selbstmorde?

10.1 Korrelation und Kausalität

Zusammenhang zwischen den Ausgaben für Haustiere (USA) und der Anzahl von Anwälten in Kalifornien

Money spent on pets (US)
correlates with
Lawyers in California



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Money spent on pets (US) *Billions of dollars (Bureau of Economic Analysis)*	39.7	41.9	44.6	46.8	49.8	53.1	56.9	61.8	65.7	67.1
Lawyers in California *Lawyers (ABA)*	128,553	131,139	132,452	134,468	136,571	139,371	141,030	145,355	148,399	149,982

Correlation: 0.998386

Abb. 70: Haustiere machen prozessierwütig?

10.1 Interpretation von Zahlen und Statistiken

Titelthema „Lügen nach Zahlen“ in der Zeit („Kann das stimmen“, 24.04.2017)

Fragwürdige Aussagen:

- Eine Million Männer gehen in Deutschland zu Prostituierten – Tag für Tag! 400 000 Prostituierte bieten auf deutschen Straßen, in Privatwohnungen und Bordellen ihre Dienste an.
- 15,7 Prozent der Deutschen sind arm, ein neuer Rekord.
- Nur 8,4 Prozent der Frauen, die ihren Vergewaltiger anzeigen, erleben die Verurteilung des Täters – vor 20 Jahren waren es noch 21,6 Prozent.
- Fast die Hälfte der Arbeitsplätze könnte durch die fortschreitende Digitalisierung aller Lebensbereiche vernichtet werden, in den nächsten 13 Jahren.
- Eine große Mehrheit der Deutschen will künftig nicht mehr vom Auto abhängig sein.

10.1 Interpretation von Zahlen und Statistiken

Titelthema „Lügen nach Zahlen“ in der Zeit („Kann das stimmen“, 24.04.2017)

Fragwürdige Aussagen:

- 300 Frauen tragen in Deutschland eine Burka.
- Um mehr als ein Fünftel ist der Anteil von Kindern mit der Aufmerksamkeitsdefizit-Störung gestiegen, innerhalb von nur vier Jahren.
- 21 000 Patienten könnten in Krankenhäusern und Heimen getötet worden sein – in einem Jahr.

10.1 Interpretation von Zahlen und Statistiken

Titelthema „Lügen nach Zahlen“ in der Zeit („Kann das stimmen“, 24.04.2017)
Fragen, die man sich stellen sollte:

1. Hat der Produzent der Zahl ein plausibles Interesse, will er ein Phänomen groß oder klein erscheinen lassen? (Und sei es bloß, um sich mit einer „Neuigkeit“ hervorzutun.)
2. Geht es um ein Phänomen, bei dem es schwierig ist, ehrliche Antworten zu bekommen?
3. Ist das Ergebnis politisch opportun?
4. Wer wurde befragt? Und: Wer nicht?
5. Ist das Phänomen klar definiert?
6. Sind auch absolute Häufigkeiten angegeben oder nur relative Veränderungen?