

Wahrscheinlichkeitsrechnung und mathematische Statistik für Informatiker

Prof. Dr. Jörg Rahnenführer

Fakultät Statistik

Technische Universität Dortmund

Wahrscheinlichkeitsrechnung und mathematische Statistik für Informatiker

Prof. Dr. Jörg Rahnenführer
Mathegebäude, Raum 720
Email: rahenfuehrer@statistik.tu-dortmund.de

- **Vorlesung (2V)**

- Do, 8:15 - 9:45 Uhr
- HG II HS 3

- **Übung (1Ü):**

- Do, 18.05-19.35 Uhr
- EF 50 HS 1
(in der Regel alle zwei Wochen)

- **Übungstermine**

- 19.10. 18.05-19.35
- 02.11. 18.05-19.35
- 16.11. 18.05-19.35
- 30.11. 18.05-19.35
- 14.12. 18.05-19.35
- 11.01. 18.05-19.35
- 01.02. 18.05-19.35

- **Klausurtermine**

- Klausur: 20.02.18, 14:00-16:00,
Audimax; HG II HS 1, HS 3; Mathe E 29
- Nachklausur: 03.04.18, 08:00-10:00
HG II HS 1, HS 3, HS 6

- **Voraussetzung zur Teilnahme an Klausur:**

- Regelmäßige Teilnahme an Übungen und selbstständige Bearbeitung der Übungsaufgaben

Übersicht

- Motivation
- Merkmale und Datentypen
- Univariate Daten
 - Tabellarische und grafische Darstellung
- Statistische Kennzahlen
 - für die Lage
 - für die Streuung
- Bivariate Daten
 - Tabellarische und grafische Darstellung
 - Zusammenhangsmaße
 - Lineare Regression
- Mengentheoretische Grundlagen
- Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume
- Zufallsvariablen und deren Verteilungen
- Wichtige Wahrscheinlichkeitsverteilungen
- Bedingte Wahrscheinlichkeiten und stochastische Unabhängigkeit
- Erwartungswert und Varianz
- Weitere wahrscheinlichkeitstheoretische Kennzahlen
- Markoffketten
- Statistische Tests
 - Normalverteilung
 - Test bei nicht normalverteilten Daten

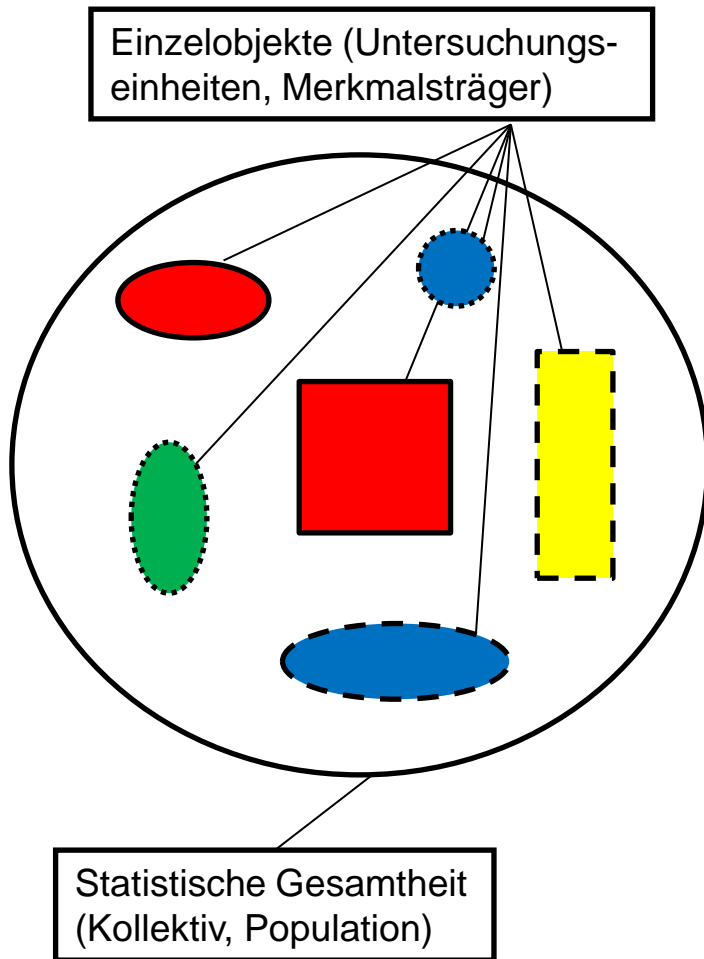
Motivation

Statistische Methoden spielen in der Informatik an vielen Stellen eine große Rolle.

Beispiele:

- Laufzeiten von Algorithmen mit stochastischem Input
 - Stochastische Algorithmen
 - Spieltheorie
 - Ausfälle von Datenverbindungen oder Hardwarekomponenten
 - Automatische Übersetzung
 - Assoziationsregeln, Bilderkennung, Signalanalyse
 - Statistische Lernverfahren
-
- Diese Vorlesung behandelt sich alle diese Themen, sondern die dazu notwendigen statistischen Grundlagen.

Merkmale und Datentypen



Merkmal	Merkmalsausprägungen	Wertebereich
Form	Ellipse, Ellipse, Ellipse, Rechteck, Rechteck, Ellipse	{Ellipse, Rechteck}
Farbe	Rot, Blau, Grün, Rot, Gelb, Blau	{Blau, Gelb, Grün, Rot}
Linienart	Durchgängig, Gepunktet, Gepunktet, Durchgängig, Gestrichelt, Gestrichelt	{Gepunktet, Gestrichelt, Durchgängig}
Breite in cm	2, 1, 1, 2, 1, 3	$(0, \infty)$
Höhe in cm	1, 1, 2, 2, 3, 1	$(0, \infty)$

Merkmale und Datentypen

Datentypen

Skalentyp	mögliche Aussagen	Im Beispiel
qualitativ		
Nominal	Gleich / Verschieden	Farbe, Form (binär, dichotom)
Ordinal	Größer / Kleiner	Linienart
quantitativ / metrisch		
Intervall	Differenzen gleich / verschieden	Breite, Höhe
Verhältnis	Verhältnisse gleich / verschieden	Breite, Höhe

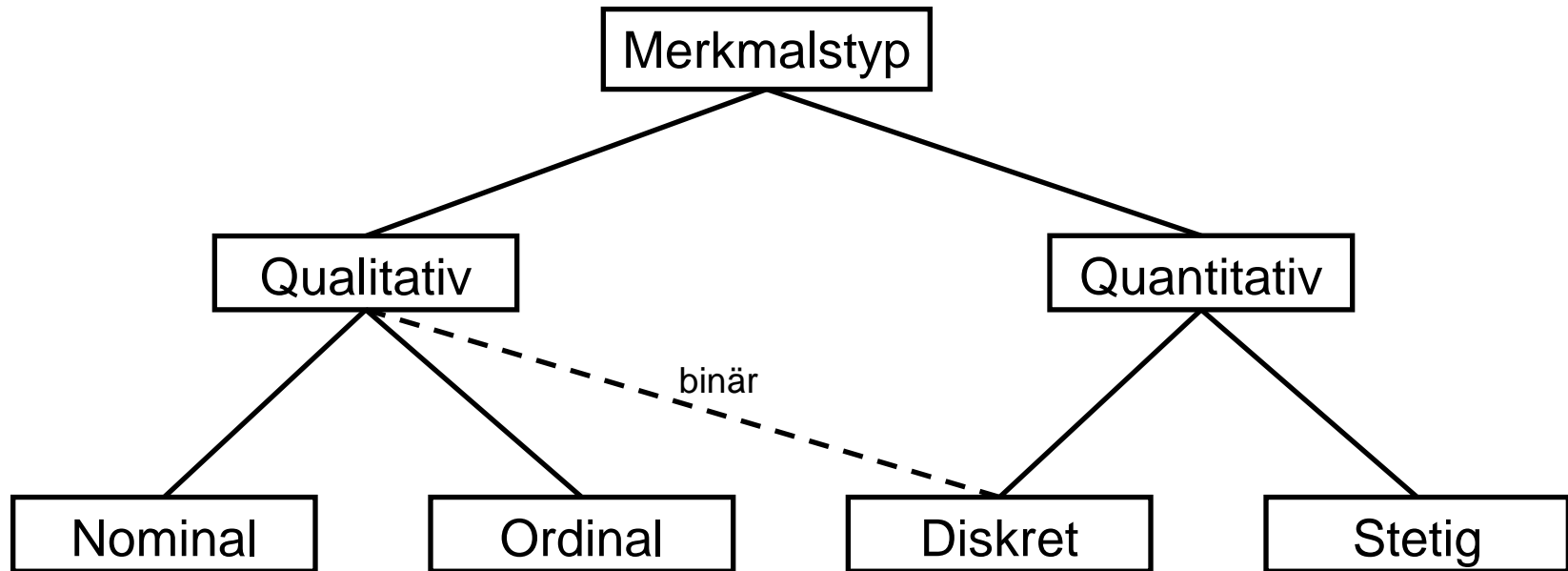
Merkmale und Datentypen

Datentypen

Merkmalstyp	Anzahl der Ausprägungen	Im Beispiel
Diskret	Endlich <i>oder</i> abzählbar unendlich viele	Form Breite, Höhe (wenn grob gemessen)
Stetig	Überabzählbar viele	Breite, Höhe (wenn beliebig fein gemessen)

Merkmale und Datentypen

Datentypen



- Qualitativ heißt immer diskret
- Skalenniveau wird von links nach rechts immer höher

Merkmale und Datentypen

- Unter Inkaufnahme von Informationsverlust können Merkmale in andere Skalenniveaus überführt und entsprechend analysiert werden
 - stetig in diskret (runden, genaue Werte gehen verloren)
 - diskret quantitativ in ordinal (Abstände gehen verloren)
 - ordinal in nominal (Ordnung geht verloren)
- Dieses Vorgehen kann generell auch sinnvoll sein (z.B. bei Linearitätsverletzung)

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten

$$M_N = \{e_1, \dots, e_N\}$$

Population bestehend aus Objekten e_1, \dots, e_N

X

Nominales bzw. ordinales Merkmal

$$x, x \in W_X$$

Merkmalsausprägungen von X

$$\begin{aligned} W_X &= \{x(j) | j = 1, \dots, J\} \\ &= \{x(1), \dots, x(J)\} \end{aligned}$$

Wertebereich von X mit
Merkmalsausprägungen $x(j), j = 1, \dots, J$

$$\begin{aligned} D_N &= \{x_n | n = 1, \dots, N\} \\ &= \{x_1, \dots, x_N\} \end{aligned}$$

Urliste aus der Messung von X in der
Population M_N , d.h. $x_n = X(e_n), n=1, \dots, N$

$$x(1) < x(2) < \dots < x(J)$$

falls X ordinal

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7
Objekte					

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

5 Variablen

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

Qualitative Daten

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

$D_{N;1}, N=12$

$X_1 = \text{Bearbeiter(in)}$

$W_{X_1} = \{\text{Kai, Miriam, Oliver, Tina}\}$

$J_1 = 4$

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

$D_{N;2}, N=12$

$X_2 = \text{Aufgabe}$

$W_{X_2} = \{\text{Abfrage}, \text{Export}, \text{Verknüpfung}\}$

$J_2 = 3$

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

$$\begin{aligned}
 &D_{N;3}, N=12 \\
 &X_3 = \text{Version} \\
 &W_{X_3} = \{1.1, 1.2, 2.0\}, 1.1 < 1.2 < 2.0 \\
 &J_3 = 3
 \end{aligned}$$

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Deskriptive Auswertung

Absolute Häufigkeit N_j von $x(j)$: $N_j = N[x(j)] = \sum_{i=1}^N d_i(j)$, $d_i(j) = I_{x(e_i)=x(j)}$

Damit gilt $\sum_{j=1}^J N_j = N$

$x_1(1)$ Kai	$x_1(2)$ Miriam	$x_1(3)$ Oliver	$x_1(4)$ Tina	Σ
II	III	IIII	III	IIII IIII II
2	3	4	3	12

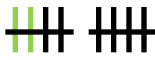
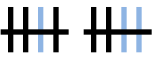
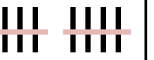


i	$x_1(e_i)$	$d_{1i}(1)$	$d_{1i}(2)$	$d_{1i}(3)$	$d_{1i}(4)$
1	Kai	1	0	0	0
2	Kai	1	0	0	0
3	Miriam	0	1	0	0
4	Tina	0	0	0	1
5	Oliver	0	0	1	0
6	Tina	0	0	0	1
7	Tina	0	0	0	1
8	Miriam	0	1	0	0
9	Miriam	0	1	0	0
10	Oliver	0	0	1	0
11	Oliver	0	0	1	0
12	Oliver	0	0	1	0
Σ		2	3	4	3

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Deskriptive Auswertung

Relative Häufigkeit f_j von $x(j)$: $f_j = \frac{N_j}{N}$

Damit gilt $\sum_{j=1}^J f_j = 1$

$x_1(1)$ Kai	$x_1(2)$ Miriam	$x_1(3)$ Oliver	$x_1(4)$ Tina	Σ
 	 	 	 	
2/12 ≈ 0.17	3/12 $= 0.25$	4/12 ≈ 0.33	3/12 $= 0.25$	12/12 $= 1$

i	$x_1(e_i)$	$d_{1i}(1)$	$d_{1i}(2)$	$d_{1i}(3)$	$d_{1i}(4)$
1	Kai	1	0	0	0
2	Kai	1	0	0	0
3	Miriam	0	1	0	0
4	Tina	0	0	0	1
5	Oliver	0	0	1	0
6	Tina	0	0	0	1
7	Tina	0	0	0	1
8	Miriam	0	1	0	0
9	Miriam	0	1	0	0
10	Oliver	0	0	1	0
11	Oliver	0	0	1	0
12	Oliver	0	0	1	0
$\Sigma/12$		0.17	0.25	0.33	0.25

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Deskriptive Auswertung

Tabellarische Darstellung absoluter und relativer Häufigkeiten

Ausprägung	Absolute Häufigkeit	Relative Häufigkeit
$x(1)$	N_1	$f_1 = N_1/N$
...
$x(J)$	N_J	$f_J = N_J/N$
	$\sum_{j=1}^J N_j = N$	$\sum_{j=1}^J f_j = 1$

Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Deskriptive Auswertung

Tabellarische Darstellung absoluter und relativer Häufigkeiten

Bearbeiter(in)		
Ausprägung	Absolute Häufigkeit	Relative Häufigkeit
Kai	2	0.17
Miriam	3	0.25
Oliver	4	0.33
Tina	3	0.25
	12	1

Aufgabe		
Ausprägung	Absolute Häufigkeit	Relative Häufigkeit
Abfrage	2	0.17
Export	6	0.5
Verknüpfung	4	0.33
	12	1

Version		
Ausprägung	Absolute Häufigkeit	Relative Häufigkeit
1.1	3	0.25
1.2	6	0.5
2.0	3	0.25
	12	1

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ diskrete Daten

$$M_N = \{e_1, \dots, e_N\}$$

Population bestehend aus Objekten e_1, \dots, e_N

X

Quantitatives Merkmal

$$x, x \in W_X$$

Merkmalsausprägungen von X

$$\begin{aligned} W_X &= \{x(j) | j = 1, \dots, J\} \\ &= \{x(1), \dots, x(J)\} \end{aligned}$$

Wertebereich von X mit
Merkmalsausprägungen $x(j), j = 1, \dots, J$

$$\begin{aligned} D_N &= \{x_n | n = 1, \dots, N\}, \\ &= \{x_1, \dots, x_N\} \end{aligned}$$

Urliste aus der Messung von X in der
Population M_N , d.h. $x_n = X(e_n)$, $n=1, \dots, N$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ diskrete Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

$D_{N;4}, N=12$

$X_4 = \text{Anzahl Clicks}$

$W_{X_4} = \{0, 1, \dots, 10, 11, 12, 13, 14, 15, 16, 17, 18, \dots, \infty\}$

$J_4 = \infty$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ diskrete Daten: Deskriptive Auswertung

Absolute Häufigkeit N_j und **relative Häufigkeit** f_j analog zu qualitativen Daten

Relative Summenhäufigkeit $s_j = \sum_{k=1}^j f_k = \frac{\#\{x_n \mid x_n \leq x(j)\}}{N}$

Ausprägung	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
$x(1)$	N_1	$f_1 = N_1/N$	f_1
$x(2)$	N_2	$f_2 = N_2/N$	$f_1 + f_2$
...	
$x(J-1)$	N_{J-1}	$f_{J-1} = N_{J-1}/N$	$f_1 + \dots + f_{J-1}$
$x(J)$	N_J	$f_J = N_J/N$	$f_1 + \dots + f_J = 1$
	$\sum_{j=1}^J N_j = N$	$\sum_{j=1}^J f_j = 1$	

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ diskrete Daten: Deskriptive Auswertung

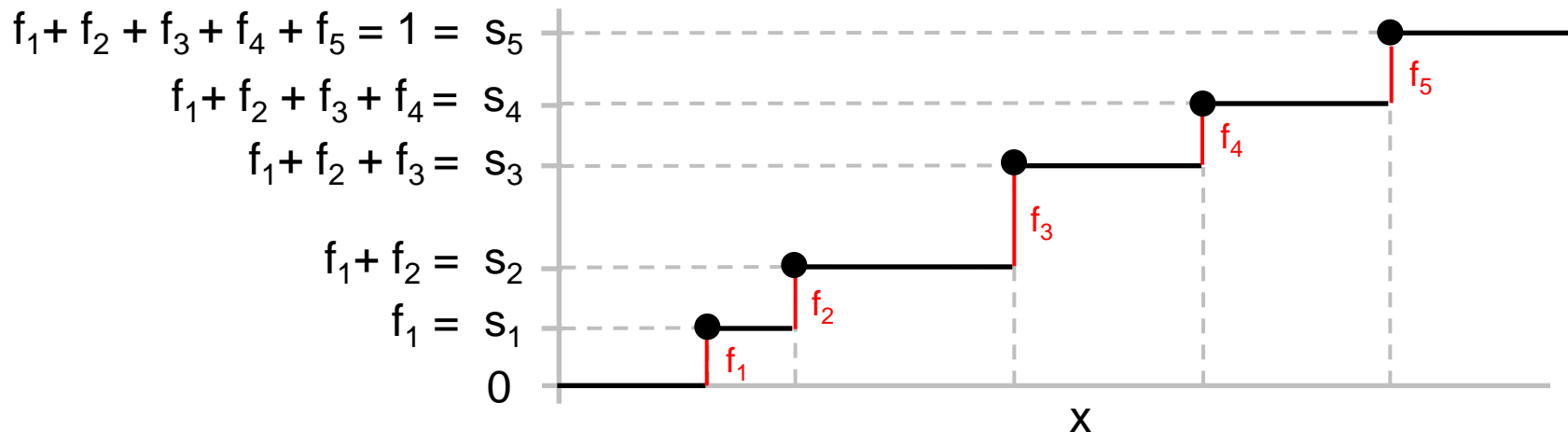
Anzahl Clicks			
Ausprägung	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
0 - 9	0	0	0
10	2	0.167	0.167
11	1	0.083	0.25
12	2	0.167	0.417
13	1	0.083	0.5
14	2	0.167	0.667
15	1	0.083	0.75
16	1	0.083	0.833
17	1	0.083	0.917
18	1	0.083	1
19 - ∞	0	0	1
	12	1	

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ diskrete Daten: Deskriptive Auswertung

Grafische Darstellung: **Empirische Verteilungsfunktion**

$$F_N(x) = \begin{cases} 0 & \text{falls } x < x(1) \\ s_j = \sum_{k=1}^j f_k, \text{ mit } j = \max \{\tilde{j} | x(\tilde{j}) \leq x\} & \text{falls } x(1) \leq x \end{cases}$$



Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten

$$M_N = \{e_1, \dots, e_N\}$$

Population bestehend aus Objekten e_1, \dots, e_N

X

Quantitatives Merkmal

$$x, x \in W_X$$

Merkmalsausprägungen von X

$$W_X = (-\infty, \infty) = \bigcup_{j=1}^J K_j$$

Klassierter (kategorisierter) Wertebereich von X

$$K_j = (v_{j-1}, v_j], j = 1, \dots, J-1$$

Merkmalsklassen mit Klassengrenzen

$$K_J = (v_{J-1}, v_J]$$

$$-\infty = v_0 < v_1 < \dots < v_{J-1} < v_J = \infty$$

$$D_N = \{x_n | n = 1, \dots, N\} = \{x_1, \dots, x_N\}$$

Urliste aus der Messung von X in der Population M_N , d.h. $x_n = X(e_n)$, $n=1, \dots, N$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeitung	Bearbeiter(in)	Aufgabe	Version	Anzahl Clicks	Bearbeitungszeit
e ₁	Kai	Export	1.1	14	8.0
e ₂	Kai	Verknüpfung	1.2	12	4.9
e ₃	Miriam	Export	1.1	12	6.6
e ₄	Tina	Verknüpfung	1.2	13	3.2
e ₅	Oliver	Export	2.0	17	3.9
e ₆	Tina	Export	1.2	11	4.5
e ₇	Tina	Verknüpfung	1.2	14	6.1
e ₈	Miriam	Export	1.2	10	3.7
e ₉	Miriam	Export	1.2	10	4.2
e ₁₀	Oliver	Abfrage	1.1	18	8.5
e ₁₁	Oliver	Verknüpfung	2.0	16	3.6
e ₁₂	Oliver	Abfrage	2.0	15	3.7

$D_{N;5}, N=12$

$X_5 = \text{Bearbeitungszeit}$

$$W_{X_5} = (-\infty, \infty) = (-\infty, 4] \cup (4, 5] \cup \dots \cup (7, 8] \cup (8, \infty) = (-\infty, 4] \cup \left(\bigcup_{j=1}^4 (j+3, j+4] \right) \cup (8, \infty)$$

$J_5 = 6$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Klassierte Häufigkeitsverteilung

Klasse K_j	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
$K_1 = (v_0, v_1]$	$N(K_1)$	$f(K_1) = N(K_1)/N$	$f(K_1)$
$K_2 = (v_1, v_2]$	$N(K_2)$	$f(K_2) = N(K_2)/N$	$f(K_1) + f(K_2)$
...	
$K_{j-1} = (v_{j-2}, v_{j-1}]$	$N(K_{j-1})$	$f(K_{j-1}) = N(K_{j-1})/N$	$f(K_1) + \dots + f(K_{j-1})$
$K_j = (v_{j-1}, v_j]$	$N(K_j)$	$f(K_j) = N(K_j)/N$	$f(K_1) + \dots + f(K_j) = 1$
	$\sum_{j=1}^J N(K_j) = N$	$\sum_{j=1}^J f(K_j) = 1$	

$$N(K_j) = \#\{x \mid x \in K_j\} = \#\{x \mid v_{j-1} < x \leq v_j\}$$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

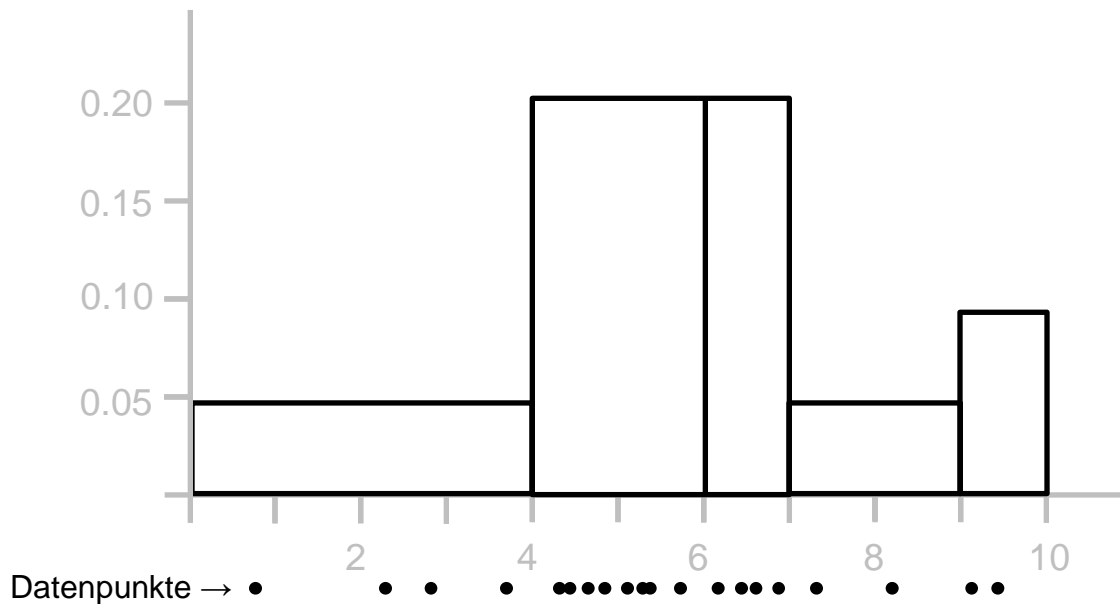
Bearbeitungszeit			
Klasse	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
$K_1 = (-\infty, 4]$	5	0.417	0.417
$K_2 = (4, 5]$	3	0.250	0.667
$K_3 = (5, 6]$	0	0.000	0.667
$K_4 = (6, 7]$	2	0.167	0.833
$K_5 = (7, 8]$	1	0.083	0.917
$K_6 = (8, \infty)$	1	0.083	1
	12	1	

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

Aufbauend auf klassierter Häufigkeitsverteilung, allerdings $v_0 \neq -\infty$ und $v_J \neq \infty$.



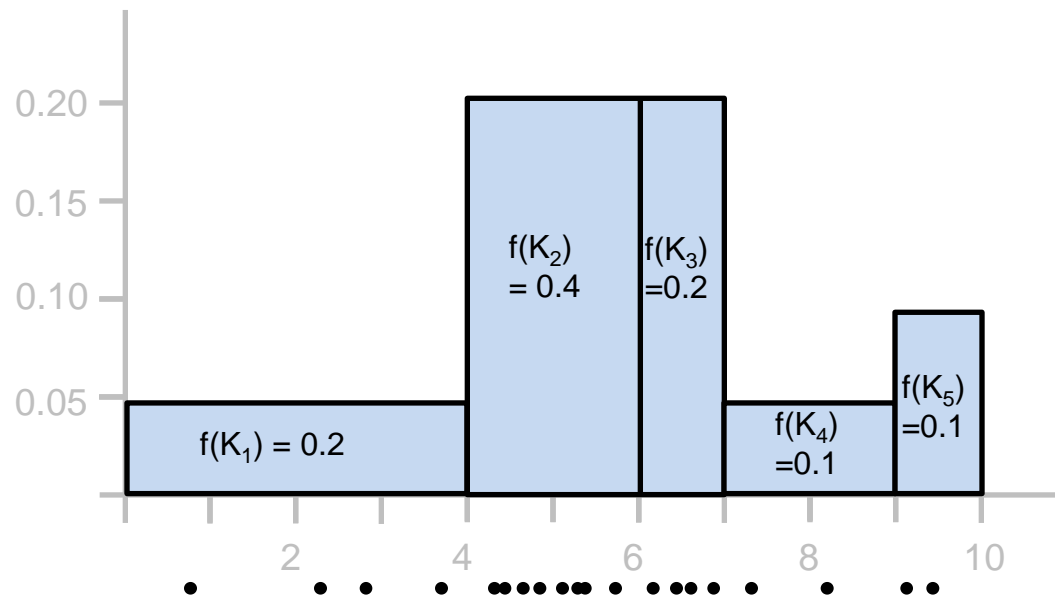
Flächen der Rechtecke zu K_j
entsprechen $f(K_j)$.
Rechteckbreiten sind gegeben
durch $b_j = v_j - v_{j-1}$.
Damit ergeben sich als
Rechteckhöhen $h_j = f(K_j)/b_j$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

Aufbauend auf klassierter Häufigkeitsverteilung, allerdings $v_0 \neq -\infty$ und $v_J \neq \infty$.



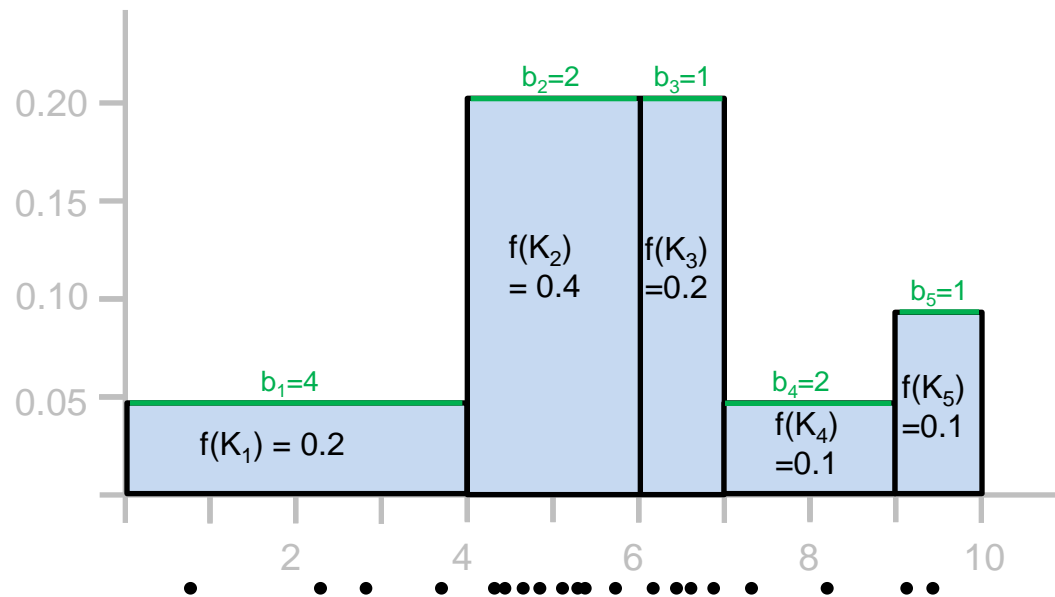
Flächen der Rechtecke zu K_j entsprechen $f(K_j)$.
 Rechteckbreiten sind gegeben durch $b_j = v_j - v_{j-1}$.
 Damit ergeben sich als Rechteckhöhen $h_j = f(K_j)/b_j$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

Aufbauend auf klassierter Häufigkeitsverteilung, allerdings $v_0 \neq -\infty$ und $v_J \neq \infty$.



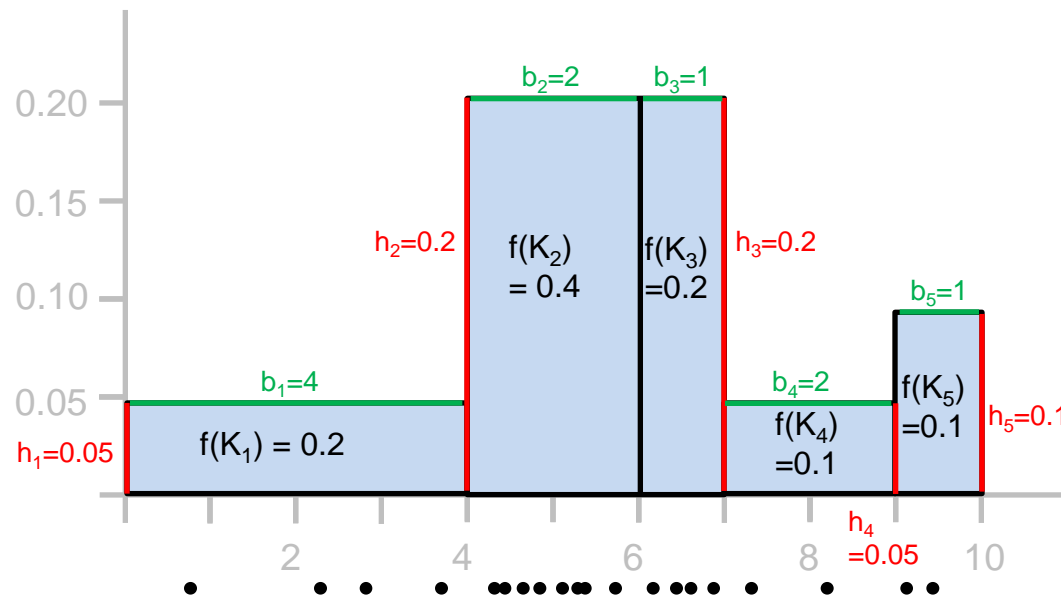
Flächen der Rechtecke zu K_j entsprechen $f(K_j)$.
 Rechteckbreiten sind gegeben durch $b_j = v_j - v_{j-1}$.
 Damit ergeben sich als Rechteckhöhen $h_j = f(K_j)/b_j$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

Aufbauend auf klassierter Häufigkeitsverteilung, allerdings $v_0 \neq -\infty$ und $v_J \neq \infty$.



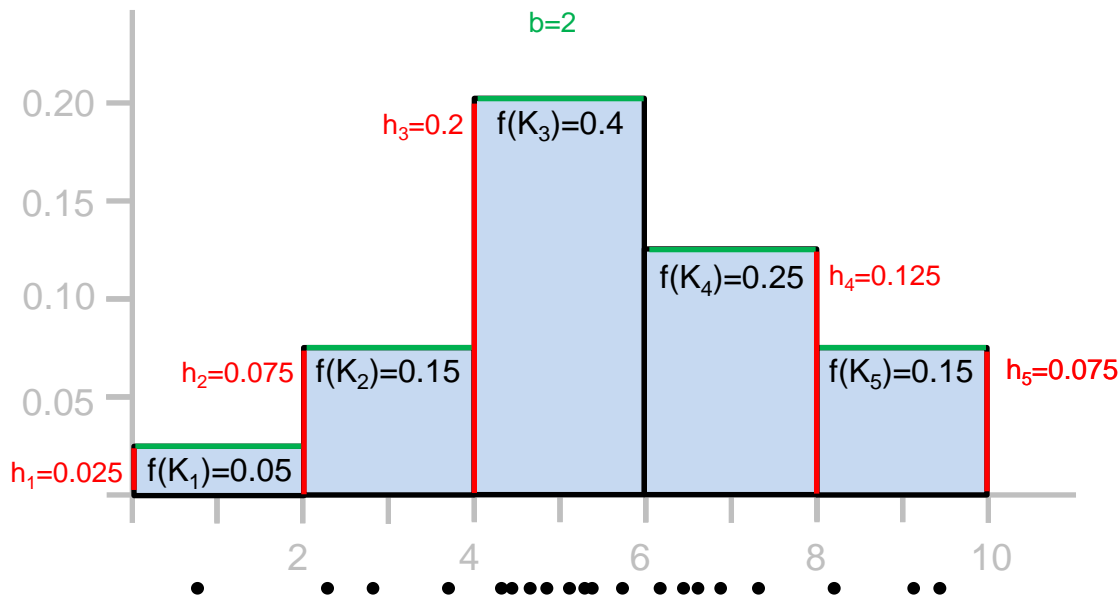
Flächen der Rechtecke zu K_j entsprechen $f(K_j)$.
 Rechteckbreiten sind gegeben durch $b_j = v_j - v_{j-1}$.
 Damit ergeben sich als Rechteckhöhen $h_j = f(K_j)/b_j$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

Üblicherweise gleiche Klassenbreiten.



Flächen der Rechtecke zu K_j entsprechen $f(K_j)$.
 Rechteckbreiten sind gegeben durch $b = v_j - v_{j-1}$.
 Damit ergeben sich als Rechteckhöhen $h_j = f(K_j)/b$

Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

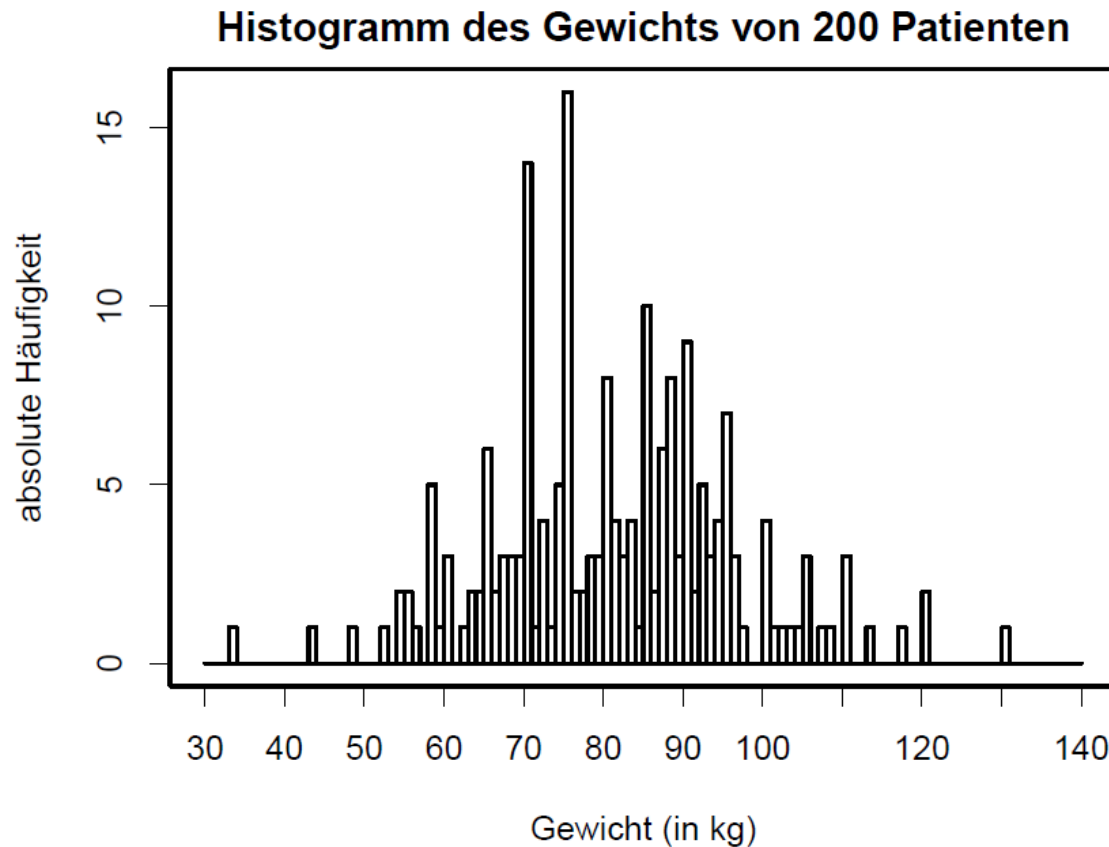
- Beispiel Patientendaten: Gewicht (in kg); NA: fehlender Wert (Not Available)
Zufällige Auswahl des Gewichts von 200 Patienten:

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								

Tabellarische und grafische Darstellung von univariaten Daten

Grafische Darstellung: **Histogramm**

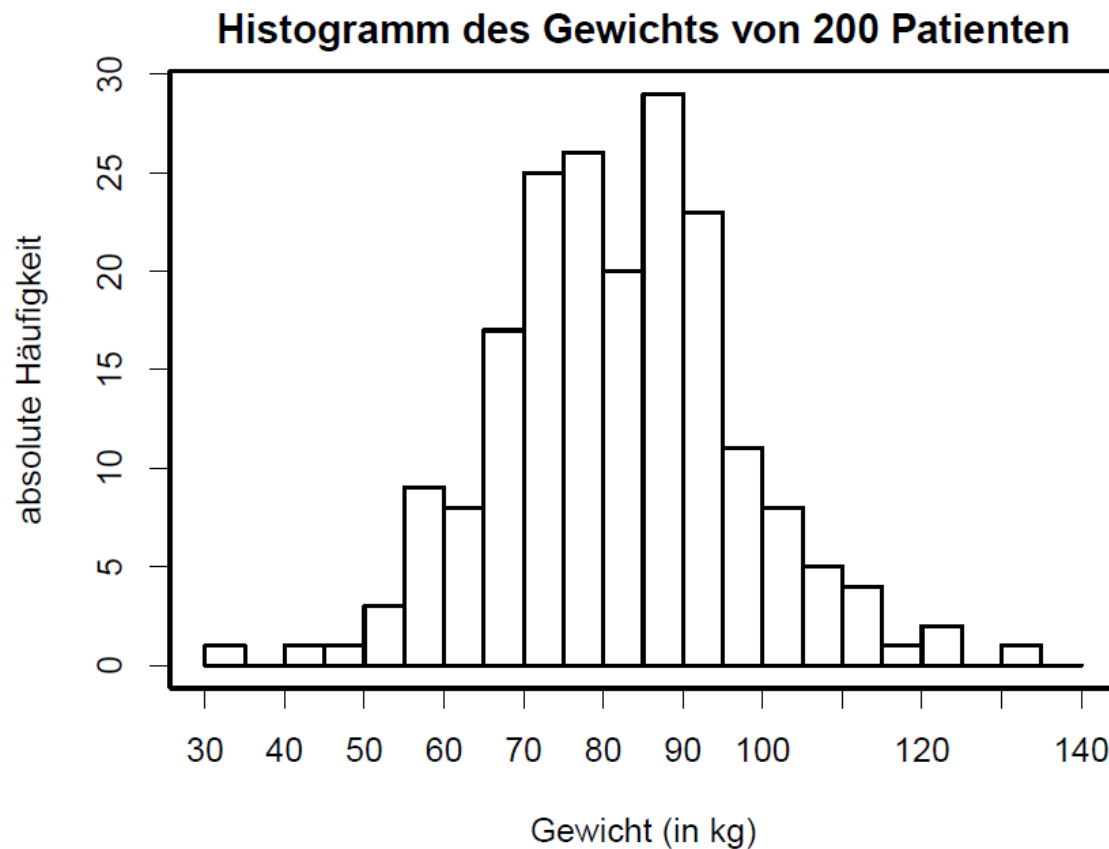
- Patientendaten: Klassenbreite 1 kg führt zu unruhigem Bild, auffällig: Häufungen bei Vielfachen von 5 kg



Tabellarische und grafische Darstellung von univariaten Daten

Grafische Darstellung: **Histogramm**

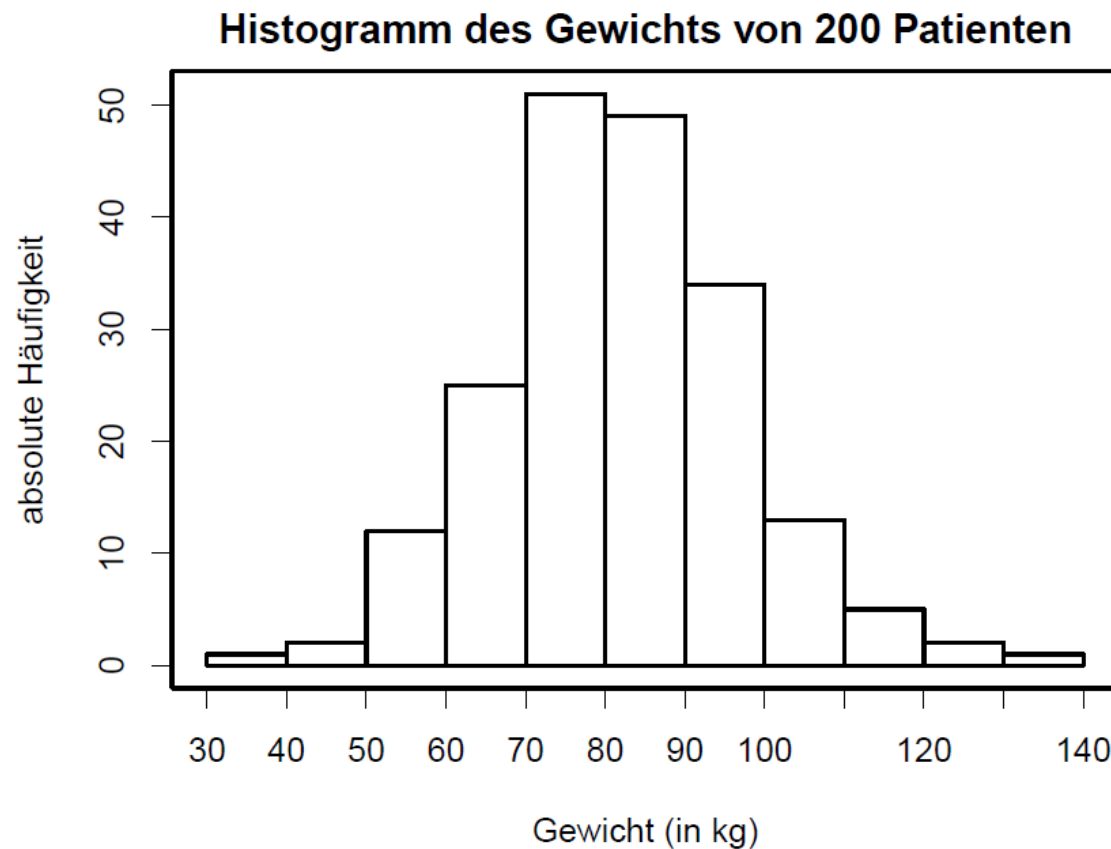
- Patientendaten: Klassenbreite 5 kg



Tabellarische und grafische Darstellung von univariaten Daten

Grafische Darstellung: **Histogramm**

- Patientendaten: Klassenbreite 10 kg



Tabellarische und grafische Darstellung von univariaten Daten

- Bei qualitativen Merkmalen ist ein sogenanntes **Stabdiagramm (Balkendiagramm)** etabliert
 - Pro Merkmalsausprägung wird ein schmaler Stab (Balken) mit der absoluten oder relativen Häufigkeit über dem Merkmalswert gezeichnet
 - Merkmalsausprägungen werden für qualitative Merkmale gleichabständig auf der x-Achse gezeichnet
 - Stäbe sind immer (im Gegensatz zu Kästen beim Histogramm) voneinander separiert!
- 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) Häufigkeiten sind
 - Kreissegmente (Winkel) sind viel schlechter vergleichbar als Stäbe/Balken, deshalb besser Stabdiagramme verwenden