Statistik für Data Science

Flurina Sophie Baumbach

Sep 15, 2025

Inhaltsverzeichnis

1 Daten

1.1 Datentypen & Messniveaus

Statistik hängt vom Datentyp ab - Unterschiedliche Methoden für verschiedene Skalen

1.1.1 Nominalskala

kategorisch, keine Ordnung

- Niedrigster Informationsgehalt
- Nominalskalierte Daten lassen sich weder in eine logische Reihenfolge sortieren, noch quantitativ differenzieren
- Merkmalsausprägungen zweier Merkmale lassen sich also lediglich in Gleichheit oder Ungleichheit unterscheiden (A $\neq$ B)
- Werte sind reine Namen / Kategorien, **keine Reihenfolge, keine Abstände**
- Beispiele: Geschlecht, Postleitzahl, HTTP-Statuscode

1.1.2 Ordinalskala

kategorisch, mit Ordnung

- Ordnet Variablen mit Ausprägungen in eine klare Rangfolge
- Die Abstände zwischen den einzelnen Rängen sind nicht interpretierbar, da sie nicht quantifiziert sind
- Beispiele: Schulnoten, Star Ratings

1.1.3 Intervallskala

metrisch, ohne echten Nullpunkt

- Es lassen sich Reihenfolgen und quantifizierbare Abstände bilden
- Sie liegt beim Skalenniveau ebenfalls über der Nominalskala und der Ordinalskala
- kein absoluter/natürlicher Nullpunkt (Verhältnisse („doppelt so viel“) ergeben keinen Sinn)
- Beispiele: Temperatur in Celsius, Zeit

1.1.4 Ratioskala

metrisch, mit absolutem Nullpunkt

- Es lassen sich sowohl Reihenfolgen als auch quantifizierbare Abstände bilden
- Alle Eigenschaften der Intervallskala + natürlicher Nullpunkt $\rightarrow$ Verhältnisse sind sinnvoll
- Beispiele: Alter in Jahren, RAM-Größe, Gewicht, Einkommen

1.2 Bias in Daten

Bias = systematische Verzerrung in Daten oder Analyse

1.2.1 Sampling Bias

Entsteht, wenn die Stichprobe nicht repräsentativ für die gesamte Population ist

1.2.2 Survivorship Bias

Entsteht, wenn man nur die „Überlebenden“ einer Gruppe betrachtet und dadurch die Analyse verzerrt wird
$\rightarrow$ Man übersieht die Fälle, die nicht mehr da sind – und genau die enthalten oft die wichtigste Information

1.2.3 Confirmation Bias

Man sucht nur nach Informationen, die die eigene Hypothese oder Erwartung bestätigen und widersprechende Daten werden ignoriert oder abgewertet

1.2.4 Publication Bias

Es werden nur „positive“ oder „signifikante“ Ergebnisse veröffentlicht, während neutrale oder negative Resultate in der Schublade verschwinden ⇒ verzerrtes Bild in der Wissenschaft oder Praxis

1.2.5 Measurement Bias

Entsteht, wenn die Messungen selbst systematisch fehlerhaft oder verzerrt sind → Es liegt also nicht an der Stichprobe oder Veröffentlichung, sondern an der Art und Weise, wie Daten erfasst werden

1.3 Missing Values

1.3.1 MCAR – Completely at Random

M**C**AR: **C**ompletely = Chaos, reiner Zufall

- Fehlende Werte treten **völlig zufällig** auf
- Es gibt keinen Zusammenhang mit beobachteten Variablen oder mit dem wahren Wert selbst
- Kein Bias → die Daten bleiben repräsentativ (Man darf die fehlenden Werte ignorieren oder simpel (ggf. durch den Median) ersetzen)

1.3.2 MAR – At Random

M**AR**: **A**t **R**andom- Abhängig von Anderen (beobachteten) Variablen

- Fehlende Werte hängen **nicht vom wahren Wert selbst** ab, sondern **von anderen beobachteten Variablen**
- Das Fehlen ist also nicht völlig zufällig, sondern erklärbar durch bekannte Informationen
- MAR-Daten sind modellierbar → man kann fortgeschrittene Imputationsmethoden nutzen, z.B. gruppenweise Mittelwert/Median, MICE (Multiple Imputation by Chained Equations) oder KNN-Imputation (K-Nearest Neighbors)

1.3.3 MNAR – Not at Random

M**N**AR: **N**ot Random = Nicht zufällig, systematisches Problem

- Das Fehlen hängt direkt vom wahren Wert selbst ab
- Menschen oder Systeme „verstecken“ bestimmte Werte systematisch → also nicht zufällig, nicht durch andere Variablen erklärbar, sondern abhängig von der fehlenden Größe selbst
- Klassische Imputation (Mittelwert, Median) führt zu Verzerrungen → man braucht spezielle Modelle oder Expertenwissen / Domain Knowledge

1.3.4 Lösungsansätze

%	MCAR	MAR	MNAR
$< 5\%$	Löschen OK	Gruppenweise Imputation	Expertenwissen (z. B. Arzt)
5–20%	Multiple Imputation	MICE / KNN	Selection Models (z. B. Heckman)
$> 20\%$	Multiple Imputation	MICE + Sensitivity	Explizite Modellierung (Warum?)

1.3.5 Strategie - Median / Mean Imputation

Fehlende Werte werden durch den Mittelwert oder Median der vorhandenen Werte ersetzt
→ Vorteile: Sehr einfach und schnell, gut, wenn nur wenige Werte fehlen (¡ 5 %) - wird bei MCAR verwendet
→ Nachteile: Unterschätzt die Varianz → Daten wirken zu gleichmäßig, kann Zusammenhänge zwischen Variablen verzerren und ignoriert Unsicherheit (alle fehlenden bekommen denselben Wert)

1.3.6 Strategie - Gruppenweise Imputation

Fehlende Werte werden **nicht global**, sondern **innerhalb von Untergruppen** ersetzt → Man teilt die Daten nach einer Gruppenvariablen (z. B. Geschlecht, Region, Altersklasse) auf und imputiert gruppenweise den Median oder Mittelwert
→ Vorteile: Berücksichtigt strukturelle Unterschiede zwischen Gruppen, weniger Verzerrung als globale Mean/Median-Imputation, einfach anzuwenden, wenn eine gute Gruppenvariable vorhanden ist.
→ Nachteile: Innerhalb der Gruppen wird die Varianz trotzdem unterschätzt, man braucht eine vollständige und sinnvolle Gruppenvariable (z. B. „Geschlecht“ darf nicht selbst viele Missing Values haben), funktioniert nur gut, wenn Gruppen groß genug sind

1.3.7 Strategie - Modellbasierte Imputation

Fehlende Werte werden mit Hilfe eines Vorhersagemodells geschätzt, das aus den vorhandenen Daten gelernt wird - statt einfach Median oder Mittelwert einzusetzen, nutzt man **Zusammenhänge zwischen Variablen**
→ Vorteile: Nutzt komplexe Zusammenhänge → realistischere Schätzungen, Erhält Varianz besser als Mean/Median-Imputation, kann auch nicht-lineare Muster erfassen
→ Nachteile: Rechenaufwendig, komplex, Gefahr von Overfitting, Qualität hängt stark von den Trainingsdaten ab

1.3.8 Vergleichstabelle Imputationsmethoden

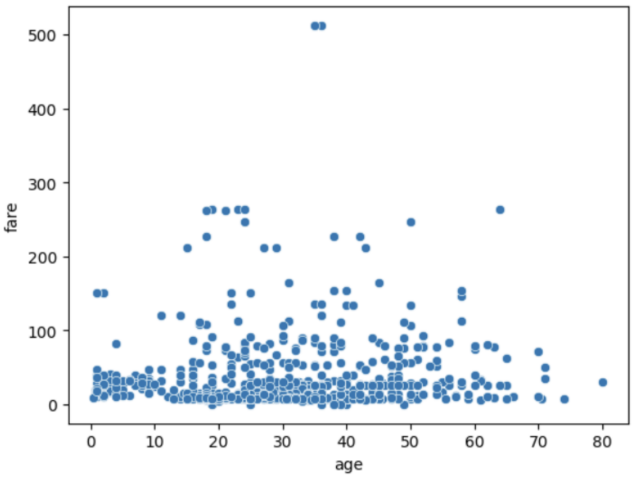
Methode	Komplexität	Varianz	Beziehungen	Wann nutzen?
Mean/Median	*	✗ Unterschätzt	✗ Ignoriert	MCAR, ¡ 5% fehlend
Gruppenweise	**	● Teilweise	✓ Innerhalb Gruppen	MAR mit klaren Gruppen
KNN	* * *	✓ Erhält	✓ Lokal	Lokale Cluster
MICE	* * **	✓ Erhält	✓ Global	MAR, viele Variablen
Deep Learning	* * * * *	✓ Erhält	✓ Komplex	Große Datensätze

2 EDA - Exploratory Data Analysis

Muster erkennen, Fehler finden, Hypothesen generieren

2.1 Scatterplot

- Stellt die Beziehung zwischen zwei metrischen Variablen in einem Koordinatensystem dar
- Jeder Datenpunkt = ein Beobachtungspaar (x,y)
- Grundlage für lineare Regression



3 Lagekennzahlen

Lagekennzahlen sollten **immer zusammen mit Streuungsmaßen** berichtet werden, da dieselbe Lage unterschiedliche Streuungen haben kann.
→ Beispiel: Median = 40 ist nicht informativ, wenn die Spannweite 39–41 oder 10–1000 sein könnte.

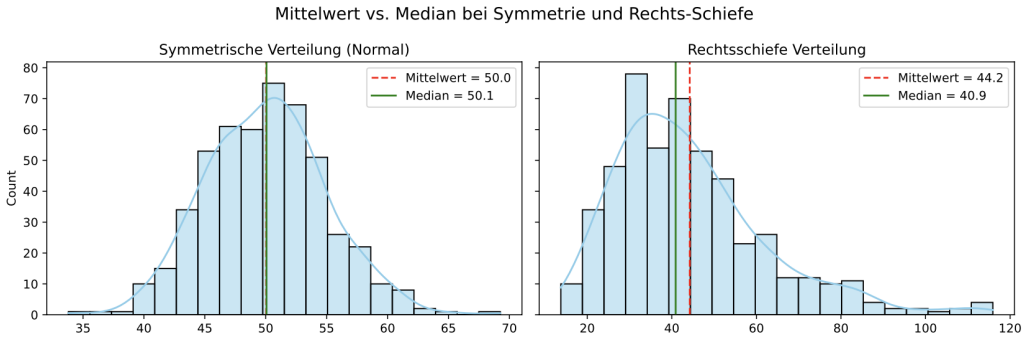
3.1 ECDF inverse zu Quantile

ECDF = Empirical Cumulative Distribution Function (empirische Verteilungsfunktion)
Sie zeigt für jeden Wert x, welcher Anteil der Daten kleiner oder gleich x ist.
$F(x) = \frac{\text{Anzahl der Werte} \leq x}{n}$, steigt monoton von 0 bis 1
Quantile sind die **inverse Sichtweise**:
Bei gegebenem Anteil p (z. B. 0.25, 0.5, 0.75) → finde den Wert Q_p, der mindestens p · 100% der Daten abdeckt.

ECDF: Zahl → Prozent
Quantil: Prozent → Zahl

3.2 Mittelwert vs. Median bei Schiefe

Rechts-Schiefe (positiv skewed): Verteilung hat einen langen rechten „Tail“ (z. B. Einkommen, Wartezeiten)
→ **Mittelwert** (Mean) liegt **größer als der Median**, weil er von den hohen Extremwerten nach rechts „gezogen“ wird
→ Median bleibt robust und zeigt das „typische“ Zentrum



Typ	Lage der Kennwerte	Beschreibung
Symmetrisch	Mean = Median = Mode	gleichmäßige Verteilung
Linksschief (negativ)	Mean < Median < Mode	langer linker Tail
Rechtsschief (positiv)	Mode < Median < Mean	langer rechter Tail

4 Streuungskennzahlen

Lagekennzahlen beschreiben nur das Zentrum
Streuungskennzahlen erfassen, wie stark die Daten um dieses Zentrum schwanken → „Breite“ oder „Variabilität“ der Verteilung

4.1 Streuungs-Profil als Routine

1. Missing prüfen
2. Lage robust (Median, IQR), dann klassisch (Mittelwert)
3. Streuung robust (IQR, MAD), dann klassisch (Varianz, Standardabweichung)
4. Ausreißerprüfung ankündigen

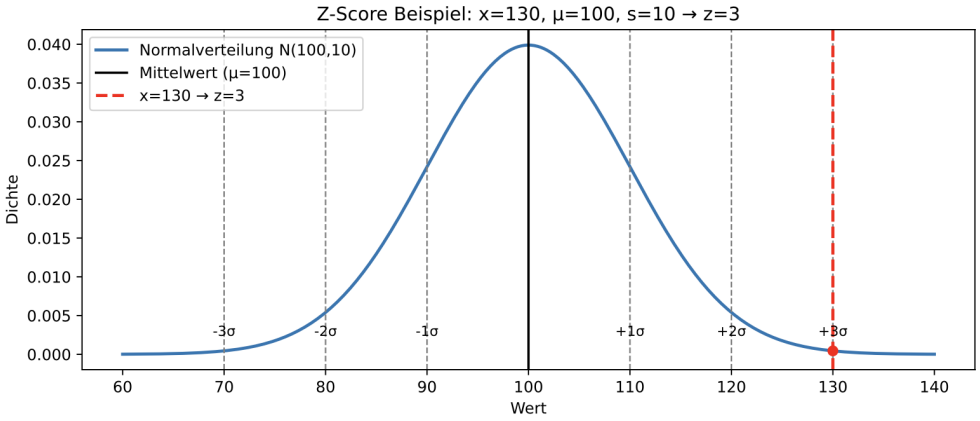
5 Ausreisser erkennen und behandeln

5.1 Einfluss von Ausreißern

- **Mean & SD kippen sehr schnell**: sie steigen stark an, wenn ein Extremwert hinzukommt
- **Robuste Kennzahlen**: Median, IQR, MAD → deshalb immer zuerst berichten, dann erst Ausreißer analysieren

5.2 Klassischer Z-Score

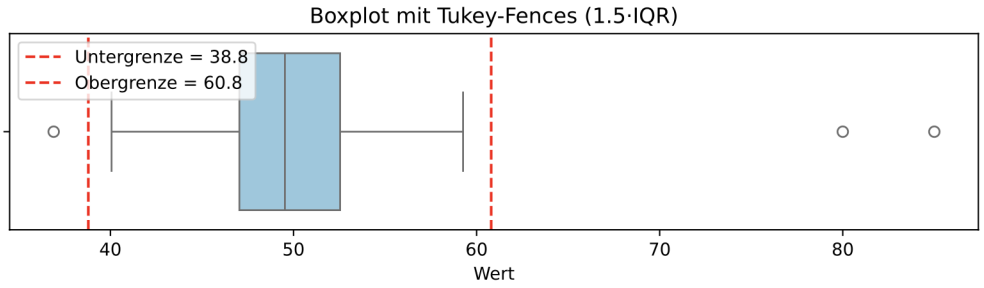
- Misst, wie viele Standardabweichungen ein Wert vom Mittelwert entfernt ist
- $z_i = \frac{x_i - \bar{x}}{s}$
- Typische Schwelle: $|z| > 3 \rightarrow$ Ausreißer
- Nur sinnvoll bei (nahezu) normalverteilten Daten, sonst irreführend



5.3 Tukey-Fences (Boxplot-Regel)

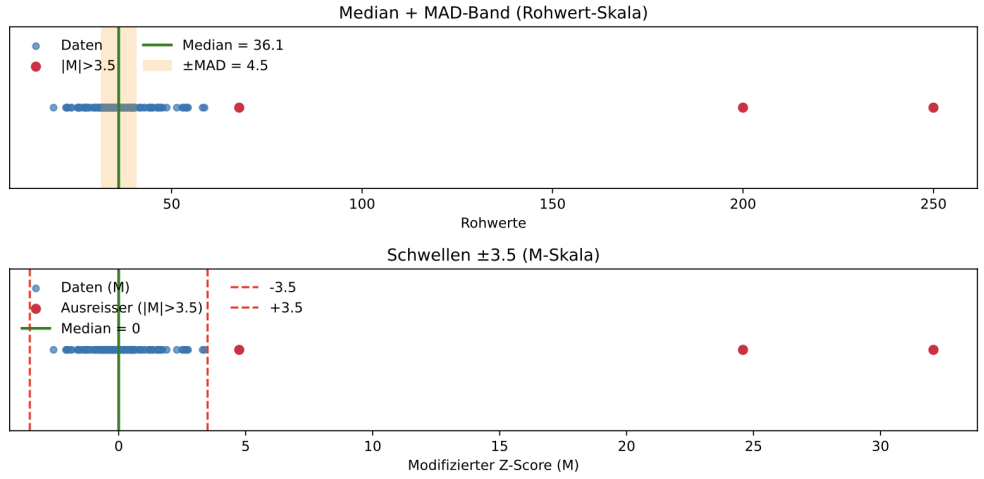
- Auf Basis des IQR: Untergrenze $= Q_1 - 1.5 \cdot IQR$, Obergrenze $= Q_3 + 1.5 \cdot IQR$

- Werte außerhalb gelten als Ausreißer-Kandidaten
- Vorteil: **robust, keine Verteilungsannahme**
- Standard in Boxplots



5.4 Modifizierter Z-Score mit MAD

- Robust gegen Schiefe & Heavy Tails
- $M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$
- Typische Schwelle: $|M| > 3.5 \rightarrow$ Ausreißer
- Sehr geeignet bei nicht-normalen Verteilungen



5.5 Strategien im Umgang

- **Markieren & berichten** (für Transparenz, nicht löschen)
- **Winsorisieren**: Extremwerte auf Grenzwert setzen (deckeln)
- **Trimmen**: Ränder entfernen (z. B. unterstes und oberstes 5 %)
- Immer **Entscheidung dokumentieren** (Regel, Schwelle, Datum, Variable, Datenversion)

5.6 Heavy Tails vs. echte Ausreißer

Heavy Tails = Verteilungen mit vielen großen Werten (z.B. Einkommen, Versicherungsfälle).
→ nicht automatisch Fehler

Echte Ausreißer = Datenfehler (Messungen, falsche Einheit etc.)

Diagnose nur mit Kombination: **Kennzahlen + Plots (Histogramm, KDE, ECDF, QQ)**

5.7 Pipeline für konsistentes Handling

Feste Reihenfolge, um Ad-hoc-Fehler zu vermeiden

- Kennzahlen berechnen (Mean, Median, SD, IQR, MAD)
- Plots prüfen (Boxplot, Histogramm, KDE, QQ-Plot)
- Regel anwenden (Z-Score, Tukey, mod. Z)
- Entscheidung dokumentieren (auch Varianten mit/ohne Ausreißer)

Zählen → Schauen → Handeln → Dokumentieren

6 Histogramm und KDE

6.1 Bin-Wahl

- **Zu wenige Bins** → glatte, aber irreführende Form (wichtige Details verschwinden)
- **Zu viele Bins** → zackig, verrauscht
- Regeln:
  - **Sturges**: konservativ, eher für normalverteilte Date
  - $\sqrt{n}$**-Regel**: einfache Faustregel
  - **Freedman–Diaconis (FD)**: robust, theoretisch fundiert, meist beste Wahl bei realen Daten

IMMER dokumentieren!

6.2 Histogramme interpretieren

- **Schiefe**: Rechts- oder linksschief erkennbar
- **Moden**: Anzahl der „Gipfel“ zeigt mögliche Mischungen/Subgruppen
- **Lücken**: können auf Clusterung oder fehlende Werte hinweisen
- **Tails**: zeigen Extremwerte oder Heavy Tails

6.3 Histogramm vs. KDE

- **Histogramm**
  - Intuitiv, zählt echte Häufigkeite
  - Stabil bei kleinen Stichproben
  - Abhängig von Bin-Wahl
- **KDE**
  - Glatt, zeigt Form der Verteilung deutlicher (Schiefe, Multimodalität)
    - abhängig von Bandbreite

7 Box- und Violinplot

7.1 Vergleich Box vs. Violin

- **Boxplot** - robust & kompakt (Median + IQR)
  - **Stärken**: Kompakt, robust, standardisiert, Ausreißer klar sichtbar
  - **Schwächen**: Zeigt keine Details der Verteilungsform

  - Boxplot mit Rohdatenpunkten (Strip/Swarm)
- **Violonplot** - formreich, zeigt Moden & Tails
  - **Stärken**: Zeigt Dichteform & Multimodalität, Quartile möglich
  - **Schwächen**: Bandbreitenabhängig, schwerer zu lesen
  - Best Practice: Violinplot mit Quartilen

7.2 Gruppenvergleiche

- Mehrere Gruppen nebeneinander vergleichen:
  - Boxplot → Fokus auf Median & IQR-Vergleich
  - Violinplot → Fokus auf Unterschiede in Form, Moden & Tails
- Interpretation:
  - Median A Median B → klare Lageunterschiede
  - Überlappung der IQRs → keine klare Trennung
  - Formunterschiede (Violin): Hinweis auf Mischungen oder Heterogenität in Gruppen

7.3 Zipfel des Violinplots

Basiert auf KDE (es wird über die Datenpunkte eine glatte Kurve gelegt, die theoretisch nie ganz aufhören) ⇒ Am unteren und oberen Rand der Violine entstehen manchmal kleine **„Zipfel“** (Bereiche, in denen die Kurve weiterläuft, obwohl dort keine echten Daten vorliegen) → **Rand-Effekt (boundary effect)**

- **Analytisch**: kein Problem – die Zipfel ändern nichts an der Aussage über Form, Schiefe oder Multimodalität
- **Visuell**: kann verwirrend wirken, weil es so aussieht, als gäbe es Werte außerhalb des tatsächlichen Datenbereichs

⇒ In Python/Seaborn (und auch in R): `sns.violinplot(data=df, x="species", y="flipper_length_mm", cut=0)`

- `cut = 0` = Kurve stoppt genau an den echten Daten → keine Zipfel mehr
- Die Zipfel verschwinden und die Violine endet genau dort, wo die Daten liegen

7.4 Stripplot und Swarmplot

Stripplot = Stellt jeden Datenpunkt als Punkt dar, optional mit leichtem Jitter (zufälliges horizontales Verschieben), um Überlagerungen zu vermeiden.

- **Vorteile**: Schnell, robust bei großem n
- **Nachteile**:Überdeckung möglich

Swarmplot = Ordnet die Punkte wie Bienenwaben an (kollisionsfrei). Dadurch wird die lokale Dichte sichtbar, keine Punkte überlappen.

- **Vorteile**: Lokale Dichte sichtbar, klar
- **Nachteile**: Langsamer, bei großem n überladen

Best Practice:

- Für kleine bis mittlere Stichproben: Swarmplot (zeigt Details)
- Für große Stichproben: Stripplot mit Transparenz (zeigt Verteilung, ohne zu überladen)

8 ECDF und QQ

8.1 ECDF (Empirical Cumulative Distribution Function)

- Empirische Verteilungsfunktion einer Stichprobe
- $\hat{F}(x) = \frac{1}{n} \sum 1\{x_i \leq x\}$
- Eigenschaften:
  - Monoton steigend von 0 bis 1
  - Jedes Datenpunkt erzeugt einen Sprung
  - Keine Bins nötig → **binfreie Darstellung**
- Ermöglicht das direkte Ablesen von **Quantilen**:
  - Median = Punkt bei $F(x) = 0.5$
  - Q1 bei $F(x) = 0.25$, Q3 bei $F(x) = 0.75$ → direkt aus der Kurve ablesbar
- Besonders hilfreich bei **kleinen Datensätzen** oder **schiefen Verteilungen**, weil Histogramme dort irreführend sein können
- Sehr geeignet für Gruppenvergleiche: mehrere ECDFs nebeneinander legen → Verschiebungen und Unterschiede sofort sichtbar

ECDF: Zahl → Prozent, Quantil: Prozent → Zahl

8.2 QQ-Plot (Quantile-Quantile-Plot)

- Vergleicht die **Quantile einer Stichprobe** mit den **Quantilen einer Referenzverteilung**
- Aufgetragen:
  - x-Achse = theoretische Quantile
  - y-Achse = empirische Quantile
- Liegen die Punkte auf einer Geraden, passt die Verteilung zur Referenz

Interpretation

- **Gerade Linie**: gute Übereinstimmung
- **Krümmung oben/unten**: Schiefe
- **Abweichungen in den Rändern**: Heavy Tails (dicke Ausläufer)
- **S-förmig**: dünnere Tails als Referenz (light tails)

QQ-Plot = Diagnose-Tool für Schiefe & Tails

8.3 Ausreißer-Analyse

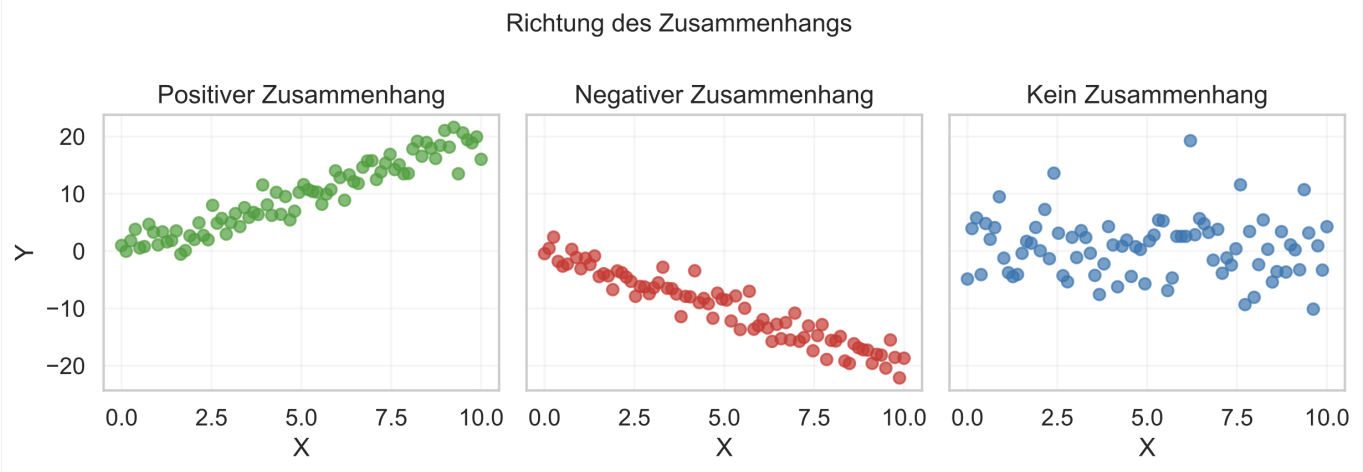
- **ECDF** → Quantile präzise ablesen, Gruppen robust vergleichen
- **QQ-Plot** → prüfen, ob Daten annähernd normalverteilt sind (wichtig für klassische Tests) oder ob Heavy Tails/Ausreißer vorliegen
- Beides ergänzt Histogramm & KDE:
  - Histogramm/KDE: visuell-intuitiv
  - ECDF/QQ: mathematisch präziser und robuster

9 Korrelation & Zusammenhang

9.0.1 Richtung und Stärke des Zusammenhangs

Der **Korrelationskoeffizient** r beschreibt Richtung und Stärke eines linearen Zusammenhangs zwischen zwei metrischen Variablen.
$r \in [-1, 1]$

- $r > 0 \rightarrow$ positiver Zusammenhang: hohe X-Werte gehen mit hohen Y-Werten einher
- $r < 0 \rightarrow$ negativer Zusammenhang: hohe X-Werte gehen mit niedrigen Y-Werten einher
- $r = 0 \rightarrow$ kein linearer Zusammenhang (aber es kann trotzdem ein nichtlinearer bestehen!)



Die Stärke wird durch den Betrag $|r|$ angegeben:

- $|r| \approx 0$: kein oder sehr schwacher linearer Zusammenhang
- $|r|$ nahe 1: starker Zusammenhang

9.0.2 Interpretation des Korrelationskoeffizienten r

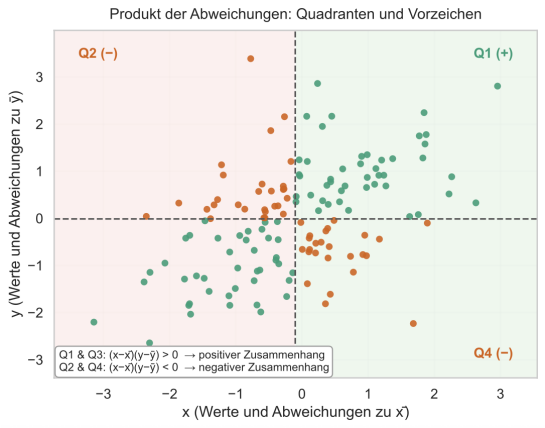
r-Wert	Richtung	Stärke	Interpretation (sprachlich)
-1.0 bis -0.9	negativ	sehr stark	Fast perfekter gegenläufiger Zusammenhang
-0.9 bis -0.7	negativ	stark	Klarer Trend: je mehr X, desto weniger Y
-0.7 bis -0.4	negativ	mittel	Merklicher, aber nicht dominanter Zusammenhang
-0.4 bis -0.2	negativ	schwach	Leichter gegenläufiger Trend, viel Rauschen
-0.2 bis 0.2	–	keiner	Praktisch kein linearer Zusammenhang
0.2 bis 0.4	positiv	schwach	Leichter gemeinsamer Trend
0.4 bis 0.7	positiv	mittel	Stabiler, aber nicht perfekter Zusammenhang
0.7 bis 0.9	positiv	stark	Klarer Trend: X und Y steigen gemeinsam
0.9 bis 1.0	positiv	sehr stark	Fast perfekter Gleichlauf

9.1 Korrelation vs. Kausalität

Korrelation beschreibt ein Muster, keine Ursache.
$X \uparrow \Rightarrow Y \uparrow \rightarrow$ also „X verursacht Y“ $\rightarrow$ **nicht zwingend richtig!**
Der beobachtete Zusammenhang kann durch Zufall, eine Drittvariable oder umgekehrte Kausalrichtung entstehen.
$\rightarrow$ Eine Konfundierung (Drittvariable) liegt vor, wenn eine dritte Variable Z sowohl X als auch Y beeinflusst.
$\rightarrow$ Nie ohne kontrolliertes **Experiment** oder **theoretisches Modell** aus Korrelation auf Kausalität schließen!

9.2 Kovarianz

Die Kovarianz misst, wie stark zwei Variablen gemeinsam von ihren Mittelwerten abweichen. → Sie zeigt, ob und wie X und Y „miteinander schwingen“: $\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$



- Quadrant I & III: beide Variablen über/unter Mittelwert → positive Produkte → positive Kovarianz
- Quadrant II & IV: eine über, die andere unter → negative Produkte → negative Kovarianz
- Wenn sich die Abweichungen zufällig ausgleichen → Produkt ≈ 0 → **kein linearer Zusammenhang**

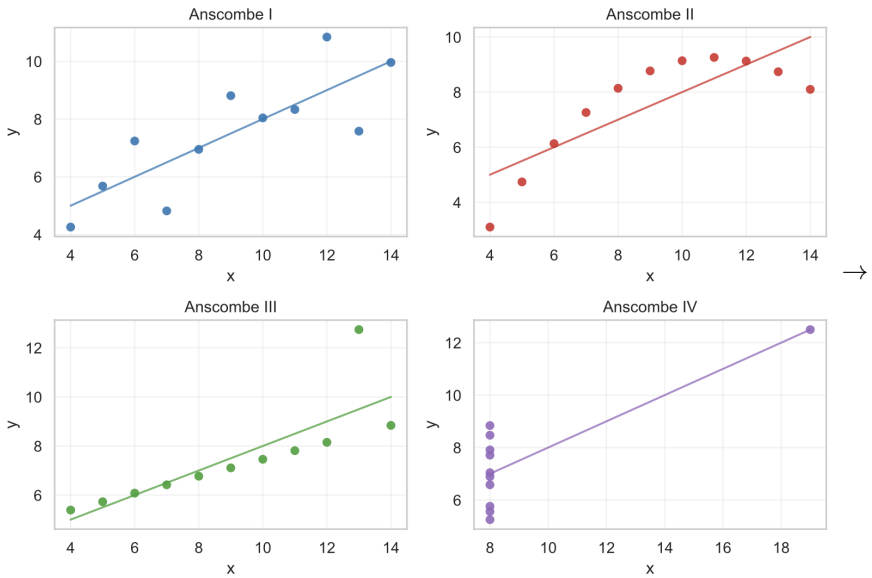
9.2.1 Grenzen der Kovarianz

- **Einheitenanhängig**:
  → z. B. Euro × Jahre oder cm × kg → schwer vergleichbar
- **Größenabhängig**:
  → größere Wertebereiche ⇒ automatisch größere Kovarianz
  → nicht normiert → keine Vergleichbarkeit zwischen Variablen oder Datensätzen
- **Lösung**: Normierung → **Pearson-Korrelation**

9.3 Anscombe's Quartett

Das Anscombe's Quartett (Francis John Anscombe, 1973) besteht aus einer Gruppe von vier Datensätzen, die grafisch sehr unterschiedlich aussehen, aber fast identische deskriptive Kennzahlen haben, z. B.:

- Gleicher Mittelwert von x und y
- Gleiche Varianz von x und y
- Gleiche Kovarianz
- Gleicher Korrelationskoeffizient r



In Zahlen sehen sie „gleich“ aus – aber **grafisch völlig unterschiedlich**! → Immer plotten!

10 Kovarianz & Pearson-Korrelation

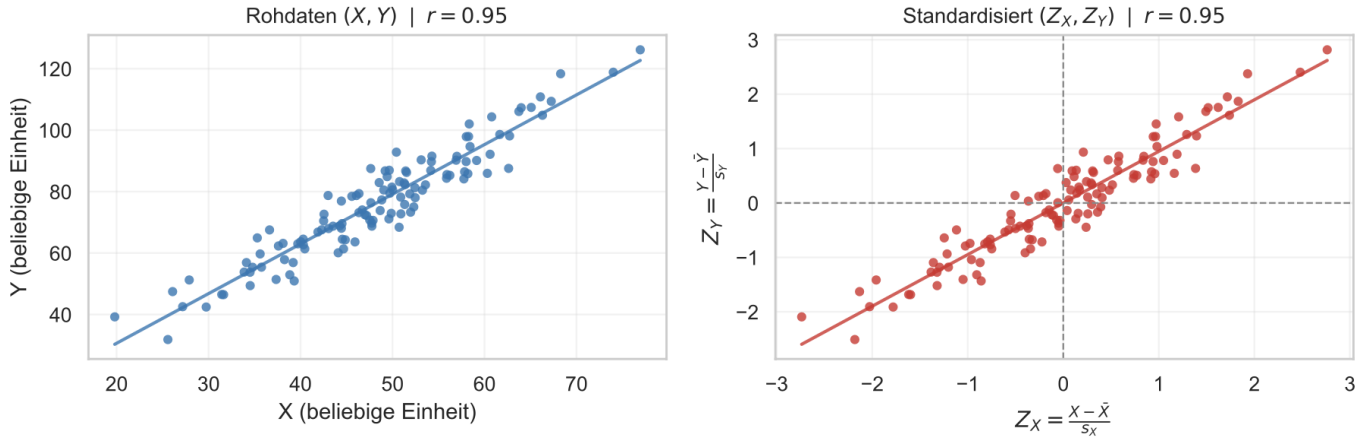
10.1 Von der Kovarianz zur Korrelation

Die Kovarianz ist nicht standardisiert und hängt von den Einheiten ab (z. B. CHF × Jahre).
→ Sie ist daher nicht vergleichbar zwischen Variablen oder Datensätzen (PROBLEM: Skalenabhängigkeit)
Lösung: Standardisierung durch Division mit den Standardabweichungen von X und Y.
$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y}$ ⇒ Das ergibt die **Pearson-Korrelation** r

10.2 Pearson-Korrelation

$r = \frac{\mathbf{Cov}(X,Y)}{s_X s_Y}$

- immer im Intervall $[-1, +1]$
- $+1$ = perfekter positiver linearer Zusammenhang
- –1 = perfekter negativer linearer Zusammenhang
- 0 = kein linearer Zusammenhang

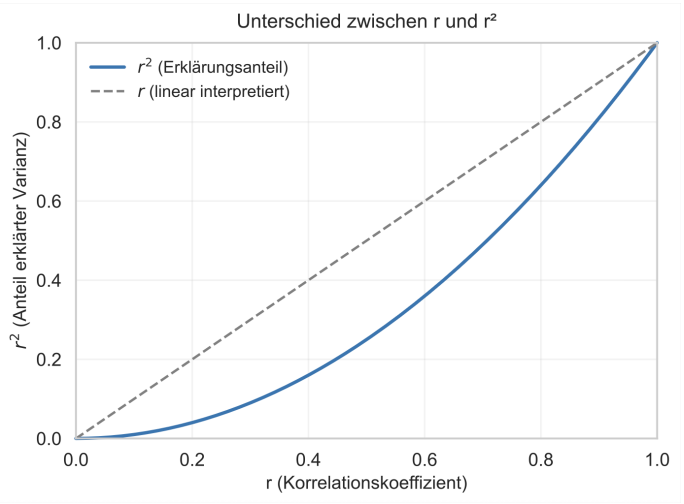


10.2.1 Wichtige Eigenschaften von Pearson r

- **Skalenunabhängig** → bleibt gleich, egal ob z. B. in CHF oder $
- **Empfindlich gegenüber Ausreißern!** → Ein einziger Extremwert kann r massiv verändern
- Nur für lineare Zusammenhänge sinnvoll → Nicht-lineare, aber monotone Beziehungen werden unterschätzt.
- Voraussetzungen:
  - metrische Variablen (Intervall-/Ratioskala)
  - annähernde Normalverteilung
  - keine starken Ausreißer

10.2.2 Anteil erklärter Varianz

r^2 als Bestimmtheitsmaß verwendet, der Korrelationskoeffizient r ist nicht der Erklärungsanteil!
r^2 = Anteil der Varianz von Y, der durch X erklärt wird.
Beispiel: $r = 0.8 \Rightarrow r^2 = 0.64 \rightarrow 64\%$ der Unterschiede in Y hängen mit X zusammen.
Aber: Das ist keine Kausalität!



11 Spearman & Kendall – Rangkorrelationen

Pearson-r misst nur lineare Zusammenhänge und reagiert empfindlich auf Ausreißer oder Schiefe. Aber in der Realität sind viele Zusammenhänge nichtlinear, aber trotzdem systematisch ⇒ Dafür brauchen wir rangbasierte Maße, die die Ordnung statt die exakten Werte vergleichen.

Spearman (ρ) und **Kendall** (τ) beruhen auf **Rängen statt Rohwerten**. → Extremwerte werden abgefedert und nur die relative Ordnung wird bewertet

11.1 Spearman-Korrelation (ρ)

Ersetzt die Werte durch **Ränge** und berechnet dann Pearson-r auf diesen Rängen: $\rho = r(\text{rank}(X), \text{rank}(Y))$
Das ergibt ein Maß für den monotonen Zusammenhang (egal ob linear oder gekrümmt).
Formel (für kleine Datensätze ohne „Ties“): $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$
mit d_i = Rangdifferenz der i-ten Beobachtung.

- $\rho = +1 \rightarrow$ perfekt gleichgerichtete Ränge
- $\rho = -1 \rightarrow$ perfekt umgekehrte Ränge
- $\rho = 0 \rightarrow$ kein monotoner Zusammenhang

Eigenschaften:

- Misst Monotonie, nicht nur Linearität
- Robust gegen Ausreißer (weil Ränge stabil bleiben)
- Keine Normalverteilungs-Annahme nötig
- Gleiche Einheit wie Pearson ($[-1, +1]$)

11.1.1 Gleiche Werte (Ties)

Wenn Werte gleich sind → d**urchschnittlicher Rang**:

Wert	Rang (ohne Ties)	Rang (mit Ties)
10	1	1
20	2	2.5
20	3	2.5
40	4	4

→ Beide „20er“ bekommen Rang $(2 + 3)/2 = 2.5$.

Das bewahrt die Ordnung und macht Spearman robust bei doppelten Werten!

11.2 Kendall's Tau (τ)

Kendall vergleicht **Paarordnungen**:
Für jedes Paar (x_i, y_i) und (x_j, y_j) prüft man:

- konkordant: beide gleich geordnet (steigend/steigend oder fallend/fallend)
- diskordant: entgegengesetzt geordnet

$\tau = \frac{\text{(konkordante Paare)} - \text{(diskordante Paare)}}{\text{Gesamtzahl der Paare}}$

Eigenschaften:

- Wertebereich $[-1, +1]$
- Etwas „gedämpfter“ als Spearman (kleinere Beträge)
- Besonders stabil bei vielen gleichen Rängen (Ties) oder kleinen Stichproben
- Misst dieselbe Idee wie Spearman, aber auf Paar-Ebene statt Rangdifferenzen

11.3 Vergleich Spearman vs. Kendall

Merkmal	**Spearman (ρ)**	**Kendall (τ)**
Idee	Pearson auf Rängen	Vergleich von Paaren
Misst	Monotone Ordnung	Konkordanz/Diskordanz
Robustheit	robust	sehr robust
Sensitivität	höher (größere Werte)	gedämpft
Eignung	große n, metrisch	kleine n, viele Ties
Bereich	$[-1, +1]$	$[-1, +1]$

12 Wann welche Korrelation?

Situation	**Empfehlung**
Lineare, metrische Beziehung	Pearson r
Monoton, aber nicht linear	Spearman ρ
Viele gleiche Werte oder kleine Stichprobe	Kendall τ
Ausreißer vorhanden	Rangmaß (ρ oder τ)
Ordinale Skalen (z. B. Zufriedenheit 1–5)	Rangmaß

- **Pearson**: misst **lineare Stärke**
- **Spearman**: misst **monotone Richtung**
- **Kendall**: misst **Rangordnung**

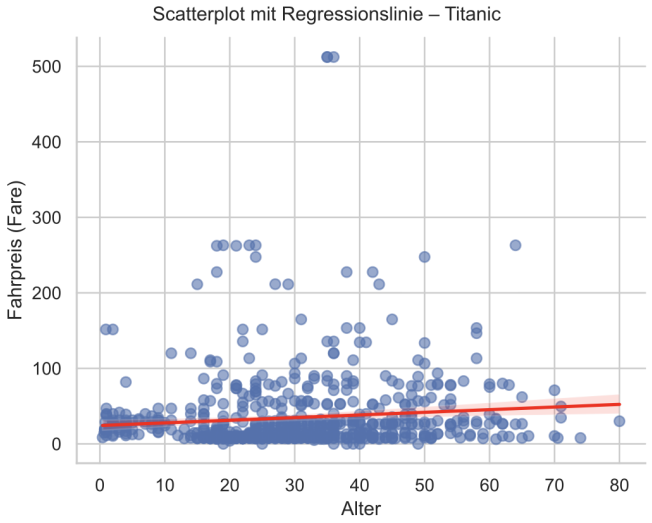
13 Visualisierung von Korrelationen

13.1 Regressionslinie im Scatterplot

Eine Regressionslinie (Trendlinie) verdeutlicht den Zusammenhang – **ohne Kausalität zu implizieren!**

- **Steigung**: Richtung des Zusammenhangs
- **Schatten (Konfidenzintervall)**: Unsicherheit
- **Nichtlinearität**: erkennbar, wenn Punkte deutlich von der Linie abweichen

Wenn die Linie gekrümmt oder unpassend erscheint → Pearson ist ungeeignet → Spearman oder Kendall nutzen



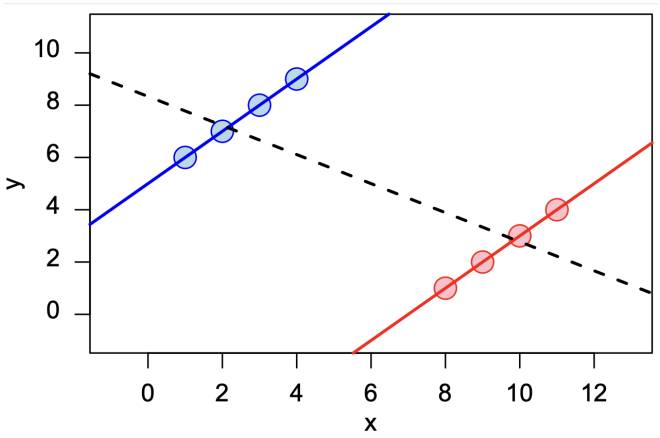
13.2 Best Practice

- Achsen immer beschriften + Einheiten angeben
- Farben sinnvoll nutzen
- Einheitliche Skalen bei Gruppenplots
- Transparenz (alpha) gegen Overplotting
- Trends immer kritisch interpretieren – **Trend ≠ Ursache!**

14 Simpson-Paradox

Korrelation ist ein mächtiges Werkzeug, aber in der Praxis häufig falsch interpretiert. Viele scheinbar „offensichtliche“ Trends verschwinden oder kehren sich sogar um, wenn man Drittvariablen oder Gruppen berücksichtigt.→ Simpson-Paradox

Definition: Das Simpson-Paradox beschreibt den Effekt, dass ein Zusammenhang zwischen zwei Variablen in aggregierten Daten anders oder sogar entgegengesetzt ist als in den einzelnen Gruppen.
→ Konfundierung: Eine Drittvariable (Z) verzerrt den scheinbaren Zusammenhang zwischen X und Y.
Ohne Kontrolle von Z ⇒ Scheinkorrelation



Praxisfallen

- Korrelation ≠ Kausalität
- Ausreißer & Nichtlinearität
- Heterogene Gruppen (Aggregationseffekte)

14.1 Umgang mit Simpson-Effekt

Stratifizierte Analysen vermeiden Fehlschlüsse. Immer auf versteckte Variablen prüfen!

- Stratifizierung: Daten nach Konfundierer aufsplitten und separate Analysen pro Gruppe durchführen
  → Vergleich Gesamt vs. innerhalb: Wenn sie widersprüchlich sind → Gefahr!
- Partielle Korrelation (lineares Herausrechnen des Z-Einflusses)
- Regressionsmodelle mit Kontrollvariablen

14.2 Partielle Korrelation

Um echte Zusammenhänge zu erkennen, musst du Drittvariablen (Z) herausrechnen → Die partielle Korrelation **misst** dann **den Zusammenhang zwischen zwei Variablen (X und Y)**

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

r_{XY}: einfache (bivariate) Korrelation zwischen X und Y

r_{XZ}, r_{YZ}: Korrelationen von X und Y mit Z

Interpretation:

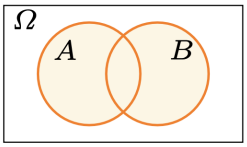
- Wenn $r_{XY \cdot Z} \approx r_{XY}$: Kaum Einfluss von Z
- Wenn $r_{XY \cdot Z} \approx 0$, obwohl r_{XY} hoch war: der scheinbare Zusammenhang X–Y war vollständig durch Z erklärt (Konfundierung!)
- $r_{XY \cdot Z} \neq r_{XY}$: zeigt, wie stark Z den ursprünglichen Zusammenhang verzerrt hat

15 Wahrscheinlichkeit und Verteilungen

15.1 Venn-DIagramme

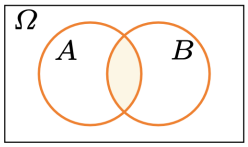
Kommutativgesetz	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Assoziativgesetz	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributivgesetz	$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$	$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

$A \cup B$



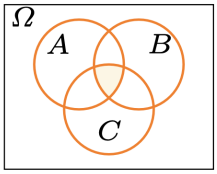
Vereinigung: Mindestens eines der Ereignisse tritt ein.

$A \cap B$



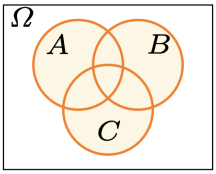
Schnitt: Beide Ereignisse treten gleichzeitig ein.

$A \cap B \cap C$



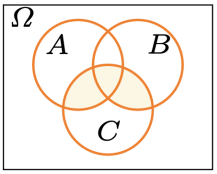
Dreifacher Schnitt: Alle drei Ereignisse treten gemeinsam ein.

$A \cup B \cup C$



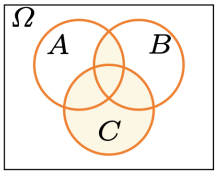
Dreifache Vereinigung: Mindestens eines der Ereignisse tritt ein.

$(A \cup B) \cap C$



Distributivbeispiel 1: Nur Teilmengen, in denen C auftritt.

$A \cap B \cup C$



Distributivbeispiel 2: A und B zusammen oder C allein.

15.2 Die drei Axiome nach Kolmogorov

1. **Nichtnegativität**: $Pr(A) \geq 0$

2. **Normierung**: $Pr(\Omega) = 1$

3. **Additivität**: Falls $A \cap B = \varnothing \Rightarrow Pr(A \cup B) = Pr(A) + Pr(B)$

15.3 Das Gesetz vom komplementären Ereignis

Jedes Ereignis hat ein Gegenteil: zusammen ergeben sie Sicherheit. Komplement: A^c = Ä tritt nicht ein"
$Pr(A^c) = 1 - Pr(A)$

15.4 Das Gesetz der Vereinigung

Berücksichtigt Überschneidungen: $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

- „Oder" ist **inklusiv**
- Wenn A und B unabhängig sind: $Pr(A \cap B) = Pr(A) \times Pr(B)$

15.5 Unabhängigkeit

Das eine Ereignis verändert nicht die Wahrscheinlichkeit des anderen.
$Pr(A \cap B) = Pr(A) \times Pr(B)$
Im Venn-Diagramm: Überlappung vorhanden, aber **Flächenverhältnis bleibt konstant**.
Unabhängigkeit $\neq$ keine Überlappung!

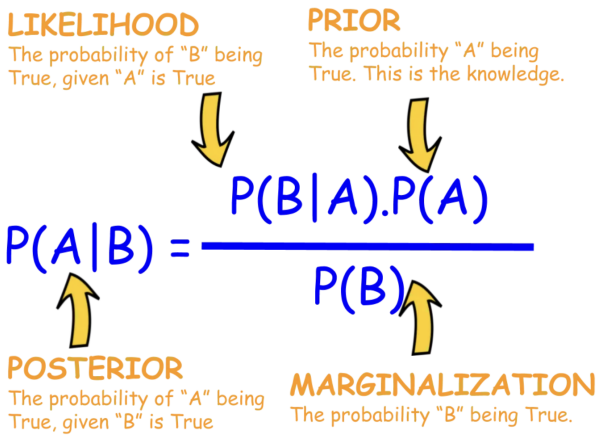
15.6 Bedingte Wahrscheinlichkeit

Wenn zusätzliche Information vorhanden ist, verändert sich die Wahrscheinlichkeit.
$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$
$Pr(A|B)$ = Wahrscheinlichkeit von A, unter der Bedingung, dass B eingetreten ist
$Pr(A|B) \neq Pr(B|A)$!

15.6.1 Das Theorem von Bayes

Bayes zeigt, **wie Vorwissen (Prior)** mit **neuer Evidenz (Daten)** kombiniert wird.
$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$

- **Prior**: Vorwissen über B
- **Likelihood**: Wie plausibel sind Daten unter B
- **Evidence**: Gesamtwahrscheinlichkeit der Beobachtung
- **Posterior**: Aktualisierte Überzeugung nach Daten



Bayesian Thinking: Wissen = dynamisch → Posterior = Prior × Evidenz

15.7 Risikomasse

Warum? Statistik untersucht nicht nur Wahrscheinlichkeiten, sondern auch Unterschiede zwischen Wahrscheinlichkeiten – also, ob ein bestimmter Faktor (E) das Risiko eines Ereignisses (D) verändert.

15.7.1 Risikodifferenz

Zeigt den absoluten Zusatznutzen oder -schaden durch den Faktor E.
$ER = Pr(D|E) - Pr(D|E^c)$ → Misst den absoluten Unterschied in Prozentpunkten

15.7.2 Relatives Risiko (RR) & Odds Ratio (OR)

Diese Maße zeigen den **Faktor**, um den ein Risiko **steigt oder sinkt**.
Relatives Risiko: $RR = \frac{Pr(D|E)}{Pr(D|E^c)}$
Gibt an, wie viel-fach höher (oder niedriger) das Risiko mit E ist.
Odds Ratio: $OR = \frac{Pr(D|E)/(1-Pr(D|E))}{Pr(D|E^c)/(1-Pr(D|E^c))}$
Misst das Verhältnis der Quoten (odds), also der Chancen für D relativ zu D^c. → Wird in der logistischen Regression verwendet (Machine-Learning-Bezug)
Bei **seltenen Ereignissen** gilt näherungsweise: $OR \approx RR$.

15.8 Zufallsvariablen & Verteilungen

- **Binomial**: Zählprozesse, n Versuche, p = Erfolgs-Wahrscheinlichkeit

- $Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- **Poisson**: Seltene Ereignisse im Intervall

- $Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$
- **Normal**: Summeneffekte / Messfehler → Symmetrisch
- **Exponential**: Zeit bis zum nächsten Ereignis
- **Uniform**: vollständige Unkenntnis – alle gleich wahrscheinlich

15.9 Erwartungswert

Der Erwartungswert $E(X)$ ist der Schwerpunkt einer Verteilung – also das mathematische Mittel aller möglichen Ausgänge.
Er beschreibt den langfristigen Durchschnitt, nicht das Ergebnis eines einzelnen Experiments.
Diskret: $E(X) = \sum x_i \, Pr(X = x_i)$
Kontinuierlich: $E(X) = \int x \, f(x) \, dx$

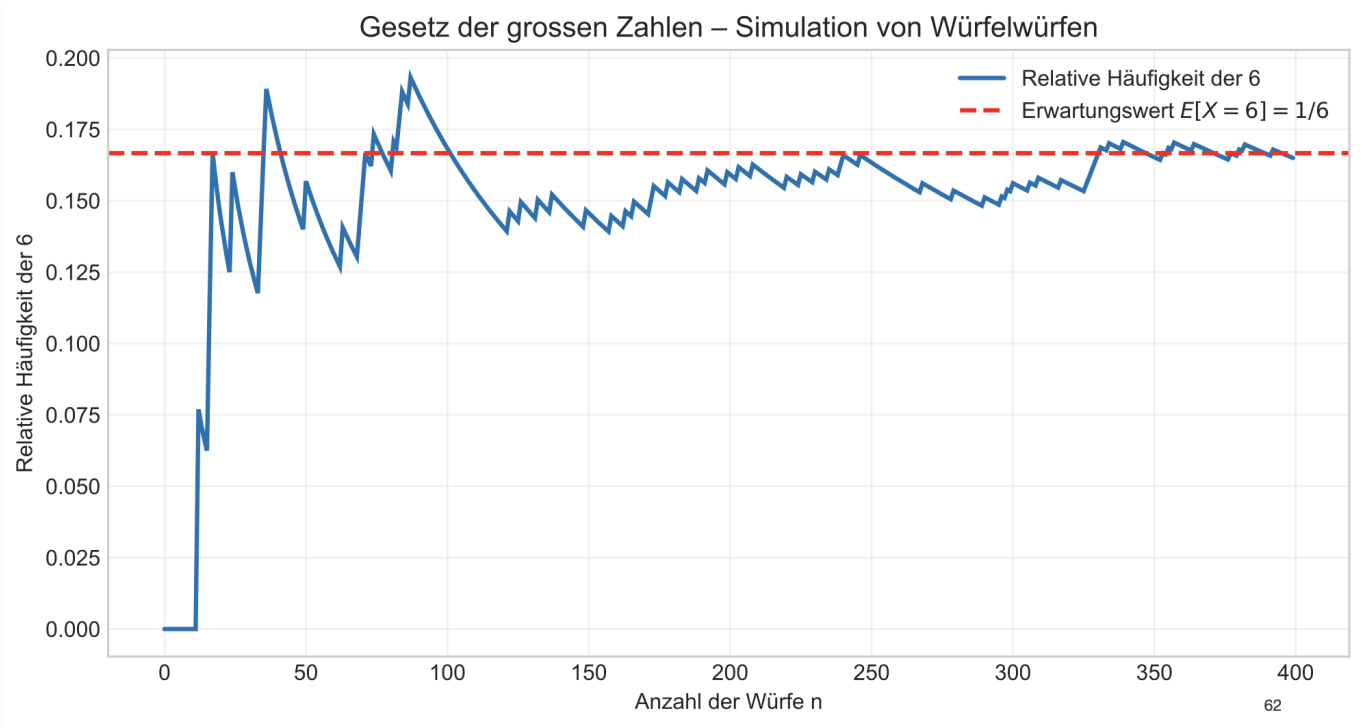
15.10 Varianz und Standardabweichung

Die Varianz misst, wie weit Zufallswerte vom Erwartungswert entfernt liegen → Streuung oder Unsicherheit der Verteilung
$Var(X) = E[(X - E(X))^2], \quad SD = \sqrt{Var(X)}$

- Kleine Varianz → Werte liegen eng um den Erwartungswert
- Große Varianz → starke Streuung, unregelmäßiger Zufall

15.11 Gesetz der grossen Zahlen (LLN)

Viele Zufälle ergeben Regelmäßigkeit.
Wenn $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind, gilt: $\bar{X}n = \frac{1}{n} \sum i = 1^n X_i \xrightarrow[n\to\infty]{} E[X]$
Der **Stichprobenmittelwert konvergiert** gegen den Erwartungswert.



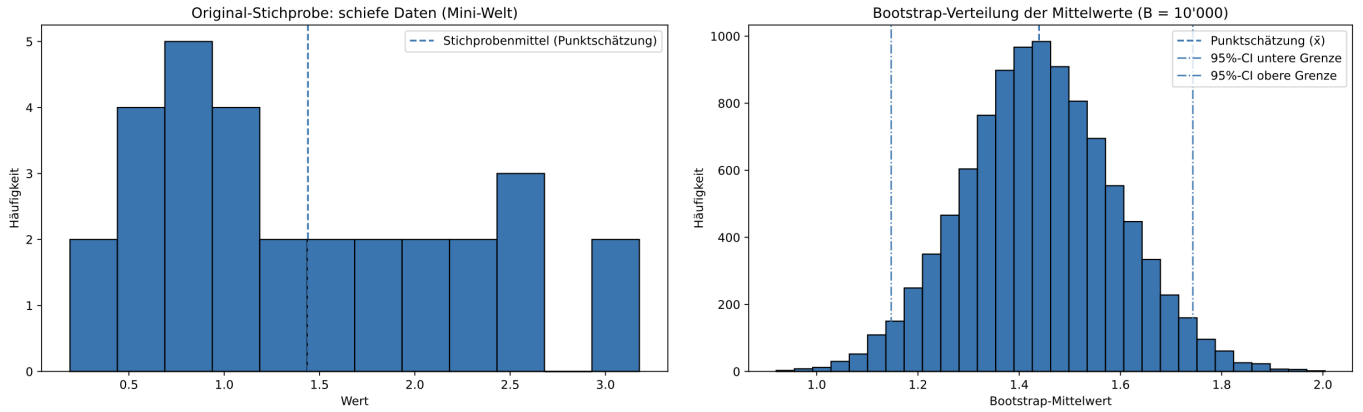
⇒ **Zufall wird vorhersagbar**

- Einzelbeobachtung = Rauschen; viele Beobachtungen = Signal
- Aggregierte Werte (Mittel, Anteile) → verlässlichere Information
- In Data Science: Durchschnitt = Signal über Rauschen
- LLN bedeutet: „Rauschen löscht sich im Schnitt aus."

16 Schätzen & Konfidenzintervalle

16.1 Bootstrap – empirische Unsicherheitsschätzung

Konfidenzintervalle und Standardfehler basierten auf theoretischen Verteilungen (z. B. Normal- oder t-Verteilung). Problem: Diese Annahmen sind oft unrealistisch – insbesondere bei: kleinen Stichproben, unbekannter oder schiefer Verteilung, nichtparametrischen Verfahren.
Bootstrap bietet eine **verteilungsfreie**, **empirische Methode**, um die Unsicherheit (z. B. Standardfehler, CI) zu schätzen. → simuliert wiederholte Stichproben aus der vorhandenen Stichprobe selbst



16.2 Margin of Error

Wenn eine Stichprobe verwendet wird, um einen Parameter der Population, wie den Anteil oder den Mittelwert, zu schätzen, ist diese Schätzung immer mit Unsicherheiten verbunden. Der Margin of Error quantifiziert diese Unsicherheit, indem er einen Bereich um die Schätzung angibt, der wahrscheinlich den tatsächlichen Wert enthält.
Der Margin of Error hängt von mehreren Faktoren ab:

- **Stichprobengröße (n)**: Je größer die Stichprobe, desto kleiner der MoE, weil die Schätzung genauer wird.
- **Stichprobenfehler (σ oder p)**: Die Streuung der Daten beeinflusst den MoE.
- **Konfidenzniveau ($1-\alpha$)**: Das ist die Wahrscheinlichkeit, mit der der wahre Wert im angegebenen Bereich liegt. Typische Werte sind 90%, 95% oder 99%.

Die allgemeine Formel für den Margin of Error bei einem Mittelwert lautet:
$MoE = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
Hierbei ist:
→ $z_{\alpha/2}$ der kritische Wert aus der Standardnormalverteilung, der vom gewünschten Konfidenzniveau abhängt (z. B. ca. 1,96 für 95%).
→ σ die Standardabweichung der Population (bei unbekannter σ wird die Stichprobenstandardabweichung verwendet).
Bei Anteils-Schätzungen (z. B. Prozentsätzen) lautet die Formel:
$MoE = z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$, wobei p der geschätzte Anteil ist.

16.3 Standardabweichung und Standardfehler

SD: Wie stark streuen die Daten innerhalb einer Stichprobe?
→ Datenebene
→ beschreibt die Variabilität in deiner Stichprobe
SE: Wie stark streut ein Schätzer (z. B. der Mittelwert) zwischen verschiedenen Stichproben?
→ Schätzer
→ beschreibt die Unsicherheit deiner Schätzung, also die Streuung von $\bar{X}$ oder $\hat{p}$ über viele gedachte Stichproben.
Für den Mittelwert: $E(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ (in der Praxis: $\widehat{SE}(\bar{X}) = \frac{s}{\sqrt{n}}$)
Für einen Anteil:$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ (in der Praxis: $\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$)

16.3.1 Quantile

Quantile geben an, wie weit wir vom Zentrum der Verteilung weggehen müssen, um z. B. 95 % abzudecken.
Konfidenzintervall (KI) = Schätzer ± (Quantil · SE).
α = Irrtumswahrscheinlichkeit (z.B. $\alpha = 0.05$ für 95 %-KI)
KI-Idee (für µ) mit z: $\bar{x} \pm z_{1-\alpha/2} \cdot SE(\bar{X}) \quad \Rightarrow \quad \bar{x} \pm 1.96 \cdot SE(\bar{X})$ bei 95 %

16.4 Typische (Stichproben) Bias-Arten

Anforderungen an die Stichprobe:

1. Zufällig (Jedes Element der Population hat die gleiche Chance, ausgewählt zu werden)→ vermeidet systematische Verzerrungen
2. Unabhängig (Der Wert einer Beobachtung beeinflusst nicht den Wert einer anderen) → Formeln für Varianz/SE stimmen
3. Repräsentativ (Die Stichprobe spiegelt die Vielfalt und Struktur der Population wider) → Schätzung ist wirklich übertragbar auf die Population

- textbfSelektionsbias: Nur bestimmte Personen werden befragt
- textbfNonresponse-Bias: Wer nicht antwortet, unterscheidet sich vom Rest
- textbfAbhängigkeiten: Daten stammen aus Clustern oder Zeitreihen

16.5 z-Verteilung

= Standardnormalverteilung mit $\mu = 0, \quad \sigma = 1$
Wir standardisieren: $z = \frac{x-\mu}{\sigma} \rightarrow$ z gibt an, wie viele Standardabweichungen ein Wert vom Mittelwert entfernt ist.

- symmetrisch, glockenförmig
- ca. 68 % der Werte zwischen -1 und +1
- ca. 95 % der Werte zwischen -1.96 und +1.96

In Konfidenzintervallen verwenden wir z-Quantile, wenn

- die Verteilung (nahezu) normal ist und
- die Populationsstreuung μ bekannt ist oder n so gross, dass $s \approx \sigma$ (CLT)

16.6 t-Verteilung

n der Praxis kennen wir μ meistens nicht, sondern schätzen sie aus der Stichprobe durch s. → Mehr Unsicherheit, insbesondere bei kleinem n.
t-Statistik: $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \rightarrow$ Wie z, aber mit s statt $\sigma \rightarrow$ zusätzliche Unsicherheit → dickere Tails.

- breitere Flanken (mehr Masse in den Rändern) als z-Verteilung
- hängt von den Freiheitsgraden $df = n - 1$ ab:
  - kleine df → sehr breit, unsicher
  - grosse df → nähert sich der z-Verteilung an

Für Konfidenzintervalle bei kleiner Stichprobe: $\bar{x} \pm t_{df,\,1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$

16.7 Punktschätzung

Idee: Schätzen liefert Zahlen, mit denen man Entscheidungen treffen kann.

16.7.1 Bias & Varianz

Bias: systematischer Fehler
Varianz: Zufallsschwankung
Beim Schätzen: Jede Stichprobe ist anders → jeder Schätzwert etwas anders ⇒ **Varianz des Schätzers**
→ Warum sind Schätzungen zufällig?

1. Zufällige Stichprobe: wer/was in die Stichprobe kommt
2. Modellunsicherheit: das angenommene Modell (z. B. Normalverteilung, Linearität) passt nie perfekt

⇒ Schätzer haben keine fixe Zahl, sondern eine Verteilung:

- Mittelwert dieser Verteilung → Bias (Abstand zur Wahrheit)
- Streuung dieser Verteilung → Varianz / SE

⇒ Mehr Daten → Varianz kleiner → stabilere Schätzungen

Modell-Bias entsteht, wenn wir bewusst ein vereinfachtes Modell verwenden

- „Modell“ = Annahmen über:
  - Verteilungsform
  - funktionale Beziehung
  - konstante Varianz, Unabhängigkeit usw.
- Diese Annahmen sind nie perfekt wahr → leichter systematischer Fehler (Modell-Bias)

→ Vereinfachte Modelle sind stabiler, interpretierbar und oft ausreichend genau für Entscheidungen

16.7.2 Gütekriterien & MSE

Gute Schätzer balancieren:

1. kleinen Bias (im Mittel nicht weit daneben)
2. kleine Varianz (Stichproben-Schätzungen schwanken wenig)

Beides zusammen fasst der Mean Squared Error (MSE): $\mathrm{MSE}(T) = \mathbb{E}\big[(T-\theta)^2\big] = \mathrm{Bias}(T)^2 + \mathrm{Var}(T)$
Ziel: **MSE minimieren** → guter Kompromiss zwischen:

- „nicht schief“ (Bias klein)
- „nicht zitterig“ (Varianz klein)

→ Ein ganz unverzerrter Schätzer ist nicht unbedingt am besten, wenn seine Varianz sehr gross ist
→ Leicht verzerrter Schätzer mit viel kleinerer Varianz kann einen kleineren MSE haben → insgesamt „besser“

17 Hypothesentests

17.1 Grundidee eines Hypothesentests

1. Formuliere H_0 (Nullhypothese):
   - „Kein Effekt“, „kein Unterschied“, „Parameter = Referenzwert“.
   - Beispiele:
     - Medikament hat keinen Effekt: $\mu_{\text{Med}} = \mu_{\text{Placebo}}$
     - Kein Gender-Pay-Gap: $\mu_{\text{Frau}} = \mu_{\text{Mann}}$
     - Anteil = 0,5: $p = 0{,}5$
2. Formuliere H_1 (Alternativhypothese):
   - „Es gibt einen Effekt / Unterschied“.
   - Beispiele: $\mu_{\text{Med}} \neq \mu_{\text{Placebo}},\ \mu_{\text{Frau}} \neq \mu_{\text{Mann}},\ p \neq 0{,}5$
   - Einseitig ($>$, $<$) oder zweiseitig ($\neq$) – je nach Fragestellung.
3. Wähle eine Teststatistik T.
   - Eine Zahl, die „Abweichung von H_0“ misst.
   - Typisch: z- oder t-Statistik, Differenz der Mittelwerte, Anteilschätzer usw.
4. Bestimme die Verteilung von T unter H_0.
   - z. B. Standardnormalverteilung oder t-Verteilung mit $\mathrm{df} = n - 1$.

5. Berechne den beobachteten Wert t_{obs} aus den Daten.
6. Vergleiche t_{obs} mit der Nullverteilung (p-Wert).
   - p-Wert = Wahrscheinlichkeit (unter H_0), einen Wert mindestens so extrem wie t_{obs} zu erhalten.
   - kleiner p-Wert $\Rightarrow$ starke Evidenz gegen H_0.
7. Entscheidungsregel:
   - Wenn $p \leq \alpha \Rightarrow H_0$ verwerfen („statistisch signifikant“).
   - Wenn $p > \alpha \Rightarrow H_0$ nicht verwerfen (Daten reichen nicht, um H_0 zu kippen).

17.2 Fehler 1. und 2. Art & Teststärke (Power)

Realität \ Entscheidung	H_0 *nicht* verwerfen	H_0 verwerfen
H_0 ist wahr	korrekt	Fehler 1. Art (α)
H_0 ist falsch	Fehler 2. Art (β)	korrekt (Power $= 1 - \beta$)

- Fehler 1. Art (α)
  - H_0 ist wahr, aber wir verwerfen sie.
  - False Positive“, Fehlalarm.
  - Signifikanzniveau α ist die erlaubte Wahrscheinlichkeit für diesen Fehler (z. B. 5 %).
  - Wird vom Forschenden festgelegt
- Fehler 2. Art (β)
  - H_0 ist falsch, aber wir verwerfen sie nicht.
  - „False Negative“, übersehener Effekt
- Teststärke / Power $= 1 - \beta$
  - Wahrscheinlichkeit, H_0 korrekt zu verwerfen, wenn sie tatsächlich falsch ist
  - Wir wollen: α klein und Power gross

18 Glossar

18.1 Lagekennzahlen

- **Mittelwert ($\bar{x}$)**: Summe aller Werte geteilt durch Anzahl.
  Eigenschaften: linear, effizient, aber **ausreißerempfindlich**
- **Median ($Q_{0.5}$)**: 50%-Quantil: teilt Daten in zwei gleich große Hälften.
  Minimiert die Summe absoluter Abweichungen, **robust gegen Ausreißer**.
  Typisch besser bei schiefen Verteilungen (z. B. Einkommen, Wartezeiten)
- **Getrimmter Mittelwert**: Entfernt einen Prozentsatz der kleinsten und größten Werte (z. B. je 10 %).
  Kompromiss: mittelt wie das Mean, aber **robuster**.
  Unterschied zur Winsorisierung: dort werden Extremwerte nicht entfernt, sondern auf den nächstinneren Wert gesetzt
- **Modus**: Am häufigsten vorkommender Wert. Besonders wichtig für **kategoriale Daten**.
  Hinweis auf Multimodalität (mehrere Peaks $\rightarrow$ mögliche Subgruppen oder Mischverteilungen)
- **Quantile und Perzentile**: Werte, die Daten in gleich große Teile teilen (z. B. Quartile = 4 Teile).
  Basis für den **Interquartilsabstand (IQR)**, nützlich bei schiefen Verteilungen. Quantile sind **robuste Maße**.

18.2 Streuungskennzahlen

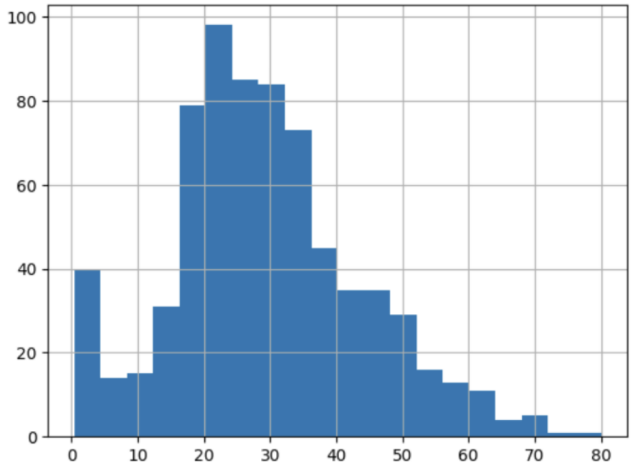
- **Varianz (s^2)**: Durchschnitt der quadrierten Abweichungen vom Mittelwert: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  quadriert → schwer interpretierbar, **ausreißerempfindlich**
- **Standardabweichung (SD, s)**: Quadratwurzel der Varianz
  Einheit = wie Originaldaten → **leichter interpretierbar**, **ausreißerempfindlich**
- **Range (Spannweite)**: Differenz zwischen Maximum und Minimum
  Extrem anfällig für Ausreißer → selten als zentrales Maß genutzt
- **Interquartilsabstand (IQR)**: Differenz zwischen dem dritten und dem ersten Quartil: $IQR = Q_3 - Q_1$
  → misst die mittleren 50 % der Daten
  Robust gegenüber wenigen Extremwerten
  Immer zusammen mit Median berichten (empfohlen bei schiefen Verteilungen)
- **Mittlere absolute Abweichung (MAD – Median Absolute Deviation)**: Median der Abstände der Werte zum Median: $MAD = \text{Median}(|x_i - \tilde{x}|)$
  Sehr robust, auch bei Schiefe und Heavy Tails
  Wird oft skaliert mit Faktor **1.4826**, um auf die gleiche Skala wie SD zu kommen (bei Normalverteilung)
  Nutzt den Median statt Mean → stabil bei Ausreißern

18.3 Ausreisser

- **Ausreisser**: Werte, die signifikant vom Rest der Daten abweichen. Sie können Messfehler, Eingabefehler oder seltene, aber valide Beobachtungen sein.
- **Tukey Fences**: Eine einfache und **robuste** Methode zur Identifizierung von Ausreisserkandidaten basierend auf dem IQR. Werte, die unter $Q_1 - 1.5 \cdot IQR$ oder über $Q_3 + 1.5 \cdot IQR$ liegen, gelten als Ausreisser.
- **Modifizierter Z-Score**: Eine **robuste** Alternative zum klassischen Z-Score, die den Median und die MAD anstelle des Mittelwerts und der Standardabweichung verwendet. Besser geeignet für schiefe Verteilungen.

18.4 Histogramm

Ein Diagramm, das die Häufigkeitsverteilung von Daten durch rechteckige Säulen darstellt. Jede Säule repräsentiert die Häufigkeit (oder Dichte) von Werten in einem bestimmten Intervall (**Bin**).
$f_j = \frac{c_j}{n \cdot h}$
c_j = Anzahl der Werte im Intervall
n = Gesamtzahl der Werte
h = Breite des Intervalls



Bin-Wahl: Die Entscheidung für die Breite der Intervalle (Bins) in einem Histogramm. Eine falsche Wahl kann die Interpretation verzerren. Regeln wie die **Freedman-Diaconis-Regel (FD)** helfen, eine optimale Bin-Breite zu finden.

18.5 Kernel Density Estimation (KDE)

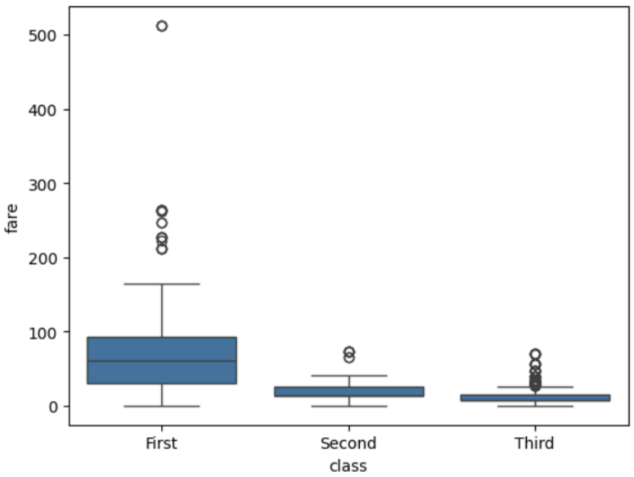
Methode zur Schätzung der Wahrscheinlichkeitsdichtefunktion einer Zufallsvariable. Erzeugt (summiert) eine glatte, kontinuierliche Kurve ohne diskrete Bins.
$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$, h = Bandbreite

- Kleine Bandbreite: viele Details, evtl. Rauschen
- Große Bandbreite: glatt, aber Details verschwinden

→ **Scott-Regel**: gut für normalverteilte Daten
→ **Silverman-Regel**: robuster bei schiefen Daten
Bandbreite (h): Schlüsselparameter bei der KDE, der die Glättung der Kurve steuert. Kleine Bandbreite → zackige Kurve; grosse Bandbreite → zu stark geglättet.

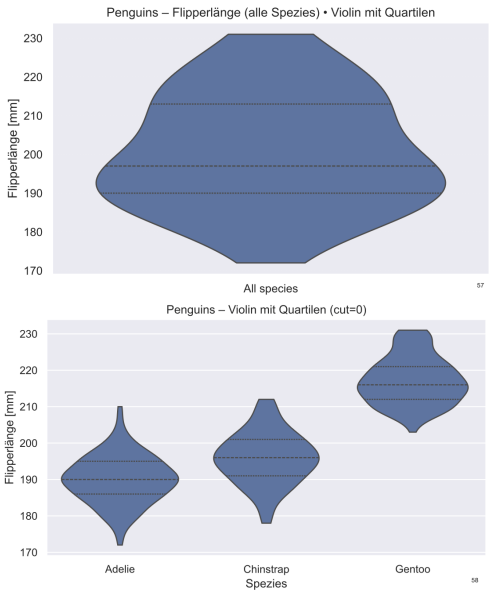
18.6 Boxplot

Darstellung, die eine Verteilung durch fünf Kennzahlen zusammenfasst: Minimum, erstes Quartil (Q_1), Median, drittes Quartil (Q_3), Maximum (oft mit Whisker, die Ausreisser ausschliessen). Bietet Überblick über Lage, Streuung und Ausreisser.
Robuste Kennzahlen:



- **Median** (dicke Linie in der Box)
- **Box** = Interquartilsabstand (Q1 bis Q3)
- **Whisker**: reichen bis zu 1.5 × IQR über Q1/Q3 hinaus
- **Punkte außerhalb** = Ausreißer-Kandidaten
- Punkte außerhalb = Ausreißer-Kandidaten
- Log-Skala: sinnvoll bei stark schiefen Daten
- whis-Parameter: alternativ zu 1.5 × IQR z. B. Perzentile (5–95 %)

18.7 Violinplot



Kombination aus Boxplot und KDE. Die Form zeigt die Dichtefunktion, während innere Markierungen (z. B. Kästen, Linien) die Quartile und den Median zeigen. Nützlich für Gruppenvergleiche.

18.8 Strip- und Swarmplots

Visualisieren individuelle Rohdatenpunkte. Ergänzen Box- oder Violinplots, um die tatsächliche Verteilung der Punkte zu zeigen.

18.9 ECDF und QQ-Plot

- **Empirical Cumulative Distribution Function (ECDF)**: Binfreie und vollständige Darstellung der kumulativen Verteilung. Zeigt für jeden Wert den Anteil der Daten $\leq$ diesem Wert. Sehr informativ, auch bei kleinen Datensätzen. Erlaubt direktes Ablesen von Quantilen.
- **Quantile-Quantile Plot (QQ-Plot)**: Diagramm, das die Quantile einer Stichprobe gegen die Quantile einer theoretischen Verteilung (z. B. Normalverteilung) aufträgt. Punkte auf einer Geraden $\rightarrow$ Verteilungen stimmen überein. Abweichungen zeigen **Schiefe** oder **schwere Tails**.
- **Schwere Tails (Heavy Tails)**: Eigenschaft einer Verteilung, bei der die Wahrscheinlichkeit für extreme Werte höher ist als bei einer Normalverteilung. Im QQ-Plot sichtbar durch Krümmung an den Enden.

18.10 Korrelation

Korrelation beschreibt, wie stark und in welche Richtung zwei Variablen miteinander zusammenhängen.
Sie misst, ob Veränderungen in einer Variable systematisch mit Veränderungen in einer anderen einhergehen.
„Ändert sich Y, wenn sich X ändert – und falls ja, in welche Richtung und wie stark?“

18.11 Zusammenhang

Ein Zusammenhang besteht, wenn sich Änderungen in X mit **systematischen Änderungen** in Y koppeln.

18.12 Kausalität

Es besteht eine eindeutige Ursache-Wirkungs-Beziehung zwischen zwei Variablen. Es liegt also eine Kausalität vor, wenn Handlung A das Ergebnis B verursacht.

18.13 Linear

Beziehung $\approx$ Gerade $\rightarrow$ Steigung konstant. $Y \approx a + bX$

18.14 Monoton

Y steigt oder fällt durchgehend, aber evtl. nicht linear (Steigung kann sich ändern, z. B. S-förmig, konkav, etc.)

18.15 Kovarianz

Misst, wie stark zwei Variablen gemeinsam variieren.

- Wenn hohe x_i mit hohen y_i einhergehen $\rightarrow$ **positiver Zusammenhang**
- Wenn hohe x_i mit niedrigen y_i einhergehen $\rightarrow$ **negativer Zusammenhang**
- Wenn kein systematisches Muster $\rightarrow$ **kein linearer Zusammenhang**

$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
Vorzeichen = Richtung, Betrag = Stärke der gemeinsamen Streuung

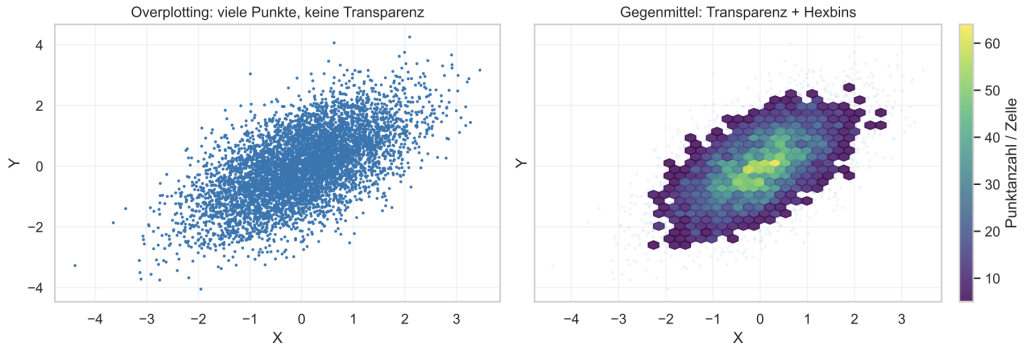
18.16 Visualisierung von Korrelationen

18.16.1 Scatterplot

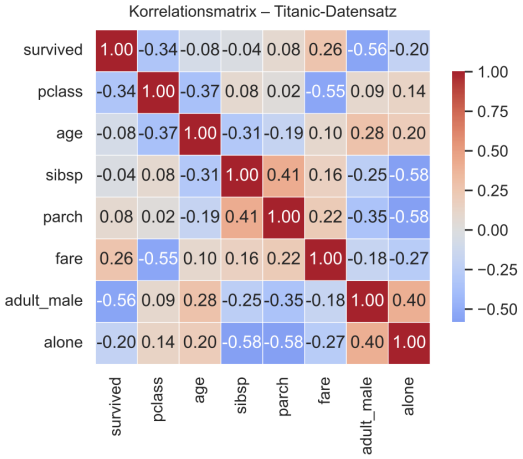
Der Scatterplot (Streudiagramm) ist das wichtigste Werkzeug für **bivariate Zusammenhänge**.

- Jeder Punkt = eine Beobachtung (x_i, y_i)
- Muster der Punkte zeigt:
  - Steigend $\rightarrow$ positiver Zusammenhang

- Fallend → negativer Zusammenhang
- Wolke ohne Trend → kein linearer Zusammenhang



18.16.2 Heatmap – Korrelationsmatrix



Heatmaps visualisieren viele Korrelationen gleichzeitig.

- Matrix aus Korrelationswerten (r-Werte) aller Variablenpaare
- Farben kodieren Stärke & Richtung:
  - Rot: positiver Zusammenhang
  - Blau: negativer Zusammenhang
  - Weiß/hell: nahe 0 → kein Zusammenhang

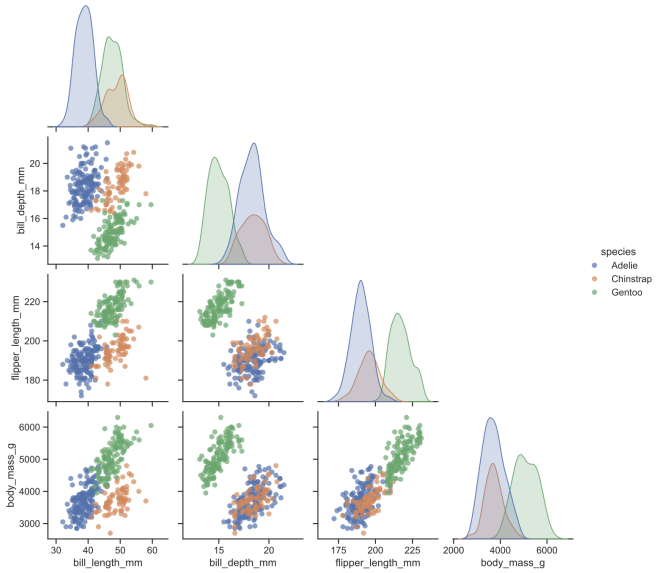
ACHTUNG: Die Farbskala kann täuschen! Kontext beachten (0.3 kann in Soziologie stark, in Physik schwach sein)

18.17 Pairplot

Kombiniert Scatterplots und univariate Histogramme für viele Variablen.

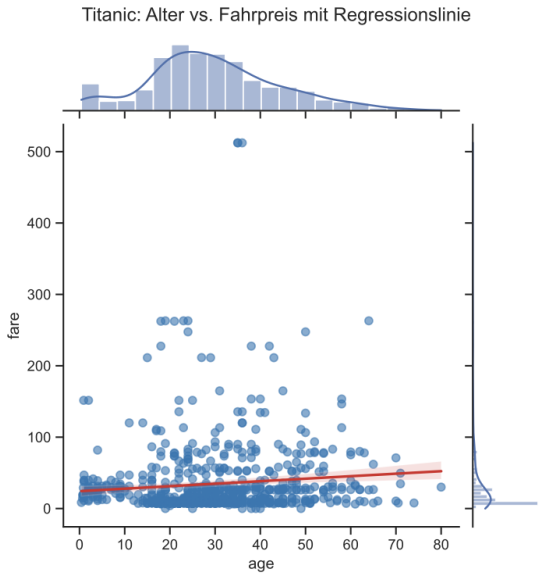
- Diagonal: Histogramme (Einzelverteilungen)
- Off-Diagonal: Scatterplots (Paarvergleiche)
- Erkennt:
  - Cluster
  - Nichtlinearität
  - Ausreißer

Sehr nützlich für erste explorative Analysen (EDA)!



18.18 Jointplot

Der Jointplot zeigt die Beziehung zweier Variablen + ihre Randverteilungen.



- Zentrum: Scatterplot mit Regressionslinie
- Ränder: Histogramme der einzelnen Variablen
- Optional kind="kde"→ zeigt Dichtekonturen statt Punkte

→ Effizient für 2 Variablen mit zusätzlichem Verteilungsüberblick.

18.19 Wahrscheinlichkeit & Verteilungen

18.19.1 Ergebnis

Ein mögliches Resultat eines Zufallsexperiments

18.19.2 Ereignis

Menge von Ergebnissen

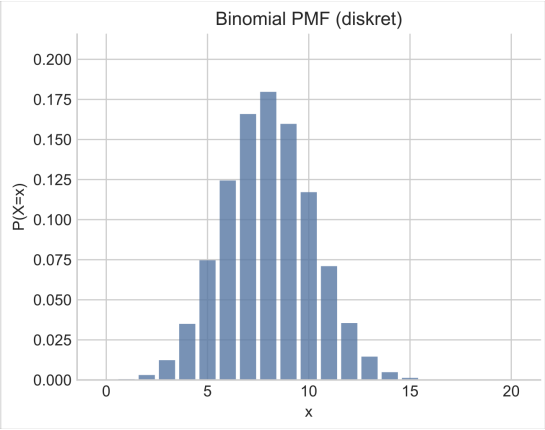
18.19.3 Ereignisraum Ω

Menge aller möglichen Ergebnisse

18.19.4 Diskrete Zufallsvariablen

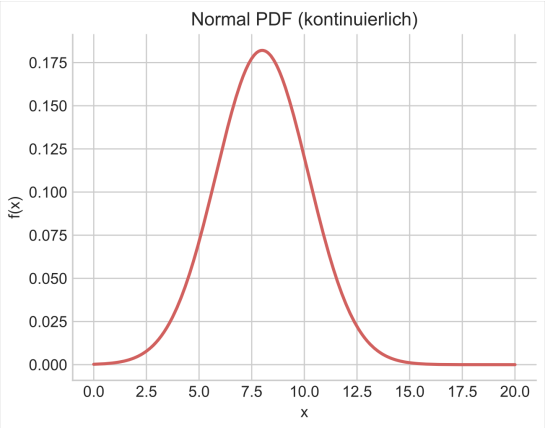
= Summen

(Nominal, Ordinal): endliche oder zählbare Werte (z. B. Augenzahl, Klicks, Fehler)
→ PMF: Probability Mass Function (Wahrscheinlichkeitsfunktion)



18.19.5 Kontinuierliche Zufallsvariablen

= I (Intervall, Ratio): beliebige reelle Werte (z. B. Messungen, Zeit, Temperatur)
→ PDF: Probability Density Function (Wahrscheinlichkeitsdichtefunktion)



18.19.6 Wahrscheinlichkeitsdichte PDF und Verteilungsfunktion CDF

PDF: Probability Density Function (Wahrscheinlichkeitsdichtefunktion): $f(x) \geq 0 \quad$ und $\quad \int f(x)\,dx = 1$
CDF: Cumulative Distribution Function (Kumulative Verteilungsfunktion): $F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(z)\,dz$

19 Schätzen & Konfindenzintervalle

19.1 Population

die Gesamtheit, über die du etwas wissen willst.
Sie hat feste, aber unbekannte Kennzahlen = Parameter:

- Mittelwert (wahrer Durchschnitt): μ
- (wahre) Varianz: σ^2
- (wahrer) Anteil / Wahrscheinlichkeit: ρ

19.2 Stichprobe

die Teilmenge, die du tatsächlich misst/befragst.

- Größe der Stichprobe: n
- Zufallsvariablen: $X_1, ..., X_n$ (jede Beobachtung als Zufallsvariable gedacht).
- ddof = 1 („Delta Degrees of Freedom“) - korrigiert die Berechnung, damit s^2 im Mittel unverzerrt ist
- Schätzer T:
  - $\bar{X}$ = Zufallsvariable „Stichprobenmittelwert“ (hängt von zufälligen Daten ab)
  - $\hat{p}$ = Zufallsvariable „Stichprobenanteil“ (konkrete Zahl aus deiner Stichprobe)
  - Allgemein: $T(X_1, ..., X_n)$
- Schätzwert (konkrete Zahlen)
  - $\bar{x}$ = beobachteter Mittelwert
  - $\hat{p}$ (gleiches Symbol, aber konkrete Zahl aus deiner Stichprobe)
  - t = konkreter Wert des Schätzers T nach Einsetzen der Daten

Großbuchstaben → Zufallsvariablen (theoretisch, „über viele Stichproben“ gedacht).
Kleinbuchstaben → Konkrete Beobachtungen aus einer Stichprobe.
Ideal: zufällig, unabhängig, repräsentativ