

Wahrscheinlichkeitsrechnung und mathematische Statistik für Informatiker

Vorlesung im Wintersemester 2018/19
an der TU Dortmund

Jun.-Prof. Dr. Andreas Groll

WiSe 18/19, Fakultät Statistik, TU Dortmund

Einleitung

1.1 WRUMS für Informatiker

Jun.-Prof. Dr. Andreas Groll

Mathegebäude, Raum 216a

E-mail: groll@statistik.tu-dortmund.de

- Vorlesung

- ▶ Termin: Do 08:30 - 10:00
- ▶ Hörsaal: HG II - HS3

1.1 WRUMS für Informatiker

Jun.-Prof. Dr. Andreas Groll

Mathegebäude, Raum 216a

E-mail: groll@statistik.tu-dortmund.de

- Vorlesung

- ▶ Termin: Do 08:30 - 10:00
- ▶ Hörsaal: HG II - HS3

Hendrik van der Wurp

Mathegebäude, Raum 226

E-mail: vanderwurp@statistik.tu-dortmund.de

- Übung (2 Gruppen)

- ▶ Termin: Do 18:00 - 19:30 (i.d.R. alle 2 Wochen)
- ▶ Gruppe 1: ab 18.10.2018
- ▶ Gruppe 2: ab 25.10.2018
- ▶ Raum: EF50 - HS1

1.1 WRUMS für Informatiker

Jun.-Prof. Dr. Andreas Groll

Mathegebäude, Raum 216a

E-mail: groll@statistik.tu-dortmund.de

- Vorlesung

- ▶ Termin: Do 08:30 - 10:00
- ▶ Hörsaal: HG II - HS3

Hendrik van der Wurp

Mathegebäude, Raum 226

E-mail: vanderwurp@statistik.tu-dortmund.de

- Übung (2 Gruppen)

- ▶ Termin: Do 18:00 - 19:30 (i.d.R. alle 2 Wochen)
- ▶ Gruppe 1: ab 18.10.2018
- ▶ Gruppe 2: ab 25.10.2018
- ▶ Raum: EF50 - HS1

- Klausurtermine

- ▶ Klausur: Mo (18.02.2019)
16-18 Uhr, Räume: Audimax & EF50 - HS3 & Mathe E28 & Mathe E29
- ▶ Nachklausur: Fr (29.03.2019)
8-10 Uhr, Räume: Audimax & Mathe E28 & Mathe E29

1.1 WRUMS für Informatiker

Jun.-Prof. Dr. Andreas Groll

Mathegebäude, Raum 216a

E-mail: groll@statistik.tu-dortmund.de

- Vorlesung

- ▶ Termin: Do 08:30 - 10:00
- ▶ Hörsaal: HG II - HS3

Hendrik van der Wurp

Mathegebäude, Raum 226

E-mail: vanderwurp@statistik.tu-dortmund.de

- Übung (2 Gruppen)

- ▶ Termin: Do 18:00 - 19:30 (i.d.R. alle 2 Wochen)
- ▶ Gruppe 1: ab 18.10.2018
- ▶ Gruppe 2: ab 25.10.2018
- ▶ Raum: EF50 - HS1

- Klausurtermine

- ▶ Klausur: Mo (18.02.2019)
16-18 Uhr, Räume: Audimax & EF50 - HS3 & Mathe E28 & Mathe E29
- ▶ Nachklausur: Fr (29.03.2019)
8-10 Uhr, Räume: Audimax & Mathe E28 & Mathe E29

- Voraussetzung für Klausur

Regelmäßige Teilnahme Ü,
selbstständige Bearbeitung der
Übungsaufgaben

1.1 WRUMS für Informatiker

Jun.-Prof. Dr. Andreas Groll

Mathegebäude, Raum 216a

E-mail: groll@statistik.tu-dortmund.de

- Vorlesung

- ▶ Termin: Do 08:30 - 10:00
- ▶ Hörsaal: HG II - HS3

Hendrik van der Wurp

Mathegebäude, Raum 226

E-mail: vanderwurp@statistik.tu-dortmund.de

- Übung (2 Gruppen)

- ▶ Termin: Do 18:00 - 19:30 (i.d.R. alle 2 Wochen)
- ▶ Gruppe 1: ab 18.10.2018
- ▶ Gruppe 2: ab 25.10.2018
- ▶ Raum: EF50 - HS1

- Klausurtermine

- ▶ Klausur: Mo (18.02.2019)
16-18 Uhr, Räume: Audimax & EF50 - HS3 & Mathe E28 & Mathe E29
- ▶ Nachklausur: Fr (29.03.2019)
8-10 Uhr, Räume: Audimax & Mathe E28 & Mathe E29

- Voraussetzung für Klausur

Regelmäßige Teilnahme Ü,
selbstständige Bearbeitung der
Übungsaufgaben

- Moodle

Website: [https://moodle.tu-dortmund.de/
course/view.php?id=13183](https://moodle.tu-dortmund.de/course/view.php?id=13183)

Passwort: wrums1819

1.2 Übersicht

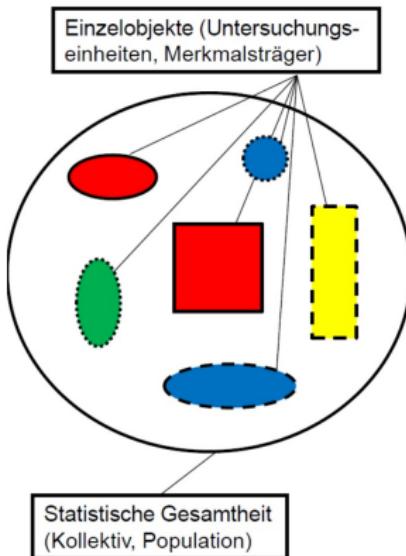
- Motivation
- Merkmale und Datentypen
- Univariate Daten
 - ▶ Tabellarische und grafische Darstellung
- Statistische Kennzahlen
 - ▶ für die Lage
 - ▶ für die Streuung
- Bivariate Daten
 - ▶ Tabellarische und grafische Darstellung
 - ▶ Zusammenhangsmaße
 - ▶ Lineare Regression
- Mengentheoretische Grundlagen
- Wahrscheinlichkeitsmaße und -räume
- 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

1.3 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 nicht alle diese Themen, sondern die dazu notwendigen statistischen Grundlagen.

Univariate Daten

2.1 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	$(0, \infty)$

2.1 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

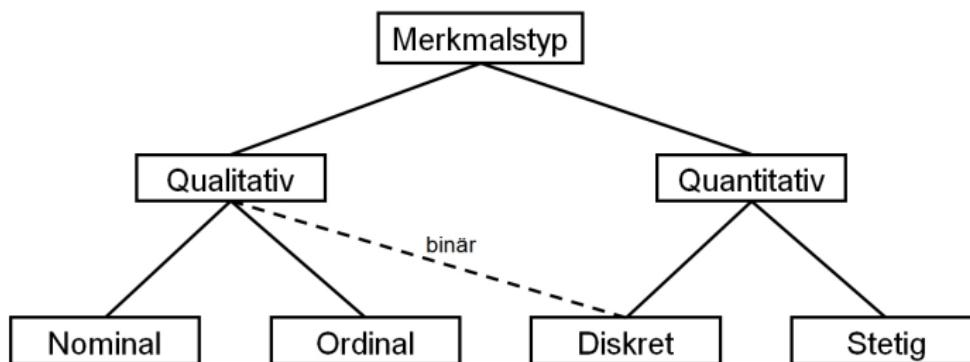
2.1 Merkmale und Datentypen

Datentypen

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

2.1 Merkmale und Datentypen

Datentypen



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

2.1 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)

2.2 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. ordinale Merkmal

$x \in W_X$ Merkmalsausprägungen von X

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

$D_N = \{x_n \mid n = 1, \dots, N\}$ Urliste aus der Messung von X in der
 $= \{x_1, \dots, x_N\}$ Population M_N , d.h. $x_n = X(e_n)$, $n = 1, \dots, N$

$x(1) < x(2) < \dots < x(J)$ falls X ordinal

2.2 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

2.2 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

2.2 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

2.2 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

2.2 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

X₁ = Bearbeiter(in)

W_{X1} = {Kai, Miriam, Oliver, Tina}

J₁ = 4

2.2 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

X₂ = Aufgabe

W_{X2} = {Abfrage, Export, Verknüpfung}

J₂ = 3

2.2 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

$D_{N,3}, N=12$
 $X_3 = \text{Version}$
 $W_{X_3} = \{1.1, 1.2, 2.0\}, 1.1 < 1.2 < 2.0$
 $J_3 = 3$

2.2 Tabellarische und grafische Darstellung von univariaten Daten

Qualitative Daten: Deskriptive Auswertung

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

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
2	3	4	3	12

i	$X_1(e_i)$	$d_{11}(1)$	$d_{11}(2)$	$d_{11}(3)$	$d_{11}(4)$
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

2.2 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	$X_1(e_i)$	$d_{11}(1)$	$d_{11}(2)$	$d_{11}(3)$	$d_{11}(4)$
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

2.2 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	$f_1 = N_1/N$
\vdots	\vdots	\vdots
$x(J)$	N_J	$f_J = N_J/N$
	$\sum_{j=1}^J N_j = N$	$\sum_{j=1}^J f_j = 1$

2.2 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

2.2 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

$x \in W_X$ Merkmalsausprägungen von X

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

$D_N = \{x_n \mid n = 1, \dots, N\}$ Urliste aus der Messung von X in der
 $= \{x_1, \dots, x_N\}$ Population M_N , d.h. $x_n = X(e_n)$, $n = 1, \dots, N$

2.2 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

X₄ = Anzahl Clicks

W_{X4} = {0, 1, ..., 10, 11, 12, 13, 14, 15, 16, 17, 18, ..., ∞}

J₄ = ∞

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

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

2.2 Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ diskrete Daten: Deskriptive Auswertung

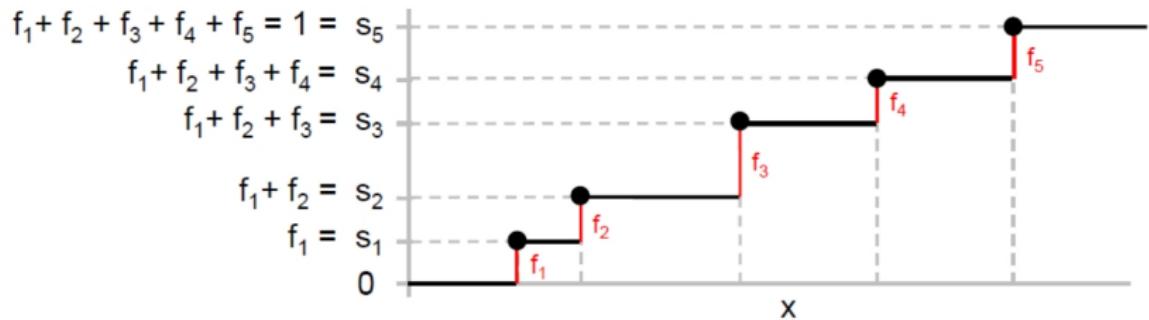
Anzahl Clicks			
Ausprägung	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
0 - 9	0	0	0
10	2	0.167	0.167
11	1	0.083	0.25
12	2	0.167	0.417
13	1	0.083	0.5
14	2	0.167	0.667
15	1	0.083	0.75
16	1	0.083	0.833
17	1	0.083	0.917
18	1	0.083	1
19 - ∞	0	0	1
	12	1	

2.2 Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ diskrete Daten: Deskriptive Auswertung

Grafische Darstellung: **Empirische Verteilungsfunktion**

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



2.2 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

$$x \in W_X$$

Merkmalsausprägungen von X

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

Klassierter (kategorisierter) Wertebereich von X

$$\begin{aligned}K_j &= (v_{j-1}, v_j], \quad j = 1, \dots, J-1 \\K_J &= (v_{J-1}, v_J)\end{aligned}$$

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

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

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

2.2 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 = Bearbeitungszeit

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

$J_5 = 6$

2.2 Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Klassierte Häufigkeitsverteilung

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

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

2.2 Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

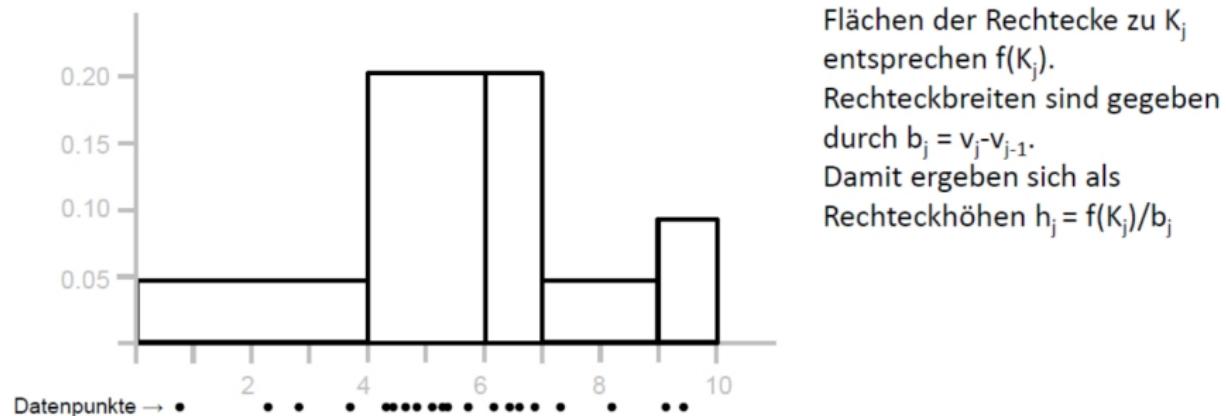
Bearbeitungszeit			
Klasse	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
$K_1 = (-\infty, 4]$	5	0.417	0.417
$K_2 = (4, 5]$	3	0.250	0.667
$K_3 = (5, 6]$	0	0.000	0.667
$K_4 = (6, 7]$	2	0.167	0.833
$K_5 = (7, 8]$	1	0.083	0.917
$K_6 = (8, \infty)$	1	0.083	1
	12	1	

2.2 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 $v_J \neq \infty$

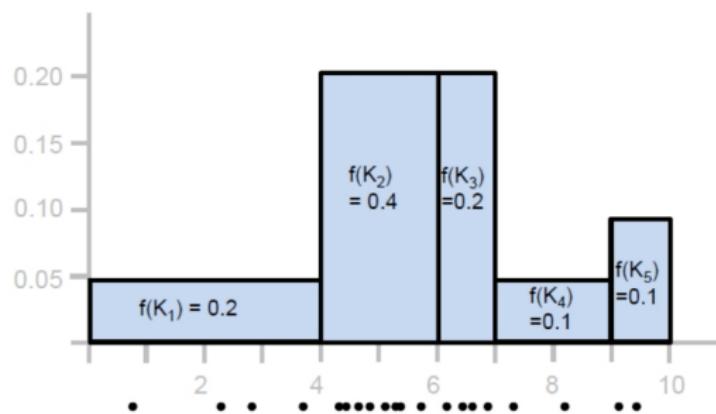


2.2 Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

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



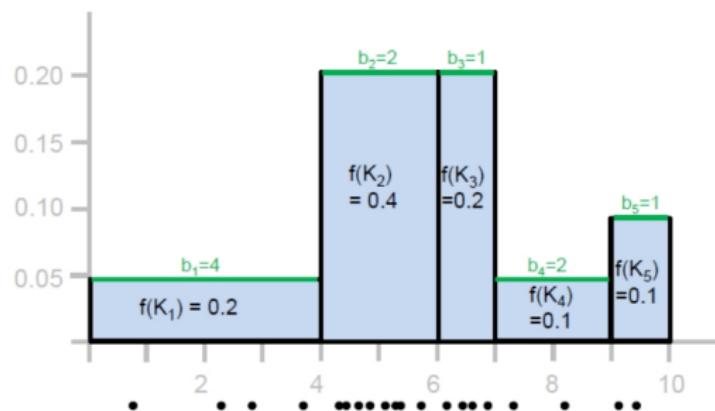
Flächen der Rechtecke zu K_j entsprechen $f(K_j)$.
Rechteckbreiten sind gegeben durch $b_j = v_j - v_{j-1}$.
Damit ergeben sich als Rechteckhöhen $h_j = f(K_j)/b_j$

2.2 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 $v_J \neq \infty$



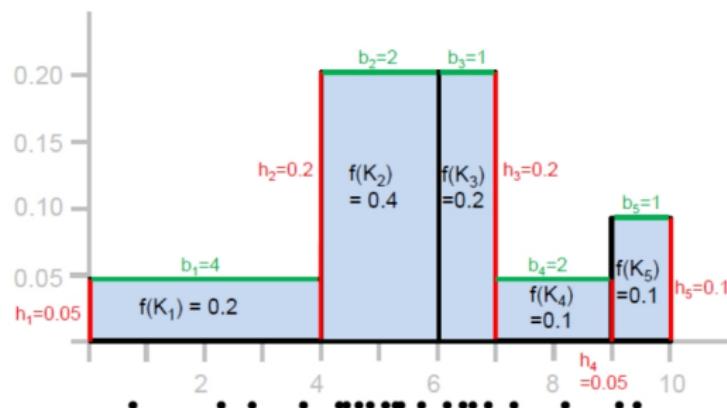
Flächen der Rechtecke zu K_j entsprechen $f(K_j)$.
Rechteckbreiten sind gegeben durch $b_j = v_j - v_{j-1}$.
Damit ergeben sich als Rechteckhöhen $h_j = f(K_j) / b_j$

2.2 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 $v_J \neq \infty$



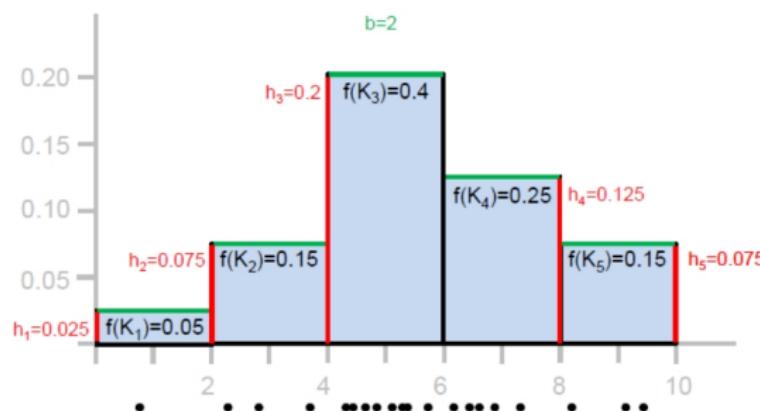
Flächen der Rechtecke zu K_j entsprechen $f(K_j)$.
 Rechteckbreiten sind gegeben durch $b_j = v_j - v_{j-1}$.
 Damit ergeben sich als Rechteckhöhen $h_j = f(K_j)/b_j$

2.2 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 $h_j = f(K_j)/b$

2.2 Tabellarische und grafische Darstellung von univariaten Daten

Quantitativ stetige Daten: Deskriptive Auswertung

Grafische Darstellung: **Histogramm**

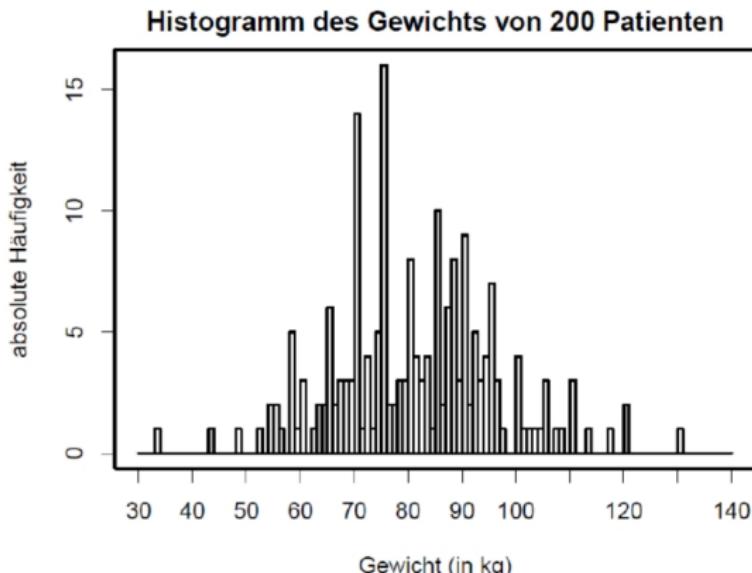
- Beispiel Patientendaten: Gewicht (in kg); NA: fehlender Wert (Not Available)
→ Zufällige Auswahl des Gewichts von 200 Patienten:

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								

2.2 Tabellarische und grafische Darstellung von univariaten Daten

Grafische Darstellung: **Histogramm**

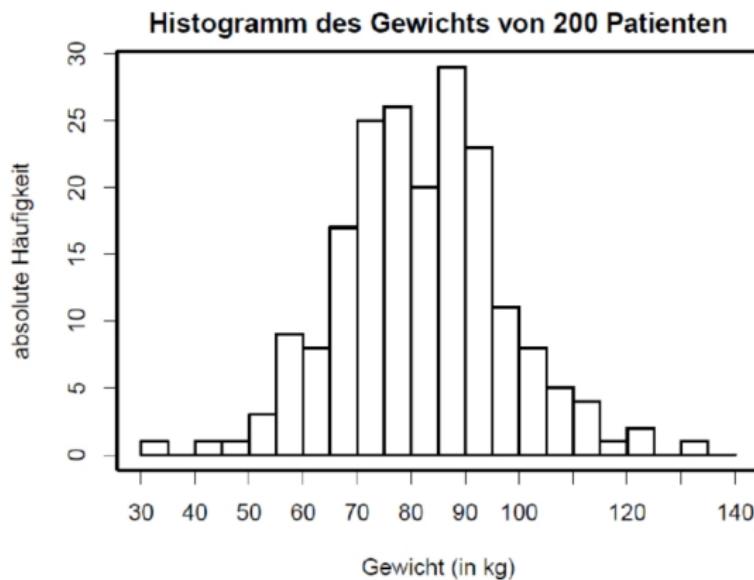
- Patientendaten: Klassenbreite 1 kg führt zu unruhigem Bild, auffällig: Häufungen bei Vielfachen von 5 kg



2.2 Tabellarische und grafische Darstellung von univariaten Daten

Grafische Darstellung: **Histogramm**

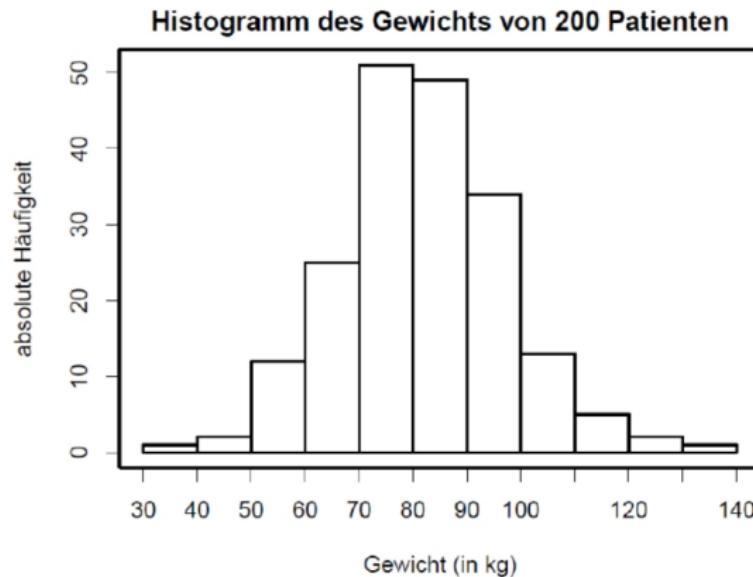
- Patientendaten: Klassenbreite 5 kg



2.2 Tabellarische und grafische Darstellung von univariaten Daten

Grafische Darstellung: **Histogramm**

- Patientendaten: Klassenbreite 10 kg



2.2 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!

Statistische Kennzahlen

3.1 Statistische Kennzahlen für die Lage

- Nach der passenden grafischen Darstellung der Werte eines Merkmals, nun (algebraische) Charakterisierungen der Verteilung solcher Werte.
- Ziel ist es, die Verteilung durch möglichst wenige Maßzahlen zu beschreiben.

① Wo liegt die Mitte der Werte?

Repräsentative Charakterisierung einer Verteilung durch eine Zahl: **Lagemaß**

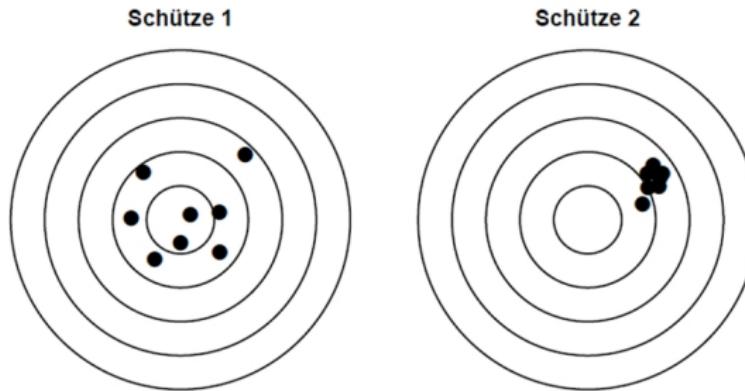
② Wie streuen die Werte um die Mitte?

Charakterisierung der Größe der Unsicherheit (=Streuung) der Merkmalswerte:
Streuungsmaß

- Später: Vergleich verschiedener Gesamtheiten miteinander mit Hilfe der Maßzahlen

3.1 Statistische Kennzahlen für die Lage

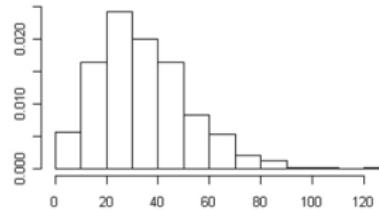
- Beispiel: Welcher Schütze schießt besser?



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

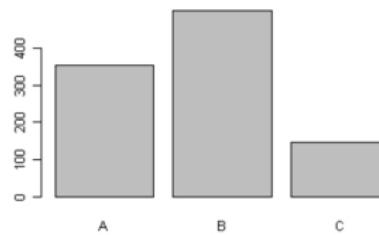
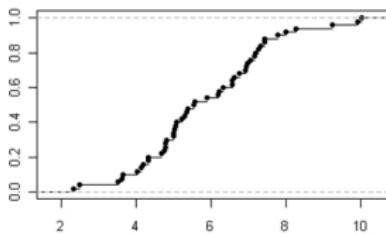
3.1 Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung



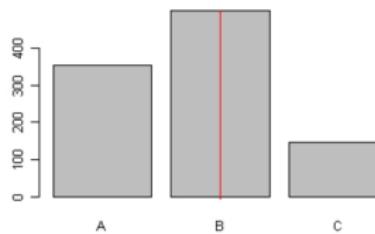
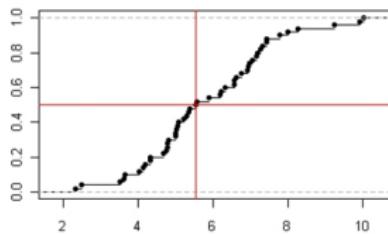
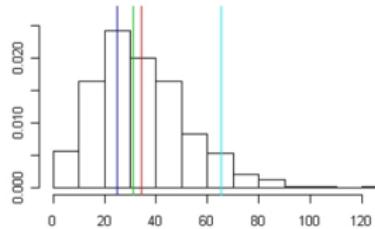
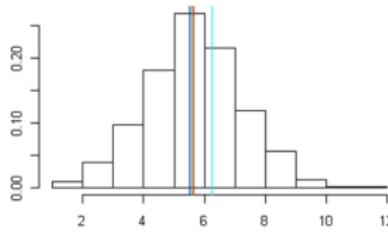
Beispiel

- Histogramm
- Empirische Verteilungsfunktion
- Stabdiagramm



3.1 Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung
Jetzt: stärkere Zusammenfassung der Daten auf ihr „Zentrum“



Farbige Linien
repräsentieren das
Zentrum

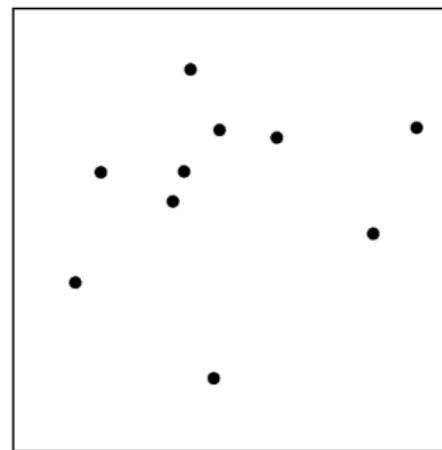
3.1 Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung

Jetzt: stärkere Zusammenfassung der Daten auf ihr „Zentrum“

Unterschiedliche Definitionen von „Zentrum“.

Allgemein: repräsentative Merkmalsausprägung, von der alle beobachteten Werte möglichst wenig abweichen



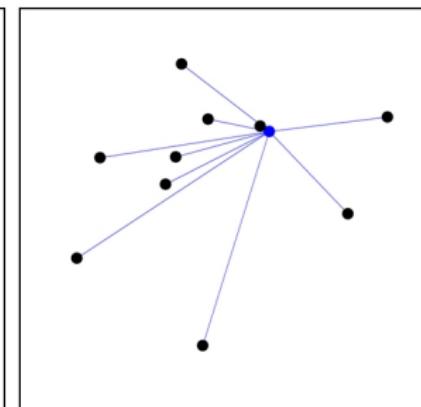
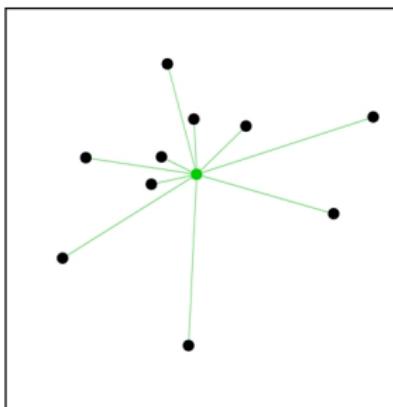
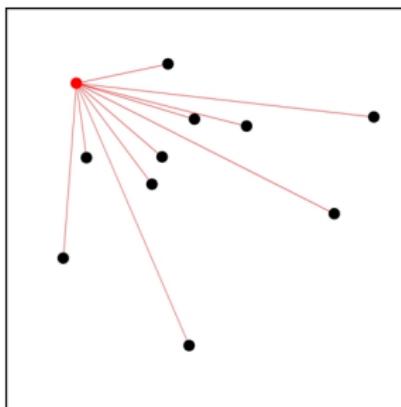
3.1 Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung

Jetzt: stärkere Zusammenfassung der Daten auf ihr „Zentrum“

Unterschiedliche Definitionen von „Zentrum“.

Allgemein: repräsentative Merkmalsausprägung, von der alle beobachteten Werte möglichst wenig abweichen



3.1 Statistische Kennzahlen für die Lage

- Charakterisierung der Merkmalswerte auf einer Gesamtheit durch eine einzige Zahl: Lagemaße
- **Lagemaß = „Mitte der Merkmalswerte“**
- Auswahl des geeigneten Lagemaßes hängt vom Skalenniveau ab
- Wichtigste Beispiele
 - ▶ **Arithmetisches Mittel:** Klassischer Mittelwert
 - ★ Reagiert am empfindlichsten auf „Ausreißer“, d.h. wenn für die Verteilung einige ungewöhnlich große oder kleine Werte vorliegen
 - ▶ **Median:** Zentralwert, mittlerer Wert in der geordneten Stichprobe
 - ★ Liegt nicht unbedingt in der Mitte der Merkmalswerte, ist aber dennoch oft ein guter „Repräsentant“
 - ★ Ist nicht unbedingt eindeutig
 - ▶ **Modalwert:** Häufigster Wert in der Stichprobe
 - ★ Ist nicht unbedingt eindeutig
 - ★ Bei stetigen Merkmalen meist erst nach Klassierung geeignet

3.1 Statistische Kennzahlen für die Lage

Lagemaße:

Arithmetisches Mittel = Mittelwert (mean)

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

Median = „Zentralwert“ = 50%-Wert: med_x

Der Median ist derjenige Wert, für den 50% der Merkmalswerte größer oder gleich und 50% kleiner oder gleich sind.

Der Median ist der mittlere Wert der Rangliste:

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

Modalwert / Modus = häufigster Wert: mod_x

Der Modalwert ist derjenige Merkmalswert, der am häufigsten vorkommt.

3.1 Statistische Kennzahlen für die Lage

- p -Quantil $Q_p = \tilde{x}_p$
 - ▶ Verallgemeinerung des Medians (50%-Wert) auf beliebige Prozentzahlen (100%-Werte)
 - ▶ Nützliche Mittel zur Beschreibung einer Rangliste $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

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

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

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

$Q_p \leq$ nächstgrößerem Merkmalswert der Gesamtheit, also

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

3.1 Statistische Kennzahlen für die Lage

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

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

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

Bezeichnung

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

3.1 Statistische Kennzahlen für die Lage

- Nominale Daten

- ▶ Gesucht: x^* , für das Abweichung zwischen x^* und x_1, \dots, x_N minimal ist
- ▶ Mit nominellen Ausprägungen kann keine sinnvolle Abweichung berechnet werden
- ▶ Dummykodierung führt auf den Modalwert $x(j^*)$

i	x_i
1	A
2	C
:	:
N	B

i	x_i	$d_i(1)$	$d_i(2)$	$d_i(3)$
1	A	1	0	0
2	C	0	0	1
:	:	:	:	:
N	B	0	1	0
\sum		N_1	N_2	N_3

3.1 Statistische Kennzahlen für die Lage

Nominale Daten

Modalwert

Beispiel Bearbeitung von Softwareaufgaben

Die Modalwerte lauten

$$x_1(j^*) = \text{Oliver}$$

$$x_2(j^*) = \text{Export}$$

$$x_3(j^*) = 1.2$$

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

3.1 Statistische Kennzahlen für die Lage

Ordinale Daten

x_1, \dots, x_N

$x_i \in W_X, i = 1, \dots, N$

$W_X = \{x(j) | j = 1, \dots, J\} = \{x(1), \dots, x(J)\}$

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

Urliste x_1, \dots, x_N

Geordnete Liste $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$

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

i	x_i
1	$x(3)$
2	$x(2)$
3	$x(1)$
4	$x(1)$
5	$x(3)$

k	$x_{(k)}$
1	$x(1)$
2	$x(1)$
3	$x(2)$
4	$x(3)$
5	$x(3)$

Geordnete Liste \rightarrow

mit $i_k = \min[\arg\min_{i^*} (x_{i^*} | i^* \in \{1, \dots, N\} \setminus \{i_1, \dots, i_{k-1}\})], k = 1, \dots, N$

$x_{(k)}$ wird k -ter **Rangwert** genannt, erster und letzter Rangwert $x_{(1)}$ und $x_{(N)}$ heißen **Minimum** und **Maximum**.

3.1 Statistische Kennzahlen für die Lage

Ordinale Daten

x_1, \dots, x_N

$x_i \in W_X, i = 1, \dots, N$

$W_X = \{x(j) | j = 1, \dots, J\} = \{x(1), \dots, x(J)\}$

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

$x_{(k)}$ wird k -ter **Rangwert** genannt, erster und letzter Rangwert $x_{(1)}$ und $x_{(n)}$ heißen Minimum und Maximum.

i	x_i	$R(x_i)$
1	$x(3)$	4.5
2	$x(2)$	3
3	$x(1)$	1.5
4	$x(1)$	1.5
5	$x(3)$	4.5

↑ Ränge

$$R(x_i) = \frac{1}{\#K^*} \sum_{k^* \in K^*} k^* \text{ mit } K^* = \{k^* | x_{(k^*)} = x_i\}$$

$R(x_i)$ ist der **Rang** von x_i

Gesucht: $x_{(k^*)}$, für das Abweichung zwischen $x_{(k^*)}$ und x_1, \dots, x_N minimal ist.

k	$x_{(k)}$
1	$x(1)$
2	$x(1)$
3	$x(2)$
4	$x(3)$
5	$x(3)$

3.1 Statistische Kennzahlen für die Lage

Ordinale Daten

Beispiel Bearbeitungen von Softwareaufgaben

i	Version _i
1	1.1
2	1.2
3	1.1
4	1.2
5	2.0
6	1.2
7	1.2
8	1.2
9	1.2
10	1.1
11	2.0
12	2.0

Geordnete
Liste →

k	Version _(k)
1	1.1
2	1.1
3	1.1
4	1.2
5	1.2
6	1.2
7	1.2
8	1.2
9	1.2
10	2.0
11	2.0
12	2.0

Ränge →

$$\frac{1}{3} \sum_{s=1}^3 s = 2$$

$$\frac{1}{6} \sum_{s=4}^9 s = 6.5$$

$$\frac{1}{3} \sum_{s=10}^{12} s = 11$$

i	Version _i	R(Version _i)
1	1.1	2
2	1.2	6.5
3	1.1	2
4	1.2	6.5
5	2.0	11
6	1.2	6.5
7	1.2	6.5
8	1.2	6.5
9	1.2	6.5
10	1.1	2
11	2.0	11
12	2.0	11

3.1 Statistische Kennzahlen für die Lage

Quantitative Daten

x_1, \dots, x_N

$x_i \in W_X, i = 1, \dots, N$

$W_X = \{x(j) | j = 1, \dots, J\} = \{x(1), \dots, x(J)\}$

bzw. $W_X = (-\infty, \infty)$

Der Median minimiert die Summe der absoluten Abweichungen

$$\Delta_a(x) = \sum_{i=1}^N |x_i - x|$$

Der Mittelwert minimiert die Summe der quadratischen Abweichungen

$$\Delta(x) = \sum_{i=1}^N (x_i - x)^2$$

3.1 Statistische Kennzahlen für die Lage

Quantitative Daten

Generell gilt: $\Delta(x) = \sum_{i=1}^N (x_i - x)^2$ ist minimal für $x = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$

Beweis $\forall x \in \mathbb{R}$:

$$\begin{aligned}\Delta(x) &= \sum_{i=1}^N (x_i - x)^2 = \sum_{i=1}^N [(x_i - \bar{x}) + (\bar{x} - x)]^2 \\ &= \sum_{i=1}^N (x_i - \bar{x})^2 + 2(\bar{x} - x) \underbrace{\sum_{i=1}^N (x_i - \bar{x})}_{=0 (*)} + \underbrace{\sum_{i=1}^N (\bar{x} - x)^2}_{=N(\bar{x}-x)^2} \\ &= \Delta(\bar{x}) + \underbrace{N(\bar{x} - x)^2}_{\geq 0} \geq \Delta(\bar{x})\end{aligned}$$

$$(*) \sum_{i=1}^N (x_i - \bar{x}) = \sum_{i=1}^N \left(x_i - \frac{1}{N} \sum_{l=1}^N x_l \right) = \sum_{i=1}^N x_i - \frac{1}{N} \sum_{i=1}^N \sum_{l=1}^N x_l = \sum_{i=1}^N x_i - \frac{1}{N} N \sum_{l=1}^N x_l = 0$$

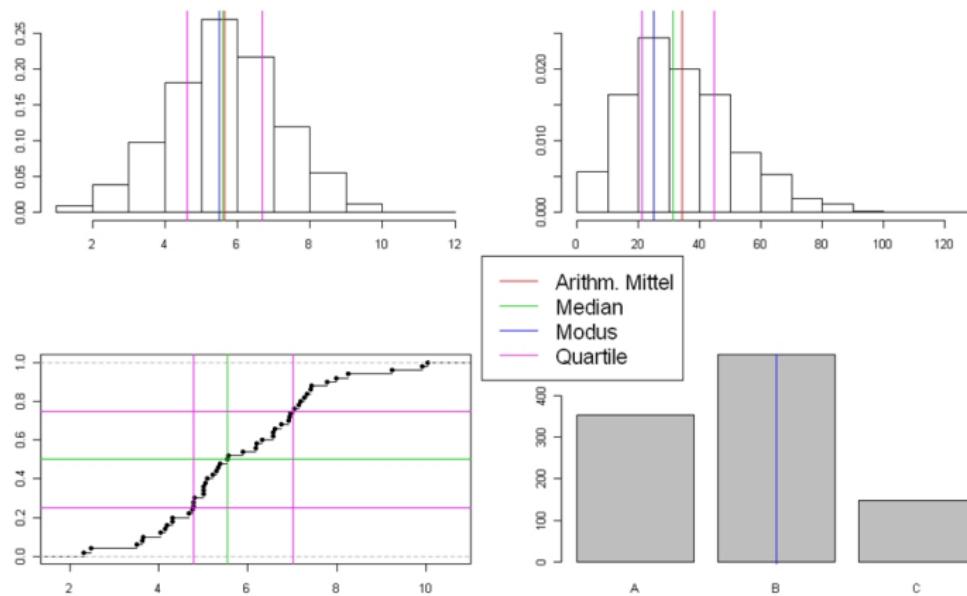
3.1 Statistische Kennzahlen für die Lage

Zusammenfassung: Welche Maßzahlen sind bei welchem Skalenniveau geeignet?

Skalenniveau → ↓ Lagemaß	Nominal	Ordinal	Quantitativ
Modus			- Informationsverlust – Nur für klassierte Daten
Median			+ Robust - Informationsverlust - Hohe Streubreite
Arithmetisches Mittel	 – Nur für J = 2		- Ausreißeranfällig + Informationsnutzung + Geringe Streubreite

3.2 Statistische Kennzahlen für die Streuung

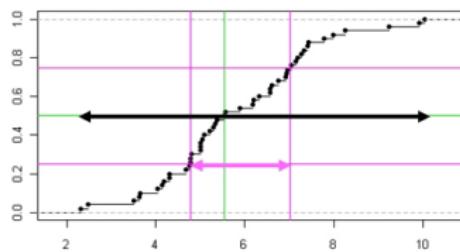
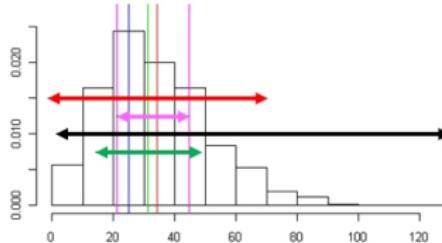
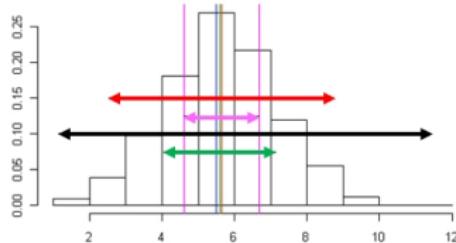
Bisher: Beschreibung von Häufigkeitsverteilung und Lage



3.2 Statistische Kennzahlen für die Streuung

Bisher: Beschreibung von Häufigkeitsverteilung und Lage

Jetzt: Beschreibung der mittleren Variation um die Lage



Allgemein: Streuung desto höher,
je schlechter sich konkrete Werte
vorhersagen lassen.

3.2 Statistische Kennzahlen für die Streuung

Streuungsmaße:

empirische Varianz und Standardabweichung

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

$$\text{var}_x = s_x^2 := \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(n-1)} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Standardabweichung: Wurzel aus der Varianz

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

Interquartilsabstand (interquartile range)

$$\text{qd}_x := q^4 - q_1$$

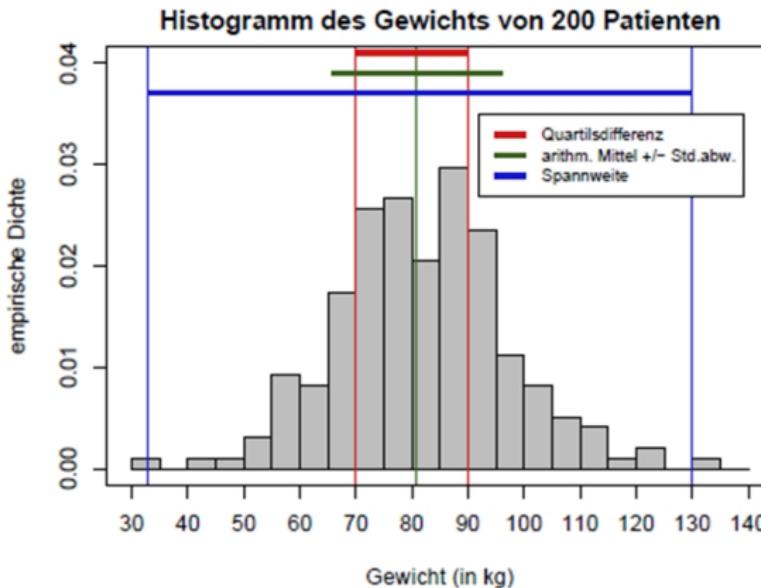
Spannweite (range)

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

3.2 Statistische Kennzahlen für die Streuung

- Beispiel 1: Gewicht von 200 Patienten

$$s_x = 15.14 \text{ kg}, \quad qd_x = 20 \text{ kg}, \quad R_x = 97 \text{ kg}$$



3.2 Statistische Kennzahlen für die Streuung

Streuungsmaße:

Variationskoeffizient (relative Standardabweichung)

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

Mittlere absolute Medianabweichung MD

(von „Mean Deviation from the Median“)

$$md_x := \frac{1}{n} \sum_{i=1}^n |x_i - \text{med}_x|$$

Mediane absolute Medianabweichung MAD

(von „Median Absolute Deviation“)

$$\text{mad}_x := \text{med}(|x_i - \text{med}_x|)$$

3.2 Statistische Kennzahlen für die Streuung

Nominale Daten

x_1, \dots, x_N

$x_i \in W_X, i = 1, \dots, N$

$W_X = \{x(j) | j = 1, \dots, J\}$
 $= \{x(1), \dots, x(J)\}$

Rechnen nur sinnvoll mit
 Dummyvariablen bzw. Häufigkeiten

i	x_i
1	A
2	C
\vdots	\vdots
N	B

i	x_i	$d_i(1)$	$d_i(2)$	$d_i(3)$
1	A	1	0	0
2	C	0	0	1
\vdots	\vdots	\vdots	\vdots	\vdots
N	B	0	1	0
\sum		N_1	N_2	N_3

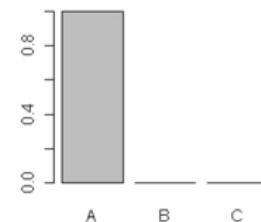
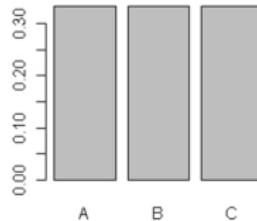
3.2 Statistische Kennzahlen für die Streuung

Nominale Daten

Allgemein: Streuung ist desto höher, je schlechter sich konkrete Werte vorhersagen lassen.

Nominale Merkmalsausprägungen lassen sich um so besser vorhersagen, je häufiger eine bestimmte Kategorie vorkommt.

Geringste Streuung , falls es ein j gibt mit $f_j = 1$. \rightarrow



\leftarrow Höchste Streuung , falls $f_j = 1/J$, $j=1,\dots,J$.

3.2 Statistische Kennzahlen für die Streuung

Nominale Daten

Geringe Streuung, falls es ein j gibt mit $f_j = 1$.

Höchste Streuung, falls $f_j = 1/J$, $j = 1, \dots, J$

D entspricht dem Anteil von Paaren mit unterschiedlichen Merkmalsausprägungen an allen aus der Urliste bildbaren Beobachtungspaaren:

Simpson's D

$$D = \frac{\#\{(i, k) \in \{1, \dots, N\} \times \{1, \dots, N\} | x_i \neq x_k\}}{N^2}$$

$$D = 1 - \sum_{j=1}^J f_j^2$$

Beispiel

i	x _i
1	A
2	B
3	A
4	C

$$\begin{aligned} D &= 1 - \left(\frac{2^2 + 1^2 + 1^2}{4^2} \right) = 1 - \frac{6}{16} = \frac{5}{8} \\ &= \frac{\#\{(1,2), (1,4), (2,1), (2,3), (2,4), (3,2), (3,4), (4,1), (4,2), (4,3)\}}{4^2} \end{aligned}$$

3.2 Statistische Kennzahlen für die Streuung

Nominale Daten

Geringste Streuung, falls es ein j gibt mit $f_j = 1$.

Höchste Streuung, falls $f_j = 1/J$, $j = 1, \dots, J$.

Simpson's D

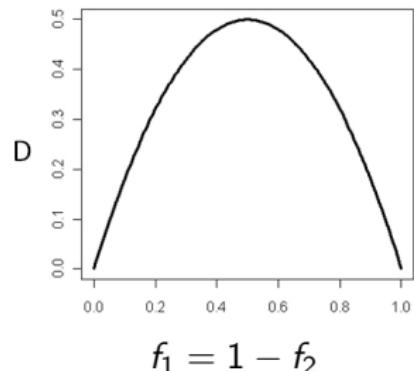
$$D = 1 - \sum_{j=1}^J f_j^2$$

$$0 \leq D \leq 1 - \frac{1}{J}$$

$D = 0$ für $\max[(f_1, \dots, f_J)] = 1$

$D = 1 - \frac{1}{J}$ für $f_1 = \dots = f_J = \frac{1}{J}$

Beispiel $J = 2$



3.2 Statistische Kennzahlen für die Streuung

Nominale Daten

Geringste Streuung, falls es ein j gibt mit $f_j = 1$.

Höchste Streuung, falls $f_j = 1/J$, $j = 1, \dots, J$.

Simpson's D_z (Normierte Version)

$$D_z = \frac{J(1 - \sum_{j=1}^J f_j^2)}{J-1}$$

$$0 \leq D_z \leq 1$$

$$D_z = 0 \text{ für } \max[(f_1, \dots, f_J)] = 1$$

$$D_z = 1 \text{ für } f_1 = \dots = f_J = \frac{1}{J}$$

