technische universität
dortmund

Lehrstuhl für
Förder- und Lagerwesen

Logistische Datenverarbeitung

Prof. Dr. Dr. h. c. Michael ten Hompel
Axel Krüger, M.Sc.

Sommersemester 2019

Lehrstuhl für Förder- und Lagerwesen
Fakultät Maschinenbau

Überblick

1. Einführung

2. Wiederholung Statistik

3. Einführung in die Programmiersprache julia

4. Datenanalyse

Kontakt

Anschrift	Technische Universität Dortmund Lehrstuhl für Förder- und Lagerwesen (FLW) LogistikCampus Joseph-von-Fraunhofer-Str. 2-4 44227 Dortmund
Name	Axel Krüger
Raum	A3.18
Telefon	+49 231 755-4832
Mail	axel2.krueger@tu-dortmund.de
Sprechstunde	flexibel nach Vereinbarung

Organisatorisches und Termine

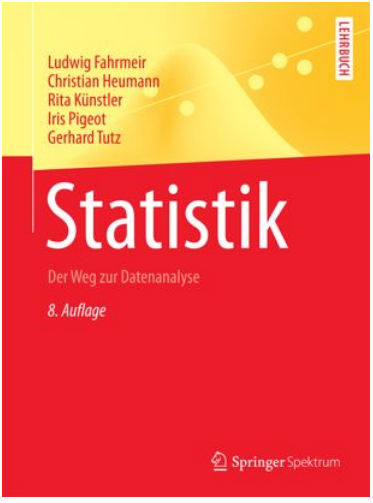
Ort	LC / A1.27
Zeit	Donnerstags um 14:15 - 15:45
Klausurtermin	wird noch bekannt gegeben
Moodleraum	https://moodle.tu-dortmund.de/enrol/index.php?id=7793
Einschreibeschlüssel	LDV4L!FE

Vorlesungs- und Übungstermine:

4. April	11. April	18. April	25. April	2. Mai
9. Mai	16. Mai	23. Mai	~~30. Mai~~	6. Juni
13. Juni	~~20. Juni~~	27. Juni	4. Juli	11. Juli

Literatur

- Titel: Statistik - der Weg zur Datenanalyse
- Autoren: Künstler, Rita ; Heumann, Christian ; Pigeot, Iris ; Fahrmeir, Ludwig ; Tutz, Gerhard
- Link: https://www.ub.tu-dortmund.de/katalog/titel/HT019079313



Literatur

- Titel: Julia Documentation
- Link: https://docs.julialang.org/en/v1/index.html

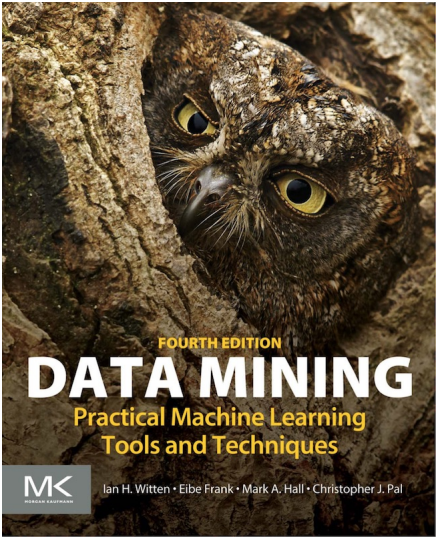
- Titel: Introducing Julia
- Link: https://en.wikibooks.org/wiki/Introducing_Julia

Literatur

- Titel: Julia for data science
- Autor: Voulgaris, Zacharias
- Link: https://www.ub.tu-dortmund.de/katalog/titel/HT019435318

Literatur

- Titel: Data mining - practical machine learning tools and techniques
- Autor: Pal, Christopher J ; Witten, Ian H ; Frank, Eibe ; Hall, Mark A
- Link: https://www.ub.tu-dortmund.de/katalog/titel/HT019136396



Lernziele

Am Ende des Semesters können Sie

- die Grundbegriffe der Statistik unterscheiden und erläutern,
- grundlegende statistische und wahrscheinlichkeitstheoretische Kennzahlen bestimmen und interpretieren,
- ausgewählte maschinelle Lernverfahren in ihren Grundzügen erklären und einsetzen,
- Aufgabentypen des maschinellen Lernens unterscheiden und für den jeweiligen Aufgabentyp geeignete Verfahren benennen,
- erlernte Modelle validieren und testen,
- Ergebnisse der Datenanalyse visualisieren,
- die Anforderungen an ein Datenbanksystem benennen und erläutern.

Einführung

Was sind Daten?

Je nach Fachbereich werden verschiedene, teilweise sehr ähnliche, Definitionen verwendet. Hier orientieren wir uns an folgender Definition:

Daten

”reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing”[1]

- Daten sind hier also eine (digitale) formalisierte Darstellung von Informationen
- Die Informationen sind wieder interpretierbar
- Die formalisierte Darstellung ist so geschaffen, dass die Informationen für Kommunikation, Interpretation und Auswertung verwendet werden können

[1] ISO/IEC 2382:2015

Wo entstehen Daten? - Wird in der Vorlesung gemeinsam erarbeitet

Anwendungen für (komplexe) Datenanalysen

- Sortierung von Briefen mittels automatisierter Erkennung von handschriftlichen Adressen
- Erkennung von Gegenständen für automatisiertes Picking
- Predictive Maintenance
- Routenplanung
- Autonome Roboter und Fahrzeuge
- Analyse von Kundenverhalten
- Prognose von Durchlaufzeiten, Mitarbeiterbedarf,...
- Anomalieerkennung in der Fertigung

Wiederholung Statistik

Objekte und Merkmale

Grundbegriffe

- *Objekte* sind Untersuchungsgegenstände/Merkmalsträger deren Eigenschaften erhoben bzw. gemessen werden
- Die *Grundgesamtheit* bezeichnet die Menge aller Objekte
- Oft ist eine Untersuchung aller Objekte nicht möglich, daher wird eine *Stichprobe*, d.h. eine (möglichst) repräsentative Auswahl von Objekten erhoben
- Die *Stichprobenumfang* ist die Anzahl der Objekte einer Stichprobe
- *Merkmale (Variablen)* sind die (beschreibenden) Eigenschaften der Objekte
- Eine *Merkmalsausprägung* ist der konkrete Wert (aus einem Wertebereich) eines Merkmals

Ein Beispiel

Studenten, die LDV besuchen:

Vorname	Name	Matrikelnummer	Studiengang	Fachsemester	Wöchentliche Lernzeit in [h]
Karl Heinz	Meier	534521	Logistik	5	28,3
Johanna	Peters	534322	Wirt.-Ing.	5	30,4
Robert	Hund	534448	Logistik	5	33,2
Thorsten	Jäger	501203	Logistik	7	29,1
Jean	Dupont	531354	Informatik	5	25,4
⋮	⋮	⋮	⋮	⋮	⋮

- Jeder Student ist ein Objekt
- Jeder Student wird durch die Merkmale Name, Vorname, Matrikelnummer, Studiengang, Fachsemester und wöchentliche Studienzeit beschrieben
- Jedes Merkmal besitzt einen Wertebereich, etwa ist die Matrikelnummer eine natürliche Zahl und die wöchentl. Lernzeit eine positive reelle Zahl

Skalenniveaus

Verschiedene Merkmale weisen verschiedene Skalen auf:

Skalentyp	Aussagen/Operationen	qualitativ/quantitativ	messb. Eigenschaften	Beispiel
Nominal	gleich, ungleich	qualitativ	Häufigkeit	Studiengänge
Ordinal	gleich, ungleich, größer, kleiner	qualitativ	Häufigkeit, Reihenfolge	Klausurnoten
Intervall	gleich, ungleich größer, kleiner, Summe, Differenz	quantitativ	Häufigkeit, Reihenfolge, Abstand	IQ-Skala
Verhältnis	gleich, ungleich größer, kleiner, Summe, Differenz, Multiplikation, Division	quantitativ	Häufigkeit, Reihenfolge, Abstand, natürlicher Nullpunkt	Preise

Unterschied zwischen *stetig* und *diskret*

Merkmalstyp	Anzahl der Ausprägungen	Beispiel
Diskret	Endlich viele	Studiengänge
	abzählbar (unendlich) viele	Kontoguthaben
Stetig	Überabzählbar viele	Körpergröße

Erläuterung:

- *Endlich viele* bedeutet, dass die Anzahl der Objekte endlich und zählbar ist
- *Abzählbar (unendlich) viele* bedeutet, dass die Anzahl weiterhin zählbar ist, die Anzahl der Objekte aber unendlich groß ist. Vergleiche die Menge $\mathbb{N}$ oder $\mathbb{Q}$
- *Überabzählbar viele* bedeutet, dass die Anzahl nicht mehr zählbar und unendlich ist. Vergleiche die Menge $\mathbb{R}$ oder das Intervall $[0, 1]$

Häufigkeiten

- Gegeben: Beobachtungen $x_1, x_2, \ldots, x_n$ eines Merkmals X
- H_a bezeichnet die *absolute Häufigkeit* einer Merkmalsausprägung a
- Es gilt $H_a = \#\{x_i | x_i = a\}$ und $\sum_a H_a = n$
- h_a bezeichnet die *relative Häufigkeit* der Merkmalsausprägung a, d.h.

$$h_a = \frac{H_a}{n}$$

- Es gilt $\sum_a h_a = 1$

Beispiel:
Ein Versandhändler hat drei verschiedene Paketgrößen (S,M,L) zur Auswahl. In der letzten Stunde hat er den Verbrauch der Paketgrößen notiert:

S, S, M, L, L, L, M, S, S, S, M, M, M, M, S, M, L, S, M, S

Damit gilt $n = 20$ und

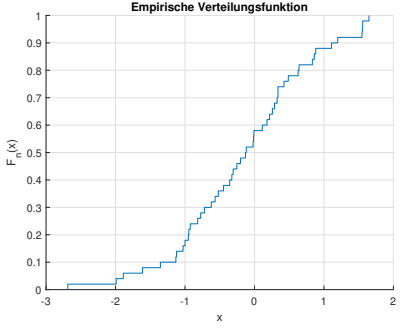
$$H_S = 8 \qquad h_S = 0.4$$
$$H_M = 8 \qquad h_M = 0.4$$
$$H_L = 4 \qquad h_L = 0.2$$

Empirische Verteilungsfunktion

Definition

Die *empirische Verteilungsfunktion* F_n ist definiert durch

$$F_n(x) = \frac{\#\{x_i | x_i \leq x\}}{n}$$



Lagemaße

- Lässt sich eine Stichprobe $x_1, x_2, \ldots, x_n$ sortieren, dann ist $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ die *sortierte* Stichprobe mit $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
- $x_{(1)} = \min\limits_{i=1,\ldots,n} x_i$ und $x_{(n)} = \max\limits_{i=1,\ldots,n} x_i$
- $\arg\max$ ist das argumentative Maximum einer Menge bzw. Funktion, d.h. der Wert bzw. die Werte an denen das Maximum angenommen wird

Def.: Lagemaße

- Modalwert: $\bar{x}_M := \arg\max\limits_a \{H_a | a \text{ ist Merkmalsausprägung}\}$
- arithmetisches Mittel: $\bar{x} := \frac{1}{n} \sum\limits_{i=1}^{n} x_i$
- Median: $\bar{x}_{med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{, falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{, falls } n \text{ gerade} \end{cases}$
- α-Quantil: $x_\alpha = \begin{cases} x_{(\lfloor n \cdot \alpha + 1 \rfloor)} & \text{, falls } n \cdot \alpha \text{ nicht ganzzahlig} \\ \frac{1}{2}\left(x_{(n \cdot \alpha)} + x_{(n \cdot \alpha + 1)}\right) & \text{, falls } n \cdot \alpha \text{ ganzzahlig} \end{cases}$ für $\alpha \in (0, 1)$

Streuungsmaße

Def.: Streuungsmaße

- Empirische Varianz: $s_x^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Empirische Standardabweichung: $s_x := \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \sqrt{s_x^2}$
- Mittlere absolute Medianabweichung: MD $:= \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}_{med}|$
- Interquartilsabstand: IQA $= x_{0.75} - x_{0.25}$
- Spannweite: $R = x_{(n)} - x_{(1)}$

Korrelationskoeffizient

Für zwei Merkmale X und Y seien $(x_1, y_1), \ldots, (x_n, y_n)$ Paare von Merkmalsausprägungen.

Def.: Stichprobenkovarianz und Korrelationskoeffizient

- Die Stichprobenkovarianz ist definiert als

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

- Der Korrelationskoeffizient ist definiert als

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

Korrelation vs. Kausalität

- Korrelation: Zwei Merkmale sind korreliert, wenn sie einen von Null signifikant verschiedenen Korrelationskoeffizienten aufweisen! Das heißt auf Datenebene besteht ein Zusammenhang
- Kausalität: Zwei Merkmale sind kausal abhängig, wenn zwischen ihnen ein Ursache-Wirkung Zusammenhang besteht

Achtung!

Korrelation $\neq$ Kausalität
Mittels statistischer Methoden kann nur eine Korrelation, nie eine Kausalität, nachgewiesen werden! (Das heißt aber noch lange nicht, dass keine Kausalität vorliegt.)

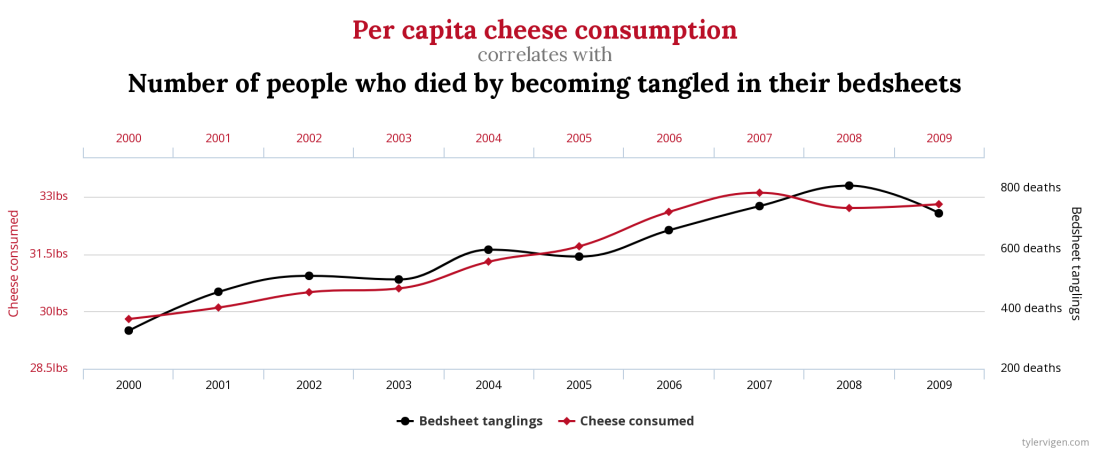


Abbildung: Beispiel für eine Korrelation, bei der kein kausaler Zusammenhang besteht ($r_{xy} = 0.947091$). Quelle: Entnommen von http://www.tylervigen.com/spurious-correlations

Mengenoperationen

Seien A und B Teilmengen von Ω, d.h. $A \subseteq \Omega$ und $B \subseteq \Omega$. Dann ist

- $A \cup B$ die *Vereinigung* der Mengen A und B, d.h. $A \cup B = \{x \in \Omega \,|x \in A \vee x \in B\}$[2]
- $A \cap B$ der *Schnitt* der Mengen A und B, d.h. $A \cap B = \{x \in \Omega \,|x \in A \wedge x \in B\}$
- $A \setminus B$ die *Differenz* von A und B („A ohne B"), d.h. $A \setminus B = \{x \in \Omega \,|x \in A \wedge x \notin B\}$
- A^C das Komplement von A, d.h. $A^C = \{x \in \Omega \,|x \notin A\}$
- $A \times B$ das *kartesische Produkt* von A und B, d.h. $A \times B = \{(x, y) \in \Omega \,|x \in A \wedge y \in B\}$

[2] Das „oder" ($\vee$) ist wie üblich nicht-exklusiv gemeint, d.h. x kann sowohl Element von A, als auch von B sein.

Venn-Diagramme (wird in der Vorlesung ergänzt)

Wahrscheinlichkeitstheorie

Begriff	Symbol	Erläuterung	Beispiel
Zufallsexperiment		Prozess zur Erhebung von Daten mit nicht vorhersagbarem Ausgang	Werfen eines 6-seitigen Würfels
Ergebnis	ω	Elementarer Ausgang eines Zufallsexperiments	3
Grundraum	Ω	Menge aller möglicher Ergebnisse	$\{1, 2, 3, 4, 5, 6\}$
Ereignis	z.B. A	Menge von Ergebnissen, also eine Teilmenge des Grundraums	$A =$ „Das Ergebnis ist gerade“ bzw. $A = \{2, 4, 6\}$
Wahrscheinlichkeits-maß	P	Gibt die Wahrscheinlichkeit von Ereignissen an[3]	$P(A) = \frac{1}{2}$

[3] P ist eine Funktion von einem geeigneten Mengensystem (einer σ-Algebra) in das Intervall $[0, 1]$.

Wichtige Rechenregeln

Seien A und B zwei Ereignisse, dann gelten die folgenden Regeln

- $A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$
- $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A^c) = 1 - P(A)$
- A und B heißen *stochastisch unabhängig* genau dann, wenn $P(A \cap B) = P(A) \cdot P(B)$ gilt[4]

[4] Gilt für zwei Ereignisse $P(A \cap B) \neq P(A) \cdot P(B)$, so wird dies nicht „stochastisch abhängig" genannt, sondern „nicht stochastisch unabhängig".

Der Satz von Bayes I

Def.: bedingte Wahrscheinlichkeit

Die *bedingte Wahrscheinlichkeit*, dass A unter der Bedingung B eintritt, ist definiert als

$$P(A|B) := \frac{P(A \cap B)}{P(B)},$$

für $P(B) > 0$.
$P(A|B)$ ist also die Wahrscheinlichkeit, dass A eintritt, wenn B bereits eingetreten ist.

Beispiel: Unter einem Würfelbecher liegt verdeckt ein Würfel. Es sei bekannt, dass die obenliegende Augenzahl ungerade sei, d.h. $B = \{1, 3, 5\}$. Wie groß ist die Wahrscheinlichkeit, dass die Augenzahl gleich 3 ist, sprich $A = \{3\}$?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{1,3,5\} \cap \{3\})}{P(\{1,3,5\})} = \frac{P(\{3\})}{P(\{1,3,5\})} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Umgekehrt gilt: $P(B|A) = 1$

Der Satz von Bayes II

Der Satz von Bayes

(1) Seien A und B zwei Ereignisse mit $P(A) > 0$ und $P(B) > 0$. Dann gilt

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

(2) Seien $B_1, \ldots, B_m$ und A Ereignisse mit $P(B_i) > 0$ für $i = 1, \ldots, m$, $P(A) > 0$, $B_i \cap B_j = \varnothing$ für $i \neq j$ und $\bigcup_{i=1}^{m} B_i = \Omega$. Dann gilt

$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{m} P(A|B_i) \cdot P(B_i)}$$

Beispiel zum Satz von Bayes

- Szenario: Ein Bauteil wird auf seine Funktionsfähigkeit untersucht. Ein Leuchtdiode an dem Prüfgerät zeigt an, ob das Bauteil den Test bestanden hat.
- Bekannt: 1% aller Bauteile sind defekt,
  Ist ein Bauteil defekt, so meldet das Prüfgerät dies in 95% der Fälle,
  In 0.5% der Fälle, in denen ein Bauteil nicht defekt ist, wird dennoch ein Defekt gemeldet
- Frage 1: Wie groß ist die Wahrscheinlichkeit, dass ein Bauteil bei dem das Prüfgerät einen Defekt meldet, auch wirklich defekt ist?
- Frage 2: Wie groß ist die Wahrscheinlichkeit, dass ein Bauteil, das als nicht defekt eingestuft wird, auch wirklich nicht defekt ist?
- Frage 3: Sind die Ereignisse „Bauteil ist defekt“ und „Prüfgerät meldet einen Defekt“ stochastisch unabhängig?

Lösung

- Definiere

$$A = \{\text{Bauteil ist defekt}\}$$
$$B = \{\text{Prüfgerät meldet einen Defekt}\}$$

- Dann gilt nach Aufgabenstellung

$$P(A) = 0.01, \quad P(B|A) = 0.95, \quad P(B|A^C) = 0.005$$

Lösung Frage 1

Frage 1: Wie groß ist die Wahrscheinlichkeit, dass ein Bauteil bei dem das Prüfgerät einen Defekt anzeigt auch wirklich defekt ist?
Gesucht ist $P(A|B)$. Es gilt

$$\begin{aligned} P(A|B) &= \frac{P(B|A) \cdot P(A)}{P(B)} \\ &= \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^C) \cdot P(A^C)} \\ &= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.005 \cdot (1 - 0.01)} \\ &= \frac{190}{289} \\ &\approx 65.74\% \end{aligned}$$

Lösung Frage 2

Wie groß ist die Wahrscheinlichkeit, dass ein Bauteil, das als nicht defekt eingestuft wird, auch wirklich nicht defekt ist? Gesucht ist $P(A^C|B^C)$. Es gilt

$$
\begin{aligned}
P(A^C|B^C) &= \frac{P(B^C|A^C) \cdot P(A^C)}{P(B^C)} \\
&= \frac{P(B^C|A^C) \cdot P(A^C)}{P(B^C|A^C) \cdot P(A^C) + P(B^C|A) \cdot P(A)} \\
&= \frac{(1 - P(B|A^C)) \cdot (1 - P(A))}{(1 - P(B|A^C)) \cdot (1 - P(A)) + (1 - P(B|A)) \cdot P(A)} \\
&= \frac{(1 - 0.005) \cdot (1 - 0.01)}{(1 - 0.005) \cdot (1 - 0.01) + (1 - 0.95) \cdot 0.01} \\
&= \frac{19701}{19711} \\
&\approx 99.99\%
\end{aligned}
$$

Lösung Frage 3

Sind die Ereignisse „Bauteil ist defekt“und „ Prüfgerät erkennt einen Defekt“ stochastisch unabhängig? Es gilt

$$\begin{aligned} P(A) &= 0.01 \\ P(B) &= P(B|A) \cdot P(A) + P(B|A^C) \cdot P(A^C) \\ &= 0.95 \cdot 0.01 + 0.005 \cdot (1 - 0.01) \\ &= 0.01445 \end{aligned}$$

und daher

$$P(A) \cdot P(B) = 0.0001445 \neq 0.0095 = P(A|B) \cdot P(B) = P(A \cap B)$$

und somit sind A und B nicht stochastisch unabhängig.[5]

[5] Äquivalent kann überprüft werden ob $P(A|B) \neq P(A)$ gilt, denn zwei Ereignisse sind gdw. stochastisch unabhängig wenn $P(A|B) = P(A)$ gilt.

Zufallsvariablen

- Eine *Zufallsvariable*, auch Zufallsgröße genannt, ist eine Vorschrift, die jedem Ergebnis $\omega \in \Omega$ eines Zufallsexperimentes einen Wert zuordnet.
- Zufallsvariablen werden üblicherweise mit lateinischen Großbuchstaben, wie $X, Y, \ldots$, bezeichnet
- Der zugeordnete Wert $X(\omega)$ wird auch Realisierung genannt
- Die Verteilungsfunktion einer Zufallsvariablen X ist definiert[6] als

$$F_X(x) := P(X \leq x)$$

- Die gemeinsame Verteilungsfunktion zweier Zufallsvariablen X und Y ist definiert[7] als

$$F_{X,Y}(x, y) := P(X \leq x, Y \leq y)$$

- Für eine diskrete Zufallsvariable ist $f(x) := P(X = x)$ die sogenannte *Zähldichte*
- Für eine stetige Zufallsvariable ist $f(x) := F'_X(x)$ (also die Ableitung der Verteilungsfunktin) die sogenannte *Dichte*

[6] $\{X \leq x\}$ ist eine abkürzende Schreibweise für $\{\omega \in \Omega | X(\omega) \leq x\}$

[7] Das Komma zwischen die beiden Mengen $\{X \leq x\}$ und $\{Y \leq y\}$ ist eine Kurzschreibweise für den Schnitt der beiden Mengen

Dichten und Verteilungsfunktionen

Eine Dichte f besitzt die folgenden definierenden Eigenschaften:

- f ist nicht-negativ, d.h. $f \geq 0$
- f ist normiert und für stetige Verteilungen integrierbar, d.h $\sum_x f(x) = 1$ für diskrete Verteilungen und $\int f(x)dx = 1$ für stetige Verteilungen
- Wichtig: Eine Dichte ist eindeutig für eine Verteilung

Eine Verteilungsfunktion F besitzt die folgenden Eigenschaften:

- F ist monoton steigend, d.h. $F(x) \leq F(y)$ für $x \leq y$
- F ist rechtsseitig stetig
- Es gilt $\lim_{x\to\infty} F(x) = 1$ und $\lim_{x\to-\infty} F(x) = 0$

Die gemeinsame Verteilung zweier Zufallsvariablen X und Y ist definiert als

$$F_{(X,Y)}(x, y) := P(X \leq x, Y \leq y)$$

Erwartungswert, Varianz und Kovarianz für diskrete Zufallsvariablen

Erwartungswert, Varianz und Kovarianz

Für zwei diskrete Zufallsvariablen X und Y ist

- der Erwartungswert definiert als

$$\mathbb{E}(X) := \sum_x x \cdot f(x),$$

- die Varianz definiert als

$$\mathsf{Var}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \sum_x (x - \mathbb{E}(X))^2 \cdot f(x)$$

Erwartungswert, Varianz und Kovarianz für stetige Zufallsvariablen

Erwartungswert und Varianz

Für zwei stetige Zufallsvariable ist

- der Erwartungswert definiert als

$$\mathbb{E}(X) := \int x \cdot f(x) dx,$$

- die Varianz definiert als

$$\mathsf{Var}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \int (x - \mathbb{E}(X))^2 \cdot f(x) dx$$

Rechenregeln für den Erwartungswert und die Varianz

Für den Erwartungswert und die Varianz zweier Zufallsvariablen X, Y gelten die folgenden sehr hilfreichen Rechenregeln:

- Transformationssatz:

$$\mathbb{E}(g(X)) = \begin{cases} \sum_x g(x) \cdot f(x) & \text{, falls } X \text{ diskret} \\ \int g(x) \cdot f(x) dx & \text{, falls } X \text{ stetig} \end{cases}$$

- Linearität:

$$\mathbb{E}(a \cdot X - b) = a \cdot \mathbb{E}(X) - b$$
$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

- Verschiebungssatz:

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

- Lineare Transformation der Varianz:

$$\text{Var}(a \cdot X - b) = a^2 \cdot \text{Var}(X)$$

Die Gleichverteilung

Gleichverteilte Zufallsvariable

- Eine diskrete Zufallsvariable X heißt *gleichverteilt* auf der Menge $T := \{x_1, \ldots, x_n\}$, wenn jede ihrer Realisierungen x_i gleichwahrscheinlich ist, das heißt

$$f_X(x_i) = \frac{1}{n} \quad \text{für } i = 1, \ldots, n.$$

  Kurz wird dies auch notiert als $X \sim \mathcal{U}(T)$. Die Menge T heißt auch Träger.
- Eine stetige Zufallsvariable X heißt *gleichverteilt* auf dem Intervall $[a, b]$, wenn für ihre Dichte gilt

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{, falls } x \in [a, b] \\ 0 & \text{, sonst} \end{cases}.$$

  Dies wird kurz notiert als $X \sim \mathcal{U}([a, b])$.

Diskrete Gleichverteilung

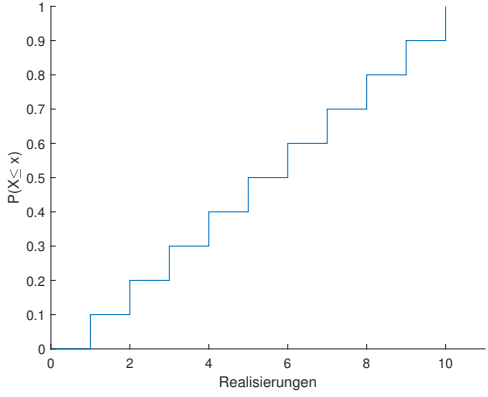


Abbildung: Verteilungsfunktion einer auf 1,2,...10 gleichverteilten Zufallsvariablen

Stetige Gleichverteilung

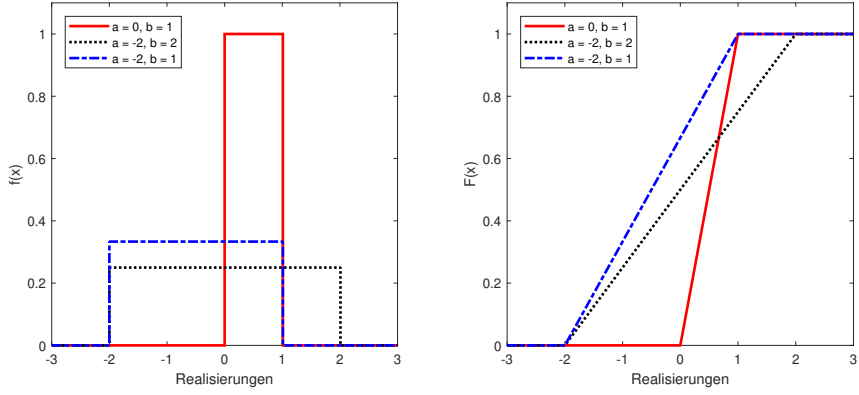


Abbildung: Dichten (links) und Verteilungsfunktionen (rechts) der stetigen Gleichverteilung mit versch. Parametern

Die Normalverteilung

Normalverteilung

Eine Zufallsvariable X heißt *normalverteilt* mit μ und σ^2 wenn

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$

gilt. Dies wird kurz durch $X \sim \mathcal{N}(\mu, \sigma^2)$ notiert.

Es gilt:

- $\mathbb{E}(X) = \mu$ und $\mathrm{Var}(X) = \sigma^2$
- $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow (a \cdot X - b) \sim \mathcal{N}(a \cdot \mu - b, a^2\sigma^2)$
- Das Integral $F(x) = \int f(x)dx$ besitzt für die Normalverteilung keine analytische Form
- Die Werte der Verteilungsfunktion F werden numerisch berechnet

Dichte und Verteilungsfunktion der Normalverteilung

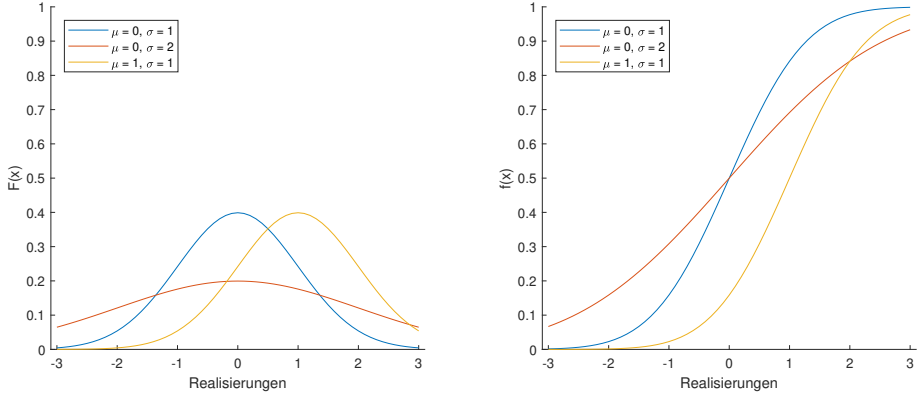


Abbildung: Dichte (links) und Verteilungsfunktion (rechts) der Normalverteilung bei versch. Parametern

Das Stab- Balken oder Säulendiagramm

- Einfach Darstellung für die Häufigkeiten verschiedener Kategorien
- Seien Merkmale $a_1, \ldots, a_n$ mit absoluten Häufigkeiten $H_{a_1}, \ldots, H_{a_n}$ und relativen Häufigkeiten $h_{a_1}, \ldots, h_{a_n}$ gegeben
- Stabdiagramm: Trage für jedes Merkmal a einen zur x-Achse senkrechten Strich (Stab) mit der Höhe H_a (oder h_a) auf
- Säulendiagramm: Wie das Stabdiagramm, nur dass hier Säulen statt Striche verwendet werden
- Balkendiagramm: Wie das Säulendiagramm, allerdings sind hier die Balken senkrecht zur y-Achse
- Bei einem Säulen- oder Balkendiagramm besitzt die Breite der Balken keine Information, d.h. die Breite ist grundsätzlich beliebig

Beispiele für Stab- Balken oder Säulendiagramme

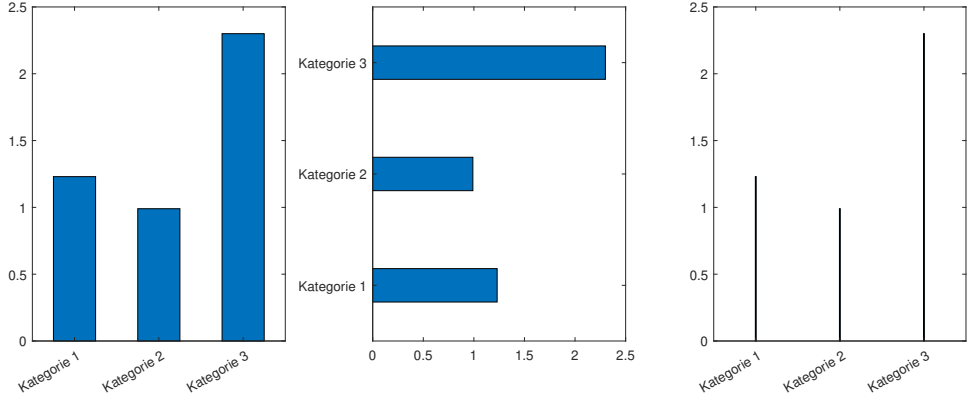


Abbildung: Beispiel für ein Säulendiagramm (links), Balkendiagramm (mittig) und Balkendiagramm (rechts)

Das Histogramm

- Bei den Balken- und Säulendiagrammen hatte die jeweilige Breite keine Bedeutung
- Ein der Verdopplung der Breite eines Balken würde jedoch optisch diesem Balken eine höhere Bedeutung zuweisen
- Wir betrachten nun viele Ausprägungen mit zumindest metrisch-skalierten Merkmalen
- Gegeben seien zunächst Gruppierungen der Merkmale in Intervalle der Form $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k)$
- Sei h_j hier die relative Häufigkeit der Beobachtung im Intervall $[c_{j-1}, c_j)$ und H_j entsprechend die absolute Häufigkeit
- Über diesen Intervallen werden nun Rechtecke mit der Breite $d_j = c_j - c_{j-1}$ und Höhe $\frac{h_j}{d_j}$ (bzw. $\frac{H_j}{d_j}$) gezeichnet
- Die Summe der Flächen aller Rechtecke ist gleich 1 (bzw. gleich n)
- Die Klassenbreiten sollten wenn möglich sinnvoll und annähernd gleich sein
- Neben der subjektiven optischen Begutachtung, existieren verschiedene Faustregeln für die Anzahl der Klassen, wie z.B.

$$k = \lfloor \sqrt{n} \rfloor, \quad k = 2\lfloor \sqrt{n} \rfloor \quad \text{oder} \quad k = \lfloor 10 \log_{10} n \rfloor$$

Beispiel für ein Histogramm

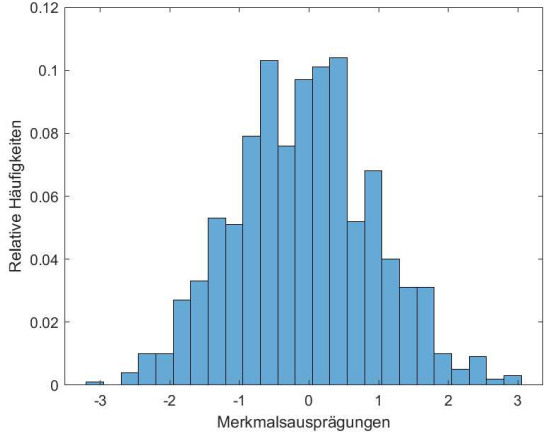


Abbildung: Histogramm für 1000 Pseudozufallszahlen über 25 äquidistante Klassen

Das Streudiagramm

Für zwei stetige Merkmale X und Y mit paarweisen Ausprägungen $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ kann das sogenannte Streudiagramm als einfache Visualisierung verwendet werden:

Streudiagramm

Die Darstellung der Messwertepaare $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ im (x,y)-Koordinatensystem heißt *Streudiagramm* (engl. scatter plot).

Beispiel für ein Streudiagramm

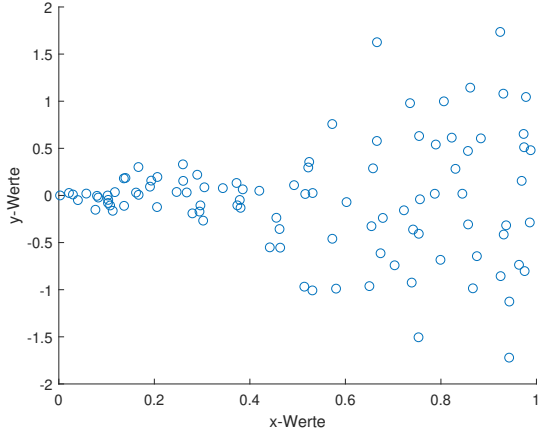


Abbildung: Beispiel für ein Streudiagramm mit 100 Wertepaaren

Der Box-Plot

In einem Box-Plot wird die 5-Punkte-Zusammenfassung $(x_{min}, x_{0.25}, x_{med}, x_{0.75}, x_{max})$ grafisch aufbereitet.

1. Anfang der Box bei $x_{0.25}$ mit beliebiger Höhe
   Ende der Box bei $x_{0.75}$ (damit ist die Länge der Box gleich dem Interquartilsabstand)
2. Die Box wird auf Höhe des Medians mit einem vertikalen Strich geteilt
3. Von den Rändern der Box werden horizontale Linien bis x_{min} bzw. x_{max} gezogen, die mit einer vertikalen Linie abgeschlossen werden

Oft wird ein Box-Plot vertikal aufgebaut, so sind Minimum und Maximum unten bzw. oben, statt links und rechts.

Beispiel für einen Box-Plot

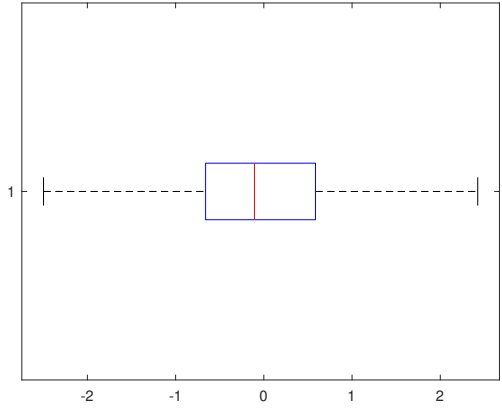


Abbildung: Beispiel für einen Boxplot

Die Heatmap

- Wir betrachten Daten, die in einer Matrix M dargestellt werden
- Der Eintrag $M_{(i,j)}$ hängt von dem i-ten Zeilenmerkmal und dem j-ten Spaltenmerkmal ab
- Die Zeilenmerkmale und Spaltenmerkmale können die gleichen oder verschiedene Merkmale sein
- Der numerische Wert $M_{(i,j)}$ wird im Vergleich zu den übrigen Werten von M farblich kodiert
- Die Kategorien werden an den Zeilen und Spalten aufgetragen und jede Zelle entsprechend ihrer Farbe eingefärbt
- Die Farbskala wird als Legende mit angegeben
- Optional können die Zellenwerte mit in die Heatmap aufgenommen werden

Beispiel I für eine Heatmap

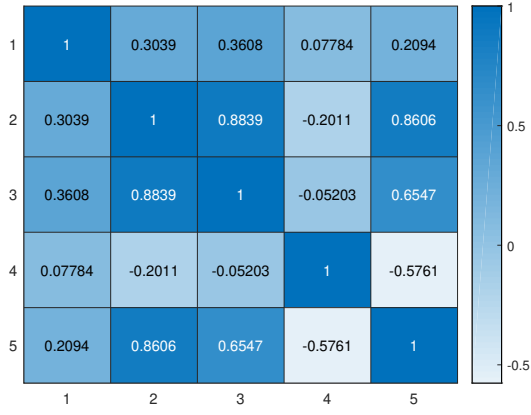


Abbildung: Beispiel für eine Heatmap einer Korrelationskoeffizientenmatrix

Beispiel II für eine Heatmap

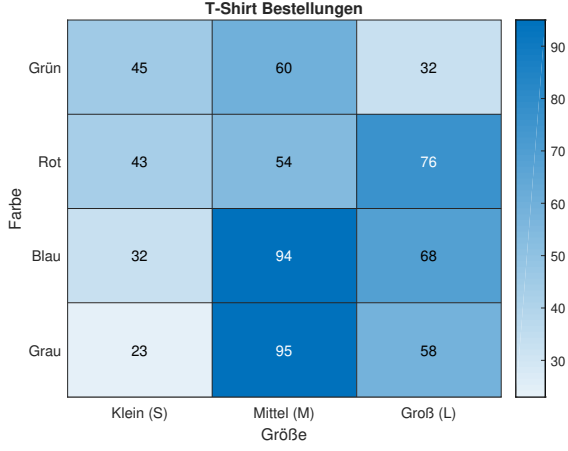


Abbildung: Beispiel für eine Heatmap zu T-Shirt Verkäufen

Regeln für gute Diagramme

1. Diagramme sollen Informationen übersichtlich darstellen, d.h. sie sollten erst bei einer hinreichend großen Anzahl von Informationen, die vermittelt werden sollen, eingesetzt werden
2. Diagramme sollen selbsterklärend und übersichtlich sein
3. Die Beschriftung eines Diagrammes ist vollständig, dazu gehören unter Berücksichtigung der Übersichtlichkeit (und sofern anwendbar)
   - Titel
   - Zeitraum
   - Ortsangabe
   - Achsenbeschriftungen
   - Angabe einer Legende
   - Quelle
4. 3D-Darstellungen nur dann, wenn das Volumen eine sinnvolle Bedeutung hat (Nicht aus optischen Gründen)
5. Sprünge in Achse müssen deutlich gekennzeichnet werden
6. Die Darstellung sollte proportional sein, insbesondere betrifft dies Höhe, Breite und Abstände