

# **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](mailto: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](mailto: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](mailto: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](mailto: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](mailto: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](mailto: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](mailto: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](mailto: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](mailto: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

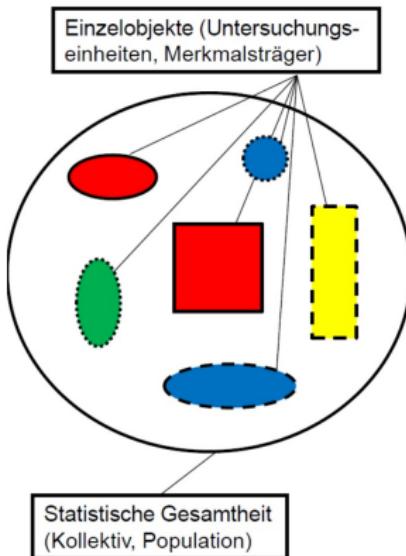
- Motivation
- Merkmale und Datentypen
- Univariate Daten
  - ▶ Tabellarische und grafische Darstellung
- Statistische Kennzahlen
  - ▶ für die Lage
  - ▶ für die Streuung
- Bivariate Daten
  - ▶ Tabellarische und grafische Darstellung
  - ▶ Zusammenhangsmaße
  - ▶ Lineare Regression
- Mengentheoretische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

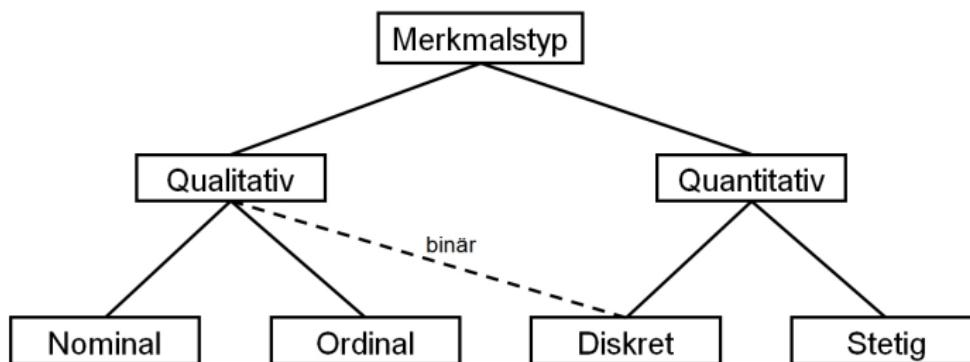
## 2.1 Merkmale und Datentypen

### 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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	Oliver	Abfrage	2.0	15	3.7

D<sub>N;1</sub>, N=12

X<sub>1</sub> = Bearbeiter(in)

W<sub>X1</sub> = {Kai, Miriam, Oliver, Tina}

J<sub>1</sub> = 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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	Oliver	Abfrage	2.0	15	3.7

D<sub>N,2</sub>, N=12

X<sub>2</sub> = Aufgabe

W<sub>X2</sub> = {Abfrage, Export, Verknüpfung}

J<sub>2</sub> = 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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	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	$\Sigma$
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
$\Sigma$		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	$\Sigma$
2/12 $\approx 0.17$	3/12 $= 0.25$	4/12 $\approx 0.33$	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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	Oliver	Abfrage	2.0	15	3.7

 $D_{N,4}, N=12$  $X_4 = \text{Anzahl Clicks}$  $W_{X4} = \{0, 1, \dots, 10, 11, 12, 13, 14, 15, 16, 17, 18, \dots, \infty\}$  $J_4 = \infty$

## 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

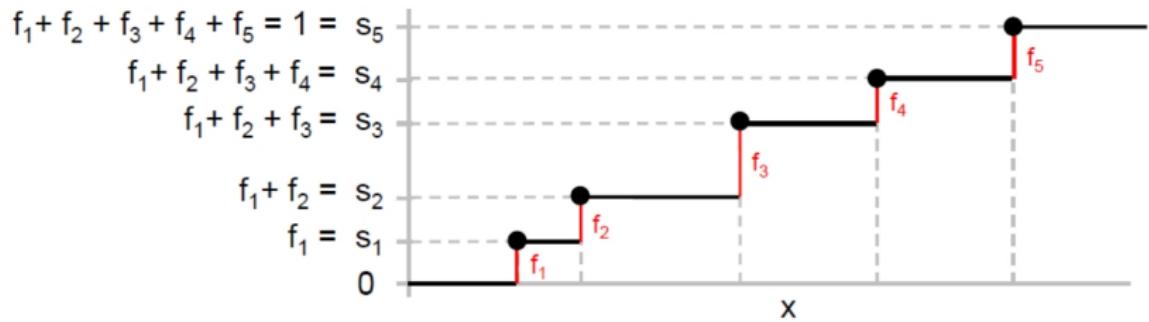
Anzahl Clicks			
Ausprägung	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
0 - 9	0	0	0
10	2	0.167	0.167
11	1	0.083	0.25
12	2	0.167	0.417
13	1	0.083	0.5
14	2	0.167	0.667
15	1	0.083	0.75
16	1	0.083	0.833
17	1	0.083	0.917
18	1	0.083	1
19 - ∞	0	0	1
	12	1	

## 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 <sub>1</sub>	Kai	Export	1.1	14	8.0
e <sub>2</sub>	Kai	Verknüpfung	1.2	12	4.9
e <sub>3</sub>	Miriam	Export	1.1	12	6.6
e <sub>4</sub>	Tina	Verknüpfung	1.2	13	3.2
e <sub>5</sub>	Oliver	Export	2.0	17	3.9
e <sub>6</sub>	Tina	Export	1.2	11	4.5
e <sub>7</sub>	Tina	Verknüpfung	1.2	14	6.1
e <sub>8</sub>	Miriam	Export	1.2	10	3.7
e <sub>9</sub>	Miriam	Export	1.2	10	4.2
e <sub>10</sub>	Oliver	Abfrage	1.1	18	8.5
e <sub>11</sub>	Oliver	Verknüpfung	2.0	16	3.6
e <sub>12</sub>	Oliver	Abfrage	2.0	15	3.7

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

$X_5$  = Bearbeitungszeit

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

$J_5 = 6$

## 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

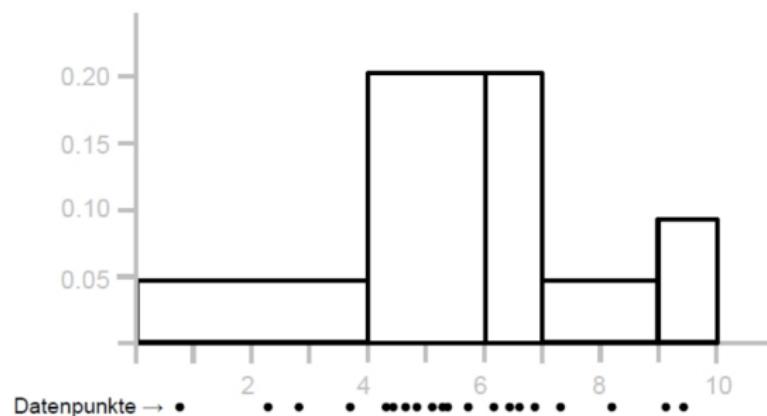
Bearbeitungszeit			
Klasse	Absolute Häufigkeit	Relative Häufigkeit	Relative Summenhäufigkeit
$K_1 = (-\infty, 4]$	5	0.417	0.417
$K_2 = (4, 5]$	3	0.250	0.667
$K_3 = (5, 6]$	0	0.000	0.667
$K_4 = (6, 7]$	2	0.167	0.833
$K_5 = (7, 8]$	1	0.083	0.917
$K_6 = (8, \infty)$	1	0.083	1
	12	1	

## 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  $v_J \neq \infty$



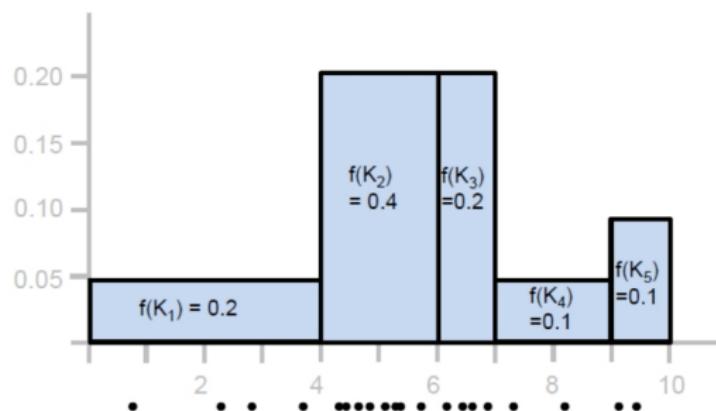
Flächen der Rechtecke zu  $K_j$  entsprechen  $f(K_j)$ .  
Rechteckbreiten sind gegeben durch  $b_j = v_j - v_{j-1}$ .  
Damit ergeben sich als Rechteckhöhen  $h_j = f(K_j)/b_j$

## 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  $v_J \neq \infty$



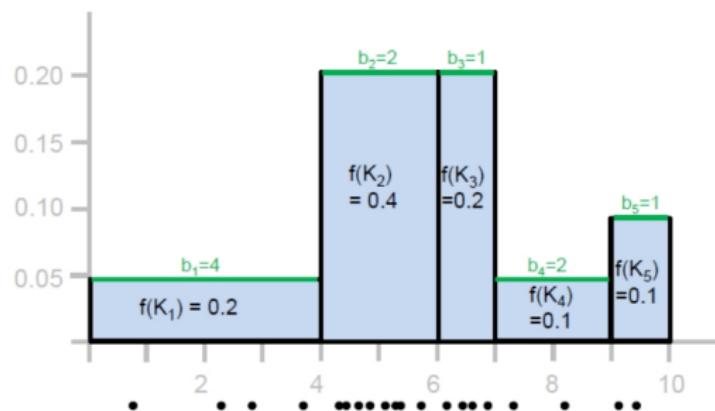
Flächen der Rechtecke zu K<sub>j</sub> entsprechen  $f(K_j)$ .  
Rechteckbreiten sind gegeben durch  $b_j = v_j - v_{j-1}$ .  
Damit ergeben sich als Rechteckhöhen  $h_j = f(K_j)/b_j$

## 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  $v_J \neq \infty$



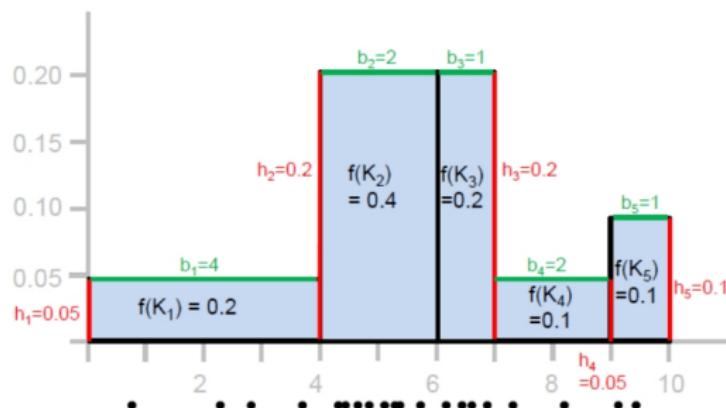
Flächen der Rechtecke zu  $K_j$  entsprechen  $f(K_j)$ .  
Rechteckbreiten sind gegeben durch  $b_j = v_j - v_{j-1}$ .  
Damit ergeben sich als Rechteckhöhen  $h_j = f(K_j) / b_j$

## 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  $v_J \neq \infty$



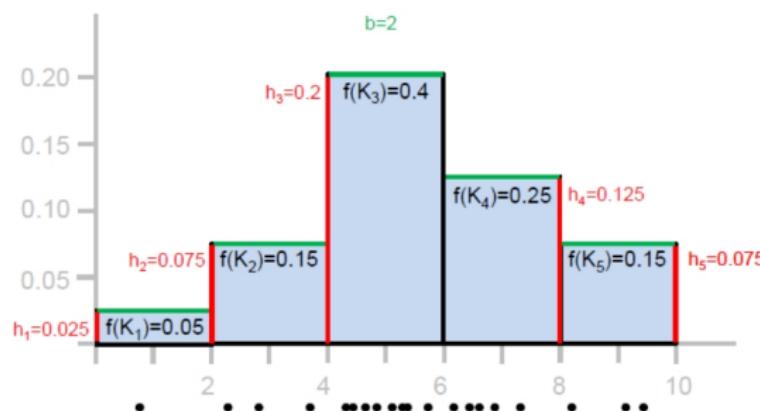
Flächen der Rechtecke zu K<sub>j</sub> entsprechen f(K<sub>j</sub>).  
Rechteckbreiten sind gegeben durch b<sub>j</sub> = v<sub>j</sub>-v<sub>j-1</sub>.  
Damit ergeben sich als Rechteckhöhen h<sub>j</sub> = f(K<sub>j</sub>)/b<sub>j</sub>

## 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:

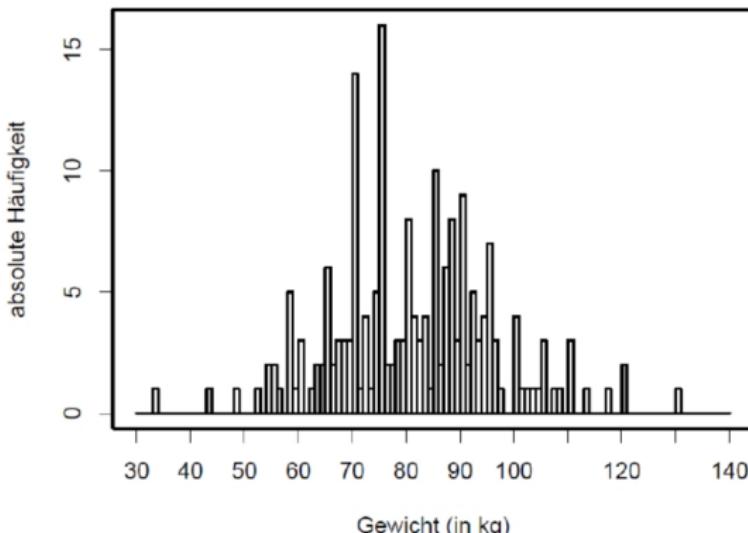
85	70	75	70	92	88	68	101	74	80	87	68	95	33	75	117
105	88	76	82	107	92	87	91	83	80	85	95	75	60	85	75
73	58	93	70	100	94	100	75	80	85	87	43	90	92	89	NA
100	96	58	72	77	83	48	74	90	58	78	75	56	70	75	70
67	95	74	88	70	68	66	102	72	74	113	72	81	75	55	60
75	90	71	93	NA	94	75	89	90	80	52	90	105	90	82	80
83	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

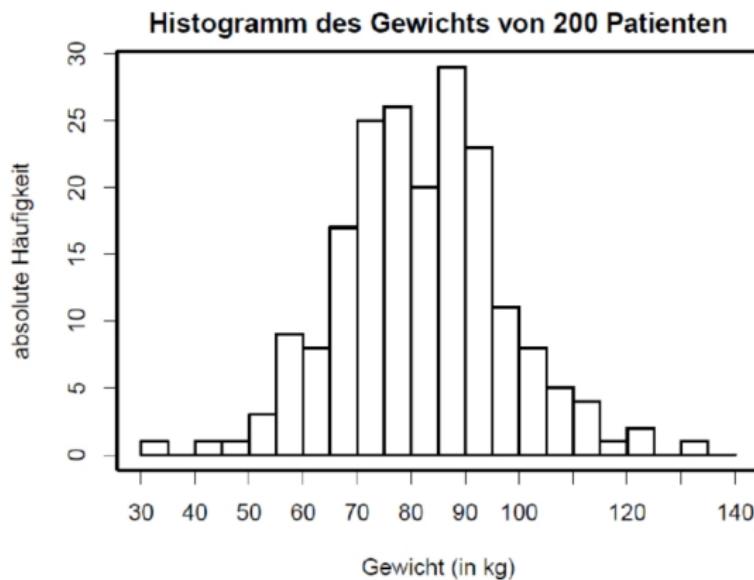
Histogramm des Gewichts von 200 Patienten



## 2.2 Tabellarische und grafische Darstellung von univariaten Daten

Grafische Darstellung: **Histogramm**

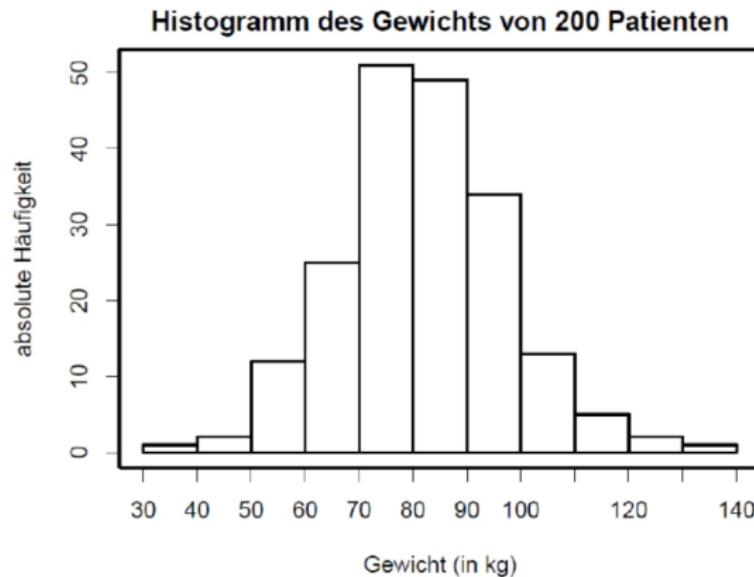
- Patientendaten: Klassenbreite 5 kg



## 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**

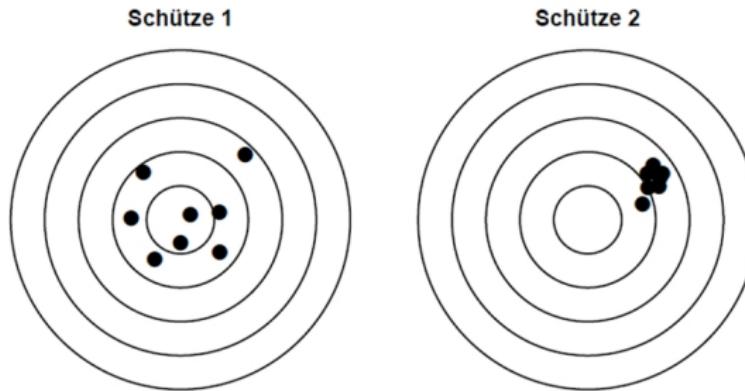
### ② Wie streuen die Werte um die Mitte?

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

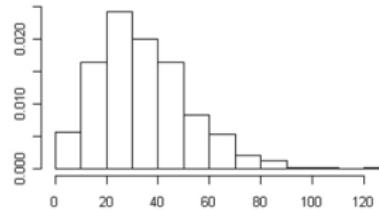
- Beispiel: Welcher Schütze schießt besser?



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

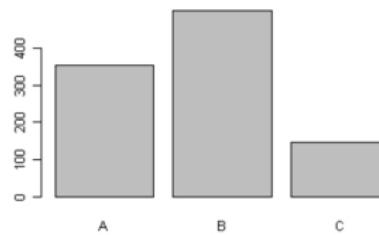
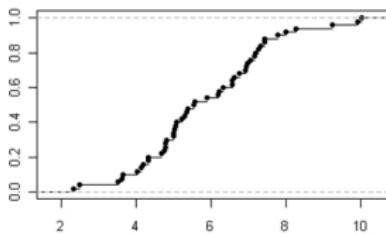
# 3.1 Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung



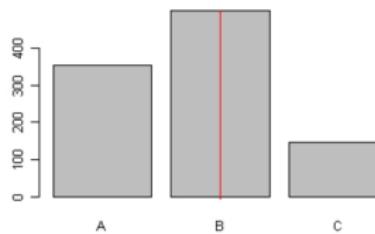
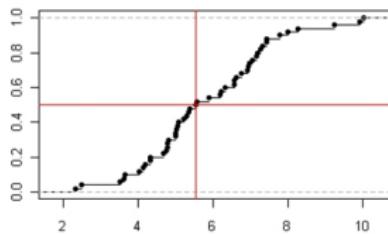
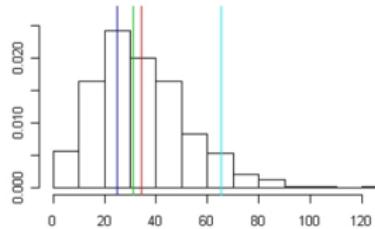
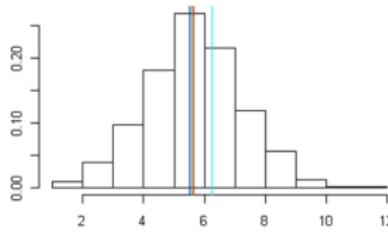
## Beispiel

- 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

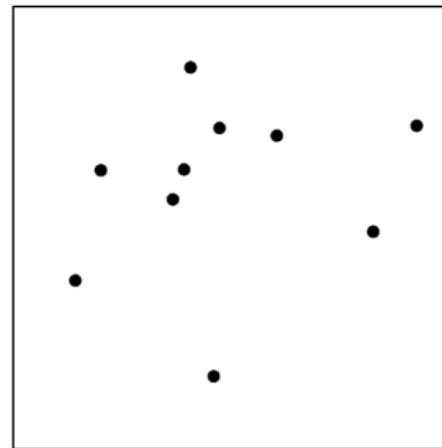
## 3.1 Statistische Kennzahlen für die Lage

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



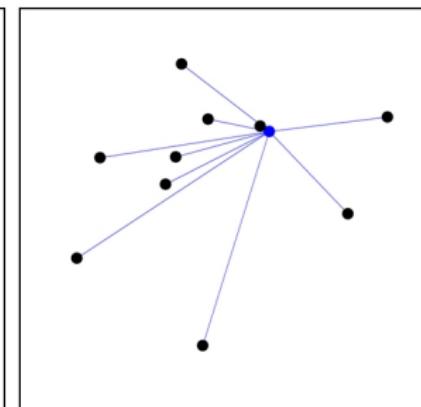
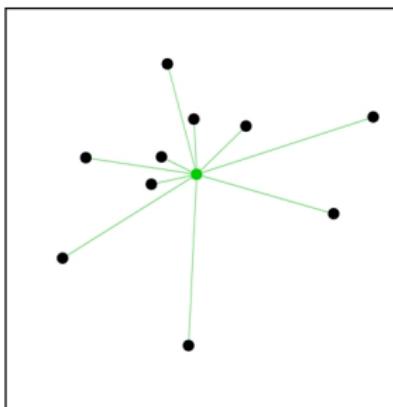
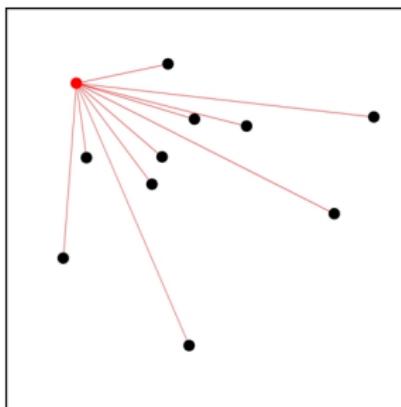
## 3.1 Statistische Kennzahlen für die Lage

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:  $\text{med}_x$

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:  $\text{mod}_x$

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  $\geq 100 \cdot p\%$  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  $x_i$

$k$	$x_{(k)}$
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 <sub>i</sub>
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 <sub>(k)</sub>
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 <sub>i</sub>	R(Version <sub>i</sub> )
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  $\forall x \in \mathbb{R}$ :

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

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

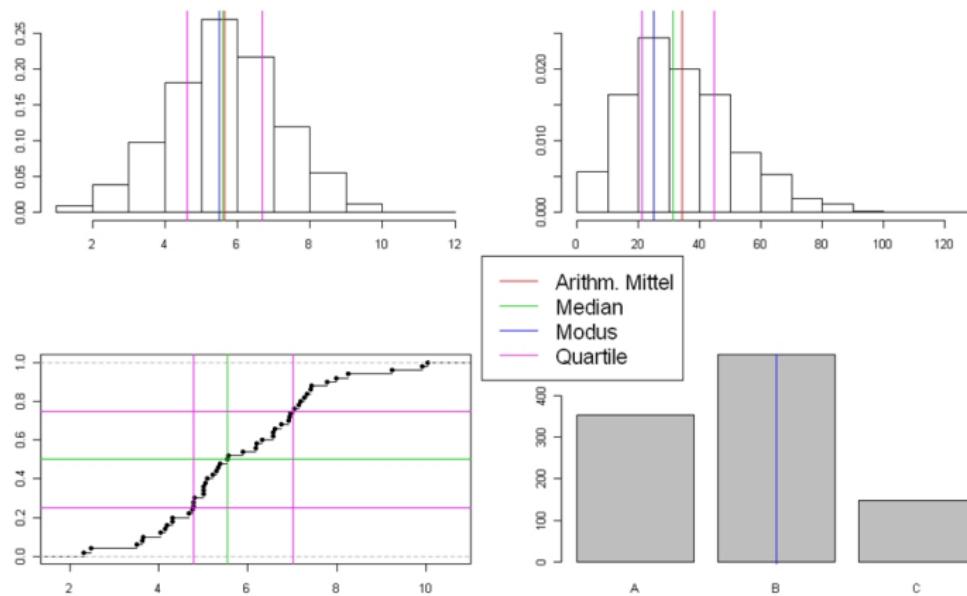
# 3.1 Statistische Kennzahlen für die Lage

Zusammenfassung: Welche Maßzahlen sind bei welchem Skalenniveau geeignet?

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

## 3.2 Statistische Kennzahlen für die Streuung

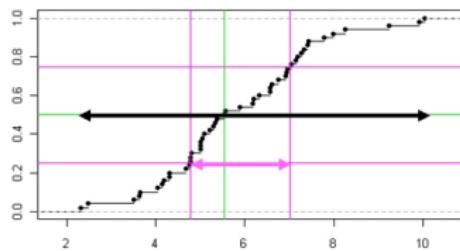
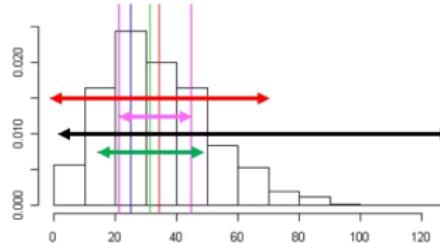
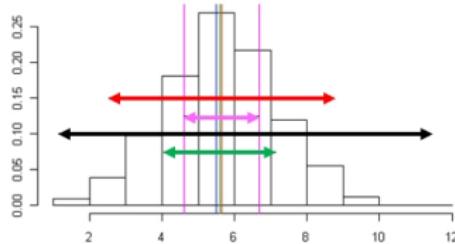
Bisher: Beschreibung von Häufigkeitsverteilung und Lage



## 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)

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

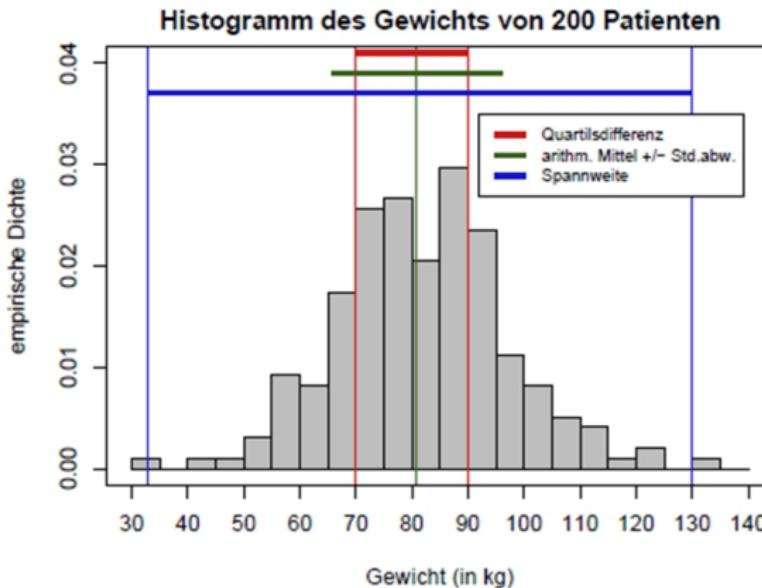
### Spannweite (range)

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

## 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
$\sum$		$N_1$	$N_2$	$N_3$

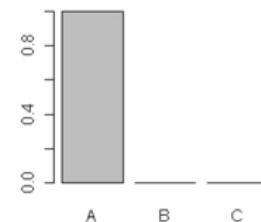
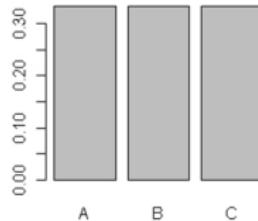
## 3.2 Statistische Kennzahlen für die Streuung

### 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 <sub>i</sub>
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  $f_j = 1/J$ ,  $j = 1, \dots, J$ .

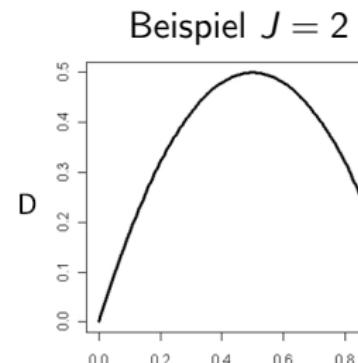
### Simpson's $D$

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

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

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

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



$$f_1 = 1 - f_2$$

## 3.2 Statistische Kennzahlen für die Streuung

### Nominale Daten

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

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

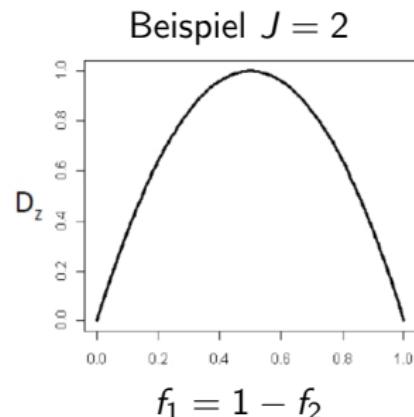
### Simpson's $D_z$ (Normierte Version)

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

$$0 \leq D_z \leq 1$$

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

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



## 3.2 Statistische Kennzahlen für die Streuung

### Nominale Daten

Informationstheorie: Ein Ereignis liefert desto mehr Information, je geringer seine Eintrittswahrscheinlichkeit ist.

Kodierung der Elementarereignisse in Bits, Beispiel **Kaffeebestellung**

1. Bit: 0 = Tasse      1 = Kännchen
2. Bit    0 = Schwarz    1 = mit Milch
3. Bit: 0 = Süßstoff    1 = Zucker

8 mögliche Bestellungen: 000, 001, 010, 011, 100, 101, 110, 111

Beträgt die Wahrscheinlichkeit einer Teilmenge dieser Bestellungen  $p = 1/8$ , wird genau eine Bestellung ausgewählt und man erhält Information über alle  $3 = -\log_2(1/8)$  Bits, falls die Teilmenge ausgewählt wird.

Wird dagegen die Menge möglicher Bestellungen auf 50%, z.B. alle Bestellungen mit Kännchen eingegrenzt, also  $p = 1/4$ , so erhält man Information über  $2 = -\log_2(1/4)$  Bits.

## 3.2 Statistische Kennzahlen für die Streuung

### Nominale Daten

Die Information einer Merkmalsausprägung  $x(j)$  in Bits kann also allgemein definiert werden durch  $-\log_2(f_j)$ .

Der Informationsgehalt des gesamten Merkmals  $x$  ergibt sich durch die **Entropie** genannte erwartete Information  $H(x)$  von  $x$ :

$$H(x) = - \sum_{j=1}^J f_j \log_2(f_j)$$

Beispiel **Kaffeebestellung**: Sei  $x_F$  die Antwort auf eine bestimmte Frage  $F$



$F$  = „Möchten Sie Ihren Kaffee mit Milch und Zucker?“

$$x_F(0) = \text{„Nein“}, x_F(1) = \text{„Ja“}, f_1 = 6/8, f_2 = 2/8,$$

$$H(x_F) = -(6/8) \cdot \log_2(6/8) - (2/8) \cdot \log_2(2/8) = 0.8113$$

## 3.2 Statistische Kennzahlen für die Streuung

### Nominale Daten

**Entropie** von  $x$ :  $H(x) = - \sum_{j=1}^J f_j \log_2(f_j)$  Beispiel **Kaffeebestellung**



$F$  = „Möchten Sie Ihren Kaffee mit Milch und Zucker?“

$$x_F(0) = \text{„Nein“}, x_F(1) = \text{„Ja“}, f_1 = 6/8, f_2 = 2/8, H(x_F) = \boxed{0.8113}$$



$F$  = „Möchten Sie Ihren Kaffee in der Tasse?“

$$x_F(0) = \text{„Nein“}, x_F(1) = \text{„Ja“}, f_1 = 4/8, f_2 = 4/8,$$

$$H(x_F) = -(4/8) \cdot \log_2(4/8) - (4/8) \cdot \log_2(4/8) = \boxed{1}$$

## 3.2 Statistische Kennzahlen für die Streuung

- Entropie gibt also die Information an, die man im Mittel durch Kenntnis der tatsächlichen Ausprägung erhält, wenn man vorher nur die Verteilung kannte. Ist diese hoch, konnte man den Wert vorher schlecht vorhersagen  $\Rightarrow$  hohe Streuung.
- Ist der Informationszugewinn gering, konnte man vorher schon gut prognostizieren.
- Beispiel „Wer wird Millionär“
  - ▶ Kandidat ist sicher = geringe Streuung, keine weitere Information durch Joker
  - ▶ Kandidat ist unsicher = hohe Streuung, erhofft Informationsgewinn durch Publikumsjoker
  - ▶ Ist hier die Streuung hoch, weiterer Informationsgewinn durch Einzelbefragungsjoker

## 3.2 Statistische Kennzahlen für die Streuung

### Nominale Daten

**Entropie** von  $x$ : 
$$H(x) = - \sum_{j=1}^J f_j \cdot \log_2(f_j)$$

Die Entropie ist ein sinnvolles Maß für die Streuung, denn sie erfüllt die Forderungen:

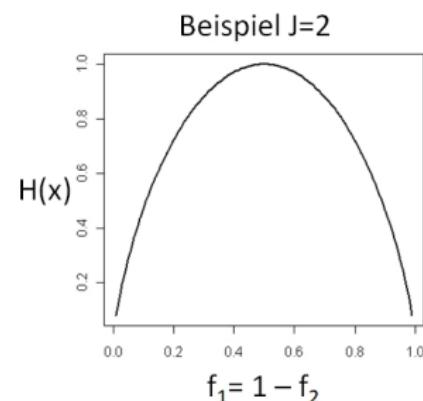
Geringe Streuung, falls es ein  $j$  gibt mit  $f_j = 1$ .

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

$$0 < H(x) \leq \log_2(J)$$

$$\lim[H(x)] = 0 \text{ für } \max[(f_1, \dots, f_J)] \rightarrow 1$$

$$H(x) = \log_2(J) \text{ für } f_1 = \dots = f_J = \frac{1}{J}$$



## 3.2 Statistische Kennzahlen für die Streuung

### Nominale Daten

**Normierte Entropie** von  $x$ :  $H_n(x) = - \sum_{j=1}^J f_j \cdot \frac{\log_2(f_j)}{\log_2(J)}$

Die Entropie ist ein sinnvolles Maß für die Streuung, denn sie erfüllt die Forderungen:

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

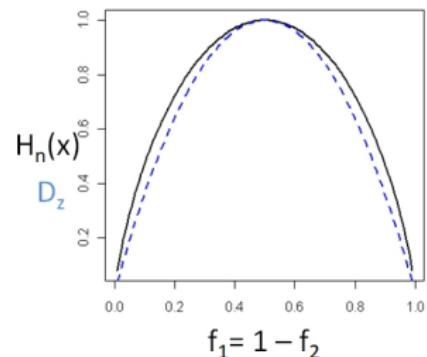
Höchste Streuung, falls  $f_j = 1/J$ ,  $j = 1, \dots, J$ .

$$0 < H_n(x) \leq 1$$

$$\lim[H_n(x)] = 0 \text{ für } \max[(f_1, \dots, f_J)] \rightarrow 1$$

$$H_n(x) = 1 \text{ für } f_1 = \dots = f_J = \frac{1}{J}$$

Beispiel  $J=2$



## 3.2 Statistische Kennzahlen für die Streuung

### Nominale Daten: Beispiel Bearbeitungen von Softwareaufgaben

Merkmal	D <sub>z</sub>	H <sub>n(x)</sub>
Bearbeiter(in)	$4/3 \cdot (1 - 0.17^2 - 0.25^2 - 0.33^2 - 0.25^2)$ = <b>0.9815</b>	$[-0.17 \cdot \log_2(0.17) - 0.25 \cdot \log_2(0.25) - 0.33 \cdot \log_2(0.33) - 0.25 \cdot \log_2(0.25)] / \log_2(4)$ = <b>0.9796</b>
Aufgabe	$3/2 \cdot (1 - 0.17^2 - 0.5^2 - 0.33^2)$ = <b>0.9167</b>	$[-0.17 \cdot \log_2(0.17) - 0.5 \cdot \log_2(0.5) - 0.33 \cdot \log_2(0.33)] / \log_2(3)$ = <b>0.9206</b>
Version	$3/2 \cdot (1 - 0.25^2 - 0.5^2 - 0.25^2)$ = <b>0.9375</b>	$[-0.25 \cdot \log_2(0.25) - 0.5 \cdot \log_2(0.5) - 0.25 \cdot \log_2(0.25)] / \log_2(3)$ = <b>0.9464</b>

Bearbeiter(in)		
Aus-prägung	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		
Aus-prägung	Absolute Häufigkeit	Relative Häufigkeit
1.1	3	0.25
1.2	6	0.5
2.0	3	0.25
	12	1

## 3.2 Statistische Kennzahlen für die Streuung

### Ordinale Daten

$x_1, \dots, x_N$

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

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

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

$i$	$x_i$
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	$x(3)$

Simpson's  $D$  und  $H(x)$  sind anwendbar, allerdings wird die Information der Kategorienordnung nicht genutzt.

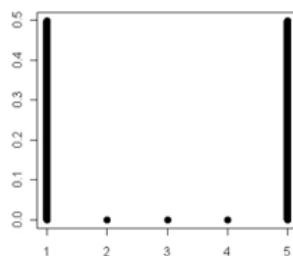
Geordnete Liste  
→

## 3.2 Statistische Kennzahlen für die Streuung

### Ordinale Daten

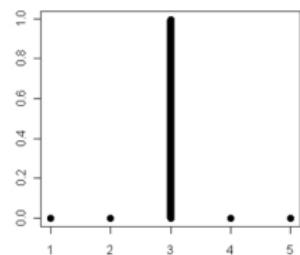
Allgemein: Streuung desto höher, je schlechter konkrete Werte sich vorhersagen lassen.

Werte lassen sich umso besser vorhersagen, je stärker sie sich um den Median verdichten.



Geringste Streuung für  $N(\tilde{x}_{0.5}) = N \rightarrow$

← Höchste Streuung für  $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$



Nicht mehr höchste Streuung bei ausgeglichener Belegung, da die Kategorien unterschiedlich weit von der Mitte entfernt sind. Höchste Streuung bei maximaler Entfernung zur Mitte, also bei gleichmäßiger Konzentration an Minimum und Maximum.

## 3.2 Statistische Kennzahlen für die Streuung

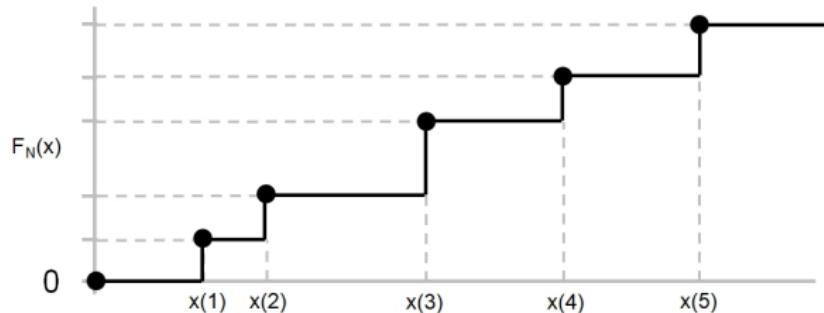
### Ordinale Daten

Geringe Streuung für  $N(\tilde{x}_{0.5}) = N$

Höchste Streuung für  $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$

### Dispersionsindex nach Leti

$$D_L = \sum_{j=1}^{J-1} F_N[x(j)] \cdot (1 - F_N[x(j)])$$



## 3.2 Statistische Kennzahlen für die Streuung

### Ordinale Daten

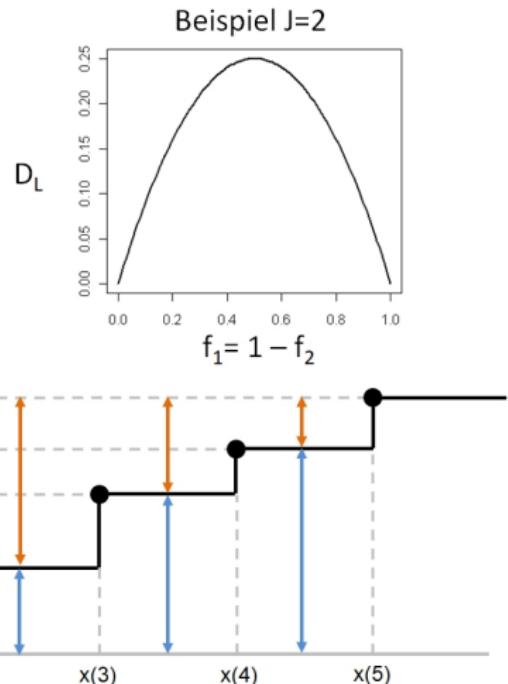
Geringe Streuung für  $N(\tilde{x}_{0.5}) = N$

Höchste Streuung für  $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$

### Dispersionsindex nach Leti

$$D_L = \sum_{j=1}^{J-1} F_N[x(j)] \cdot (1 - F_N[x(j)])$$

$$0 \leq D_L \leq \frac{J-1}{4}$$



## 3.2 Statistische Kennzahlen für die Streuung

### Ordinale Daten

Geringe Streuung für  $N(\tilde{x}_{0.5}) = N$

Höchste Streuung für  $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$

### Normierter Dispersionsindex nach Leti

$$D_{LZ} = \frac{4}{J-1} \sum_{j=1}^{J-1} F_N[x(j)] \cdot (1 - F_N[x(j)])$$

$$0 \leq D_{LZ} \leq 1$$

Für  $J = 2$  gilt  $D_Z = D_{LZ}$ ,  
d.h. normierte Versionen von  
Simpson und Leti sind  
äquivalent.

Beweis:

$$\begin{aligned} D_{LZ} &= \\ &= \frac{4}{2-1} \sum_{j=1}^1 F_N[x(j)](1 - F_N[x(j)]) \\ &= 4 \cdot (f_1(1-f_1)) = 2 \cdot (2f_1 - 2f_1^2) \\ &= 2(1 - f_1^2 - 1 + 2f_1 - f_1^2) \\ &= 2(1 - [f_1^2 + (1 - f_1)^2]) \\ &= 2\left(1 - \sum_{j=1}^2 f_j^2\right) \\ &= \frac{2\left(1 - \sum_{j=1}^2 f_j^2\right)}{2-1} = D_Z \quad \square \end{aligned}$$

## 3.2 Statistische Kennzahlen für die Streuung

### 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.  $W_x = (-\infty, \infty)$

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

Werte lassen sich umso besser vorhersagen, je stärker sie sich um das jeweilige Lagemaß verdichten.

## 3.2 Statistische Kennzahlen für die Streuung

### Quantitative Daten

Werte lassen sich umso besser vorhersagen, je stärker sie sich um das jeweilige Lagemaß verdichten.

Lagemaß: Arithmetisches Mittel

Streuungsmaß:

### **Varianz (mittlere quadratische Abweichung)**

$$s_x^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2 \quad (\text{bzw. } d_x^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2)$$

### **Standardabweichung**

$$s = \sqrt{s_x^2} = \sqrt{\frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2}$$

## 3.2 Statistische Kennzahlen für die Streuung

### Quantitative Daten

Von Streuungsparametern abgeleitete Größen für verhältnisskalierte Merkmale

### Quartilskoeffizient

$$Q_{\text{koeff}} = \frac{2Q}{\tilde{x}_{0.25} + \tilde{x}_{0.75}} = \frac{2(\tilde{x}_{0.75} - \tilde{x}_{0.25})}{\tilde{x}_{0.25} + \tilde{x}_{0.75}} = (\tilde{x}_{0.75} - \tilde{x}_{0.25}) \Bigg/ \left( \frac{\tilde{x}_{0.25} + \tilde{x}_{0.75}}{2} \right)$$

### Variationskoeffizient

$$V_x = \frac{s_x}{\bar{x}}$$

## 3.2 Statistische Kennzahlen für die Streuung

### Quantitative Daten: Berechnung der Varianz aus Häufigkeitsverteilung

$$s_x^2 = \frac{N}{N-1} \sum_{j=1}^J f_j \cdot [x(j) - \sum_{k=1}^J f_k \cdot x(k)]^2 = \frac{N}{N-1} \sum_{j=1}^J f_j \cdot (x(j) - \bar{x})^2$$

Beweis:

$$\begin{aligned} s_x^2 &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N-1} \sum_{i=1}^N (x_{(i)} - \bar{x})^2 = \frac{(x_{(1)} - \bar{x})^2}{N-1} + \dots + \frac{(x_{(N)} - \bar{x})^2}{N-1} \\ &= \underbrace{\frac{(x(1) - \bar{x})^2}{N-1} + \dots + \frac{(x(1) - \bar{x})^2}{N-1}}_{f_1 \cdot N\text{-mal}} + \dots + \underbrace{\frac{(x(J) - \bar{x})^2}{N-1} + \dots + \frac{(x(J) - \bar{x})^2}{N-1}}_{f_J \cdot N\text{-mal}} \\ &= \frac{N}{N-1} f_1 \cdot (x(1) - \bar{x})^2 + \dots + \frac{N}{N-1} f_J \cdot (x(J) - \bar{x})^2 = \frac{N}{N-1} \sum_{j=1}^J f_j \cdot (x(j) - \bar{x})^2 \quad \square \end{aligned}$$

## 3.2 Statistische Kennzahlen für die Streuung

### Quantitative Daten: Varianz von Lineartransformationen

$$y = ax + b \Rightarrow s_y^2 = a^2 s_x^2$$

Beweis:

$$\begin{aligned} s_y^2 &= \frac{1}{N-1} \sum_{n=1}^N (y_n - \bar{y})^2 = \frac{1}{N-1} \sum_{n=1}^N [ax_n + b - (a\bar{x} + b)]^2 \\ &= \frac{1}{N-1} \sum_{n=1}^N (ax_n - a\bar{x})^2 \\ &= a^2 \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2 = \boxed{a^2 s_x^2} \quad \square \end{aligned}$$

$$\bar{y} = \overline{ax + b} = \frac{1}{N} \sum_{n=1}^N (ax_n + b) = a \frac{1}{N} \sum_{n=1}^N x_n + \frac{bN}{N} = a\bar{x} + b$$

## 3.2 Statistische Kennzahlen für die Streuung

Quantitative Daten: **Verschiebungssatz von Steiner**

$$d_x^2 = \left( \frac{1}{N} \sum_{n=1}^N (x_n - b)^2 \right) \quad \text{speziell für } b = 0 : d_x^2 = \bar{x^2} - \bar{x}^2$$

Beweis:

$$\begin{aligned} d_x^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2 = \frac{1}{N} \sum_{n=1}^N [(x_n - b) + (b - \bar{x})]^2 \\ &= \frac{1}{N} \sum_{n=1}^N [(x_n - b)^2 + 2(x_n - b)(b - \bar{x}) + (b - \bar{x})^2] \\ &= \frac{1}{N} \sum_{n=1}^N (x_n - b)^2 + 2(b - \bar{x}) \frac{1}{N} \sum_{n=1}^N (x_n - b) + \frac{1}{N} \sum_{n=1}^N (b - \bar{x})^2 \\ &= \frac{1}{N} \sum_{n=1}^N (x_n - b)^2 - 2(\bar{x} - b)^2 + (\bar{x} - b)^2 = \boxed{\frac{1}{N} \sum_{n=1}^N (x_n - b)^2 - (\bar{x} - b)^2} \quad \square \end{aligned}$$

## 3.2 Statistische Kennzahlen für die Streuung

### Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$$s_{x_4}^2 = \frac{2 \cdot (10 - 13.5)^2 + 1 \cdot (11 - 13.5)^2 + 2 \cdot (12 - 13.5)^2}{11} + \frac{1 \cdot (13 - 13.5)^2 + 2 \cdot (14 - 13.5)^2 + 1 \cdot (15 - 13.5)^2}{11} + \frac{1 \cdot (16 - 13.5)^2 + 1 \cdot (17 - 13.5)^2 + 1 \cdot (18 - 13.5)^2}{11} = 7$$

$$V_4 = \frac{\sqrt{7}}{13.5} = 0.196$$

$$\bar{x}_4 = 13.5$$

k	Anzahl Clicks <sub>(k)</sub>
1	10
2	10
3	11
4	12
5	12
6	13
7	14
8	14
9	15
10	16
11	17
12	18

$$Q_{\text{koeff};4} = \frac{2 \cdot 4}{11.5 + 15.5} = 0.296$$

$$\tilde{x}_{4;0.25} = 11.5$$

$$Q_4 = 4$$

$$R_4 = 8$$

$$\tilde{x}_{4;0.75} = 15.5$$

## 3.2 Statistische Kennzahlen für die Streuung

### Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$$\begin{aligned}s_{x_5}^2 &= \frac{(3.2 - 5.075)^2 + (3.6 - 5.075)^2 + 2 \cdot (3.7 - 5.075)^2}{11} \\&+ \frac{(3.9 - 5.075)^2 + (4.2 - 5.075)^2 + (4.5 - 5.075)^2}{11} \\&+ \frac{(4.9 - 5.075)^2 + (6.1 - 5.075)^2 + (6.6 - 5.075)^2}{11} \\&+ \frac{(8.0 - 5.075)^2 + (8.5 - 5.075)^2}{11} = 3.24\end{aligned}$$

$$\bar{x}_5 = 5.075$$

$$V_5 = \frac{\sqrt{3.24}}{5.075} = 0.355$$

k	Bearbeitungszeit <sub>(k)</sub>
1	3.2
2	3.6
3	3.7
4	3.7
5	3.9
6	4.2
7	4.5
8	4.9
9	6.1
10	6.6
11	8.0
12	8.5

$$Q_{\text{koeff};5} = \frac{2 \cdot 2.65}{3.7 + 6.35} = 0.527$$

$$\tilde{x}_{5;0.25} = 3.7$$

$$Q_5 = 2.65$$

$$R_5 = 5.3$$

$$\tilde{x}_{5;0.75} = 6.35$$

## 3.2 Statistische Kennzahlen für die Streuung

Zusammenfassung: Welche Maßzahlen sind bei welchem Skalenniveau geeignet?

Skalenniveau → ↓ Streuungsmaß	Nominal	Ordinal	Quantitativ
Simpson's D/ Entropie		– Informationsverlust	– Nur für klassierte Daten
Leti's D	– Nur für J = 2		– Nur für klassierte Daten
MAD/ Spannweite/ Quartilsdifferenz		– Geringe Aussagekraft für kleine J	+ Robust – Informationsverlust – Hohe Streubreite
Varianz/ Standardabweichung Variationskoeffizient	– Nur für J = 2		– Ausreißeranfälligkeit + Informationsnutzung + Geringe Streubreite

# Bivariate Daten

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

Bisher: Betrachtung einzelner Merkmale  $X$

Jetzt: Betrachtung von Merkmalspaaren ( $X, Y$ )

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

X=Aufgabe

Y=Anzahl Clicks

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

## Nominale Daten

Univariante Urlisten:

$$x_1, \dots, x_N$$

$$y_1, \dots, y_N$$

Univariante Wertebereiche:

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

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

$$\begin{aligned} W_Y &= \{Y(k) | k = 1, \dots, K\} = \\ &\{y(1), \dots, y(K)\} \end{aligned}$$

i	x <sub>i</sub>	y <sub>i</sub>
1	A	D
2	C	E
...	...	
N	B	E

Bivariate Urliste:

$$(x_1, y_1), \dots, (x_N, y_N)$$

Bivariater Wertebereich:

$$\begin{aligned} (x_i, y_i) &\in W_{XY} = W_X \times W_Y = \\ &\{(x[1], y[1]), \dots, (x[1], y[K]), (x[2], y[1]), \\ &\dots, (x[J], y[K])\} \end{aligned}$$

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

## Nominale Daten

$$x_1, \dots, x_N; y_1, \dots, y_N \\ x_i \in W_X, y_i \in W_Y$$

$$(x_1, y_1), \dots, (x_N, y_N) \\ (x_i, y_i) \in W_{XY} = W_X \times W_Y$$

$$d_i(j) = I_{x(e_i)=x(j)} \\ r_i(k) = I_{y(e_i)=y(k)}$$

i	x <sub>i</sub>	y <sub>i</sub>
1	A	D
2	C	E
...	...	
N	B	E

Dummykodierung →

i	x <sub>i</sub>	y <sub>i</sub>	d <sub>i</sub> (1)	d <sub>i</sub> (2)	d <sub>i</sub> (3)	r <sub>i</sub> (1)	r <sub>i</sub> (2)
1	A	D	1	0	0	1	0
2	C	E	0	0	1	0	1
...	...	...	...	...	...	...	...
N	B	E	0	1	0	0	1
Σ			N <sub>1•</sub>	N <sub>2•</sub>	N <sub>3•</sub>	N <sub>•1</sub>	N <sub>•2</sub>

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

### Nominale Daten

Häufigkeitsverteilung eines bivariaten Merkmals

$$(x_i, y_i) \in W_{XY} = W_X \times W_Y, i = 1, \dots, N$$

$$W_{XY} = \{(x[j], y[k]) | j = 1, \dots, J; k = 1, \dots, K\}$$

$$= \left\{ \begin{array}{ccc} (x[1], y[1]) & \dots & (x[1], y[K]) \\ (x[2], y[1]) & \dots & (x[2], y[K]) \\ \vdots & \ddots & \vdots \\ (x[J], y[1]) & \dots & (x[J], y[K]) \end{array} \right\}$$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

### Nominale Daten

#### Gemeinsame absolute Häufigkeitsverteilung von $x$ und $y$

$$N_{jk} = N((x[j], y[k])), \quad j = 1, \dots, J; k = 1, \dots, K$$

$$\begin{pmatrix} N_{11} & \dots & N_{1K} \\ N_{21} & \dots & N_{2K} \\ \vdots & \ddots & \vdots \\ N_{J1} & \dots & N_{JK} \end{pmatrix}$$

#### Gemeinsame relative Häufigkeitsverteilung von $x$ und $y$

$$f_{jk} = \frac{N_{jk}}{N}, \quad j = 1, \dots, J; k = 1, \dots, K$$

$$\begin{pmatrix} f_{11} & \dots & f_{1k} \\ f_{21} & \dots & f_{2k} \\ \vdots & \ddots & \vdots \\ f_{j1} & \dots & f_{jk} \end{pmatrix}$$

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Häufigkeitsverteilung eines bivariaten Merkmals

i	x <sub>i</sub>	y <sub>i</sub>	d <sub>i</sub> (1)	d <sub>i</sub> (2)	d <sub>i</sub> (3)	r <sub>i</sub> (1)	r <sub>i</sub> (2)
1	A	D	1	0	0	1	0
2	C	E	0	0	1	0	1
...	...	...	...	...	...	...	...
N	B	E	0	1	0	0	1
$\Sigma$			$N_{1\bullet}$	$N_{2\bullet}$	$N_{3\bullet}$	$N_{\bullet 1}$	$N_{\bullet 2}$

$$\begin{aligned}
 N_{jk} &= N((x[j], y[k])) \\
 &= \sum_{i \in \{I | r_i(k) = 1\}} d_i(j) = \sum_{i \in \{I | d_i(j) = 1\}} r_i(k) \\
 &= \sum_{i=1}^N d_i(j) \cdot r_i(k)
 \end{aligned}$$

$$\begin{aligned}
 N_{j\bullet} &= \sum_{i=1}^N d_i(j) = \sum_{i \in \{I | r_i(1) = 1\}} d_i(j) + \sum_{i \in \{I | r_i(2) = 1\}} d_i(j) + \dots + \sum_{i \in \{I | r_i(K) = 1\}} d_i(j) \\
 &= \sum_{k=1}^K \sum_{l=1}^N d_l(j) \cdot r_l(k) = \sum_{k=1}^K N_{jk}
 \end{aligned}$$

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

## Kontingenztafel

Absolute Häufigkeiten

		Y				
		y(1)	y(2)	...	y(K)	$\Sigma$
X	x(1)	$N_{11}$	$N_{12}$	...	$N_{1K}$	$N_{1\cdot}$
	x(2)	$N_{21}$	$N_{22}$	...	$N_{2K}$	$N_{2\cdot}$
	...	...	...	...	...	...
	x(J)	$N_{J1}$	$N_{J2}$	...	$N_{JK}$	$N_{J\cdot}$
$\Sigma$		$N_{\cdot 1}$	$N_{\cdot 2}$	...	$N_{\cdot K}$	N

$$N_{j\cdot} = \sum_{k=1}^K N_{jk}$$

$$N_{\cdot k} = \sum_{j=1}^J N_{jk}$$

$$N = \sum_{j=1}^J \sum_{k=1}^K N_{jk}$$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

**Kontingenztafel**

**Gemeinsame absolute Häufigkeitsverteilung von X und Y**

		Y				
		y(1)	y(2)	...	y(K)	$\Sigma$
X	x(1)	$N_{11}$	$N_{12}$	...	$N_{1K}$	$N_{1\cdot}$
	x(2)	$N_{21}$	$N_{22}$	...	$N_{2K}$	$N_{2\cdot}$
	...	...	...	...	...	...
	x(J)	$N_{J1}$	$N_{J2}$	...	$N_{JK}$	$N_{J\cdot}$
$\Sigma$		$N_{\cdot 1}$	$N_{\cdot 2}$	...	$N_{\cdot K}$	N

$$N_{j\cdot} = \sum_{k=1}^K N_{jk}$$

$$N_{\cdot k} = \sum_{j=1}^J N_{jk}$$

$$N = \sum_{j=1}^J \sum_{k=1}^K N_{jk}$$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

### Kontingenztafel

		Y				
		y(1)	y(2)	...	y(K)	$\Sigma$
X	x(1)	$N_{11}$	$N_{12}$	...	$N_{1K}$	$N_{1\cdot}$
	x(2)	$N_{21}$	$N_{22}$	...	$N_{2K}$	$N_{2\cdot}$
	...	...	...	...	...	...
	x(J)	$N_{J1}$	$N_{J2}$	...	$N_{JK}$	$N_{J\cdot}$
$\Sigma$		$N_{\cdot 1}$	$N_{\cdot 2}$	...	$N_{\cdot K}$	N

**Absolute Randhäufigkeitsverteilung von X**

**Absolute Randhäufigkeitsverteilung von Y**

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

### Kontingenztafel

		Y				$\Sigma$
		y(1)	y(2)	...	y(K)	
X	x(1)	$N_{11}/N$	$N_{12}/N$	...	$N_{1K}/N$	$N_{1\cdot}/N$
	x(2)	$N_{21}/N$	$N_{22}/N$	...	$N_{2K}/N$	$N_{2\cdot}/N$
	...	...	...	...	...	...
	x(J)	$N_{J1}/N$	$N_{J2}/N$	...	$N_{JK}/N$	$N_{J\cdot}/N$
$\Sigma$	$N_{\cdot 1}/N$	$N_{\cdot 2}/N$	...	$N_{\cdot K}/N$	$N/N$	

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

## Kontingenztafel

### Relative Häufigkeiten

		Y				
		y(1)	y(2)	...	y(K)	$\Sigma$
X	x(1)	$f_{11}$	$f_{12}$	...	$f_{1K}$	$f_{1\cdot}$
	x(2)	$f_{21}$	$f_{22}$	...	$f_{2K}$	$f_{2\cdot}$
	...	...	...	...	...	...
	x(J)	$f_{J1}$	$f_{J2}$	...	$f_{JK}$	$f_{J\cdot}$
$\Sigma$		$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot K}$	1

$$f_{\cdot k} = \sum_{j=1}^J f_{jk}$$

$$f_{j\cdot} = \sum_{k=1}^K f_{jk}$$

$$1 = \sum_{j=1}^J \sum_{k=1}^K f_{jk}$$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

**Kontingenztafel**

**Gemeinsame relative Häufigkeitsverteilung  $f_{XY}$  von  $X$  und  $Y$**

		Y				$\Sigma$
		y(1)	y(2)	...	y(K)	
X	x(1)	$f_{11}$	$f_{12}$	...	$f_{1K}$	$f_{1\cdot}$
	x(2)	$f_{21}$	$f_{22}$	...	$f_{2K}$	$f_{2\cdot}$
	...	...	...	...	...	...
	x(J)	$f_{J1}$	$f_{J2}$	...	$f_{JK}$	$f_{J\cdot}$
	$\Sigma$	$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot K}$	1

$$f_{XY} = \{f_{jk} \mid j=1, \dots, J; k=1, \dots, K\}$$

$$f_{j\cdot} = \sum_{k=1}^K f_{jk}$$

$$f_{\cdot k} = \sum_{j=1}^J f_{jk}$$

$$1 = \sum_{j=1}^J \sum_{k=1}^K f_{jk}$$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

### Kontingenztafel

		Y					$f_{x\bullet} = \{f_{j\bullet} \mid j = 1, \dots, J\}$
		y(1)	y(2)	...	y(K)	$\Sigma$	
X	x(1)	$f_{11}$	$f_{12}$	...	$f_{1K}$	$f_{1\cdot}$	Relative Randhäufigkeitsverteilung $f_{x\bullet}$ von X
	x(2)	$f_{21}$	$f_{22}$	...	$f_{2K}$	$f_{2\cdot}$	
	...	...	...	...	...	...	
	x(J)	$f_{J1}$	$f_{J2}$	...	$f_{JK}$	$f_{J\cdot}$	
$\Sigma$		$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot K}$	1	
Relative Randhäufigkeitsverteilung $f_{\cdot Y}$ von Y							$f_{\cdot Y} = \{f_{\cdot k} \mid k = 1, \dots, K\}$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

### Kontingenztafel

Wie lautet die Verteilung von  $Y$  im Teildatensatz, für den  $X = x(2)$  gilt?

		Y				
		y(1)	y(2)	...	y(K)	$\Sigma$
X	x(1)					
	x(2)	$N_{21}$	$N_{22}$	...	$N_{2K}$	$N_{2\cdot}$
...						
x(J)						

Dieser Datensatz hat Umfang  $N_2$ .  
Absolute Häufigkeitsverteilung:

$$N_{y;k|2} = N_{2k}, \quad k=1, \dots, K$$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

### Kontingenztafel

Wie lautet die Verteilung von  $Y$  im Teildatensatz, für den  $X = x(2)$  gilt?

		Y				
		y(1)	y(2)	...	y(K)	$\Sigma$
X	x(1)					
	x(2)	$N_{21}/N_{2\bullet}$	$N_{22}/N_{2\bullet}$	...	$N_{2K}/N_{2\bullet}$	$N_{2\bullet}/N_{2\bullet}$
	...					
	x(J)					

Dieser Datensatz hat Umfang  $N_2$ .

Relative Häufigkeitsverteilung:

$$f_{y;k|2} = N_{y;2k}/N_{2\bullet} = f_{2k}/f_{2\bullet}, \quad k=1, \dots, K$$

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

### Kontingenztafel

Bedingte Verteilung von  $Y$  gegeben  $X$ , d.h. von  $Y|X = k$

		Y				
		y(1)	y(2)	...	y(K)	$\Sigma$
X	x(1)	$f_{11}/f_{1\bullet}$	$f_{12}/f_{1\bullet}$	...	$f_{1K}/f_{1\bullet}$	1
	x(2)	$f_{21}/f_{2\bullet}$	$f_{22}/f_{2\bullet}$	...	$f_{2K}/f_{2\bullet}$	1
	...	...	...	...	...	...
	x(J)	$f_{J1}/f_{J\bullet}$	$f_{J2}/f_{J\bullet}$	...	$f_{JK}/f_{J\bullet}$	1
	$\Sigma$					J

**Bedingte Verteilung**  
von  $Y$  gegeben  $X=x(2)$

$$f_{y;k|2} = f_{2k}/f_{2\bullet}$$

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Darstellung einer bivariaten Häufigkeitsverteilung

## Kontingenztafel

		Y				$\Sigma$
		y(1)	y(2)	...	y(K)	
X	x(1)	$f_{11}/f_{1\bullet}$	$f_{12}/f_{1\bullet}$	...	$f_{1K}/f_{1\bullet}$	1
	x(2)	$f_{21}/f_{2\bullet}$	$f_{22}/f_{2\bullet}$	...	$f_{2K}/f_{2\bullet}$	1
	...	...	...	...	...	...
	x(J)	$f_{J1}/f_{J\bullet}$	$f_{J2}/f_{J\bullet}$	...	$f_{JK}/f_{J\bullet}$	1
$\Sigma$						J

**Bedingte Verteilung**  
 $f_{Y|X}$  von Y gegeben X

$$f_{y;klj} = f_{jk}/f_{j\bullet}$$

$$f_{Y|X} = \{f_{y;klj} \mid j=1, \dots, J; k=1, \dots, K\}$$

**Bedingte Verteilung**  
 $f_{X|Y}$  von X gegeben Y

$$f_{x;jlk} = f_{jk}/f_{\bullet k}$$

$$f_{X|Y} = \{f_{x;jlk} \mid j=1, \dots, J; k=1, \dots, K\}$$

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

Be-arbeiter(in)	Aufgabe
Kai	Export
Kai	Verknüpfung
Miriam	Export
Tina	Verknüpfung
Oliver	Export
Tina	Export
Tina	Verknüpfung
Miriam	Export
Miriam	Export
Oliver	Abfrage
Oliver	Verknüpfung
Oliver	Abfrage

Absolute Häufigkeiten

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bear-bei-ter(in)	Kai	0	1	1	2
	Miriam	0	3	0	3
	Oliver	2	1	1	4
	Tina	0	1	2	3
	$\Sigma$	2	6	4	12

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

Be-arbeiter(in)	Aufgabe
Kai	Export
Kai	Verknüpfung
Miriam	Export
Tina	Verknüpfung
Oliver	Export
Tina	Export
Tina	Verknüpfung
Miriam	Export
Miriam	Export
Oliver	Abfrage
Oliver	Verknüpfung
Oliver	Abfrage

Relative Häufigkeiten

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bear-bei-ter(in)	Kai	0	1/12	1/12	2/12
	Miriam	0	3/12	0	3/12
	Oliver	2/12	1/12	1/12	4/12
	Tina	0	1/12	2/12	3/12
	$\Sigma$	2/12	6/12	4/12	1

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeiter(in)	Aufgabe
Kai	Export
Kai	Verknüpfung
Miriam	Export
Tina	Verknüpfung
Oliver	Export
Tina	Export
Tina	Verknüpfung
Miriam	Export
Miriam	Export
Oliver	Abfrage
Oliver	Verknüpfung
Oliver	Abfrage

Relative Häufigkeiten Aufgabe bedingt auf Bearbeiter(in)

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbeiter(in)	Kai	0	1/2	1/2	1
	Miriam	0	1	0	1
	Oliver	2/4	1/4	1/4	1
	Tina	0	1/3	2/3	1
	$\Sigma$				4

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeiter(in)	Aufgabe
Kai	Export
Kai	Verknüpfung
Miriam	Export
Tina	Verknüpfung
Oliver	Export
Tina	Export
Tina	Verknüpfung
Miriam	Export
Miriam	Export
Oliver	Abfrage
Oliver	Verknüpfung
Oliver	Abfrage

Relative Häufigkeiten Bearbeiter(in) bedingt auf Aufgabe

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbeiter(in)	Kai	0	1/6	1/4	
	Miriam	0	1/2	0	
	Oliver	1	1/6	1/4	
	Tina	0	1/6	1/2	
$\Sigma$		1	1	1	3

# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Beispiel in R:

Überlebende der Titanic

Der **Mosaikplot**

Code in R:

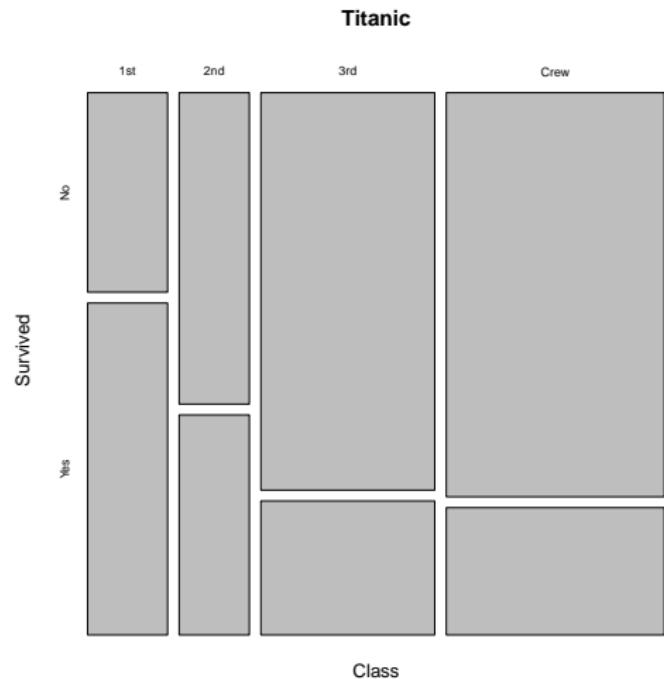
```
mosaicplot(~ Class + Survived,  
          data = Titanic)
```

Rechteckbreiten entsprechen  $f_{\bullet c}$

Rechteckhöhen entsprechen  $f_{s|c}$

Rechteckflächen entsprechen

$$f_{sc} = f_{s|c} \cdot f_{\bullet c}$$



# 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Nominale Daten:** Beispiel in R:

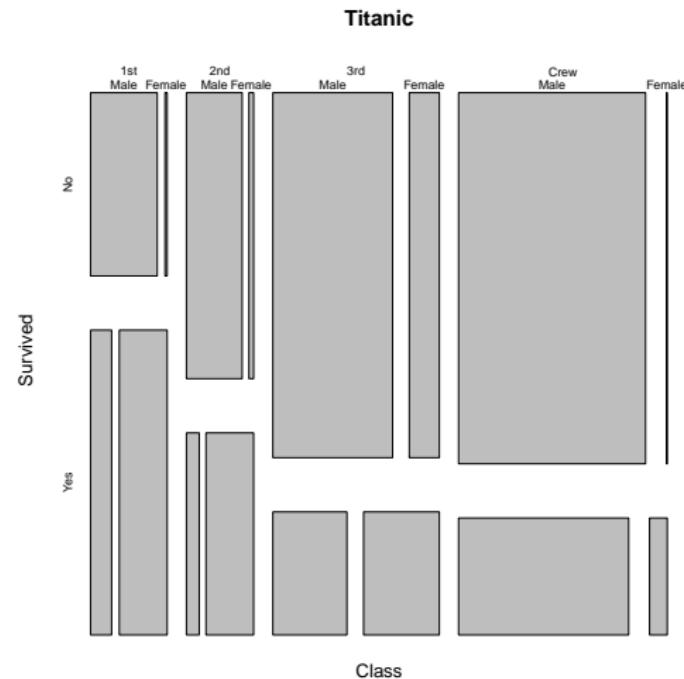
Überlebende der Titanic

Der **Mosaikplot**

Code in R:

```
mosaicplot(~ Class + Survived +  
Sex, data = Titanic)
```

Zusätzliche Einteilung der Flächen nach Geschlecht



## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

### Ordinale Daten:

- Kontingenztafeln und Mosaikplots mit geordneten Kategorien

### Quantitative Daten:

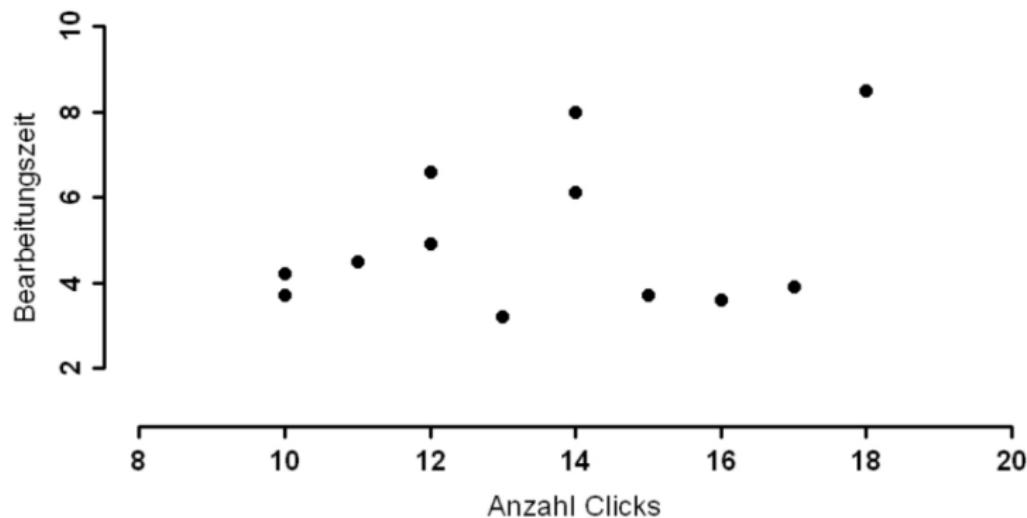
- Kontingenztafeln und Mosaikplots mit klassierten Daten
- Streudiagramme!

## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Quantitative Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

### Streudiagramm

Darstellung der Punktpaare  $(x_i, y_i)$  in einem kartesischen Koordinatensystem

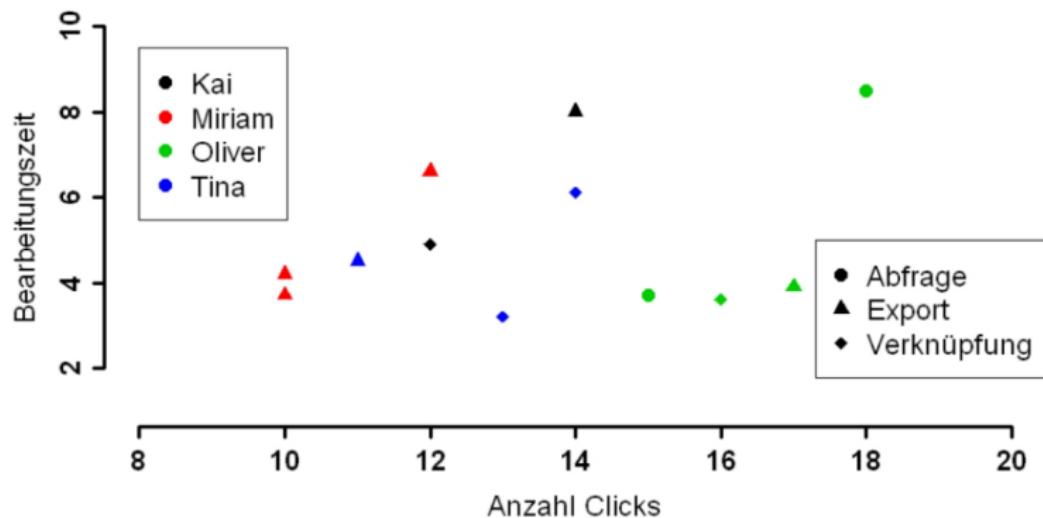


## 4.1 Bivariate Daten: Tabellarische und grafische Darstellungen

**Quantitative Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

### Streudiagramm

Darstellung der Punktepaare  $(x_i, y_i)$  in einem kartesischen Koordinatensystem



## 4.2 Bivariate Daten: Zusammenhangsmaße

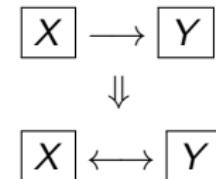
- Erinnerung: Allgemeine Eigenschaft der Streuung univariater Daten: Streuung von  $X$  desto höher, je schlechter sich konkrete Werte vorhersagen lassen.
  - ▶ Bisher: Vorhersage der Werte von  $X$  durch einzelne Lageparameter.
  - ▶ Jetzt: **Vorhersage der Werte von  $Y$  unter Verwendung der Werte von  $X$ .**
- Allgemein: Zusammenhang ( = **Korrelation**) zwischen  $X$  und  $Y$  desto größer, je besser sich der Wert von  $Y$  unter Kenntnis des Werts von  $X$  vorhersagen lässt (oder umgekehrt).
- **Wichtige Unterscheidung**
  - ▶ **Korrelation** bedeutet nicht notwendig **Kausalität** (Beziehung zwischen *Ursache* und *Wirkung* oder *Aktion* und *Reaktion*)

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Korrelation und Kausalität

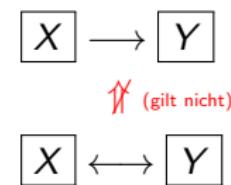
Es gilt:

$X$  ist Ursache von  $Y \Rightarrow X$  und  $Y$  korrelieren



Aber:

$X$  und  $Y$  korrelieren  $\not\Rightarrow X$  ist Ursache von  $Y$



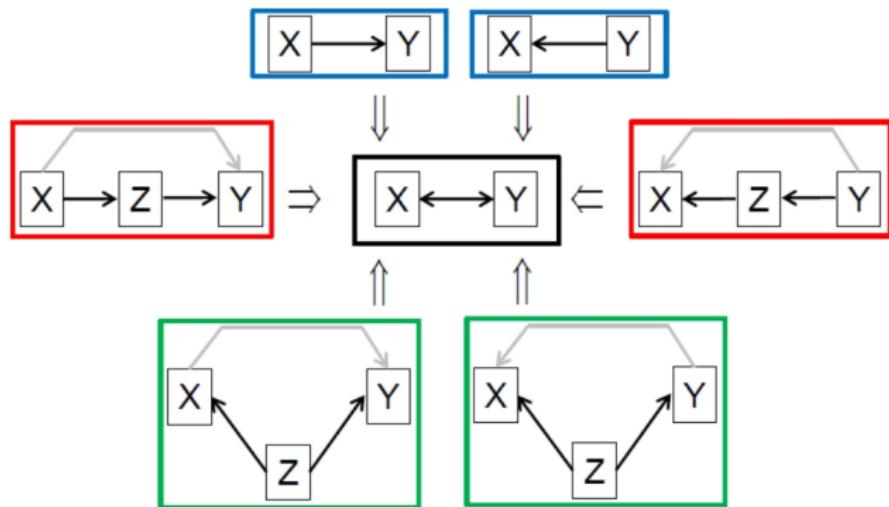
## 4.2 Bivariate Daten: Zusammenhangsmaße

### Korrelation und Kausalität

$X$  ist Ursache von  $Y \Rightarrow X$  und  $Y$  korrelieren

$X$  und  $Y$  korrelieren  $\not\Rightarrow X$  ist Ursache von  $Y$

Verschiedene  
Korrelationsquellen  
möglich



## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten

Zusammenhang (=Korrelation) zwischen  $X$  und  $Y$  desto größer, je besser sich der Wert von  $Y$  unter Kenntnis des Werts von  $X$  vorhersagen lässt (oder umgekehrt).

	$Y$					
	$y(1)$	$y(2)$	$\dots$	$y(K)$	$\Sigma$	
$X$	$x(1)$	$f_{y;1 1}$	$f_{y;2 1}$	$\dots$	$f_{y;K 1}$	1
	$x(2)$	$f_{y;1 2}$	$f_{y;2 2}$	$\dots$	$f_{y;K 2}$	1
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$x(J)$	$f_{y;1 J}$	$f_{y;2 J}$	$\dots$	$f_{y;K J}$	1
	$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot K}$		

Wert von  $Y$  lässt sich bei Kenntnis von  $X$  umso besser vorhersagen, je stärker die bedingte Verteilung  $f_{Y|X}$  von  $Y$  gegeben  $X$  von der Randverteilung  $f_{\cdot Y}$  von  $Y$  abweicht.

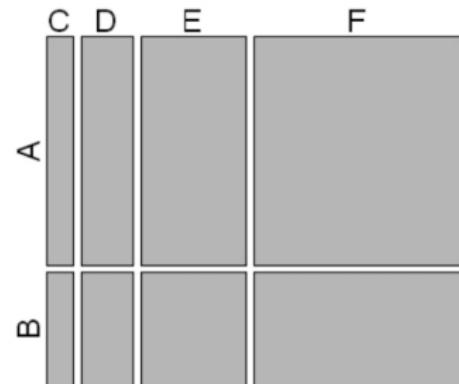
## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten

Wert von  $Y$  lässt sich bei Kenntnis von  $X$  umso besser vorhersagen, je stärker die bedingte Verteilung  $f_{Y|X}$  von  $Y$  gegeben  $X$  von der Randverteilung  $f_{\bullet Y}$  von  $Y$  abweicht.

	$Y$				
	$y(1)$	$y(2)$	$\dots$	$y(K)$	$\Sigma$
$x(1)$	$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet K}$	1
$x(2)$	$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet K}$	1
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x(J)$	$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet K}$	1
	$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet K}$	

Zusammenhang minimal, falls  
 $f_{y;k|j} = f_{\bullet j}$  für alle  $j \in \{1, \dots, J\}$   
und  $k \in \{1, \dots, K\}$



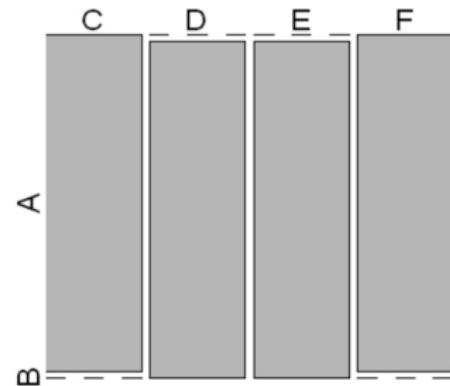
## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten

Wert von  $Y$  lässt sich bei Kenntnis von  $X$  umso besser vorhersagen, je stärker die bedingte Verteilung  $f_{Y|X}$  von  $Y$  gegeben  $X$  von der Randverteilung  $f_{\bullet Y}$  von  $Y$  abweicht.

	$Y$				
	$y(1)$	$y(2)$	$\dots$	$y(K)$	$\Sigma$
$x(1)$	0	1	$\dots$	0	1
$x(2)$	0	0	$\dots$	1	1
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x(J)$	1	0	$\dots$	0	1
	$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet K}$	

Zusammenhang maximal, falls es für alle  $j \in \{1, \dots, J\}$  ein  $k \in \{1, \dots, K\}$  mit  $f_{y;k|j} = 1$  gibt



## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten

Wert von  $Y$  lässt sich bei Kenntnis von  $X$  umso besser vorhersagen, je stärker die bedingte Verteilung  $f_{Y|X}$  von  $Y$  gegeben  $X$  von der Randverteilung  $f_{\bullet Y}$  von  $Y$  abweicht.

	$Y$					
	$y(1)$	$y(2)$	$\dots$	$y(K)$	$\Sigma$	
$X$	$x(1)$	$f_{y;1 1}$	$f_{y;2 1}$	$\dots$	$f_{y;K 1}$	1
	$x(2)$	$f_{y;1 2}$	$f_{y;2 2}$	$\dots$	$f_{y;K 2}$	1
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$x(J)$	$f_{y;1 J}$	$f_{y;2 J}$	$\dots$	$f_{y;K J}$	1
		$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet K}$	

Ein Maß, dass desto größer wird, je größer die Abweichung der bedingten Verteilung  $f_{Y|X}$  von der Randverteilung  $f_{\bullet Y}$  ist, ist also ein sinnvolles Zusammenhangsmaß.

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten

Ein Maß, dass desto größer wird, je größer die Abweichung der bedingten Verteilung  $f_{Y|X}$  von der Randverteilung  $Y_{\bullet Y}$  ist, ist also ein sinnvolles Zusammenhangsmaß.

	Y				
	y(1)	y(2)	...	y(K)	$\Sigma$
x(1)	$f_{0;11}$	$f_{0;12}$	...	$f_{0;1K}$	$f_{1\bullet}$
x(2)	$f_{0;21}$	$f_{0;22}$	...	$f_{0;2K}$	$f_{2\bullet}$
...	...	...	...	...	...
x(J)	$f_{0;J1}$	$f_{0;J2}$	...	$f_{0;JK}$	$f_{J\bullet}$
$\Sigma$	$f_{\bullet 1}$	$f_{\bullet 2}$	...	$f_{\bullet K}$	1

Wären bedingte und Randverteilung identisch, so würde ein Anteil von von  $f_{0;jk} = f_{\bullet k} \cdot f_{j\bullet}$  an den N Daten in Kategorie  $(x(j), y(k))$  fallen.

Dieser Fall wird als **empirische Unabhängigkeit** von X und Y bezeichnet.

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten

Ein Maß, dass desto größer wird, je größer die Abweichung der bedingten Verteilung  $f_{Y|X}$  von der Randverteilung  $Y \bullet Y$  ist, ist also ein sinnvolles Zusammenhangsmaß.

	Y				
	y(1)	y(2)	...	y(K)	$\Sigma$
x(1)	$v_{11}$	$v_{12}$	...	$v_{1K}$	$N_{1\bullet}$
x(2)	$v_{21}$	$v_{22}$	...	$v_{2K}$	$N_{2\bullet}$
...	...	...	...	...	...
x(J)	$v_{J1}$	$v_{J2}$	...	$v_{JK}$	$N_{J\bullet}$
$\Sigma$	$N_{\bullet 1}$	$N_{\bullet 2}$	...	$N_{\bullet K}$	N

Somit würden bei Unabhängigkeit

$$v_{jk} = f_{\bullet k} \cdot f_{j\bullet} \cdot N = \frac{N_{\bullet k} \cdot N_{j\bullet} \cdot N}{N \cdot N} = \frac{N_{\bullet k} \cdot N_{j\bullet}}{N}$$

Beobachtungen in Kategorie  $(x(j), x(k))$  erwartet.

## 4.2 Bivariate Daten: Zusammenhangsmaße

**Nominale Daten** Je größer die beobachteten Anzahlen  $N_{jk}$  von den erwarteten  $v_{jk}$  abweichen, desto mehr unterscheiden sich bedingte und Randverteilungen. Ein Maß, dass auf der quadratischen Abweichung der erwarteten von den beobachteten Häufigkeiten basiert, ist die  **$\chi^2$ -Größe**

$$\chi^2 = \sum_{j=1}^J \sum_{k=1}^K \frac{(N_{jk} - v_{jk})^2}{v_{jk}}, \quad v_{jk} = \frac{N_{j\bullet} N_{\bullet k}}{N}$$

		Y				$\Sigma$
		$y(1)$	$y(2)$	...	$y(K)$	
X	$x(1)$	$(N_{11}-v_{11})^2$	$(N_{12}-v_{12})^2$	...	$(N_{1K}-v_{1K})^2$	$N_{1\bullet}$
	$x(2)$	$(N_{21}-v_{21})^2$	$(N_{22}-v_{22})^2$	...	$(N_{2K}-v_{2K})^2$	$N_{2\bullet}$
	...	...	...	...	...	...
	$x(J)$	$(N_{J1}-v_{J1})^2$	$(N_{J2}-v_{J2})^2$	...	$(N_{JK}-v_{JK})^2$	$N_{J\bullet}$
$\Sigma$		$N_{\bullet 1}$	$N_{\bullet 2}$	...	$N_{\bullet K}$	N

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten: die $\chi^2$ -Größe

$$\chi^2 = \sum_{j=1}^J \sum_{k=1}^K \frac{(N_{jk} - v_{jk})^2}{v_{jk}}, \quad v_{jk} = \frac{N_{j\bullet} N_{\bullet k}}{N}$$

Die  $\chi^2$ -Größe erfüllt die Forderung, desto größer zu werden, je größer die Abweichung der bedingten Verteilung  $f_{Y|X}$  von der Randverteilung  $f_{\cdot Y}$  ist.

$$\begin{aligned} \boxed{\chi^2} &= \sum_{j=1}^J \sum_{k=1}^K \frac{\left(N_{jk} - \frac{N_{j\bullet} N_{\bullet k}}{N}\right)^2 N}{N_{j\bullet} N_{\bullet k}} = \sum_{j=1}^J \sum_{k=1}^K \frac{(f_{jk} N - f_{j\bullet} f_{\bullet k} N)^2}{f_{j\bullet} f_{\bullet k} N} \\ &= \sum_{j=1}^J \sum_{k=1}^K \frac{N(f_{jk} - f_{j\bullet} f_{\bullet k})^2}{f_{j\bullet} f_{\bullet k}} = \sum_{j=1}^J \sum_{k=1}^K \frac{N f_{j\bullet}^2 \left(\frac{f_{jk}}{f_{j\bullet}} - f_{\bullet k}\right)^2}{f_{j\bullet} f_{\bullet k}} \\ &= \boxed{\sum_{j=1}^J \sum_{k=1}^K \frac{N f_{j\bullet} (f_{y;k|j} - f_{\bullet k})^2}{f_{\bullet k}}} \end{aligned}$$

## 4.2 Bivariate Daten: Zusammenhangsmaße

**Nominale Daten:** die  $\chi^2$ -Größe

$$\chi^2 = \sum_{j=1}^J \sum_{k=1}^K \frac{(N_{jk} - v_{jk})^2}{v_{jk}} = N \left( \sum_{j=1}^J \sum_{k=1}^K \frac{N_{jk}^2}{N_{j\bullet} N_{\bullet k}} - 1 \right), \quad v_{jk} = \frac{N_{j\bullet} N_{\bullet k}}{N}$$

Es gilt:  $0 \leq \chi^2 \leq N(\min\{J, K\} - 1)$

Beweis:

$0 \leq \chi^2$  klar wegen  $N_{j\bullet} > 0, N_{\bullet k} > 0, (N_{jk} - v_{jk})^2 \geq 0$

$0 = \chi^2$ , wenn  $N_{jk} = v_{jk}$ , d.h. wenn alle bedingten Häufigkeiten den unter Unabhängigkeit erwarteten Häufigkeiten entsprechen.

## 4.2 Bivariate Daten: Zusammenhangsmaße

**Nominale Daten:** die  $\chi^2$ -Größe

$$\chi^2 = \sum_{j=1}^J \sum_{k=1}^K \frac{(N_{jk} - v_{jk})^2}{v_{jk}} = N \left( \sum_{j=1}^J \sum_{k=1}^K \frac{N_{jk}^2}{N_{j\bullet} N_{\bullet k}} - 1 \right), \quad v_{jk} = \frac{N_{j\bullet} N_{\bullet k}}{N}$$

Wann gilt:  $\boxed{\chi^2 = N(\min\{J, K\} - 1)}$  ?

**Beweisskizze:** Sei o.B.d.A.  $K \leq J$ .

Dann gilt für alle  $k = 1, \dots, K$  und  $j = 1, \dots, J$  mit  $N_{jk} > 0$ :

$$\sum_{j=1}^J \sum_{k=1}^K \frac{N_{jk}^2}{N_{j\bullet} N_{\bullet k}} = K \Leftrightarrow \frac{N_{jk}}{N_{j\bullet}} = 1 \quad \text{für ein } k_j,$$

d.h.  $\chi^2$  wird maximal, wenn es zu jedem  $j$  ein  $k_j$  mit  $f_{y;k_j|j} = 1$  gibt.

## 4.2 Bivariate Daten: Zusammenhangsmaße

**Nominale Daten:** die  $\chi^2$ -Größe

$$\chi^2 = \sum_{j=1}^J \sum_{k=1}^K \frac{(N_{jk} - v_{jk})^2}{v_{jk}} = N \left( \sum_{j=1}^J \sum_{k=1}^K \frac{N_{jk}^2}{N_{j\bullet} N_{\bullet k}} - 1 \right), \quad v_{jk} = \frac{N_{j\bullet} N_{\bullet k}}{N}$$

Es gilt:  $0 \leq \chi^2 \leq N(\min\{J, K\} - 1)$

**(Korrigierter) Kontingenzkoeffizient nach Pearson:**

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N \min\{J, K\} - 1}} \in [0, 1]$$

Eliminiert Abhängigkeit des Koeffizienten vom Stichprobenumfang  $N$  und von der Dimension  $\min\{J, K\}$ .

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten: Beispiel Bearbeitungen von Softwareaufgaben

 $N_{jk}$ 

b		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbei- ter(in)	Kai	0	1	1	2
	Miriam	0	3	0	3
	Oliver	2	1	1	4
	Tina	0	1	2	3
	$\Sigma$	2	6	4	12

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten: Beispiel Bearbeitungen von Softwareaufgaben

$V_{jk}$

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbei- ter(in)	Kai	0 $2 \cdot 2 / 12 = 1/3$	1 $2 \cdot 6 / 12 = 1$	1 $2 \cdot 4 / 12 = 2/3$	2
	Miriam	0 $3 \cdot 2 / 12 = 1/2$	3 $3 \cdot 6 / 12 = 3/2$	0 $3 \cdot 4 / 12 = 1$	3
	Oliver	2 $4 \cdot 2 / 12 = 2/3$	1 $4 \cdot 6 / 12 = 2$	1 $4 \cdot 4 / 12 = 4/3$	4
	Tina	0 $3 \cdot 2 / 12 = 1/2$	1 $3 \cdot 6 / 12 = 3/2$	2 $3 \cdot 4 / 12 = 1$	3
	$\Sigma$	2	6	4	12

## 4.2 Bivariate Daten: Zusammenhangsmaße

**Nominale Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

$$(N_{jk} - v_{jk})^2$$

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbei- ter(in)	Kai	0 $(0-1/3)^2=1/9$	1 $(1-1)^2=0$	1 $(1-2/3)^2=1/9$	2
	Miriam	0 $(0-1/2)^2=1/4$	3 $(3-3/2)^2=9/4$	0 $(0-1)^2=1$	3
	Oliver	2 $(2-2/3)^2=16/9$	1 $(1-2)^2=1$	1 $(1-4/3)^2=1/9$	4
	Tina	0 $(0-1/2)^2=1/4$	1 $(1-3/2)^2=1/4$	2 $(2-1)^2=1$	3
	$\Sigma$	2	6	4	12

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten: Beispiel Bearbeitungen von Softwareaufgaben

$$(N_{jk} - v_{jk})^2 / v_{jk}$$

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbei- ter(in)	Kai	0 $1 \cdot 3 / (9 \cdot 1) = 1/3$	1 $0/1=0$	1 $1 \cdot 3 / (9 \cdot 2) = 1/6$	2
	Miriam	0 $1 \cdot 2 / (4 \cdot 1) = 1/2$	3 $9 \cdot 2 / (4 \cdot 3) = 3/2$	0 $1/1=1$	3
	Oliver	2 $16 \cdot 3 / (9 \cdot 2) = 8/3$	1 $1/2$	1 $1 \cdot 3 / (9 \cdot 4) = 1/12$	4
	Tina	0 $1 \cdot 2 / (4 \cdot 1) = 1/2$	1 $1 \cdot 2 / (4 \cdot 3) = 1/6$	2 $1/1=1$	3
	$\Sigma$	2	6	4	12

## 4.2 Bivariate Daten: Zusammenhangsmaße

**Nominale Daten:** Beispiel **Bearbeitungen von Softwareaufgaben**

$$\begin{aligned}\chi^2 &= \sum_{j=1}^J \sum_{k=1}^K \frac{(N_{jk} - v_{jk})^2}{v_{jk}} \\ &= \frac{1}{12}(4 + 6 + 32 + 6 + 0 + 18 + 6 + 2 + 2 + 12 + 1 + 12) = \frac{101}{12} \approx 8.417\end{aligned}$$

$$(N_{jk} - v_{jk})^2 / v_{jk}$$

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbei- ter(in)	Kai	1/3	0	1/6	2
	Miriam	1/2	3/2	1	3
	Oliver	8/3	1/2	1/12	4
	Tina	1/2	1/6	1	3
$\Sigma$		2	6	4	12

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Nominale Daten: Beispiel Bearbeitungen von Softwareaufgaben

$$\chi^2 = \frac{101}{12}, \quad C = \sqrt{\frac{\chi^2}{\chi^2 + N} \frac{\min\{J, K\}}{\min\{J, K\} - 1}} = \sqrt{\frac{101 \cdot 12}{12 \cdot 245} \cdot \frac{3}{2}} = \sqrt{\frac{303}{490}} \approx 0.786$$

$$(N_{jk} - v_{jk})^2 / v_{jk}$$

		Aufgabe			$\Sigma$
		Abfrage	Export	Verknüpfung	
Bearbei- ter(in)	Kai	1/3	0	1/6	2
	Miriam	1/2	3/2	1	3
	Oliver	8/3	1/2	1/12	4
	Tina	1/2	1/6	1	3
	$\Sigma$	2	6	4	12

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Ordinale Daten

Allgemein: Zusammenhang (=Korrelation) zwischen  $Y$  und  $X$  desto größer, je besser sich der Wert von  $Y$  unter Kenntnis des Werts von  $X$  vorhersagen lässt (oder umgekehrt).

Wert von  $Y$  lässt sich bei Kenntnis von  $X$  umso besser vorhersagen, je mehr ein hoher Wert von  $X$  einen hohen Wert von  $Y$  impliziert (**positiver Zusammenhang**) bzw. je mehr ein hoher Wert von  $X$  einen niedrigen Wert von  $Y$  impliziert (**negativer Zusammenhang**).

Ein sinnvolles Zusammenhangsmaß für ordinale Daten sollte also im Absolutwert hoch sein, wenn hohe **Ränge** von  $X$  mit hohen bzw. niedrigen **Rängen** von  $Y$  einhergehen und niedrig, wenn Paare von hohen und hohen, hohen und niedrigen, niedrigen und hohen sowie niedrigen und niedrigen **X- und Y-Rängen** in gleichem Maße auftreten.

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Quantitative Daten

Allgemein: Zusammenhang (=Korrelation) zwischen  $Y$  und  $X$  desto größer, je besser sich der Wert von  $Y$  unter Kenntnis des Werts von  $X$  vorhersagen lässt (oder umgekehrt).

Wert von  $Y$  lässt sich bei Kenntnis von  $X$  umso besser vorhersagen, je mehr ein hoher **Wert** von  $X$  einen hohen **Wert** von  $Y$  impliziert (positiver Zusammenhang) bzw. je mehr ein hoher **Wert** von  $X$  einen niedrigen **Wert** von  $Y$  impliziert (negativer Zusammenhang).

Ein sinnvolles Zusammenhangsmaß für ordinale Daten sollte also im Absolutwert hoch sein, wenn hohe **Werte** von  $X$  mit hohen bzw. niedrigen **Werten** von  $Y$  einhergehen und niedrig, wenn Paare von hohen und hohen, hohen und niedrigen, niedrigen und hohen sowie niedrigen und niedrigen **X- und Y-Werten** in gleichem Maße auftreten.

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Quantitative Daten

Allgemein: Zusammenhang (=Korrelation) zwischen  $Y$  und  $X$  desto größer, je besser sich der Wert von  $Y$  unter Kenntnis des Werts von  $X$  vorhersagen lässt (oder umgekehrt).

$$\text{Kovarianz: } s_{xy} = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})$$

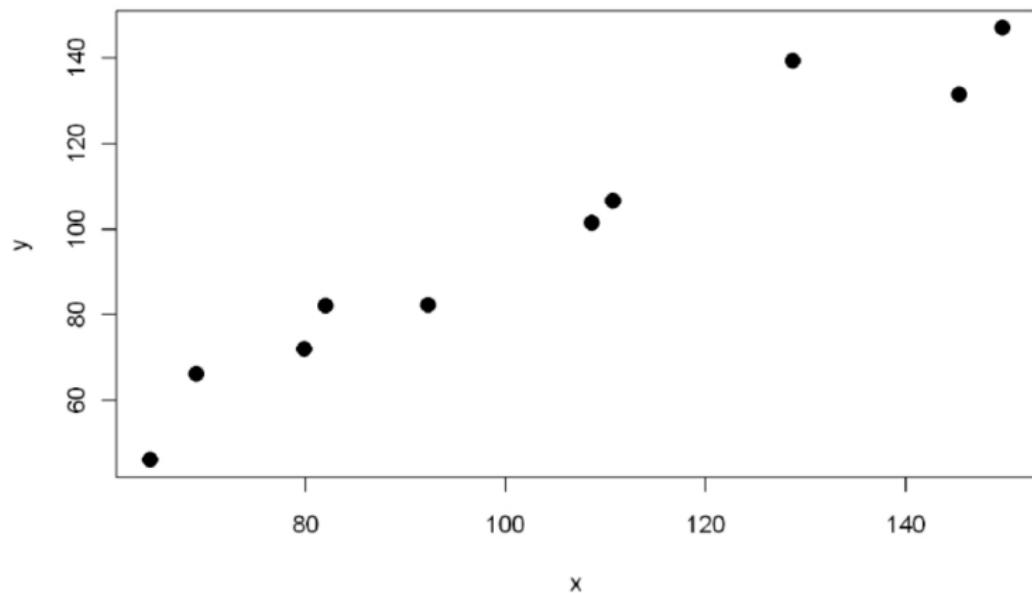
$s_{xy} > 0$ , wenn hohe Werte von  $X$  in hohem Maße mit hohen Werten von  $Y$  einhergehen (Positive Korrelation)

$s_{xy} < 0$ , wenn hohe Werte von  $X$  in hohem Maße mit niedrigen Werten von  $Y$  einhergehen (Negative Korrelation)

$s_{xy} = 0$ , wenn hohe Werte von  $X$  in gleichem Maße mit hohen Werten wie mit niedrigen Werten von  $Y$  einhergehen (Unkorreliertheit)

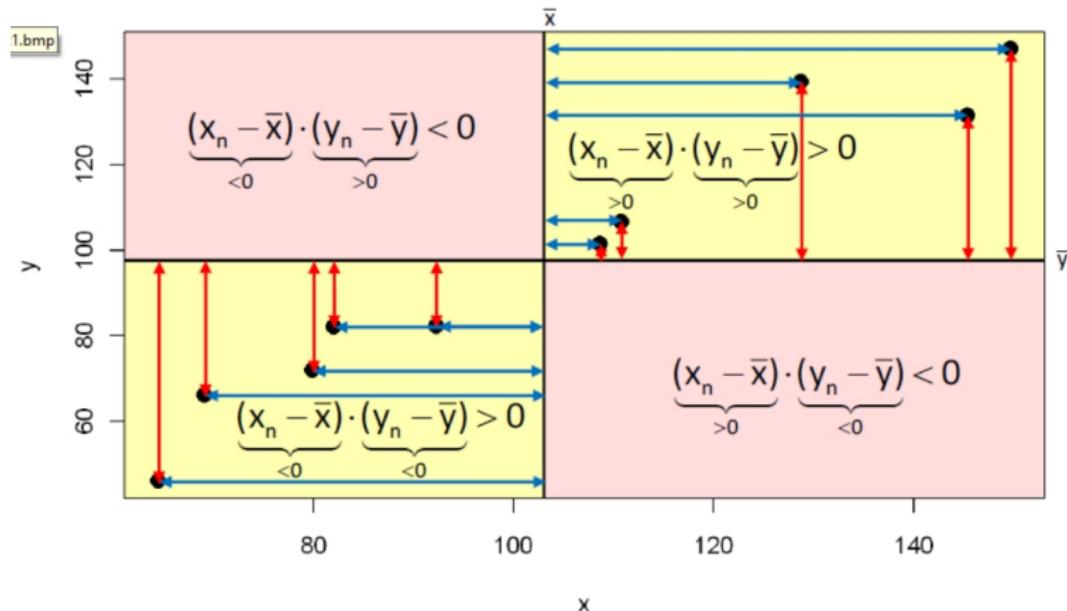
## 4.2 Bivariate Daten: Zusammenhangsmaße

**Quantitative Daten: Kovarianz:**  $s_{xy} = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})$



## 4.2 Bivariate Daten: Zusammenhangsmaße

Quantitative Daten: Kovarianz:  $s_{xy} = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})$



## 4.2 Bivariate Daten: Zusammenhangsmaße

### Quantitative Daten:

#### Kovarianz:

$$s_{xy} = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}) = \frac{1}{N-1} \left( \sum_{n=1}^N x_n y_n - N \bar{x} \bar{y} \right) = \frac{N}{N-1} (\bar{xy} - \bar{x} \cdot \bar{y})$$

Beweis analog zu Beweis von  $d_x^2 = \bar{x^2} - \bar{x}^2$ :

$$\begin{aligned} s_{xy}^2 &= \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}) = \frac{1}{N-1} \sum_{n=1}^N (x_n y_n - x_n \bar{y} - \bar{x} y_n + \bar{x} \cdot \bar{y}) \\ &= \frac{1}{N-1} \sum_{n=1}^N x_n y_n - \frac{1}{N-1} \left( \sum_{n=1}^N x_n \right) \bar{y} - \bar{x} \frac{1}{N-1} \left( \sum_{n=1}^N y_n \right) + \frac{N}{N-1} \bar{x} \cdot \bar{y} \\ &= \frac{N}{N-1} \bar{xy} - \frac{N}{N-1} \bar{x} \cdot \bar{y} - \frac{N}{N-1} \bar{x} \cdot \bar{y} + \frac{N}{N-1} \bar{x} \cdot \bar{y} = \boxed{\frac{N}{N-1} (\bar{xy} - \bar{x} \cdot \bar{y})} \quad \square \end{aligned}$$

## 4.2 Bivariate Daten: Zusammenhangsmaße

Quantitative Daten: Kovarianz:  $-s_x s_y \leq s_{xy} \leq s_x s_y$

Beweis: Spezialfall der Cauchy-Schwarz-Ungleichung:

Für  $(a_n, b_n) \in \mathbb{R}^2$  gilt:

$$\left( \sum_{n=1}^N a_n b_n \right)^2 \leq \sum_{n=1}^N a_n^2 \cdot \sum_{n=1}^N b_n^2 \implies \left( \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}) \right)^2 \leq \sum_{n=1}^N (x_n - \bar{x})^2 \cdot \sum_{n=1}^N (y_n - \bar{y})^2$$

$$\Leftrightarrow -\sqrt{\sum_{n=1}^N (x_n - \bar{x})^2 \cdot \sum_{n=1}^N (y_n - \bar{y})^2} \leq \left( \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}) \right) \leq \sqrt{\sum_{n=1}^N (x_n - \bar{x})^2 \cdot \sum_{n=1}^N (y_n - \bar{y})^2}$$

$$\Leftrightarrow -\sqrt{\frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N-1}} \sqrt{\frac{\sum_{n=1}^N (y_n - \bar{y})^2}{N-1}} \leq \frac{\left( \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}) \right)}{N-1} \leq \sqrt{\frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N-1}} \sqrt{\frac{\sum_{n=1}^N (y_n - \bar{y})^2}{N-1}}$$

$$\Leftrightarrow -s_x s_y \leq s_{xy} \leq s_x s_y \quad \square$$

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Quantitative Daten: Korrelationskoeffizient nach Bravais-Pearson

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad -s_x s_y \leq s_{xy} \leq s_x s_y \implies -1 \leq r_{xy} \leq 1$$

Gleichheitsbedingung bei der Cauchy-Schwarz-Ungleichung:

Für  $(a_n, b_n) \in \mathbb{R}^2$  gilt:

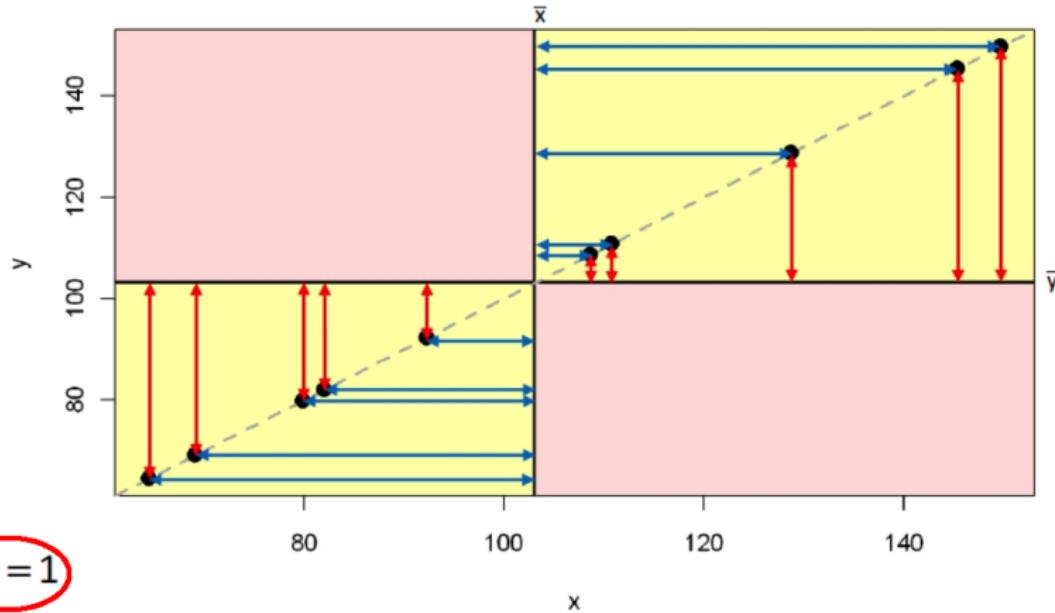
$$\left( \sum_{n=1}^N a_n b_n \right)^2 = \sum_{n=1}^N a_n^2 \cdot \sum_{n=1}^N b_n^2 \Leftrightarrow \text{es gibt eine Konstante } d \text{ mit } b_n = d \cdot a_n \forall n$$

$$\begin{aligned} \implies r_{xy} \in \{-1, 1\} &\Leftrightarrow (y_n - \bar{y}) = d \cdot (x_n - \bar{x}) \\ &\Leftrightarrow y_n = c + d \cdot x_n \quad \text{mit } c = \bar{y} - d \bar{x} \end{aligned}$$

Das heißt,  $|r_{xy}|$  ist genau dann 1, wenn alle  $x_n$  und  $y_n$  auf einer Geraden liegen.

## 4.2 Bivariate Daten: Zusammenhangsmaße

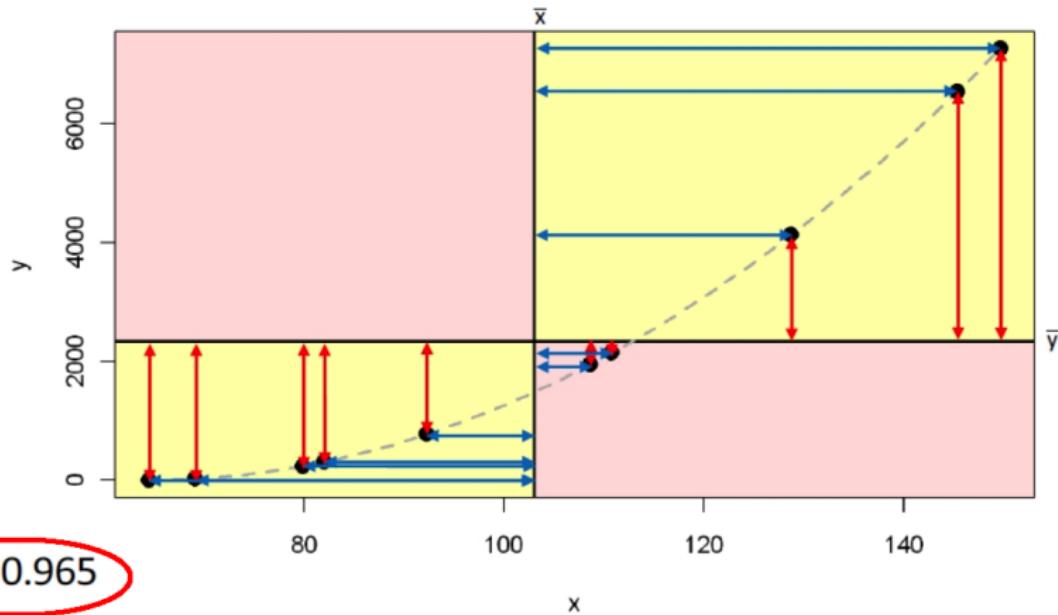
Quantitative Daten: Kovarianz:  $s_{xy} = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(c + dx_n - c + d\bar{x})$



## 4.2 Bivariate Daten: Zusammenhangsmaße

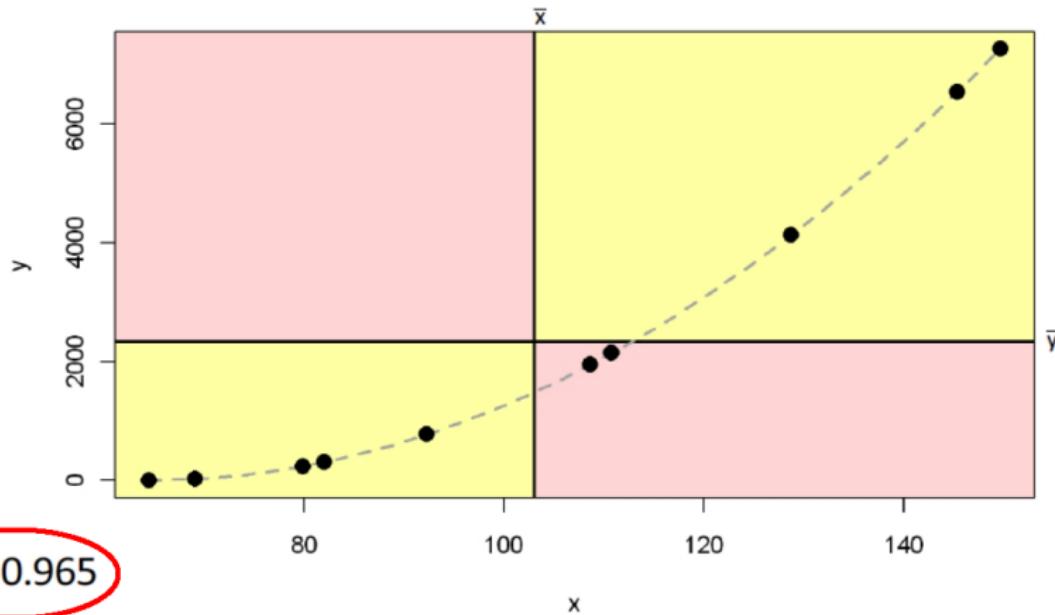
### Quantitative Daten: Korrelationskoeffizient nach Bravais-Pearson

Nicht-linearer monotoner Zusammenhang



## 4.2 Bivariate Daten: Zusammenhangsmaße

**Ordinal/Quantitative Daten:** Nicht-linearer monotoner Zusammenhang  
Übergang zu Rängen

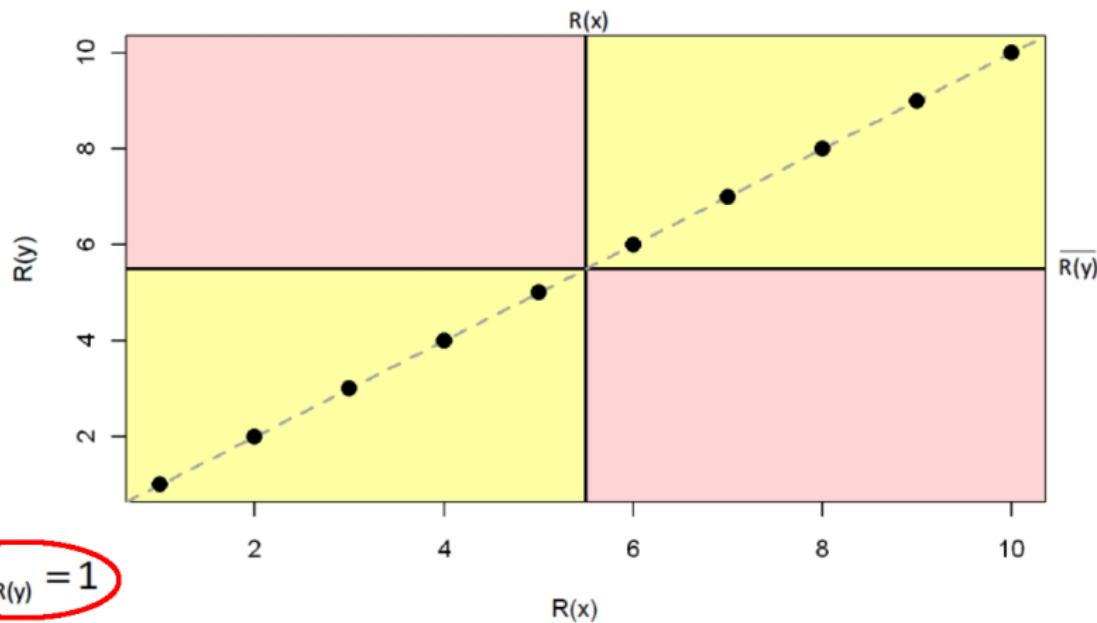


$$r_{xy} = 0.965$$

## 4.2 Bivariate Daten: Zusammenhangsmaße

**Ordinal/Quantitative Daten:** Nicht-linearer monotoner Zusammenhang

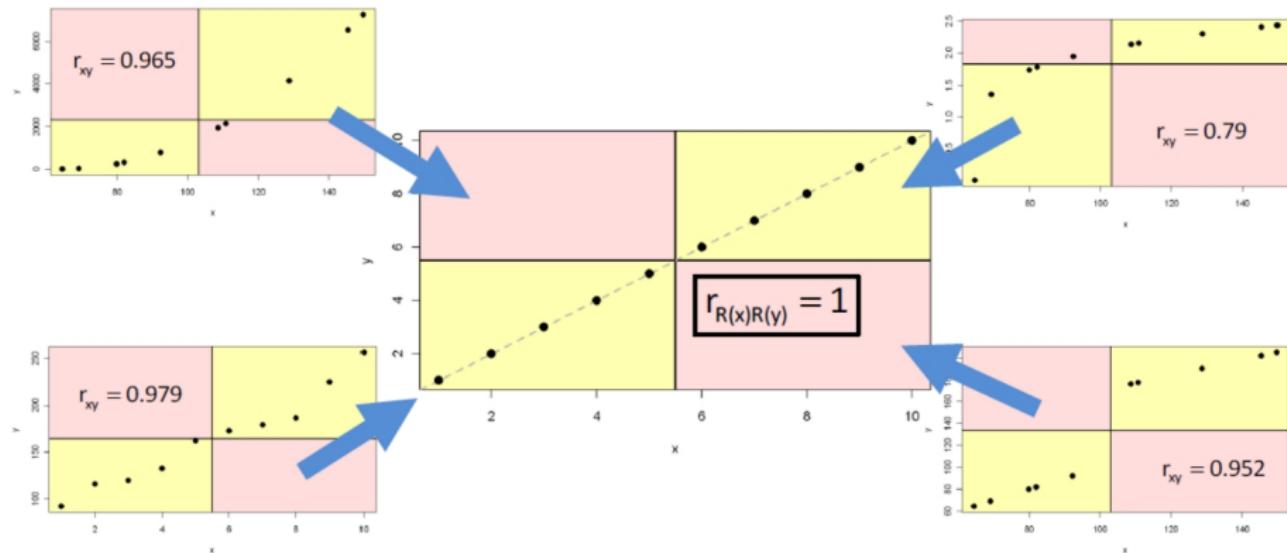
Übergang zu Rängen



## 4.2 Bivariate Daten: Zusammenhangsmaße

### Ordinalale/Quantitative Daten

Absolute Korrelation von Rängen bei monotonem Zusammenhang immer 1



## 4.2 Bivariate Daten: Zusammenhangsmaße

### Ordinale/Quantitative Daten

Falls  $X$  und  $Y$  mindestens ordinale Skalenniveau haben, so wird der Bravais-Pearson-Korrelationskoeffizient der Ränge  $R(X)$  und  $R(Y)$  von  $X$  und  $Y$  der **Spearman'sche Rangkorrelationskoeffizient**  $r_{xy}^{\text{Sp}}$  von  $X$  und  $Y$  genannt:

$$r_{xy}^{\text{Sp}} = r_{R(X)R(Y)} = \frac{s_{R(X)R(Y)}}{s_{R(X)}s_{R(Y)}} = \frac{\sum_{n=1}^N (R(x_n) - \overline{R(x)}) (R(y_n) - \overline{R(y)})}{\sqrt{\sum_{n=1}^N (R(x_n) - \overline{R(x)})^2 \sum_{n=1}^N (R(y_n) - \overline{R(y)})^2}}$$

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Ordinale/Quantitative Daten

#### Spearman'scher Rangkorrelationskoeffizient

Falls keine Bindungen auftreten, d.h.  $R(x_j) \neq R(x_k)$  und  $R(y_j) \neq R(y_k)$  für alle  $j \neq k$ , so gilt:

$$r_{xy}^{\text{Sp}} = 1 - \frac{6}{N(N^2 - 1)} \sum_{n=1}^N (R(x_n) - R(y_n))^2$$

#### Beweisansatz:

$$\sum_{n=1}^N R(x_n) = \sum_{n=1}^N R(y_n) = \sum_{n=1}^N n = \frac{N(N+1)}{2}$$

$$\text{und } \sum_{n=1}^N R(x_n)^2 = \sum_{n=1}^N R(y_n)^2 = \sum_{n=1}^N n^2 = \frac{N(N+1)(2N+1)}{6}$$

## 4.2 Bivariate Daten: Zusammenhangsmaße

### Ordinale/Quantitative Daten: Beispiel Bearbeitungen von Softwareaufgaben

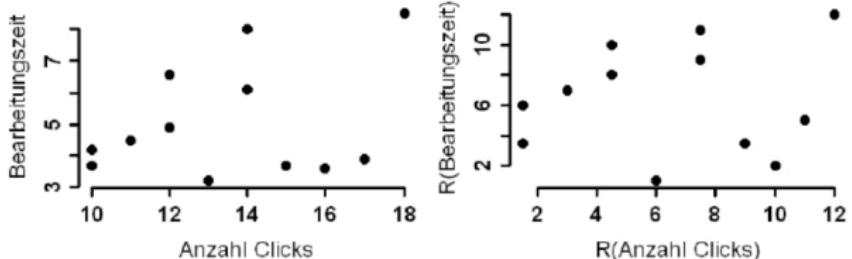
Anzahl Clicks	
Rang	
14	7.5
12	4.5
12	4.5
13	6
17	11
11	3
14	7.5
10	1.5
10	1.5
18	12
16	10
15	9
Bearbeitungszeit	
Rang	
8.0	11
4.9	8
6.6	10
3.2	1
3.9	5
4.5	7
6.1	9
3.7	3.5
4.2	6
8.5	12
3.6	2
3.7	3.5

$$\bar{x}_4 = 13.5$$

$$s_{x_4}^2 = 7$$

$$\bar{x}_5 = 5.075$$

$$s_{x_5}^2 = 3.24$$



$$r_{x_4 x_5} = [(0.5 \cdot 2.925) + (-1.5 \cdot -0.175) + (-1.5 \cdot 1.525) + (-0.5 \cdot -1.875) + (3.5 \cdot -1.175) + (-2.5 \cdot -0.575) + (0.5 \cdot 1.025) + (-3.5 \cdot -1.375) + (-3.5 \cdot -0.875) + (4.5 \cdot 3.425) + (2.5 \cdot -1.475) + (1.5 \cdot -1.375)] / (11 \cdot \sqrt{7 \cdot 3.24}) = 0.301$$

$$r_{x_4 x_5}^{Sp} = [(1 \cdot 4.5) + (-2 \cdot 1.5) + (-2 \cdot 3.5) + (-0.5 \cdot -5.5) + (4.5 \cdot -1.5) + (-3.5 \cdot 0.5) + (1 \cdot 2.5) + (-5 \cdot -3) + (-5 \cdot -0.5) + (5.5 \cdot 5.5) + (3.5 \cdot -4.5) + (2.5 \cdot -3)] / (39.4525) = 0.111$$

## 4.3 Bivariate Daten: Lineare Regression

### Korrelation und Linearität:

Der Korrelationskoeffizient ist auch deshalb so beliebt, weil er ein *Maß für die Linearität eines Zusammenhangs* darstellt.

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

## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten: Erinnerung

Allgemein: Zusammenhang (=Korrelation) zwischen  $Y$  und  $X$  desto größer, je besser sich der Wert von  $Y$  unter Kenntnis des Werts von  $X$  **vorhersagen** lässt (oder umgekehrt).

Bravais-Pearson-Korrelationskoeffizient misst linearen Zusammenhang.

Wie lässt sich der lineare Zusammenhang zur Vorhersage nutzen?

## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten

$$|r_{xy}| = 1 \Leftrightarrow y_n = c + dx_n \text{ für } n=1, \dots, N$$

Für beliebiges  $(j, k)$  mit  $j \neq k$ :

$$y_j = c + dx_j$$

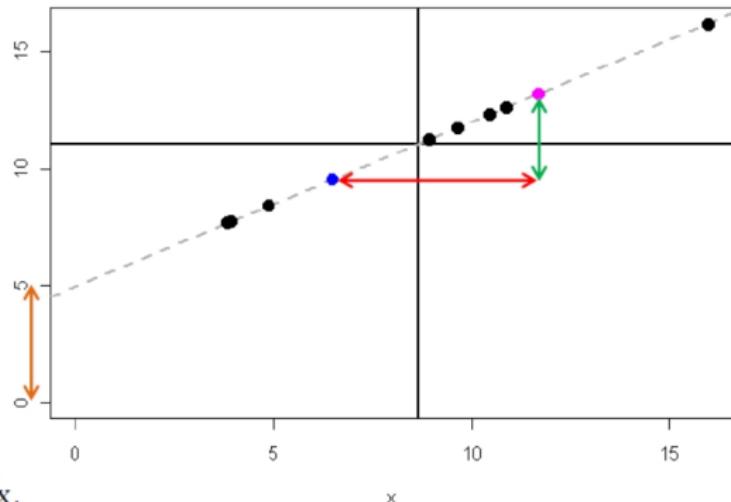
$$y_k = c + dx_k$$

$$\begin{aligned} \Rightarrow y_j - y_k &= (c + dx_j) - (c + dx_k) \\ &= d(x_j - x_k) \end{aligned}$$

$$\Leftrightarrow d = \frac{(y_j - y_k)}{(x_j - x_k)}$$

$$y_k = c + dx_k$$

$$\begin{aligned} \Leftrightarrow c &= y_k - dx_k \\ &= y_k - \frac{(y_j - y_k)}{(x_j - x_k)} x_k \end{aligned}$$



Perfekte Vorhersage durch Einsetzen in die Gleichung.

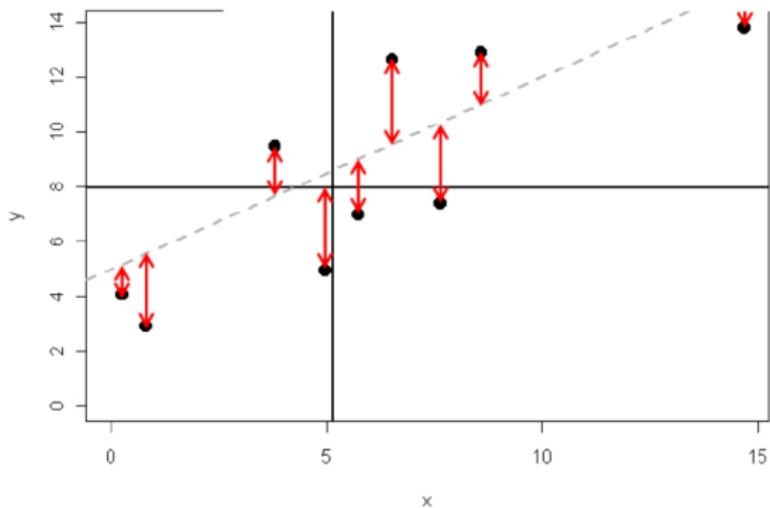
## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten

$$0 < |r_{xy}| < 1 \Leftrightarrow y_n = [c + dx_n] + \boxed{\epsilon_n} \text{ für } n=1, \dots, N$$

Vorhersagefehler

$$\boxed{\epsilon_n} = y_n - [c + dx_n]$$



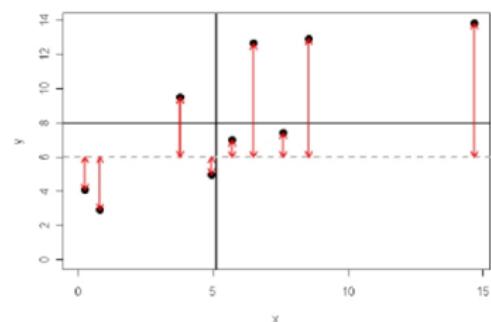
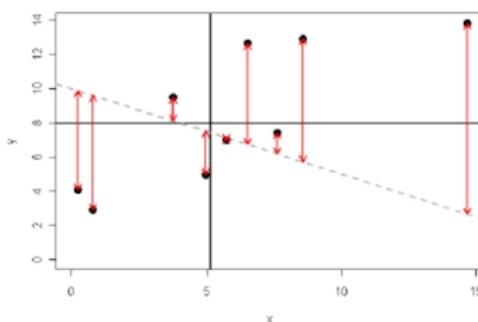
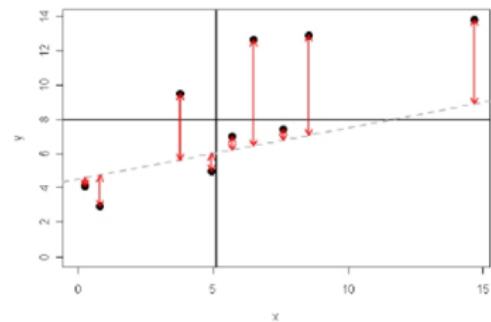
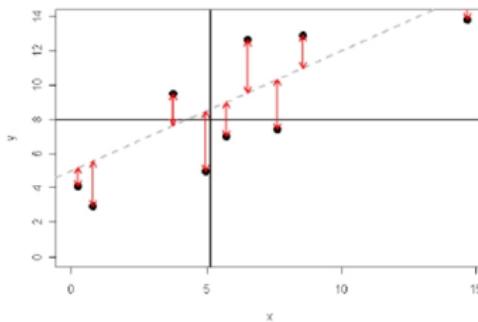
## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten: Methode der kleinsten Quadrate

Koeffizienten  $c$  und  $d$  so bestimmen, dass Fehlerquadratsumme

$$Q(c, d) = \sum_{n=1}^N \varepsilon_n^2$$

minimal wird.



## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten: Methode der kleinsten Quadrate

Die Fehlerquadratsumme  $Q(c, d) = \sum_{n=1}^N (Y_n - c - dx_n)^2$  ist minimal für

$$d = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad c = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}$$

Beweis:

$$\frac{\partial}{\partial c} Q(c, d) = \sum_{n=1}^N 2(c + dx_n - y_n) = 2Nc + 2dN\bar{x} - 2N\bar{y} \stackrel{!}{=} 0 \Leftrightarrow c + d\bar{x} - \bar{y} = 0$$

$$\frac{\partial}{\partial d} Q(c, d) = \sum_{n=1}^N 2(c + dx_n - y_n)x_n = 2Nc\bar{x} + 2d \sum_{n=1}^N x_n^2 - 2 \sum_{n=1}^N x_n y_n \stackrel{!}{=} 0$$

$$\Leftrightarrow cN\bar{x} + d \sum_{n=1}^N x_n^2 - \sum_{n=1}^N x_n y_n = 0$$

## 4.3 Bivariate Daten: Lineare Regression

### Beweis (Fortsetzung)

$$(1) c + d\bar{x} - \bar{y} = 0 \Leftrightarrow c = \bar{y} - d\bar{x} \quad (2) cN\bar{x} + d \sum_{n=1}^N x_n^2 - \sum_{n=1}^N x_n y_n = 0$$

$$(1) \text{ in } (2) (\bar{y} - d\bar{x})N\bar{x} + d \sum_{n=1}^N x_n^2 - \sum_{n=1}^N x_n y_n = 0$$

$$\Leftrightarrow d \left( \sum_{n=1}^N x_n^2 - N\bar{x}^2 \right) = \sum_{n=1}^N x_n y_n - N\bar{x} \cdot \bar{y}$$

$$\Leftrightarrow d = \frac{\sum_{n=1}^N x_n y_n - N\bar{x} \cdot \bar{y}}{\left( \sum_{n=1}^N x_n^2 - N\bar{x}^2 \right)} = \frac{\frac{N}{N-1}(\bar{xy} - \bar{x} \cdot \bar{y})}{\frac{N}{N-1} \left( \frac{1}{N} \sum_{n=1}^N x_n^2 - \bar{x}^2 \right)} = \frac{s_{xy}}{s_x^2} \quad (3)$$

$$(3) \text{ in } (1) c = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}$$

## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten: Methode der kleinsten Quadrate

#### Beweis (Fortsetzung)

$$\frac{\partial}{\partial c} Q(c, d) = 2Nc + 2dN\bar{x} - 2N\bar{y}, \quad \frac{\partial}{\partial d} Q(c, d) = 2Nc\bar{x} + 2d \sum_{n=1}^N x_n^2 - 2 \sum_{n=1}^N x_n y_n$$

$$\frac{\partial^2}{\partial c \partial c} Q(c, d) = 2N, \quad \frac{\partial^2}{\partial c \partial d} Q(c, d) = 2 \sum_{n=1}^N x_n, \quad \frac{\partial^2}{\partial d \partial d} Q(c, d) = 2 \sum_{n=1}^N x_n^2$$

$$\det \begin{bmatrix} 2N & 2 \sum_{n=1}^N x_n \\ 2 \sum_{n=1}^N x_n & 2 \sum_{n=1}^N x_n^2 \end{bmatrix} = 4N \sum_{n=1}^N x_n^2 - 4 \left( \sum_{n=1}^N x_n \right)^2 = 4(N-1)Ns_x^2 > 0 \quad \square$$

## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten: Methode der kleinsten Quadrate

Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme

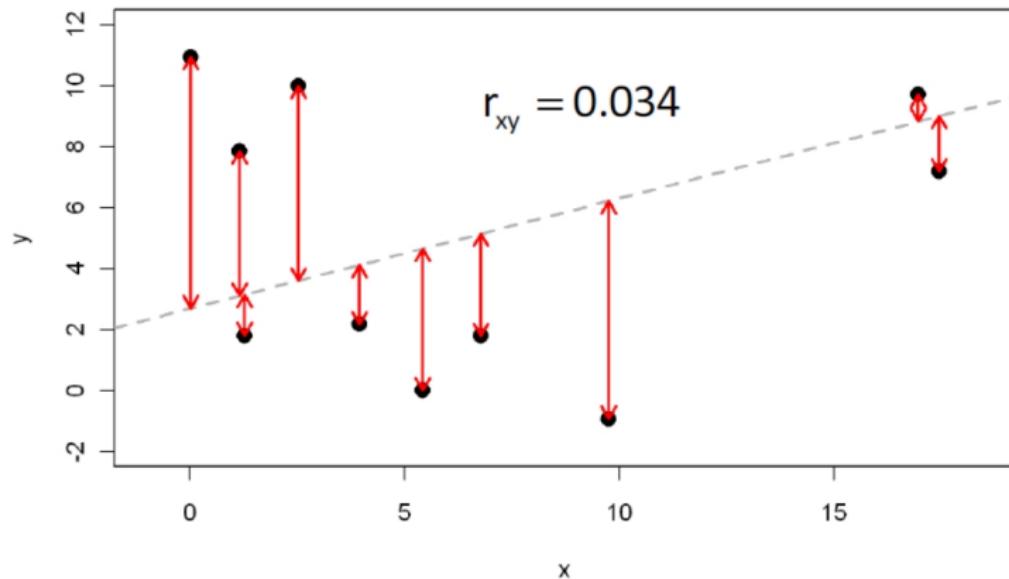
$$d = \frac{s_{xy}}{s_x^2} \text{ und } c = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}$$

$$\begin{aligned} \sum_{n=1}^N \varepsilon_n^2 &= \sum_{n=1}^N \left( y_n - \left( \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right) - \frac{s_{xy}}{s_x^2} x_n \right)^2 = \sum_{n=1}^N \left( y_n - \left( \bar{y} - r_{xy} \frac{s_y}{s_x} \bar{x} \right) - r_{xy} \frac{s_y}{s_x} x_n \right)^2 \\ &= \sum_{n=1}^N \left( (y_n - \bar{y}) - r_{xy} \frac{s_y}{s_x} (x_n - \bar{x}) \right)^2 \\ &= \sum_{n=1}^N \left( (y_n - \bar{y})^2 - 2r_{xy} \frac{s_y}{s_x} (y_n - \bar{y})(x_n - \bar{x}) + \left( r_{xy} \frac{s_y}{s_x} \right)^2 (x_n - \bar{x})^2 \right) \\ &= (N-1) \cdot \left( s_y^2 - 2r_{xy} \frac{s_y}{s_x} s_{xy} + \left( r_{xy} \frac{s_y}{s_x} \right)^2 s_x^2 \right) = (N-1) \cdot (s_y^2 - 2r_{xy}^2 s_y^2 + r_{xy}^2 s_y^2) \\ &= (N-1) \cdot (s_y^2 - r_{xy}^2 s_y^2) \quad \square \end{aligned}$$

## 4.3 Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

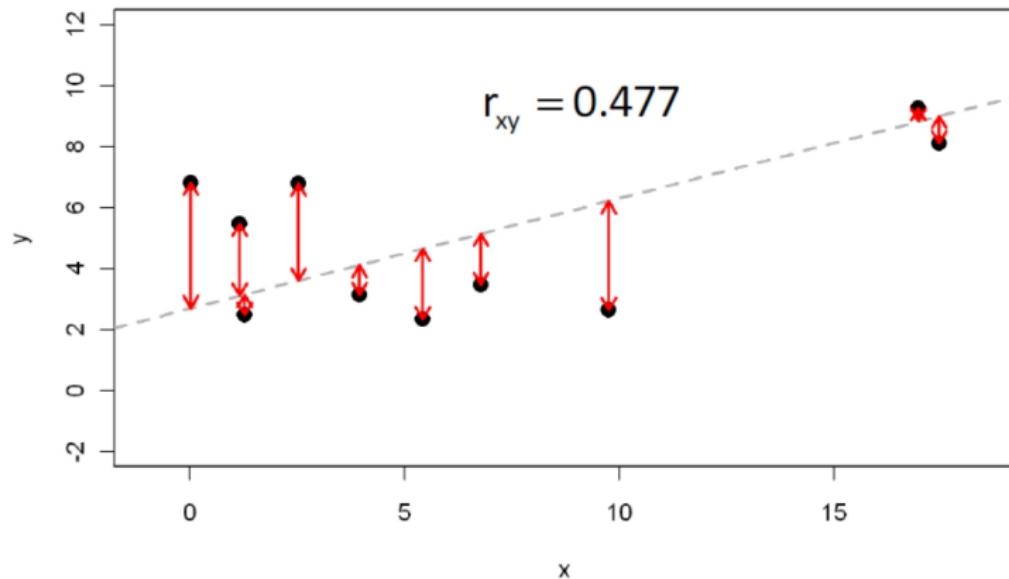
Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



## 4.3 Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

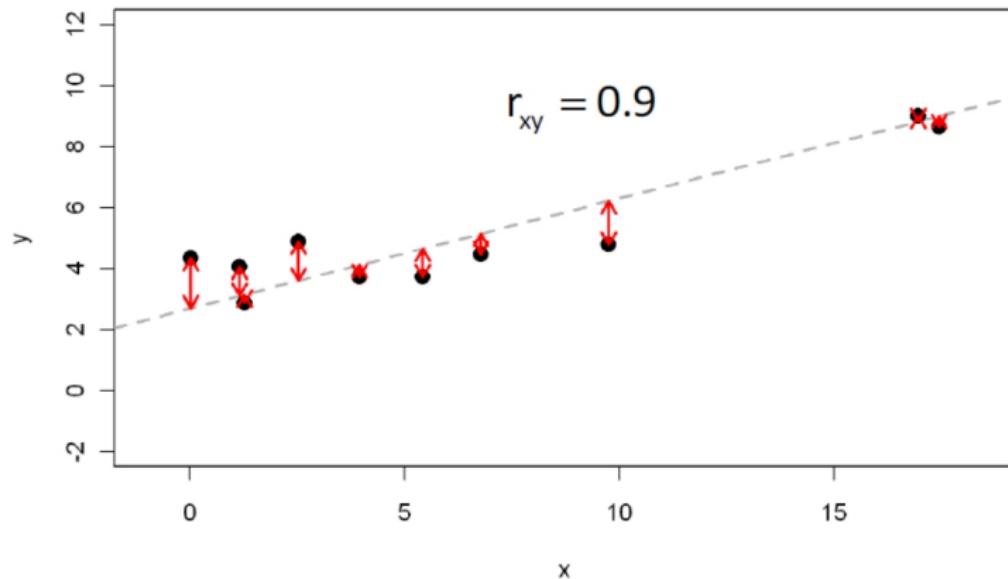
Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



## 4.3 Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

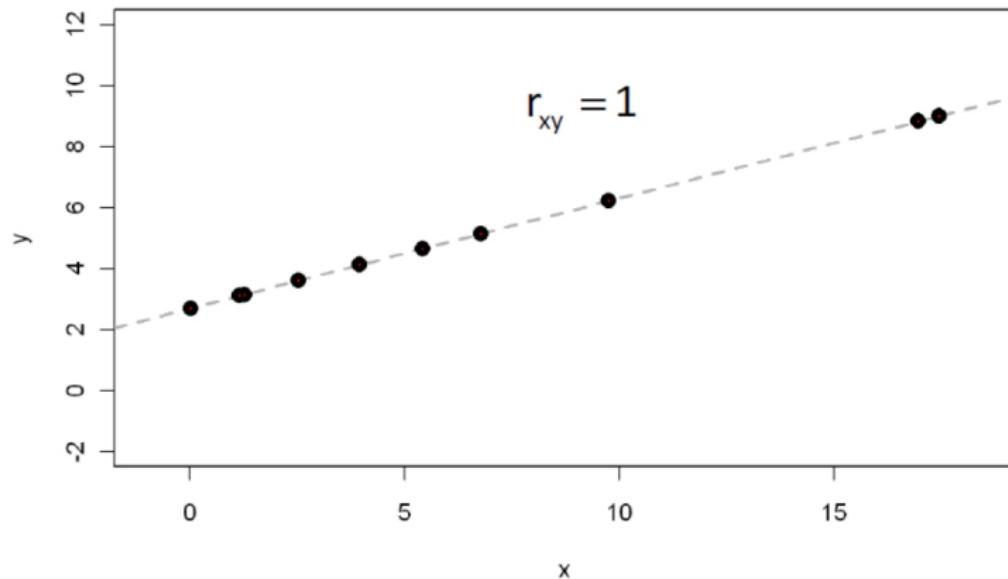
Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



## 4.3 Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



## 4.3 Bivariate Daten: Lineare Regression

### Quantitative Daten: Beispiel Bearbeitungen von Softwareaufgaben

Anzahl Clicks	Bearbeitungszeit	$c+dx_4$	$\epsilon$
14	8.0	5.177	2.823
12	4.9	4.768	0.132
12	6.6	4.768	1.832
13	3.2	4.973	-1.773
17	3.9	5.791	-1.891
11	4.5	4.564	-0.064
14	6.1	5.177	0.922
10	3.7	4.359	-0.659
10	4.2	4.359	-0.159
18	8.5	5.995	2.505
16	3.6	5.586	-1.986
15	3.7	5.382	-1.682

$$\bar{x}_4 = 13.5$$

$$S_{x_4}^2 = 7$$

$$\bar{x}_5 = 5.075$$

$$S_{x_5}^2 = 3.24$$

$$r_{x_4x_5} = 0.301$$

$$x_5 = c + dx_4 + \epsilon$$

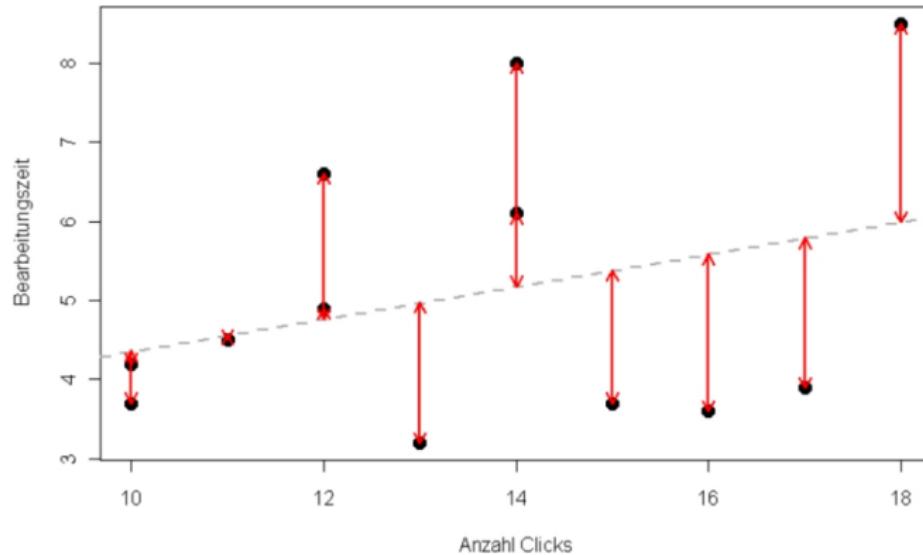
$$\begin{aligned}
 c &= \bar{x}_5 - r_{x_4x_5} \frac{s_{x_5}}{s_{x_4}} \bar{x}_4 \\
 &= 5.075 - 0.301 \sqrt{\frac{3.24}{7}} 13.5 \\
 &= 2.314
 \end{aligned}$$

$$\begin{aligned}
 d &= r_{x_4x_5} \frac{s_{x_5}}{s_{x_4}} = 0.301 \sqrt{\frac{3.24}{7}} \\
 &= 0.205
 \end{aligned}$$

## 4.3 Bivariate Daten: Lineare Regression

Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$$x_5 = 2.314 + 0.205x_4 + \varepsilon$$



## 4.3 Bivariate Daten: Lineare Regression

### Zusammenfassung

Skalenniveau → ↓ Zusammenhangsmaß	Nominal	Ordinal	Quantitativ
$\chi^2$ -Größe/ Kontingenzkoeffizient nach Pearson		 – Informationsverlust	 – Nur für klassierte Daten
Rangkorrelationskoeffizient nach Spearman	 – Nur für J = 2		 + Robust + Allg. Zusammenhang – Informationsverlust
Korrelationskoeff. nach Bravais-Pearson/lin. Regr.	 – Nur für J = 2		 – Ausreißeranfälligkeit – Lin. Zusammenhang + Informationsnutzung

# Wahrscheinlichkeitstheorie

## 5.0 Wahrscheinlichkeitstheorie

### Wahrscheinlichkeitstheorie

- Teilgebiet der Mathematik, dass sich mit Wahrscheinlichkeiten und der Analyse zufälliger Prozesse beschäftigt
- Mathematische Abstraktion von nicht-deterministischen Ereignissen

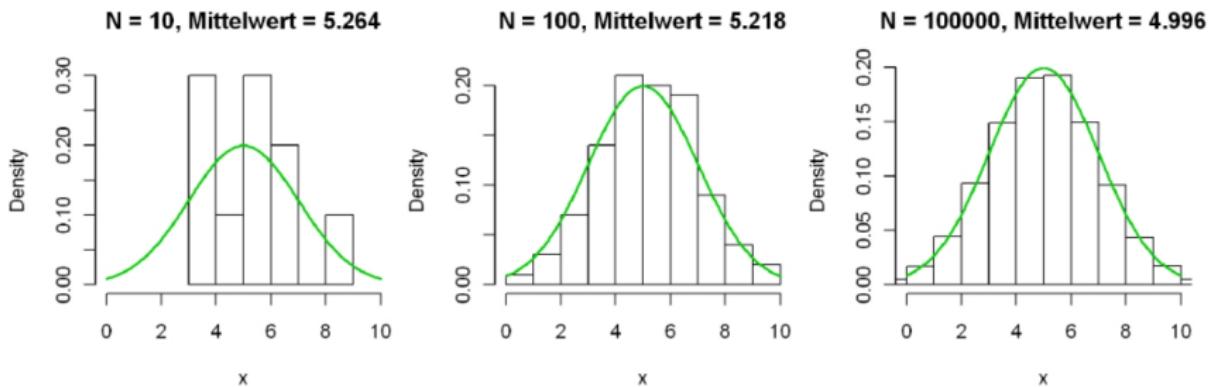
### Unterscheidung von Empirie und Theorie

- Bisher: reine Beschreibung von Lage, Streuung und Zusammenhang von Daten ohne Berücksichtigung ihrer Entstehung
- Jetzt: Interpretation der Daten als Realisationen von Zufallsvariablen und Beschreibung von deren Wahrscheinlichkeitsverteilungen
- Auf Basis dieser Wahrscheinlichkeitsverteilungen lassen sich dann Aussagen über nicht betrachtete oder zukünftige Daten machen

## 5.0 Beschreibung des Zufalls

Vergleich unterschiedlich großer Stichproben aus der gleichen Wahrscheinlichkeitsverteilung

Beispiel Normalverteilung: Gesetz der großen Zahlen



# 5.1 Mengentheoretische Grundlagen

## Elementare Begriffe

**Zufallsexperiment** Datenerhebungsprozess mit nicht vorhersagbarem Ausgang

**Ergebnis**  $\omega$  Elementarer Ausgang eines Zufallsexperiments

**Grundraum**  $\Omega$  Menge aller möglichen Ergebnisse  
 $\Omega = \{\omega | \omega \text{ ist Ergebnis des Zufallsexperiments}\}$

**Ereignis**  $A$  Menge von Ergebnissen, d.h. Teilmenge von  $\Omega$

**Elementarereignis** Einelementiges Ereignis

## 5.1 Mengentheoretische Grundlagen

### Beispiel

**Zufallsexperiment**      Einfacher Würfelwurf

**Ergebnisse**                 $\omega_1 = 1, \omega_2 = 2, \omega_3 = 3, \omega_4 = 4, \omega_5 = 5, \omega_6 = 6$

**Grundraum**                 $\Omega = \{1, 2, 3, 4, 5, 6\}$

**Ereignisse**                 $A = \{2, 4, 6\}, B = \{1, 3, 5\}, C = \{1, 2, 3, 4, 5\},$   
 $D = \{3, 4, 5, 6\}, E = \{2, 3, 5\}, F = \{1, 2, 3, 4, 5, 6\} = \Omega$

**Elementarereignisse**     $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$

# 5.1 Mengentheoretische Grundlagen

## Beispiele

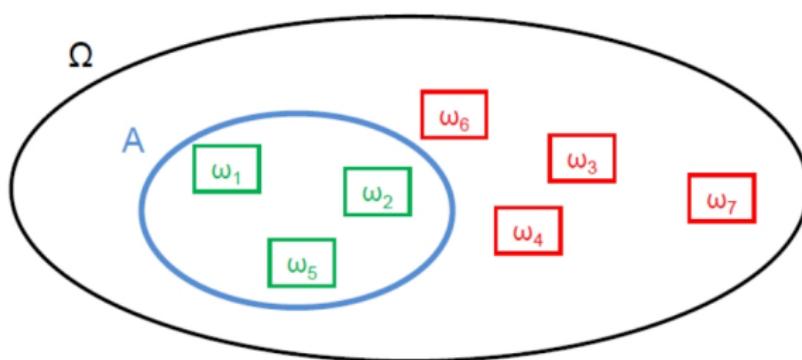
Experiment	Grundraum $\Omega$	Ergebnis $\omega$
Roulette	$\{0, 1, \dots, 36\}$	Zahlenfeld der Kugel
Würfeln: Warten auf 6	$\mathbb{N} \cup \{\infty\}$	Anzahl Würfe bis zur ersten 6
6 aus 49	$\{(\omega_1, \dots, \omega_6) \mid 1 \leq \omega_1 < \dots < \omega_6 \leq 49\}$	Geordnete Nummern der gezogenen Kugeln
Einzelne Serveranfrage	$[t_{\min}, t_{\max}]$	Anfragezeitpunkt $t$
Mausaktivität	$\{\omega : [t_{\min}, t_{\max}] \rightarrow (1, \dots, 600) \times (1, \dots, 800) \times \{0, 1, 2\}\}$	Koordinaten und Clickzustand (nicht, links, rechts) des Mauszeigers zu jeder Zeit
Wartezeit bis zur nächsten Serveranfrage	$[0, \infty)$	Zeit zwischen zwei Anfragen

# 5.1 Mengentheoretische Grundlagen

## Bezeichnungen

$$\omega \in A$$

Ergebnis  $\omega$  ist im Ereignis  $A$  enthalten

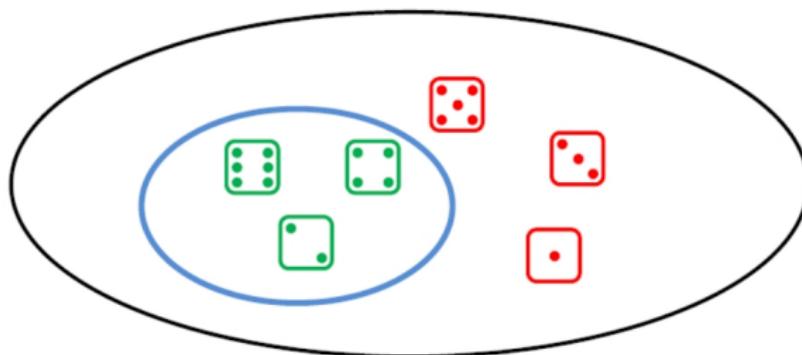


## 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

$$2 \in \{2, 4, 6\}$$

Augenzahl 2 ist gerade Zahl



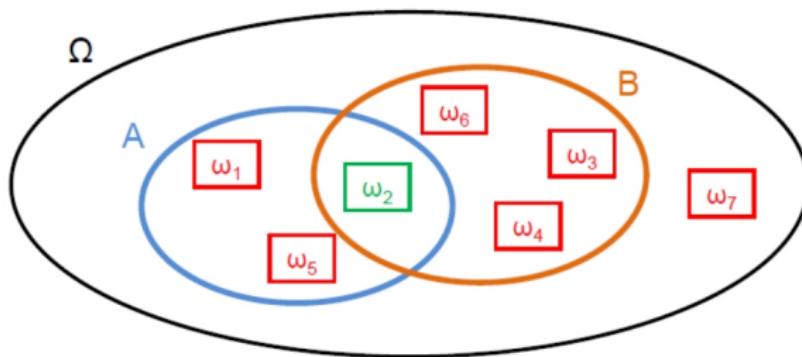
# 5.1 Mengentheoretische Grundlagen

## Bezeichnungen

### Schnittereignis zweier Mengen

$$A \cap B = \{\omega \in \Omega \mid \omega \in A \text{ und } \omega \in B\}$$

Ergebnis  $\omega$  ist in Ereignis A **und** Ereignis B enthalten



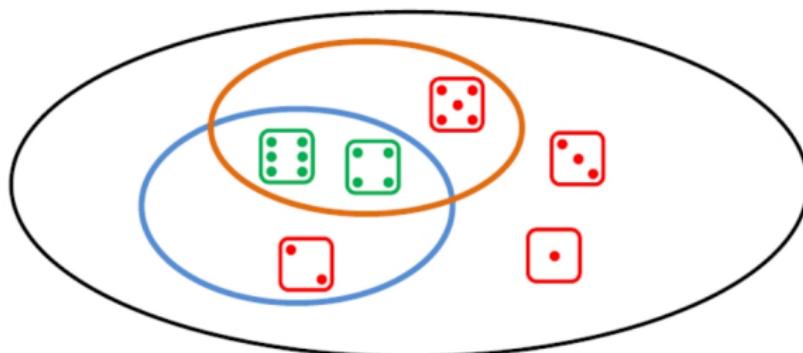
## 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

**Schnittereignis** zweier Mengen

$$4 \in \{2, 4, 6\} \cap \{4, 5, 6\}$$

Augenzahl 4 ist gerade Zahl und größer als 3



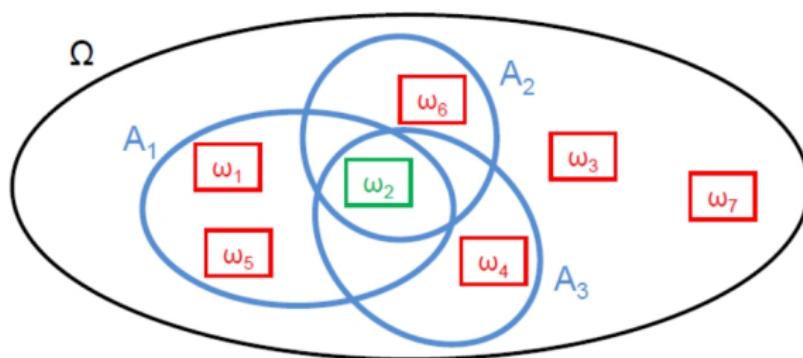
# 5.1 Mengentheoretische Grundlagen

## Bezeichnungen

**Schnittereignis** beliebig vieler Mengen

$$\bigcap_{i \in I} A_i = \{\omega \in \Omega \mid \omega \in A_i \text{ für } i \in I\}$$

Ergebnis  $\omega$  ist in allen Ereignissen  $A_i$  enthalten



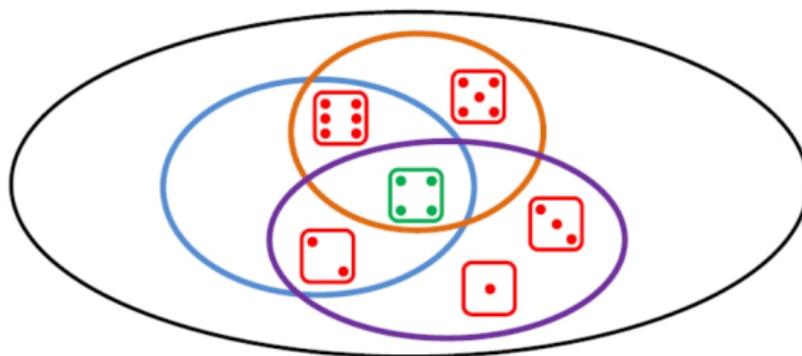
## 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

**Schnittereignis** beliebig vieler Mengen

$$4 \in \{2, 4, 6\} \cap \{4, 5, 6\} \cap \{1, 2, 3, 4\}$$

Augenzahl 4 ist **gerade Zahl** und **größer als 3** und **kleiner als 5**



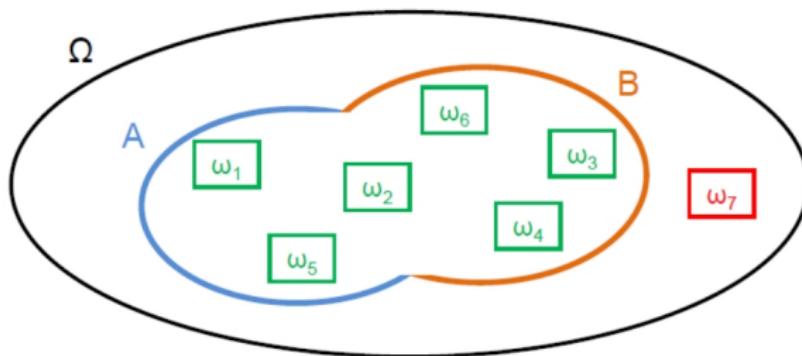
# 5.1 Mengentheoretische Grundlagen

## Bezeichnungen

**Vereinigungseignis** zweier Mengen

$$A \cup B = \{\omega \in \Omega | \omega \in A \text{ und/oder } \omega \in B\}$$

Ergebnis  $\omega$  ist in Ereignis A **und/oder** Ereignis B enthalten



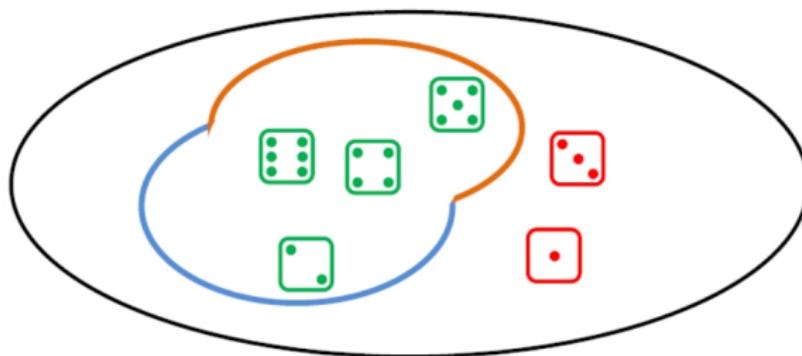
## 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

**Vereinigungssereignis** zweier Mengen

$$2 \in \{2, 4, 6\} \cup \{4, 5, 6\}$$

Augenzahl 2 ist gerade Zahl



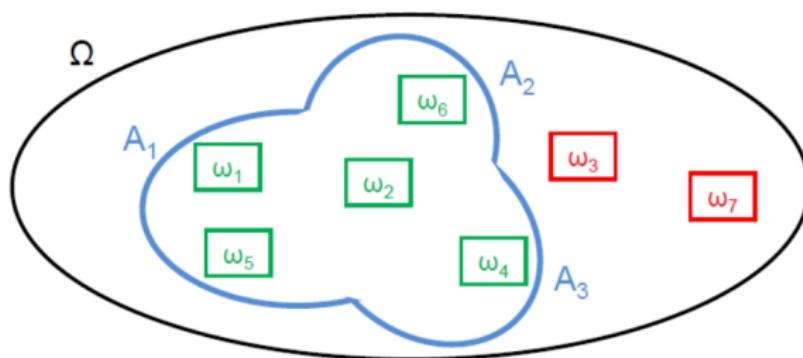
# 5.1 Mengentheoretische Grundlagen

## Bezeichnungen

**Vereinigungseignis** beliebig vieler Mengen

$$\bigcup_{i \in I} = \{\omega \in \Omega \mid \omega \in A_i \text{ für mindestens ein } i \in I\}$$

Ergebnis  $\omega$  ist in mindestens einem  $A_i$  enthalten



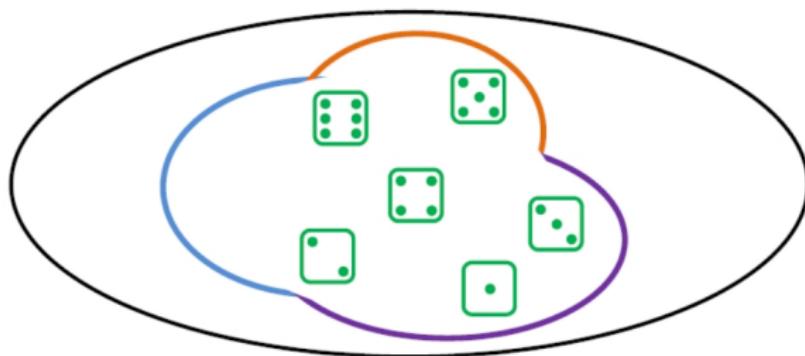
## 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

**Vereinigungseignis** beliebig vieler Mengen

$$5 \in \{2, 4, 6\} \cup \{4, 5, 6\} \cup \{1, 2, 3, 4\}$$

Augenzahl 5 ist größer als 3



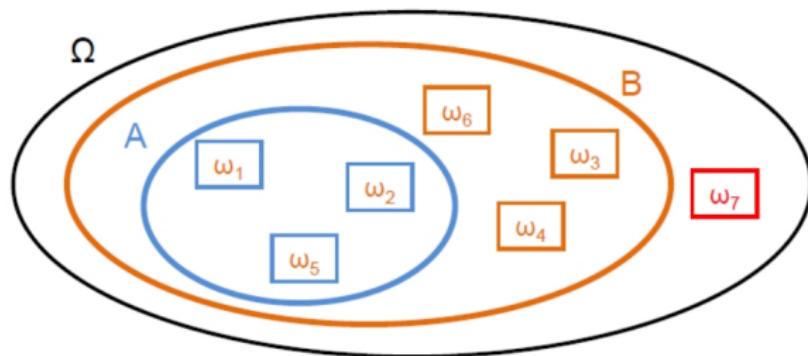
# 5.1 Mengentheoretische Grundlagen

## Bezeichnungen

### Teilereignis

$$A \subset B \text{ (bzw. } A \subseteq B)$$

Ereignis A ist in Ereignis B enthalten, aus Ereignis A folgt Ereignis B



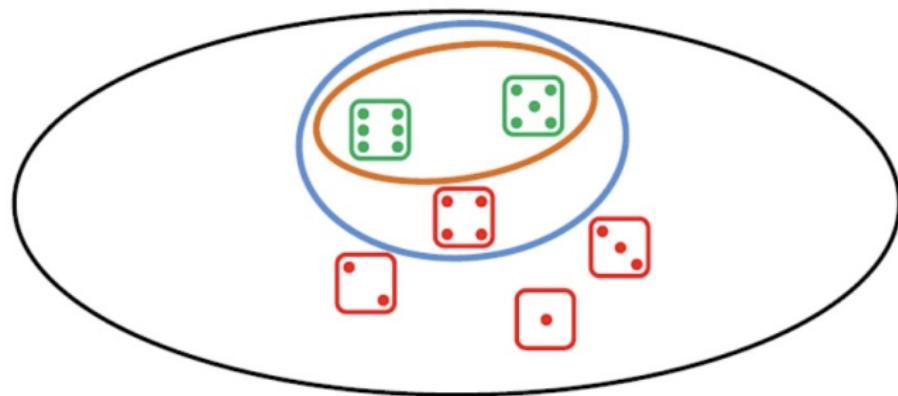
## 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

**Teilereignis**

$$\{5, 6\} \subset \{4, 5, 6\}$$

Augenzahl 5 ist größer als 4 und damit auch größer als 3.



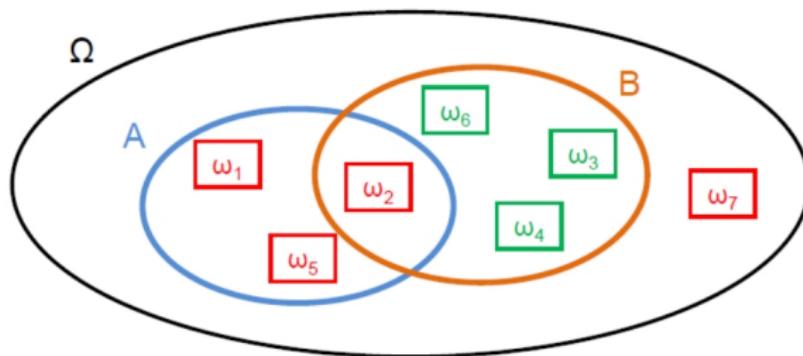
# 5.1 Mengentheoretische Grundlagen

## Bezeichnungen

### Differenzereignis

$$B \setminus A = \{\omega \in \Omega \mid \omega \in B \text{ und } \omega \notin A\}$$

Ergebnis  $\omega$  ist in Ereignis  $B$ , aber nicht in Ereignis  $A$  enthalten



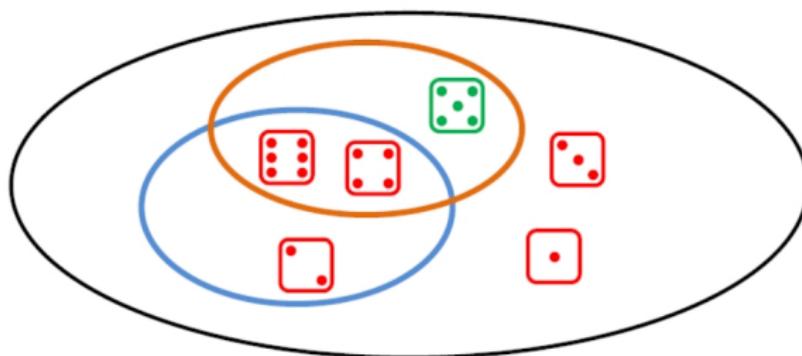
# 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

**Differenzereignis**

$$5 \in \{4, 5, 6\} \setminus \{2, 4, 6\}$$

Augenzahl 5 ist größer als 3, aber nicht gerade Zahl



# 5.1 Mengentheoretische Grundlagen

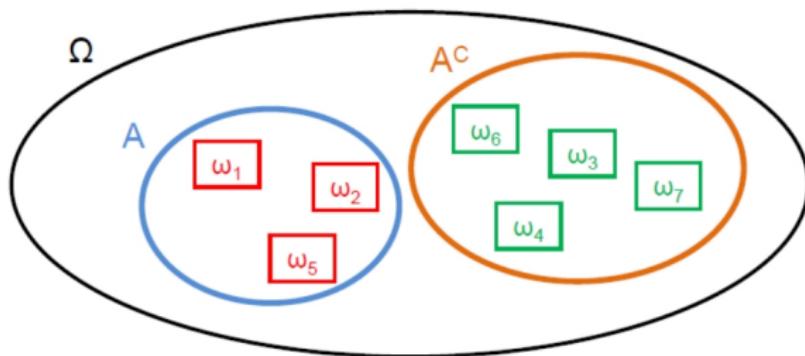
## Bezeichnungen

### Komplementärereignis

$$A^c = \Omega \setminus A = \{\omega \in \Omega | \omega \notin A\}$$

Ergebnis  $\omega$  ist in Ereignis  $A^c$  enthalten  $\Leftrightarrow$  Ergebnis  $\omega$  ist nicht in Ereignis  $A$  enthalten

Das Ereignis  $A^c$  heißt **Komplement** bzw. **Gegenereignis** von bzw. zu  $A$



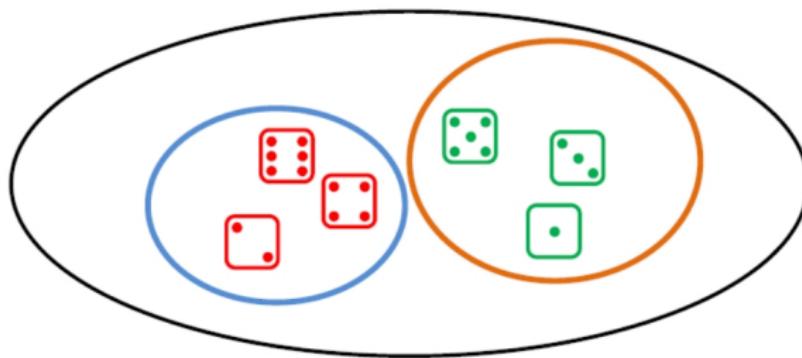
## 5.1 Mengentheoretische Grundlagen

Bezeichnungen: Beispiel Würfelwurf

### Komplementärereignis

$$5 \in \{2, 4, 6\}^c$$

Augenzahl 5 ist nicht gerade Zahl



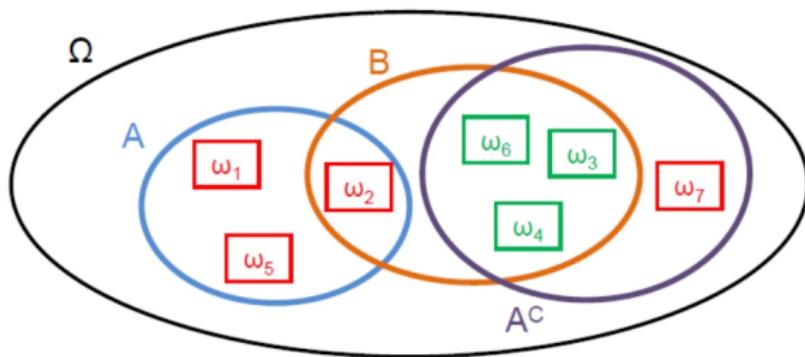
# 5.1 Mengentheoretische Grundlagen

## Regeln

### Differenzereignis und Komplementärereignis

$$B \setminus A = \{\omega \in \Omega \mid \omega \in B \text{ und } \omega \notin A\} = B \cap A^c = B \setminus (A \cap B)$$

Ergebnis  $\omega$  ist in Ereignis  $B$ , aber nicht in Ereignis  $A$  enthalten

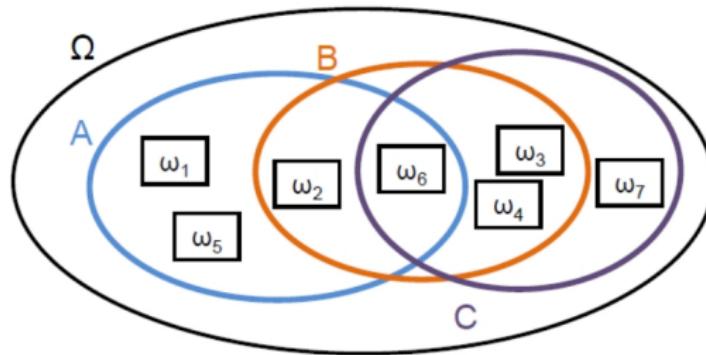


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

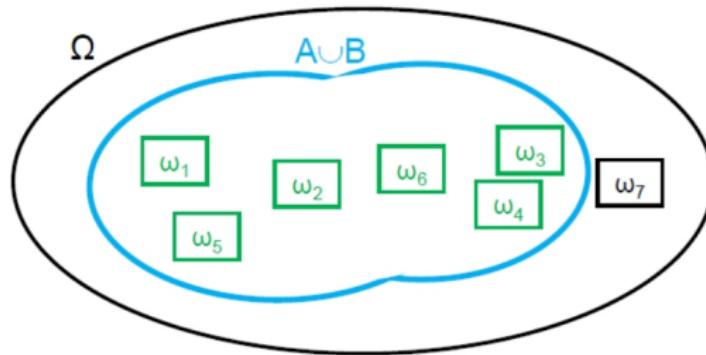


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

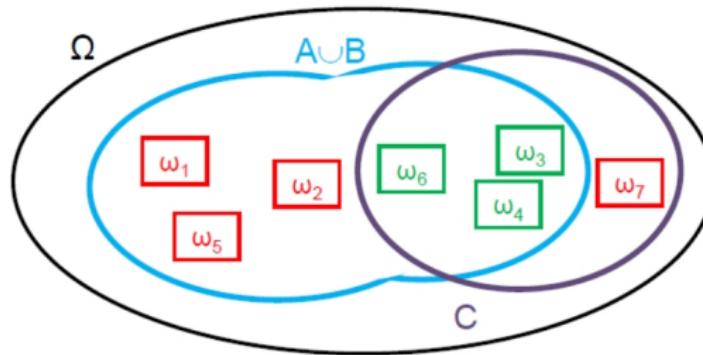


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

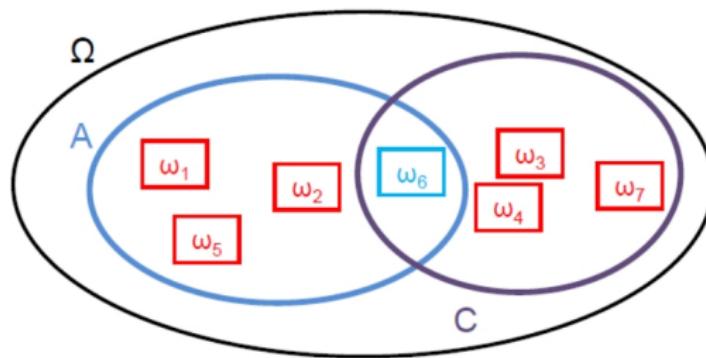


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

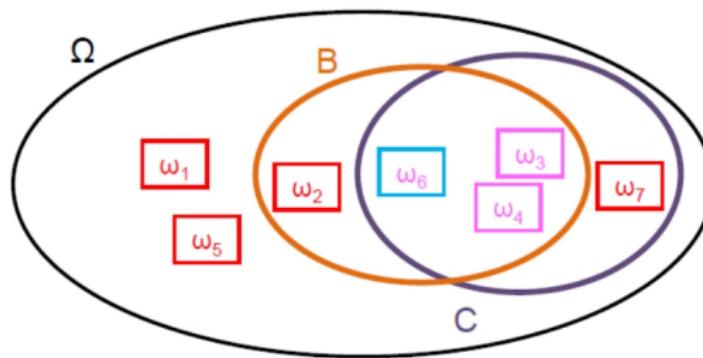


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

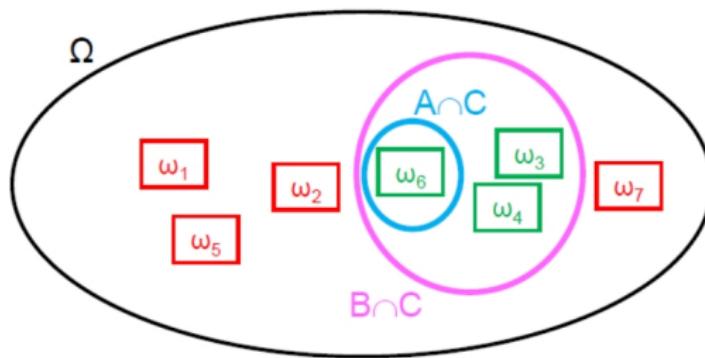


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

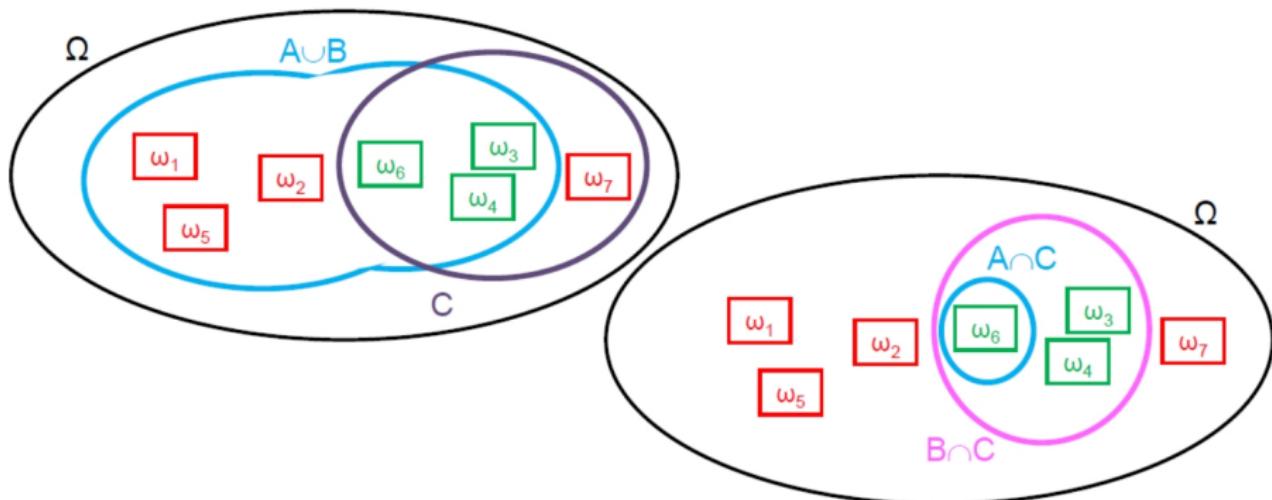


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

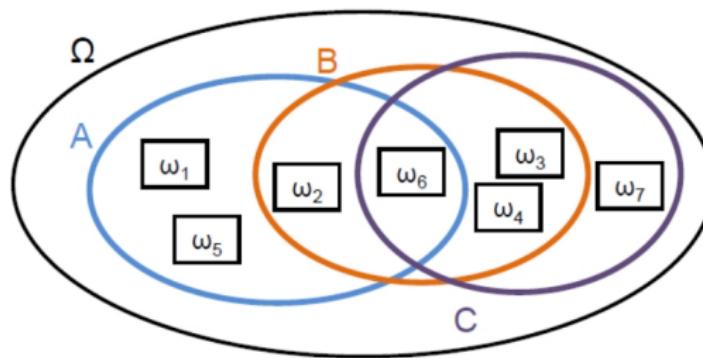


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

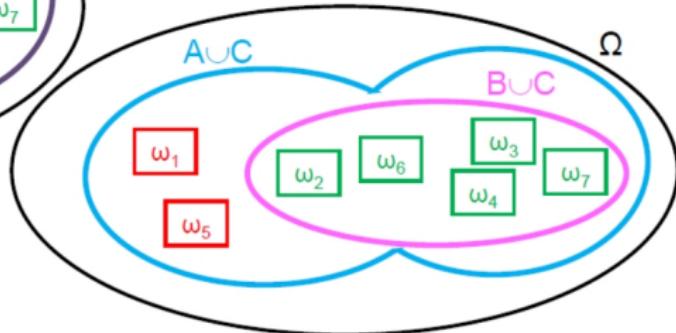
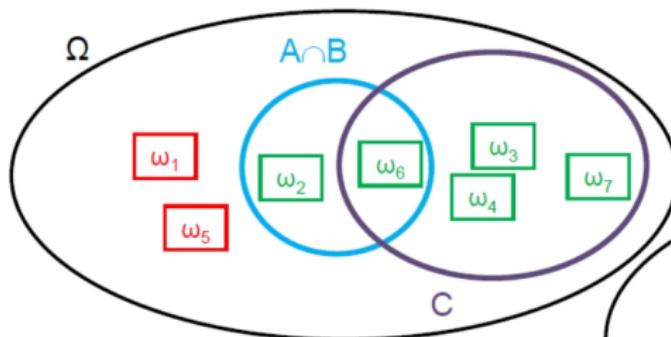


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Distributivgesetz

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

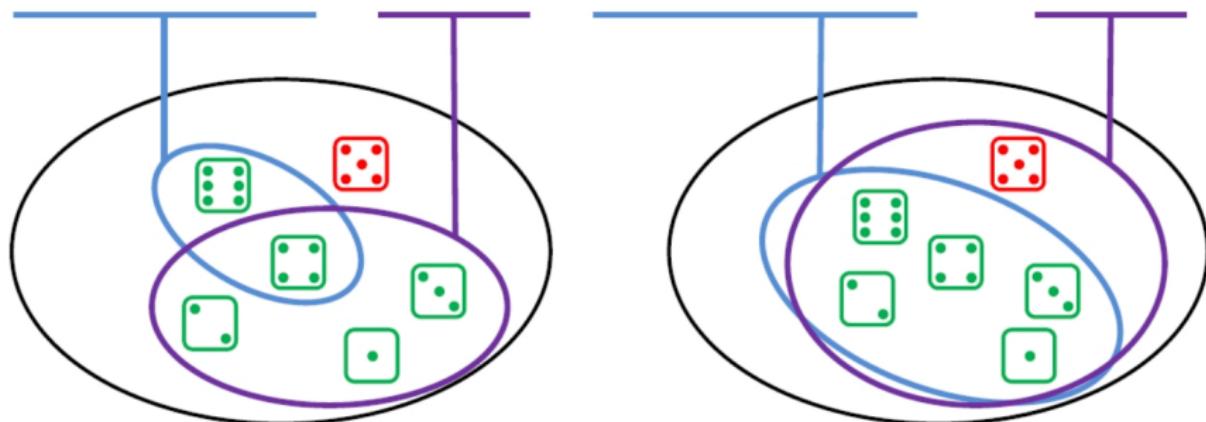


## 5.1 Mengentheoretische Grundlagen

Regeln: Beispiel Würfelwurf

**Distributivgesetz**

$$(\{2,4,6\} \cap \{4,5,6\}) \cup \{1,2,3,4\} = (\{2,4,6\} \cup \{1,2,3,4\}) \cap (\{4,5,6\} \cup \{1,2,3,4\})$$

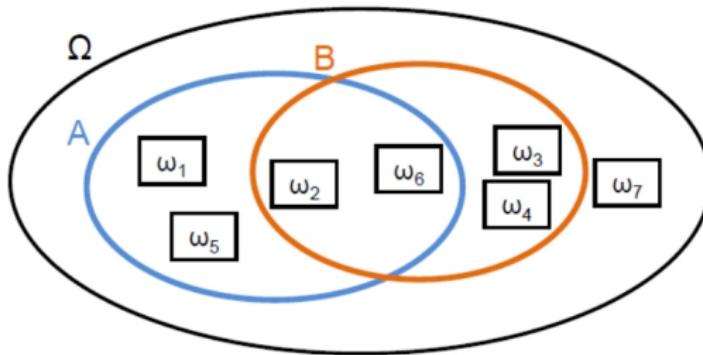


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Regeln von de Morgan

$$(A \cap B)^c = A^c \cup B^c , \quad (A \cup B)^c = A^c \cap B^c$$

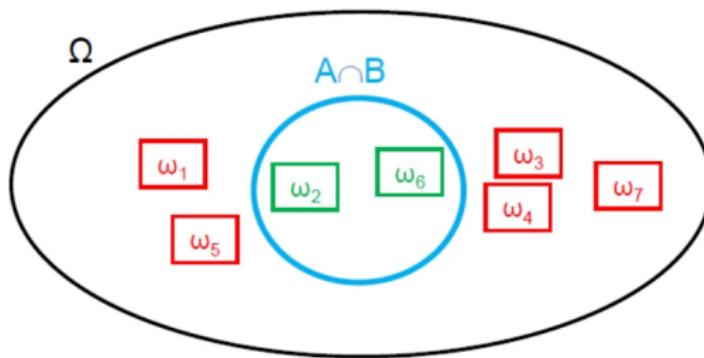


## 5.1 Mengentheoretische Grundlagen

### Regeln

#### Regeln von de Morgan

$$(A \cap B)^c = A^c \cup B^c$$

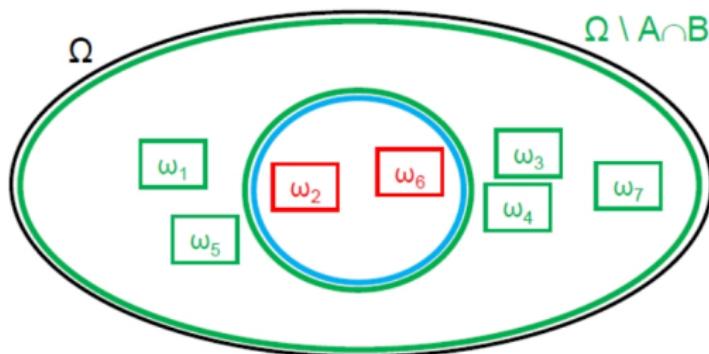


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Regeln von de Morgan

$$(A \cap B)^c = A^c \cup B^c$$

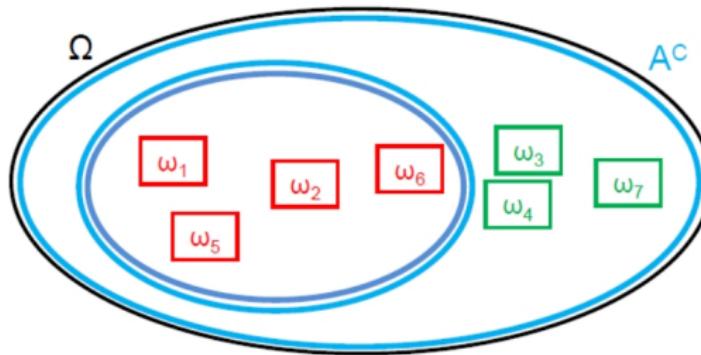


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Regeln von de Morgan

$$(A \cap B)^c = A^c \cup B^c$$

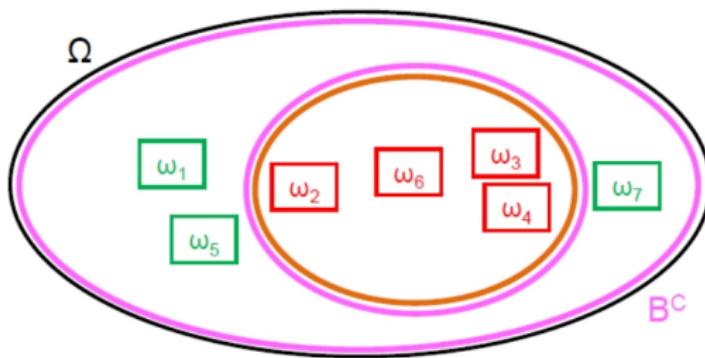


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Regeln von de Morgan

$$(A \cap B)^c = A^c \cup B^c$$

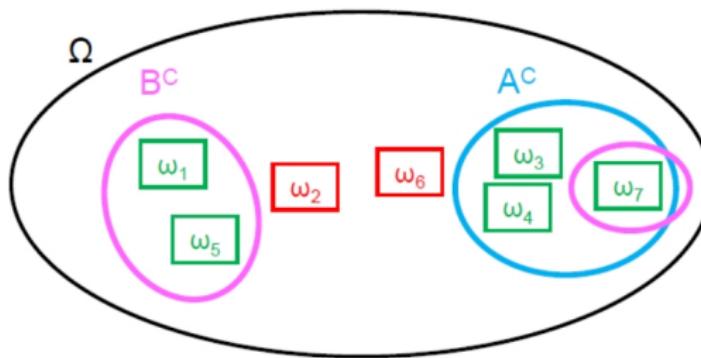


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Regeln von de Morgan

$$(A \cap B)^c = A^c \cup B^c$$

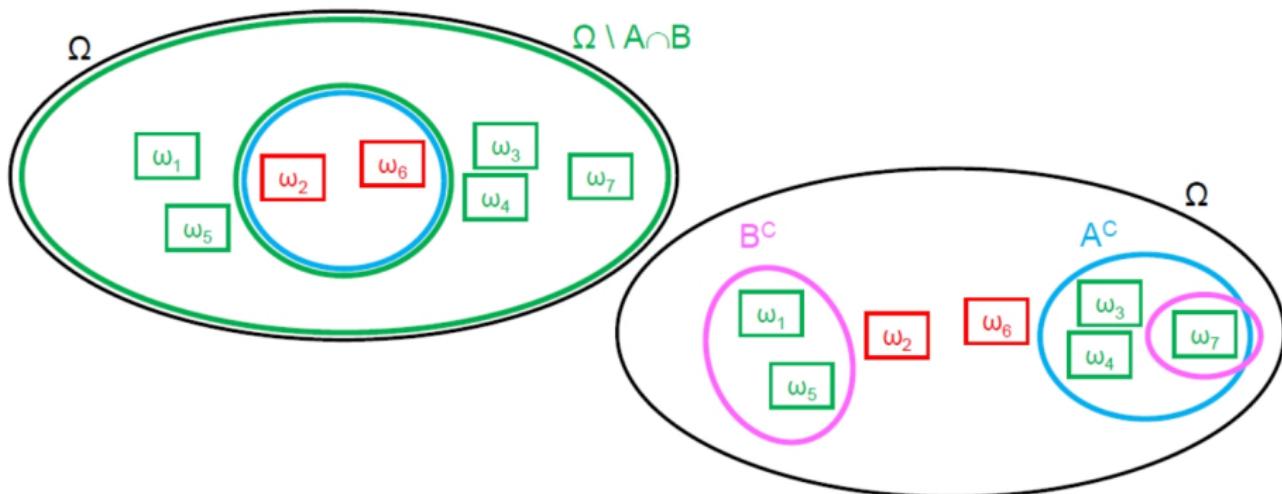


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Regeln von de Morgan

$$(A \cap B)^c = A^c \cup B^c$$

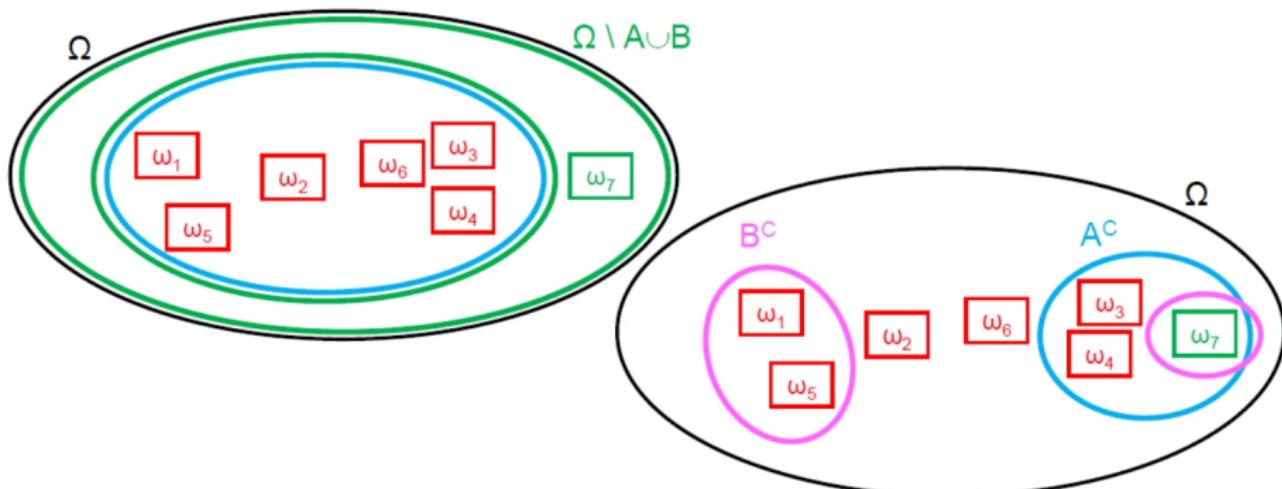


# 5.1 Mengentheoretische Grundlagen

## Regeln

### Regeln von de Morgan

$$(A \cup B)^c = A^c \cap B^c$$



## 5.1 Mengentheoretische Grundlagen

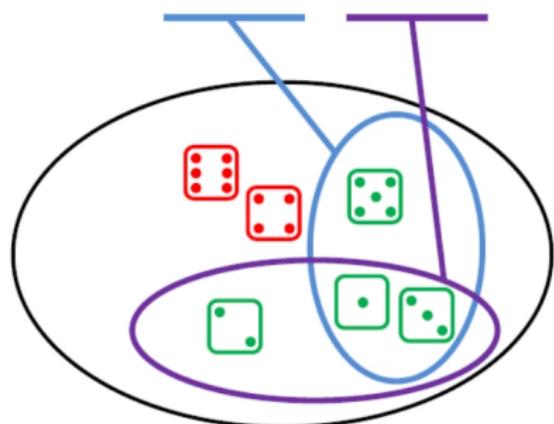
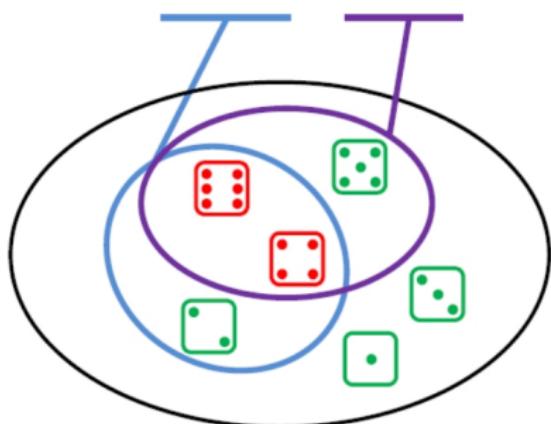
Regeln: Beispiel Würfelwurf

**Regeln von de Morgan**

$$(\{2,4,6\} \cap \{4,5,6\})^c$$

=

$$\{2,4,6\}^c \cup \{4,5,6\}^c$$



# 5.1 Mengentheoretische Grundlagen

## Beispiel Mausaktivität

### Ereignisbeispiele

$$A \supset \{\omega_1, \omega_3, \omega_4\}$$

„Letzter Click auf LU“

$$B \supset \{\omega_3, \omega_4\}$$

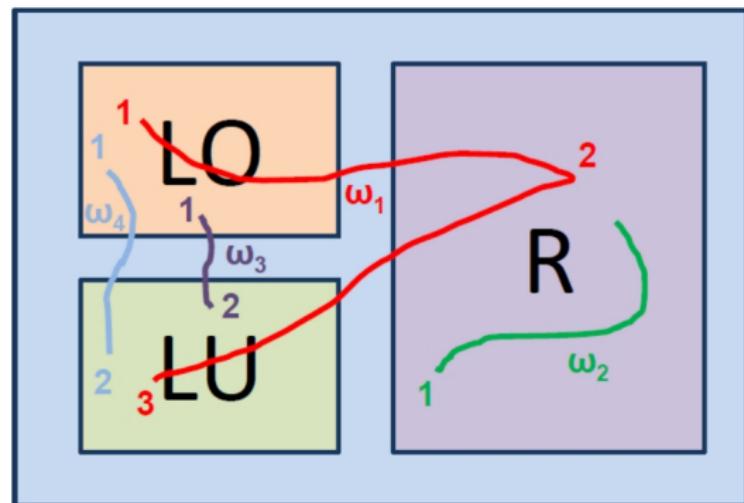
„Mauszeiger immer in linker Hälfte“

$$C \supset \{\omega_2\}$$

„Nur einmal geklickt“

$$D \supset \{\omega_1, \omega_2\}$$

„R wurde geklickt“



# 5.1 Mengentheoretische Grundlagen

## Zusammenfassung Bezeichnungen

Mathematische Schreibweise	Ausformulierte Schreibweise
$\omega \in A$	Ergebnis $\omega$ ist in Ereignis $A$ enthalten
$A \cap B$	<b>Schnittereignis:</b> Menge aller Ergebnisse, die in $A$ und $B$ enthalten sind
$A \cap B = \emptyset$	$A$ und $B$ sind <b>disjunkt:</b> es gibt kein Ergebnis, das in $A$ und $B$ enthalten ist
$A \cup B$	<b>Vereinigungsergebnis:</b> Menge aller Ergebnisse, die in $A$ <b>und/oder</b> $B$ enthalten sind
$A \subseteq B$	$A$ ist <b>Teilereignis</b> von $B$ : Alle in $A$ enthaltenen Ergebnisse sind auch in $B$ enthalten
$B \setminus A$	<b>Differenzereignis:</b> Menge der Ergebnisse, die in $B$ , aber nicht in $A$ enthalten sind
$A^c = \Omega \setminus A$	<b>Komplementärereignis:</b> Menge aller Ergebnisse, die nicht in $A$ enthalten sind

# 5.1 Mengentheoretische Grundlagen

## Zusammenfassung Regeln

Mathematische Schreibweise	Ausformulierte Schreibweise
$(A \cup B) \cap C = (A \cap B) \cup (B \cap C)$	<b>Distributivgesetze</b> Die Schnittmenge einer zwei Mengen $A$ und $B$ vereinigenden Menge mit einer weiteren Menge $C$ ist gleich der Vereinigung der beiden aus $C$ und jeweils einer der beiden Mengen $A$ und $B$ gebildeten Schnittmengen.
$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$	Die Vereinigung der Schnittmenge zweier Mengen $A$ und $B$ mit einer weiteren Menge $C$ ist gleich der Schnittmenge der beiden aus $C$ und jeweils einer der beiden Mengen $A$ und $B$ gebildeten Vereinigungen

# 5.1 Mengentheoretische Grundlagen

## Zusammenfassung Regeln

Mathematische Schreibweise	Ausformulierte Schreibweise
$(A \cap B)^c = A^c \cup B^c$	<b>Regeln von de Morgan</b> Das Komplementärereignis der Schnittmenge zweier Mengen ist gleich der Vereinigung der Komplementärereignisse der zwei Mengen.
$(A \cup B)^c = A^c \cap B^c$	Das Komplementärereignis der Vereinigung zweier Mengen ist gleich der Schnittmenge der Komplementärereignisse der zwei Mengen

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

Zuletzt: Interpretation von Daten als Realisationen von Zufallsvariablen.

Mengentheoretische Grundlagen zur Ordnung von Ergebnissen und Ereignissen

Ergebnis und Ereignis $\omega \in A$	Teilereignis $A \subseteq B$	Vereinigungseignis $A \cup B$
Schnittereignis $A \cap B$	Differenzereignis $B \setminus A$	Distributivgesetze $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
Disjunkte Ereignisse $A \cap B = \emptyset$	Komplementärereignis $A^c = \Omega \setminus A$	Regeln von de Morgan $(A \cap B)^c = A^c \cup B^c$ $(A \cup B)^c = A^c \cap B^c$

Jetzt: Zuordnung von Wahrscheinlichkeiten zu Ergebnissen und Ereignissen

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Kolmogorov-Axiome, Wahrscheinlichkeitsmaß

Seien  $\Omega$  ein Grundraum und  $\mathcal{A}$  die Menge aller Ereignisse über  $\Omega$ . Dann heißt die Abbildung

$$P : \mathcal{A} \rightarrow [0, 1], \quad A \mapsto P(A),$$

ein **Wahrscheinlichkeitsmaß**, falls sie folgende Eigenschaften (**Kolmogorov-Axiome**) besitzt:

- ①  $0 \leq P(A) \leq 1$  für jedes Ereignis  $A \in \mathcal{A}$
- ②  $P(\Omega) = 1$
- ③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarweise disjunkten Ereignisse  $A_i \in \mathcal{A}$

Der Wert  $P(A)$  für ein Ereignis  $A$  heißt **Wahrscheinlichkeit** von  $A$ .

Das Tripel  $(\Omega, \mathcal{A}, P)$  heißt **Wahrscheinlichkeitsraum**

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

**Kolmogorov-Axiome, Wahrscheinlichkeitsmaß:** Beispiel Würfelwurf

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\mathcal{A} = \left\{ \emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \right. \\ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}, \{1, 6\}, \{2, 3\}, \{2, 4\}, \{2, 5\}, \{2, 6\}, \{3, 4\}, \{3, 5\}, \{3, 6\}, \\ \{4, 5\}, \{4, 6\}, \{5, 6\}, \{1\}, \\ \{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 2, 6\}, \{1, 3, 4\}, \{1, 3, 5\}, \{1, 3, 6\}, \{1, 4, 5\}, \{1, 4, 6\}, \{1, 5, 6\}, \\ \{2, 3, 4\}, \{2, 3, 5\}, \{2, 3, 6\}, \{2, 4, 5\}, \{2, 4, 6\}, \{2, 5, 6\}, \{3, 4, 5\}, \{3, 4, 6\}, \{3, 5, 6\}, \{4, 5, 6\}, \\ \{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \{1, 2, 3, 6\}, \{1, 2, 4, 5\}, \{1, 2, 4, 6\}, \{1, 2, 5, 6\}, \{1, 2, 5, 6\}, \{1, 3, 4, 5\}, \\ \{1, 3, 4, 6\}, \{1, 3, 5, 6\}, \{1, 4, 5, 6\}, \{2, 3, 4, 5\}, \{2, 3, 4, 6\}, \{2, 4, 5, 6\}, \{3, 4, 5, 6\}, \\ \{1, 2, 3, 4, 5\}, \{1, 2, 3, 4, 6\}, \{1, 2, 4, 5, 6\}, \{1, 3, 4, 5, 6\}, \{2, 3, 4, 5, 6\}, \\ \left. \{1, 2, 3, 4, 5, 6\} \right\}$$

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = 1/6$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

**Kolmogorov-Axiome, Wahrscheinlichkeitsmaß:** Beispiel Würfelwurf  
(Fortsetzung)

①  $0 \leq P(A) \leq 6/6 = 1$  für alle  $A \in \mathcal{A}$

②  $P(\Omega) = 1$

③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarw. disj. Ereignisse  $A_i \in \mathcal{A}$ ,

insb. für  $A_i = \{i\}$ ,  $i = 1, \dots, 6$ ,  $A_i = \emptyset$ ,  $i > 6$ :  $P\left(\bigcup_{i=1}^{\infty} \{i\}\right) = P\left(\bigcup_{i=1}^6 \{i\}\right)$

$$= P(\Omega) = \sum_{i=1}^6 P(i) = 6/6 = 1$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Kolmogorov-Axiome, Wahrscheinlichkeitsmaß:

Beispiel Mausaktivität, interpolierte x-Position des Mauszeigers zu stetiger Zeit  $t$

$$\Omega = [1, 800]$$

$$\mathcal{A} = \left\{ \emptyset, \{(a, b) | a \in \Omega, b \in \Omega, a \leq b\}, \left\{ \bigcup_{c=1}^2 (a_c, b_c) | a_c \in \Omega, b_c \in \Omega, a_c \leq b_c \right\}, \dots, \right.$$
$$\left. \left\{ \bigcup_{c=1}^{\infty} (a_c, b_c) | a_c \in \Omega, b_c \in \Omega, a_c \leq b_c \right\}, \{[a, b]|\dots\}, \dots \{([a, b]|\dots\}, \{([a, b]|\dots\}, \dots \right\}$$

$$P([a, b]) = (b - a)/799, \quad a \leq b$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Kolmogorov-Axiome, Wahrscheinlichkeitsmaß:

Beispiel Mausaktivität, interpolierte  $x$ -Position des Mauszeigers zu stetiger Zeit  $t$  (Fortsetzung)

①  $0 \leq P(A) \leq 799/799 = 1$  für alle  $A \in \mathcal{A}$

②  $P(\Omega) = 1$

③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarw. disj. Ereignisse  $A_i \in \mathcal{A}$ ,

z.B. für

$$A_1 = 1, A_2 = (1, 400), A_3 = 400, A_4 = (400, 800), A_5 = 800, A_i = \emptyset, i > 5 :$$

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i\right) &= P(1 \cup (1, 400) \cup 400 \cup (400, 800) \cup 800) = P(\Omega) = \sum_{i=1}^5 P(i) \\ &= \frac{0+399+0+400+0}{799} = 1 \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

$P : \mathcal{A} \rightarrow [0, 1], \quad A \mapsto P(A)$

- ①  $0 \leq P(A) \leq 1$  für jedes Ereignis  $A \in \mathcal{A}$
- ②  $P(\Omega) = 1$
- ③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarweise disjunkten Ereignisse  $A_i \in \mathcal{A}$

**(i)**  $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Beweis:

Setze  $A_1 = A, A_2 = B, A_i = \emptyset$ , für  $i > 2$

$$\begin{aligned}
 P(A \cup B) &= P\left(\bigcup_{i=1}^{\infty} A_i\right) \stackrel{3.}{=} \sum_{i=1}^{\infty} P(A_i) = P(A_1) + \sum_{i=2}^{\infty} P(A_i) \\
 &= P(A_1) + \left(\bigcup_{i=2}^{\infty} A_i\right) = \boxed{P(A) + P(B)} \quad \square
 \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

$P : \mathcal{A} \rightarrow [0, 1], \quad A \mapsto P(A)$

- ①  $0 \leq P(A) \leq 1$  für jedes Ereignis  $A \in \mathcal{A}$
- ②  $P(\Omega) = 1$
- ③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarweise disjunkten Ereignisse  $A_i \in \mathcal{A}$

(ii)  $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$

Beweis:

$$P(B) = P((B \setminus A) \cup A) \stackrel{(i)}{=} P(B \setminus A) + P(A) \Rightarrow P(B \setminus A) = P(B) - P(A) \quad \square$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

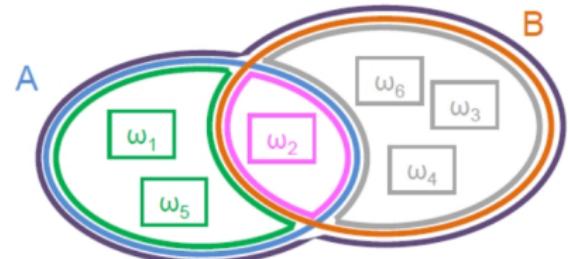
$P : \mathcal{A} \rightarrow [0, 1], \quad A \mapsto P(A)$

- ①  $0 \leq P(A) \leq 1$  für jedes Ereignis  $A \in \mathcal{A}$
- ②  $P(\Omega) = 1$
- ③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarweise disjunkten Ereignisse  $A_i \in \mathcal{A}$

$$\text{(iii)} \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Beweis:

$$A \cup B = [A \setminus (A \cap B)] \cup [B \setminus (A \cap B)] \cup [A \cap B]$$



## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

Beweis: (Fortsetzung)

$$A \cup B = [A \setminus (A \cap B)] \cup [B \setminus (A \cap B)] \cup [A \cap B]$$

$$\Rightarrow P([A \cup B]) = P([A \setminus (A \cap B)] \cup [B \setminus (A \cap B)] \cup [A \cap B])$$

$$\stackrel{(i)}{=} P([A \setminus (A \cap B)]) + P([B \setminus (A \cap B)]) + P(A \cap B)$$

$$\stackrel{(ii)}{=} P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B)$$

$$= P(A) + P(B) - P(A \cap B) \quad \square$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

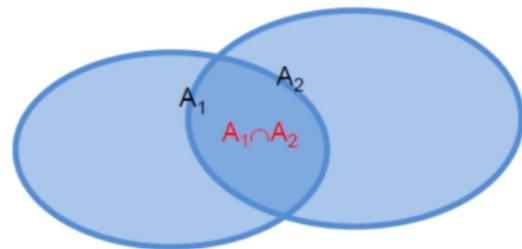
(iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(iv) **Poincaré-Sylvesterformel**

$$P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$$

Am Beispiel  $N = 2$ :

$$\begin{aligned} P(A_1 \cup A_2) &= (-1)^{1+1} \cdot P(A_1) + (-1)^{1+1} \cdot P(A_2) + (-1)^{2+1} \cdot P(A_1 \cap A_2) \\ &= \boxed{P(A_1)} + \boxed{P(A_2)} - \boxed{P(A_1 \cap A_2)} \end{aligned}$$



## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

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

(iv) **Poincaré-Sylvesterformel**

$$P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$$

Für  $N = 3$ :

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= (-1)^{1+1} \cdot P(A_1) + (-1)^{1+1} \cdot P(A_2) + (-1)^{1+1} \cdot P(A_3) \\ &\quad + (-1)^{2+1} \cdot P(A_1 \cap A_2) + (-1)^{2+1} \cdot P(A_1 \cap A_3) + (-1)^{2+1} \cdot P(A_2 \cap A_3) \\ &\quad + (-1)^{3+1} \cdot P(A_1 \cap A_2 \cap A_3) \\ &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &\quad + P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

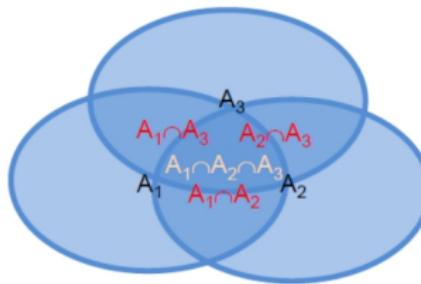
(iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(iv) **Poincaré-Sylvesterformel**

$$P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$$

Für  $N = 3$ :  $P(A_1 \cup A_2 \cup A_3) =$

$$\boxed{P(A_1)} + \boxed{P(A_2)} + \boxed{P(A_3)} - \boxed{P(A_1 \cap A_2)} - \boxed{P(A_1 \cap A_3)} - \boxed{P(A_2 \cap A_3)} \\ + \boxed{P(A_1 \cap A_2 \cap A_3)}$$



## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

$P : \mathcal{A} \rightarrow [0, 1], \quad A \mapsto P(A)$

- ①  $0 \leq P(A) \leq 1$  für jedes Ereignis  $A \in \mathcal{A}$
- ②  $P(\Omega) = 1$
- ③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarweise disjunkten Ereignisse  $A_i \in \mathcal{A}$

$$\text{(v)} \quad P(A^c) = 1 - P(A)$$

Beweis:

$$P(A^c) = P(\Omega \setminus A) \stackrel{(ii)}{=} P(\Omega) - P(A) = \boxed{1 - P(A)} \quad \square$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

$P : \mathcal{A} \rightarrow [0, 1], \quad A \mapsto P(A)$

- ①  $0 \leq P(A) \leq 1$  für jedes Ereignis  $A \in \mathcal{A}$
- ②  $P(\Omega) = 1$
- ③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarweise disjunkten Ereignisse  $A_i \in \mathcal{A}$

**(vi)**  $P(\emptyset) = 0$

Beweis:

$$P(\emptyset) = P(\Omega^c) \stackrel{(v)}{=} 1 - P(\Omega) = 0 \quad \square$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

$P : \mathcal{A} \rightarrow [0, 1]$ ,  $A \mapsto P(A)$

- ①  $0 \leq P(A) \leq 1$  für jedes Ereignis  $A \in \mathcal{A}$
- ②  $P(\Omega) = 1$
- ③  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  für alle paarweise disjunkten Ereignisse  $A_i \in \mathcal{A}$

**(vii)**  $A \subseteq B \Rightarrow P(A) \leq P(B)$

Beweis:

$$\begin{aligned} A \subseteq B &\Rightarrow P(B \setminus A) = P(B) - P(A) \\ &\Rightarrow P(A) = P(B) - \underbrace{P(B \setminus A)}_{\geq 0} \leq P(B) \quad \square \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Diskreter Wahrscheinlichkeitsraum

Seien  $\Omega = \{\omega_1, \omega_2, \dots\}$  ein **endlicher** oder **abzählbarer unendlicher** Grundraum und  $P$  ein Wahrscheinlichkeitsmaß auf  $\Omega$ . Dann heißt  $(\Omega, \mathcal{A}, P)$  **diskreter Wahrscheinlichkeitsraum**.

Für beliebiges Ereignis  $A \in \mathcal{A}$  gilt dann nach (i):

$$P(A) = P\left(\bigcup_{i:\omega_i \in A} \{\omega_i\}\right) = \sum_{i:\omega_i \in A} P(\{\omega_i\})$$

### Laplace-Raum

Treten die Elemente von endlichem  $\Omega = \{\omega_1, \dots, \omega_{|\Omega|}\}$  aus einem diskreten Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  alle mit der selben Wahrscheinlichkeit auf, d.h. gilt  $P(\{\omega_i\}) = 1/|\Omega|$  für  $i = 1, \dots, |\Omega|$ , so wird  $(\Omega, \mathcal{A}, P)$  auch **Laplace-Raum** genannt und die Wahrscheinlichkeit für ein Ereignis  $A \in \mathcal{A}$  kann durch  $P(A) = |A|/|\Omega|$  angegeben werden.

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen; Diskreter

**Wahrscheinlichkeitsraum:** Beispiel: *Bearbeitungen von Softwareaufgaben*

Bearbeitung	Bearbeiter(in)	Aufgabe	Version
e <sub>1</sub>	Kai	Export	1.1
e <sub>2</sub>	Kai	Verknüpfung	1.2
e <sub>3</sub>	Miriam	Export	1.1
e <sub>4</sub>	Tina	Verknüpfung	1.2
e <sub>5</sub>	Oliver	Export	2.0
e <sub>6</sub>	Tina	Export	1.2
e <sub>7</sub>	Tina	Verknüpfung	1.2
e <sub>8</sub>	Miriam	Export	1.2
e <sub>9</sub>	Miriam	Export	1.2
e <sub>10</sub>	Oliver	Abfrage	1.1
e <sub>11</sub>	Oliver	Verknüpfung	2.0
e <sub>12</sub>	Oliver	Abfrage	2.0

Zufällige Auswahl einer Bearbeitung

→ Ergebnis  $\omega \in \{e_1, \dots, e_{12}\} = \Omega$

Elementarwahrscheinlichkeiten

$$P(\{e_i\}) = 1/12, i = 1, \dots, 12$$

Ereignisse

- ① Bearbeiter männlich

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\}$$

- ② Gestellte Aufgabe Export

$$A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\}$$

- ③ Verwendete Version 2.0

$$A_3 = \{e_5, e_{11}, e_{12}\}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

- (i)  $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$       (v)  $P(A^c) = 1 - P(A)$
- (ii)  $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$       (vi)  $P(\emptyset) = 0$
- (iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$       (vii)  $A \subseteq B \Rightarrow P(A) \leq P(B)$
- (iv)  $P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$

$$\omega \in \{e_1, \dots, e_{12}\} = \Omega$$

$$P(\{e_i\}) = 1/12, i = 1, \dots, 12$$

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\}$$

$$A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\}$$

$$A_3 = \{e_5, e_{11}, e_{12}\}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

- (i)  $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- (ii)  $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$
- (iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- (iv)  $P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$
- (v)  $P(A^c) = 1 - P(A)$
- (vi)  $P(\emptyset) = 0$
- (vii)  $A \subseteq B \Rightarrow P(A) \leq P(B)$

$$\omega \in \{e_1, \dots, e_{12}\} = \Omega$$

$$P(\{e_i\}) = 1/12, i = 1, \dots, 12$$

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\}$$

$$A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\}$$

$$A_3 = \{e_5, e_{11}, e_{12}\}$$

$$\begin{aligned}
 P(A_1) &= P(\{e_1\} \cup \{e_2\} \cup \{e_5\} \cup \{e_{10}\} \cup \{e_{11}\} \cup \{e_{12}\}) \\
 &= P(\{e_1\}) + P(\{e_2\}) + P(\{e_5\}) \\
 &\quad + P(\{e_{10}\}) + P(\{e_{11}\}) + P(\{e_{12}\}) = 6/12 = 1/2 \\
 P(A_2) &= P(\{e_1\} \cup \{e_3\} \cup \{e_5\} \cup \{e_6\} \cup \{e_8\} \cup \{e_9\}) \\
 &= P(\{e_1\}) + P(\{e_3\}) + P(\{e_5\}) \\
 &\quad + P(\{e_6\}) + P(\{e_8\}) + P(\{e_9\}) = 6/12 = 1/2 \\
 P(A_3) &= P(\{e_5\} \cup \{e_{11}\} \cup \{e_{12}\}) \\
 &= P(\{e_5\}) + P(\{e_{11}\}) + P(\{e_{12}\}) = 3/12 = 1/4
 \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

$$(i) A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B) \quad (v) P(A^c) = 1 - P(A)$$

$$(ii) A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A) \quad (vi) P(\emptyset) = 0$$

$$(iii) P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (vii) A \subseteq B \Rightarrow P(A) \leq P(B)$$

$$(iv) P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$$

$$\omega \in \{e_1, \dots, e_{12}\} = \Omega \\ P\{\{e_i\}\} = 1/12, i = 1, \dots, 12$$

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\}$$

$$A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\}$$

$$A_3 = \{e_5, e_{11}, e_{12}\}$$

$$P(A_1) = 1/2$$

$$P(A_2) = 1/2$$

$$P(A_3) = 1/4$$

Wahrscheinlichkeit für eine Bearbeitung, die von einem Mann mit einer anderen Version als 2.0 durchgeführt wurde

$$A_3 = \{e_5, e_{11}, e_{12}\} \subset \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\} = A_1 \\ \Rightarrow (ii) P(A_1 \setminus A_3) = P(A_1) - P(A_3) = 1/2 - 1/4 = 1/4 \\ \Rightarrow (vii) 1/4 = P(A_3) \leq P(A_1) = 1/2$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

- (i)  $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- (ii)  $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$
- (iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- (v)  $P(A^c) = 1 - P(A)$
- (vi)  $P(\emptyset) = 0$
- (vii)  $A \subseteq B \Rightarrow P(A) \leq P(B)$

$$(iv) P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$$

$$\omega \in \{e_1, \dots, e_{12}\} = \Omega$$

$$P(\{e_i\}) = 1/12, i = 1, \dots, 12$$

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\}$$

$$A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\}$$

$$A_3 = \{e_5, e_{11}, e_{12}\}$$

$$P(A_1) = 1/2$$

$$P(A_2) = 1/2$$

$$P(A_3) = 1/4$$

Wahrscheinlichkeit für eine Bearbeitung, die Aufgabe Export hatte und/oder von einem Mann durchgeführt wurde

$$\begin{aligned}
 P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\
 &= 1/2 + 1/2 - P(\{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\} \cap \{e_1, e_3, e_5, e_6, e_8, e_9\}) \\
 &= 1 - P(\{e_1, e_5\}) = 1 - P(\{e_1\} \cup \{e_5\}) \\
 &= 1 - (P(\{e_1\}) + P(\{e_5\})) = 1 - 2/12 \\
 &= 10/12 = 5/6
 \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

- (i)  $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$     (v)  $P(A^c) = 1 - P(A)$
- (ii)  $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$     (vi)  $P(\emptyset) = 0$
- (iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$     (vii)  $A \subseteq B \Rightarrow P(A) \leq P(B)$

(iv)  $P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$

$$\omega \in \{e_1, \dots, e_{12}\} = \Omega$$

$$P(\{e_i\}) = 1/12, i = 1, \dots, 12$$

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\}$$

$$A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\}$$

$$A_3 = \{e_5, e_{11}, e_{12}\}$$

$$P(A_1) = 1/2$$

$$P(A_2) = 1/2$$

$$P(A_3) = 1/4$$

W'keit für eine Bearbeitung, die Aufgabe Export hatte und/oder von einem Mann und /oder mit Version 2.0 durchgeführt wurde

$$\begin{aligned}
 & P(A_1 \cup A_2 \cup A_3) \\
 &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) \\
 &\quad - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \\
 &= 1/2 + 1/2 - 1/4 - P(\{e_1, e_5\}) - P(\{e_5, e_{11}, e_{12}\}) \\
 &\quad - P(\{e_5\}) + P(\{e_5\}) \\
 &= 15/12 - 2/12 - 3/12 - 1/12 + 1/12 = 10/12 = 5/6
 \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

- (i)  $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- (ii)  $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$
- (iii)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- (iv)  $P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$
- (v)  $P(A^c) = 1 - P(A)$
- (vi)  $P(\emptyset) = 0$
- (vii)  $A \subseteq B \Rightarrow P(A) \leq P(B)$

$$\omega \in \{e_1, \dots, e_{12}\} = \Omega$$

$$P(\{e_i\}) = 1/12, i = 1, \dots, 12$$

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\}$$

$$A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\}$$

$$A_3 = \{e_5, e_{11}, e_{12}\}$$

$$P(A_1) = 1/2$$

$$P(A_2) = 1/2$$

$$P(A_3) = 1/4$$

W'keit für eine Bearbeitung, die weder Aufgabe Export hatte noch von einem Mann noch mit Version 2.0 durchgeführt wurde

$$\text{Mit (v): } P([A_1 \cup A_2 \cup A_3]^c) = 1 - P(A_1 \cup A_2 \cup A_3)$$

$$= 1 - 5/6 = 1/6$$

$$\begin{aligned} \text{Mit de Morgan: } P([A_1 \cup A_2 \cup A_3]^c) &= P(A_1^c \cap A_2^c \cap A_3^c) \\ &= P(\{e_3, e_4, e_6, e_7, e_8, e_9\} \cap \{e_2, e_4, e_7, e_{10}, e_{11}, e_{12}\}) \\ &\quad \cap \{e_1, e_2, e_3, e_4, e_6, e_7, e_8, e_9, e_{10}\}) \\ &= P(\{e_4, e_7\}) = 2/12 = 1/6 \end{aligned}$$

## 5.2 Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsräume

### Eigenschaften von Wahrscheinlichkeitsmaßen

$$(i) A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B) \quad (v) P(A^c) = 1 - P(A)$$

$$(ii) A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A) \quad (vi) P(\emptyset) = 0$$

$$(iii) P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (vii) A \subseteq B \Rightarrow P(A) \leq P(B)$$

$$(iv) P\left(\bigcup_{n=1}^N A_n\right) = \sum_{m=1}^N (-1)^{m+1} \sum_{1 \leq n_1 < \dots < n_m \leq N} P(A_{n_1} \cap \dots \cap A_{n_m})$$

$$\omega \in \{e_1, \dots, e_{12}\} = \Omega \\ P(\{e_i\}) = 1/12, i = 1, \dots, 12$$

$$A_1 = \{e_1, e_2, e_5, e_{10}, e_{11}, e_{12}\} \\ A_2 = \{e_1, e_3, e_5, e_6, e_8, e_9\} \\ A_3 = \{e_5, e_{11}, e_{12}\}$$

$$P(A_1) = 1/2 \\ P(A_2) = 1/2 \\ P(A_3) = 1/4$$

W'keit für eine Bearbeitung, die mit Version 2.0 von einer Frau durchgeführt wurde

$$P(A_1^c \cap A_3) = P(\{e_3, e_4, e_6, e_7, e_8, e_9\} \cap \{e_5, e_{11}, e_{12}\}) \\ = P(\emptyset) = 0$$

# Zufallsvariablen und deren Verteilung

## 6.0 Zufallsvariablen und deren Verteilung

### Erinnerung

<b>Zufallsexperiment</b>	Datenerhebungsprozess mit nicht vorhersagbarem Ausgang
<b>Ergebnis</b> $\omega$	Elementarer Ausgang eines Zufallsexperiments
<b>Grundraum</b> $\Omega$	Menge aller möglichen Ergebnisse $\Omega = \{\omega \mid \omega \text{ ist Ergebnis des Zufallsexperiments}\}$

### Zufallsvariable

Eine Abbildung, die jedem Ergebnis eines Zufallsexperiments eine reelle Zahl zuordnet, wird **Zufallsvariable (ZV)** genannt. Ein konkreter Wert  $x = X(\omega)$  heißt **Realisation** der Zufallsvariable  $X$ .

$$X : \Omega \rightarrow \mathbb{R} \quad \omega \mapsto X(\omega)$$

## 6.0 Zufallsvariablen und deren Verteilung

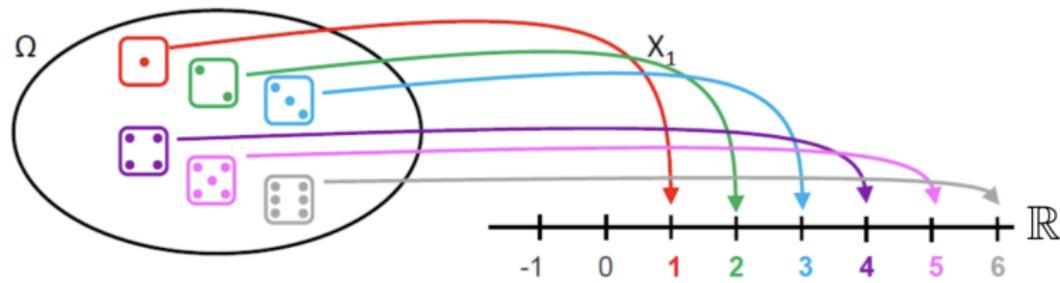
### Zufallsvariable

Eine Abbildung, die jedem Ergebnis eines Zufallsexperiments eine reelle Zahl zuordnet, wird **Zufallsvariable** genannt. Ein konkreter Wert  $x = X(\omega)$  heißt **Realisation** der Zufallsvariable  $X$ .

$$X : \Omega \rightarrow \mathbb{R} \quad \omega \mapsto X(\omega)$$

### Beispiel Würfelwurf

Zufallsvariable Augenzahl:  $X_1(\omega) = \omega$



## 6.0 Zufallsvariablen und deren Verteilung

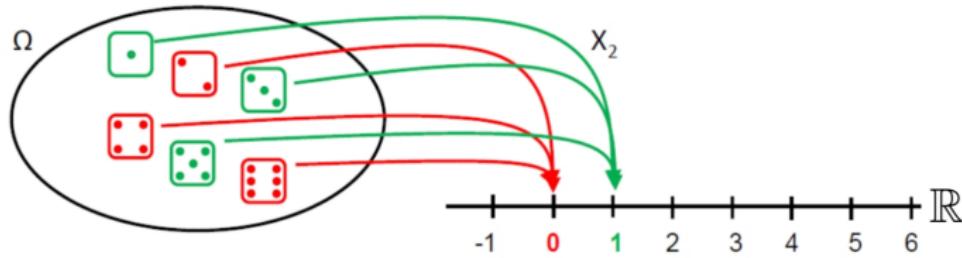
### Zufallsvariable

Eine Abbildung, die jedem Ergebnis eines Zufallsexperiments eine reelle Zahl zuordnet, wird **Zufallsvariable** genannt. Ein konkreter Wert  $x = X(\omega)$  heißt **Realisation** der Zufallsvariable  $X$ .

$$X : \Omega \rightarrow \mathbb{R} \quad \omega \mapsto X(\omega)$$

**Beispiel Würfelwurf:**  $X_2(\omega_i) = 1$ , falls  $i$ -ter Wurf Kopf,  $X_2(\omega_i) = 0$  sonst

Zufallsvariable Gerade/Ungerade:  $X_2(\omega) \in \{0, 1\}$



## 6.0 Zufallsvariablen und deren Verteilung

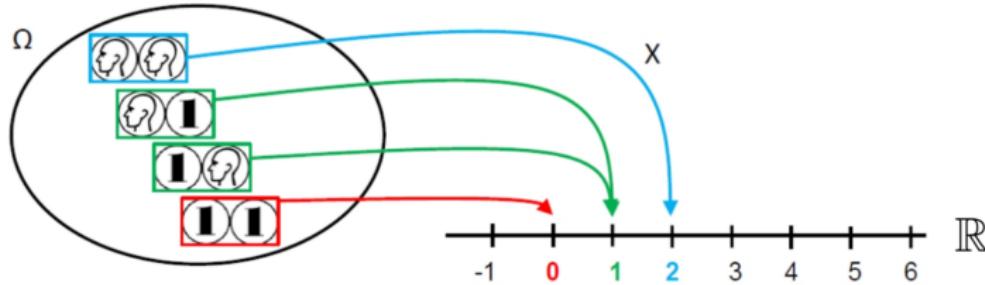
## Zufallsvariable

Eine Abbildung, die jedem Ergebnis eines Zufallsexperiments eine reelle Zahl zuordnet, wird **Zufallsvariable** genannt. Ein konkreter Wert  $x = X(\omega)$  heißt **Realisation** der Zufallsvariable  $X$ .

$$X : \Omega \rightarrow \mathbb{R} \quad \omega \mapsto X(\omega)$$

## Beispiel zweifacher Münzwurf

Zufallsvariable Anzahl Kopf:  $X((\omega_1, \omega_2)) = \omega_1 + \omega_2$



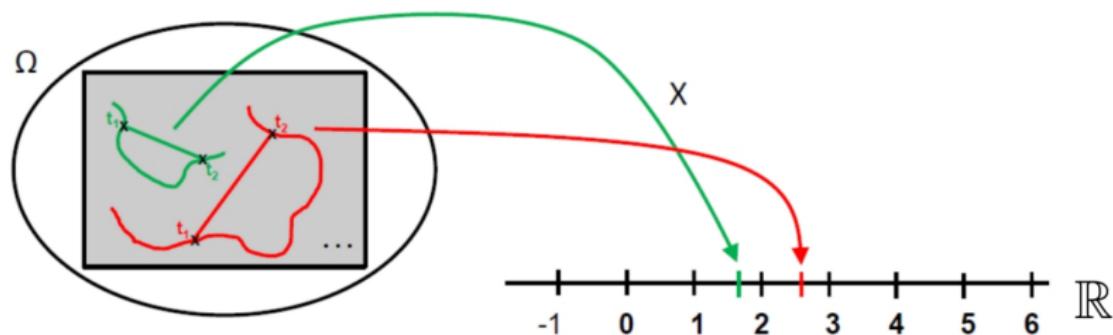
## 6.0 Zufallsvariablen und deren Verteilung

**Beispiel Mausaktivität:**  $\omega(t) = [x(t), y(t), c(t)]$

ZV: Distanz zwischen ersten 2 Mausclicks

$$X(\omega) = \sqrt{[x(t_2) - x(t_1)]^2 + [y(t_2) - y(t_1)]^2}$$

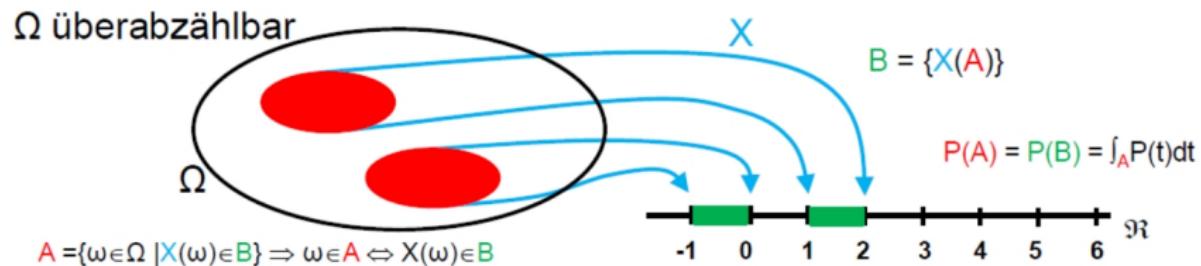
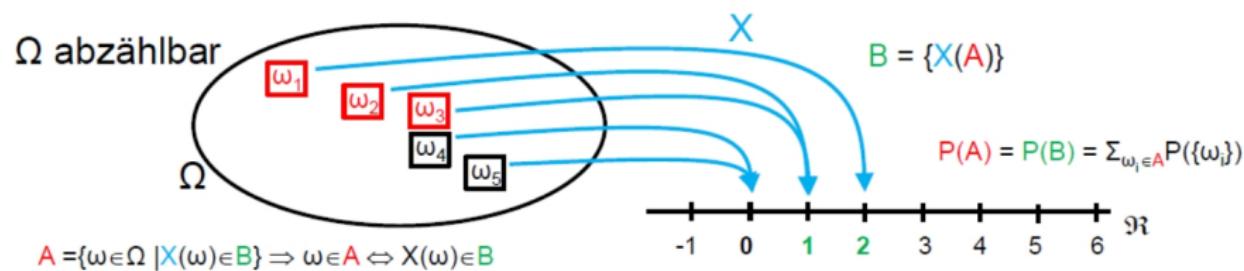
$$t_1 = \min(t | c(t) > 0) \quad t_2 = \min(t | c(t) > 0, t > t_1)$$



## 6.0 Zufallsvariablen und deren Verteilungen

## Verteilung eindimensionaler Zufallsvariablen

Die durch die Zufallsvariable definierte Abbildung von beliebigem Grundraum  $\Omega$  auf die reellen Zahlen erlaubt die Zuordnung von Wahrscheinlichkeiten zu Teilmengen von  $\mathbb{R}$ .



## 6.0 Zufallsvariablen und deren Verteilungen

## Verteilung eindimensionaler Zufallsvariablen

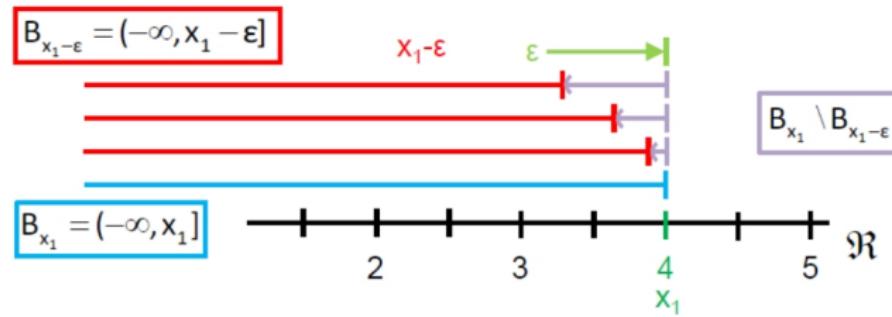
Die **Wahrscheinlichkeitsverteilung** oder kurz **Verteilung** einer Zufallsvariablen  $X$  ist definiert durch

$$P^X(B) = P(X \in B) = P(\{\omega \in \Omega | X(\omega) \in B\}), B \subseteq \mathbb{R}$$

Diese Verteilung ist eindeutig definiert, wenn  $P^X(B_x)$  für jedes Intervall der Form  $B_x = (-\infty, x]$  bekannt ist:

$$B = \{x_1\} = \lim_{\epsilon \downarrow 0} (\{B_{x_1} \setminus B_{x_1 - \epsilon}\}) \Rightarrow P^X(B) = \lim_{\epsilon \downarrow 0} [P^X(B_{x_1}) - P^X(B_{x_1 - \epsilon})],$$

$$\text{da } B_{x_1-\epsilon} \subset B_{x_1}$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

Die **Wahrscheinlichkeitsverteilung** oder kurz **Verteilung** einer Zufallsvariablen  $X$  ist definiert durch

$$P^X(B) = P(X \in B) = P(\{\omega \in \Omega | X(\omega) \in B\}), B \subseteq \mathbb{R}$$

Diese Verteilung ist eindeutig definiert, wenn  $P^X(B_x)$  für jedes Intervall der Form  $B_x = (-\infty, x]$  bekannt ist:

$$\boxed{B = \{x_1\}} = \lim_{\epsilon \downarrow 0} (\{B_{x_1} \setminus B_{x_1 - \epsilon}\}) \Rightarrow P^X(B) = \lim_{\epsilon \downarrow 0} [P^X(B_{x_1}) - P^X(B_{x_1 - \epsilon})],$$

da  $B_{x_1 - \epsilon} \subset B_{x_1}$

$$x_1 \neq \dots \neq x_k : \boxed{B = \{x_1, \dots, x_k\}} = \bigcup_{i=1}^k \lim_{\epsilon \downarrow 0} (\{B_{x_i} \setminus B_{x_i - \epsilon}\})$$

$$\Rightarrow P^X(B) = \sum_{i=1}^k \lim_{\epsilon \downarrow 0} [P^X(B_{x_i}) - P^X(B_{x_i - \epsilon})]$$

## 6.0 Zufallsvariablen und deren Verteilung

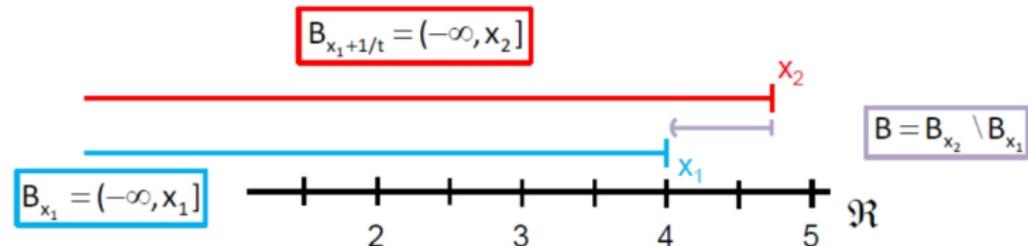
### Verteilung eindimensionaler Zufallsvariablen

Die **Wahrscheinlichkeitsverteilung** oder kurz **Verteilung** einer Zufallsvariablen  $X$  ist definiert durch

$$P^X(B) = P(X \in B) = P(\{\omega \in \Omega | X(\omega) \in B\}), B \subseteq \mathbb{R}$$

Diese Verteilung ist eindeutig definiert, wenn  $P^X(B_x)$  für jedes Intervall der Form  $B_x = (-\infty, x]$  bekannt ist:

$$x_1 < x_2 : B = (x_1, x_2] = B_{x_2} \setminus B_{x_1} \Rightarrow P^X(B) = P^X(B_{x_2}) - P^X(B_{x_1}), \text{ da } B_{x_1} \subset B_{x_2}$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

**Wahrscheinlichkeitsverteilung:**  $P^X(B) = P(X \in B) = P(\{\omega \in \Omega | X(\omega) \in B\})$ ,  $B \subseteq \mathbb{R}$

→ eindeutig definiert, wenn  $P^X(B_x)$  für jedes Intervall der Form  $B_x = (-\infty, x]$  bekannt:

$$\boxed{B = \{x_1\}} = \lim_{\epsilon \downarrow 0} (\{B_{x_1} \setminus B_{x_1 - \epsilon}\}) \Rightarrow P^X(B) = \lim_{\epsilon \downarrow 0} [P^X(B_{x_1}) - P^X(B_{x_1 - \epsilon})],$$

da  $B_{x_1 - \epsilon} \subset B_{x_1}$

$$x_1 \neq \dots \neq x_k : \boxed{B = \{x_1, \dots, x_k\}} = \bigcup_{i=1}^k \lim_{\epsilon \downarrow 0} (\{B_{x_i} \setminus B_{x_i - \epsilon}\})$$

$$\Rightarrow P^X(B) = \sum_{i=1}^k \lim_{\epsilon \downarrow 0} [P^X(B_{x_i}) - P^X(B_{x_i - \epsilon})]$$

$$x_1 < x_2 : \boxed{B = (x_1, x_2]} = B_{x_2} \setminus B_{x_1} \Rightarrow P^X(B) = P^X(B_{x_2}) - P^X(B_{x_1}), \text{ da } B_{x_1} \subset B_{x_2}$$

Beliebige Ereignisse lassen sich dann aus den halboffenen Intervallen durch Schnitte und Vereinigungen konstruieren.

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

Die Funktion  $F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit

$$F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}$$

wird **Verteilungsfunktion** genannt.

Die Entsprechung der Verteilungsfunktion in der deskriptiven Statistik ist die empirische Verteilungsfunktion, bei der an die Stelle von Wahrscheinlichkeiten kumulierte relative Häufigkeiten treten.

$$F_N(x) = \begin{cases} 0 & \text{falls } x < x(1) \\ s_j = \frac{\#\{x_n | x_n \leq x(j)\}}{N} \text{ mit } j = \max\{\tilde{j} | x(\tilde{j}) \leq x\} & \text{falls } x(1) \leq x \end{cases}$$

$$= \frac{\#\{x_n | x_n \leq x\}}{N}$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit

$$F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}$$

### Eigenschaften der Verteilungsfunktion

$$(A) \lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1$$

Beweis:

$$\begin{aligned} \lim_{x \rightarrow -\infty} F(x) &= \lim_{x \rightarrow -\infty} P(\{\omega \in \Omega | X(\omega) \in (-\infty, x] \cap \mathbb{R}\}) \\ &= P(\{\omega \in \Omega | X(\omega) \in \{-\infty\} \cap \mathbb{R}\}) = P(\{\omega \in \Omega | X(\omega) = \emptyset\}) \\ &\stackrel{(*)}{=} P(\emptyset) = 0 \quad (*) \quad [\omega \in \Omega \Rightarrow X(\omega) \in \mathbb{R}] \Leftrightarrow [X(\omega) \notin \mathbb{R} \Rightarrow \omega \notin \Omega] \end{aligned}$$

$$\begin{aligned} \lim_{x \rightarrow \infty} F(x) &= \lim_{x \rightarrow \infty} P(\{\omega \in \Omega | X(\omega) \in (-\infty, x] \cap \mathbb{R}\}) \\ &= P(\{\omega \in \Omega | X(\omega) \in \mathbb{R}\}) = P(\Omega) = 1 \quad \square \end{aligned}$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit

$$F(X) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}$$

### Eigenschaften der Verteilungsfunktion

(A)  $\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1$

(B)  $x < y \Rightarrow F(x) < F(y)$

Beweis:

$$F(x) = P(A) \text{ mit } A = \{\omega \in \Omega | X(\omega) \leq x\}$$

$$F(y) = P(B) \text{ mit } B = \{\omega \in \Omega | X(\omega) \leq y\}$$

$$\boxed{x < y} \Rightarrow A \subseteq B \Rightarrow P(A) \leq P(B) \Leftrightarrow \boxed{F(x) \leq F(y)}$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit

$$F(X) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}$$

### Eigenschaften der Verteilungsfunktion

(A)  $\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1$

(C)  $\lim_{x \downarrow z} F(x) = F(z)$

(B)  $x < y \Rightarrow F(x) < F(y)$

Beweis:

Setze  $A_n = \{\omega \in \Omega | X(\omega) \in (-\infty, z + 1/n]\}, \quad A_0 = \Omega$

$$\Rightarrow A = \bigcap_{n=1}^{\infty} A_n = \{\omega \in \Omega | X(\omega) \in (-\infty, z]\}, \quad A_n \subset A_{n-1}, \quad A_{n-1}^c \subset A_n^c, \quad n = 1, 2, \dots$$

## 6.0 Zufallsvariablen und deren Verteilung

### Beweis (Fortsetzung)

$$\begin{aligned} F(z) &= P(A) = P\left(\bigcap_{n=1}^{\infty} A_n\right) = 1 - P\left(\bigcup_{n=1}^{\infty} A_n^c\right) = 1 - \sum_{n=1}^{\infty} P(A_n^c \setminus A_{n-1}^c) \\ &= 1 - \lim_{N \uparrow \infty} \sum_{n=1}^N P(A_n^c \setminus A_{n-1}^c) = 1 - \lim_{N \uparrow \infty} P(A_N^c) = \lim_{N \uparrow \infty} P(A_N) \\ &= \boxed{\lim_{x \downarrow z} F(x)} \quad \square \end{aligned}$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit

$$F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}$$

### Eigenschaften der Verteilungsfunktion

$$(A) \lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1$$

$$(C) \lim_{x \downarrow z} F(x) = F(z)$$

$$(B) x < y \Rightarrow F(x) < F(y)$$

$$(D) P(a < X \leq b) = F(b) - F(a)$$

Beweis:

Setze  $A = \{\omega \in \Omega | X(\omega) \in (-\infty, a]\}$  und  $B = \{\omega \in \Omega | X(\omega) \in (-\infty, b]\}$

$$\begin{aligned} \Rightarrow P(a < X \leq b) &= P(\{\omega \in \Omega | X(\omega) \in (a, b]\}) = P(B \setminus A) \underset{A \subseteq B}{=} P(B) - P(A) \\ &= P(X \leq b) - P(X \leq a) = F(b) - F(a) \quad \square \end{aligned}$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit

$$F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}$$

### Eigenschaften der Verteilungsfunktion

- |  |   |
|--|---|
| (A) $\lim_{x \rightarrow -\infty} F(x) = 0$ , $\lim_{x \rightarrow \infty} F(x) = 1$ | (C) $\lim_{x \downarrow z} F(x) = F(z)$ |
| (B) $x < y \Rightarrow F(x) < F(y)$  | (D) $P(a < X \leq b) = F(b) - F(a)$     |
| (E) $P(X > a) = 1 - F(a)$  |   |

Beweis:

Setze  $A = \{\omega \in \Omega | X(\omega) \leq a\} \Rightarrow A^c = \{\omega \in \Omega | X(\omega) > a\}$

$$\Rightarrow P(X > a) = P(A^c) = 1 - P(A) = \boxed{1 - F(a)} \quad \square$$

## 6.0 Zufallsvariablen und deren Verteilung

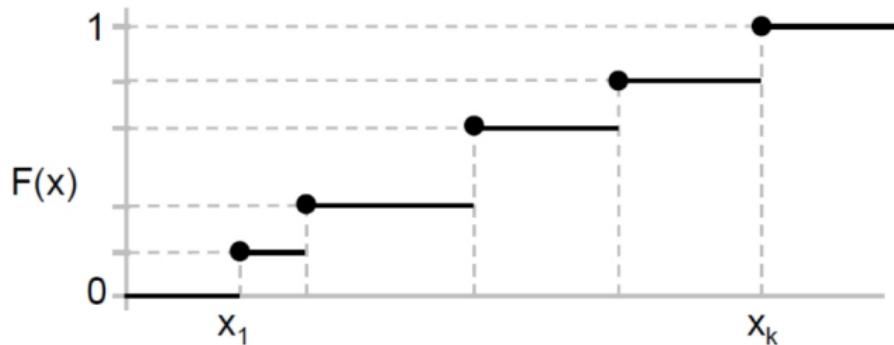
### Verteilung eindimensionaler Zufallsvariablen

#### Spezialfall diskrete Verteilungsfunktion ( $\Omega$ abzählbar)

$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\}$  mit  $-\infty < x_1 < \dots < x_k < \infty$ ,  $k \leq n$

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit

$F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\})$ ,  $x \in \mathbb{R}$



# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung eindimensionaler Zufallsvariablen

### Spezialfall diskrete Verteilungsfunktion ( $\Omega$ abzählbar)

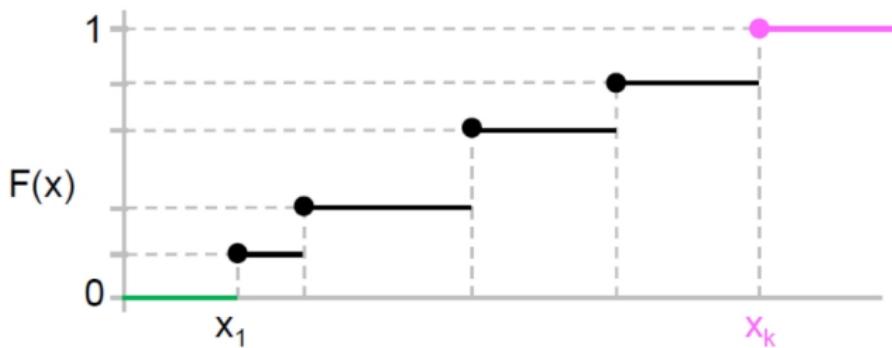
$$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\} \text{ mit } -\infty < x_1 < \dots < x_k < \infty, k \leq n$$

(A)  $\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1$

$$F(x) = P(A_x) \text{ mit } A_x = \{\omega \in \Omega \mid X(\omega) \in (-\infty, x] \cap \{x_1, \dots, x_k\}\}$$

$$x < x_1 \Rightarrow A_x = \emptyset \Rightarrow P(A_x) = 0$$

$$x \geq x_k \Rightarrow A_x = \Omega \Rightarrow P(A_x) = 1$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

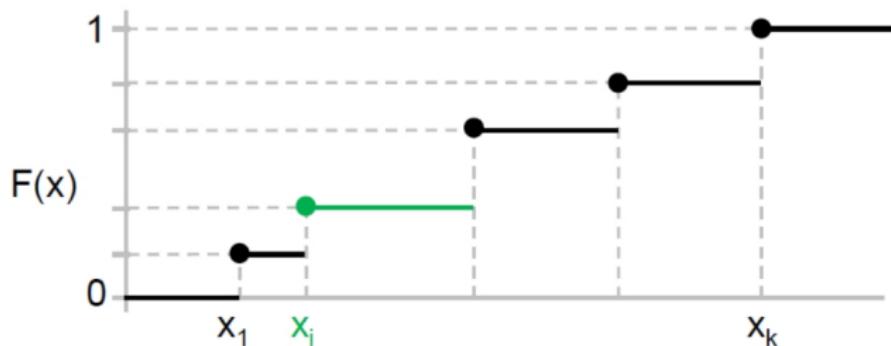
**Spezialfall diskrete Verteilungsfunktion** ( $\Omega$  abzählbar)

$$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\} \text{ mit } -\infty < x_1 < \dots < x_k < \infty, k \leq n$$

(C)  $\lim_{x \downarrow z} F(x) = F(z)$

$$F(x) = P(A_x) \text{ mit } A_x = \{\omega \in \Omega \mid X(\omega) \in (-\infty, x] \cap \{x_1, \dots, x_k\}\}$$

$$\begin{aligned} i=1, \dots, n-1: x_i \leq x < x_{i+1} \Rightarrow A_x &= \{\omega \in \Omega \mid X(\omega) \in \{x_1, \dots, x_i\}\} \\ \Rightarrow P(A_x) &= F(x_i) \end{aligned}$$



# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung eindimensionaler Zufallsvariablen

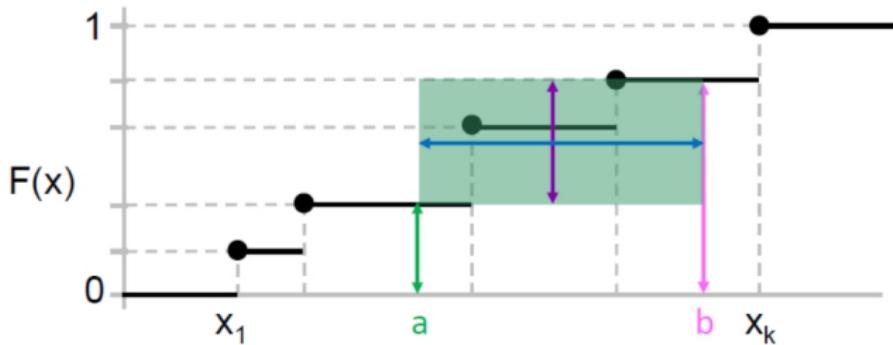
### Spezialfall diskrete Verteilungsfunktion ( $\Omega$ abzählbar)

$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\}$  mit  $-\infty < x_1 < \dots < x_k < \infty$ ,  $k \leq n$

$$(D) P(a < X \leq b) = F(b) - F(a)$$

$$A_b \setminus A_a = \{\omega \in \Omega \mid X(\omega) \in \{x_1, \dots, x_k\}, a < X(\omega) \leq b\}$$

$$P(a < X \leq b) = P(A_b \setminus A_a) = F(b) - F(a)$$



# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung eindimensionaler Zufallsvariablen

### Spezialfall diskrete Verteilungsfunktion ( $\Omega$ abzählbar)

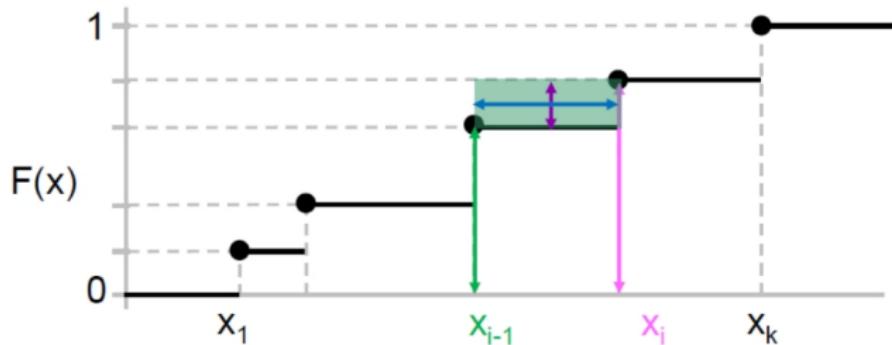
$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\}$  mit  $-\infty < x_1 < \dots < x_k < \infty$ ,  $k \leq n$

$$(D) \quad i=1, \dots, n: P(x_{i-1} < X \leq x_i)$$

$$(x_0 = -\infty)$$

$$A_{x_i} \setminus A_{x_{i-1}} = \{\omega \in \Omega \mid X(\omega) \in \{x_i\}\}$$

$$P(x_{i-1} < X \leq x_i) = P(A_{x_i} \setminus A_{x_{i-1}}) = P(X = x_i) = p_i$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

**Spezialfall diskrete Verteilungsfunktion** ( $\Omega$  abzählbar)

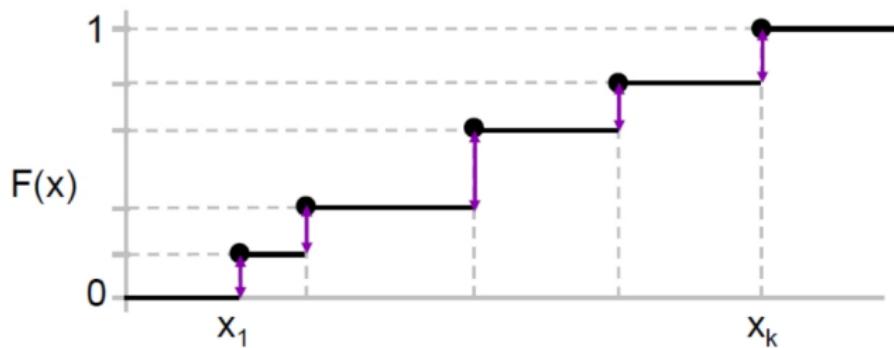
$$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\} \text{ mit } -\infty < x_1 < \dots < x_k < \infty, k \leq n$$

(D)  $i=1, \dots, n: P(x_{i-1} < X \leq x_i)$

$$(x_0 = -\infty)$$

$$A_{x_i} \setminus A_{x_{i-1}} = \{\omega \in \Omega \mid X(\omega) \in \{x_i\}\}$$

$$P(x_{i-1} < X \leq x_i) = P(A_{x_i} \setminus A_{x_{i-1}}) = P(X = x_i) = p(x_i)$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

**Spezialfall diskrete Verteilungsfunktion** ( $\Omega$  abzählbar)

$$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\} \text{ mit } -\infty < x_1 < \dots < x_k < \infty, k \leq n$$

Die Funktion:  $p : \mathbb{R} \rightarrow [0, 1]$  mit  $p(x) = P(X = x)$  heißt **Zähldichte von  $X$** .



## 6.0 Zufallsvariablen und deren Verteilung

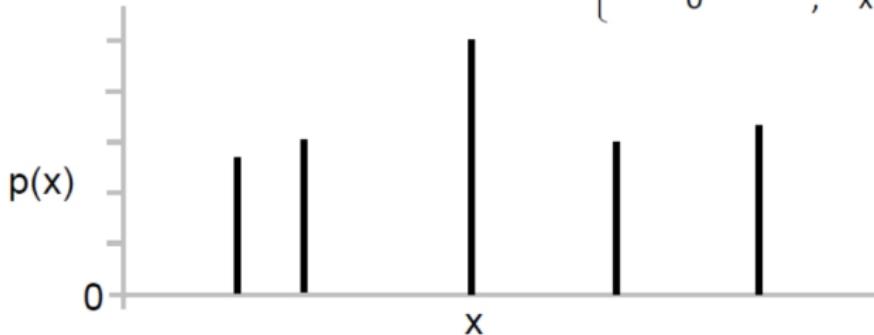
### Verteilung eindimensionaler Zufallsvariablen

**Spezialfall diskrete Verteilungsfunktion** ( $\Omega$  abzählbar)

$$\Omega = \{\omega_1, \dots, \omega_n\} \Rightarrow X \in \{x_1, \dots, x_k\} \text{ mit } -\infty < x_1 < \dots < x_k < \infty, k \leq n$$

Die Funktion:  $p : \mathbb{R} \rightarrow [0, 1]$  mit  $p(x) = P(X = x)$  heißt **Zähldichte von X**.

$$p(x) = \begin{cases} F(x_i) - F(x_{i-1}) & , \quad x \cap \{x_1, \dots, x_k\} = \{x_i\} \\ 0 & , \quad x \cap \{x_1, \dots, x_k\} = \emptyset \end{cases}$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen, diskrete Verteilungsfunktion

Beispiel: Anzahl Kopf beim **5-fachen Münzwurf**

#### Zähldichte

$x$	0	1	2	3	4	5
$A_x = \{\omega \in \Omega \mid X(\omega) = x\}$						
$p(x) = P(X=x)$ $=  A_x  /  \Omega $	1/32	5/32	10/32	10/32	5/32	1/32

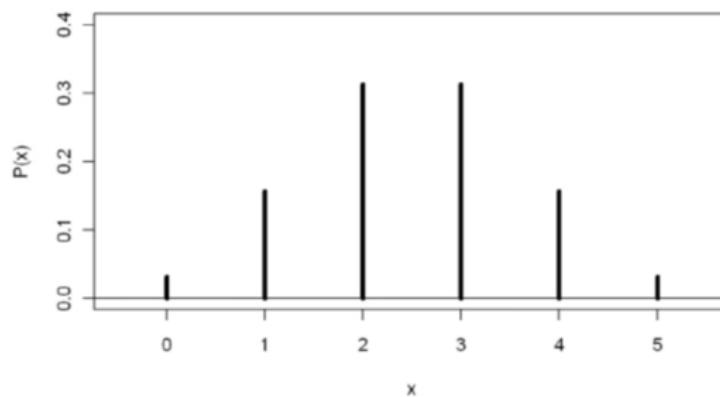
## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen, diskrete Verteilungsfunktion

Beispiel: Anzahl Kopf beim **5-fachen Münzwurf**

#### Zähldichte

$x$	0	1	2	3	4	5
$p(x)=P(X=x)$	1/32	5/32	10/32	10/32	5/32	1/32



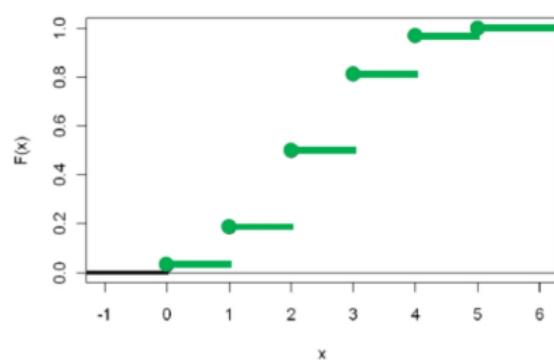
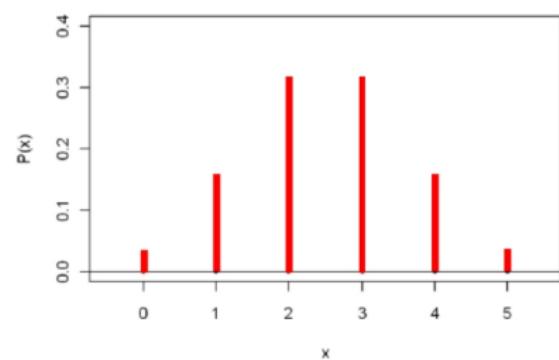
# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung eindimensionaler Zufallsvariablen, diskrete Verteilungsfunktion

Beispiel: Anzahl Kopf beim **5-fachen Münzwurf**

### Zähldichte und Verteilungsfunktion

$x$	0	1	2	3	4	5
$p(x) = P(X=x)$	$1/32$	$5/32$	$10/32$	$10/32$	$5/32$	$1/32$
$F(x) = P(X \leq x) = \sum_{i=0}^x p(i)$	$1/32$	$6/32$	$16/32$	$26/32$	$31/32$	$32/32$



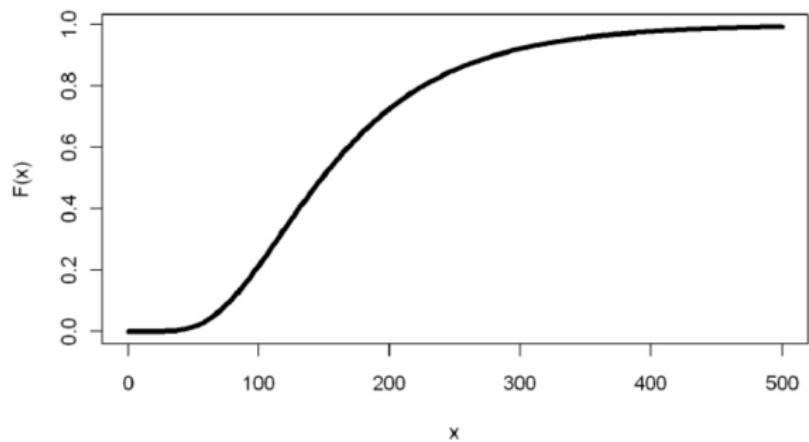
## 6.0 Zufallsvariablen und deren Verteilung

**Verteilung eindimensionaler Zufallsvariablen**

**Spezialfall stetige Verteilungsfunktion ( $\Omega$  überabzählbar)**

$\omega \in \Omega : X(\omega) \in B, B \subseteq \mathbb{R}$

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit  $F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), x \in \mathbb{R}$



## 6.0 Zufallsvariablen und deren Verteilung

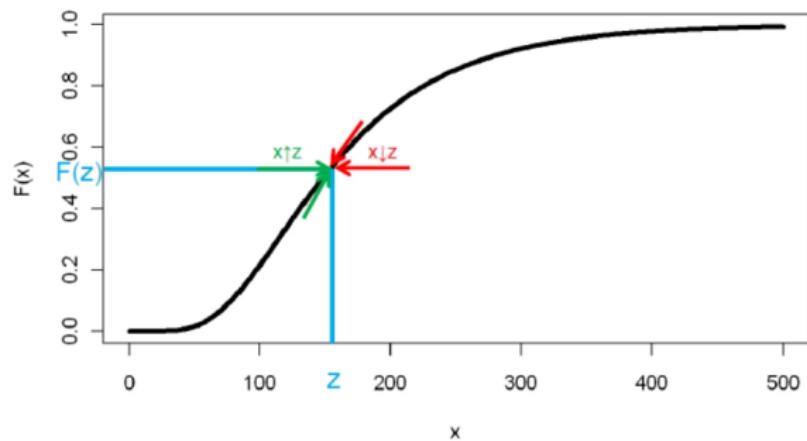
**Verteilung eindimensionaler Zufallsvariablen**

**Spezialfall stetige Verteilungsfunktion** ( $\Omega$  überabzählbar)

$\omega \in \Omega : X(\omega) \in B, B \subseteq \mathbb{R}$

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit  $F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), x \in \mathbb{R}$

$$(C) \lim_{x \downarrow z} F(x) = F(z) = \lim_{x \uparrow z} F(x)$$



## 6.0 Zufallsvariablen und deren Verteilung

## Verteilung eindimensionaler Zufallsvariablen

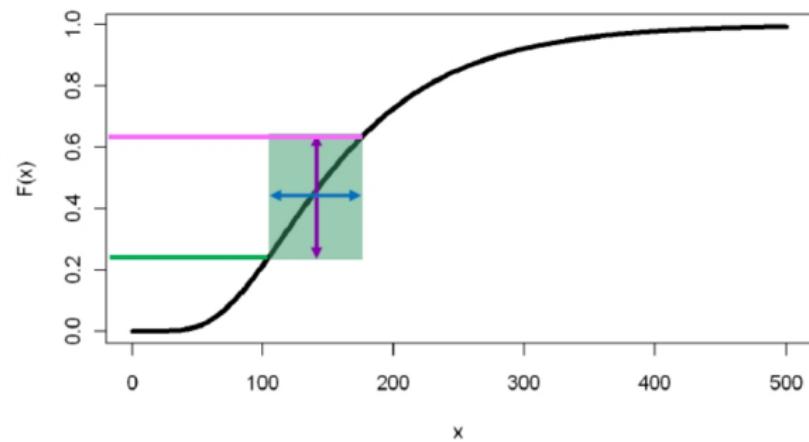
### **Spezialfall stetige Verteilungsfunktion ( $\Omega$ überabzählbar)**

$$\omega \in \Omega : X(\omega) \in B, \quad B \subseteq \mathbb{R}$$

$$F = F^X : \mathbb{R} \rightarrow [0, 1] \text{ mit } F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), \quad x \in \mathbb{R}$$

$$(C) \lim_{x \rightarrow z} F(x) = F(z) = \lim_{x \uparrow z} F(x)$$

$$(D) P(a < X \leq b) = F(b) - F(a)$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

#### Spezialfall stetige Verteilungsfunktion ( $\Omega$ überabzählbar)

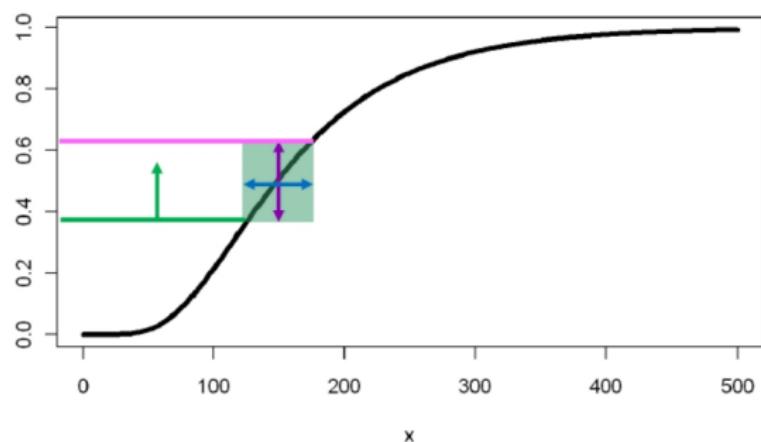
$\omega \in \Omega : X(\omega) \in B, B \subseteq \mathbb{R}$

$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit  $F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), x \in \mathbb{R}$

$$(C) \lim_{x \downarrow z} F(x) = F(z) = \lim_{x \uparrow z} F(x)$$

$$(D) P(a < X \leq b) = F(b) - F(a)$$

$$\begin{aligned} P(X = b) &= \lim_{a \uparrow b} P(a < X \leq b) \\ &= F(b) - \lim_{a \uparrow b} F(a) = F(b) - F(b) = 0 \end{aligned}$$



## 6.0 Zufallsvariablen und deren Verteilung

**Verteilung eindimensionaler Zufallsvariablen**

**Spezialfall stetige Verteilungsfunktion** ( $\Omega$  überabzählbar)

$\omega \in \Omega : X(\omega) \in B, B \subseteq \mathbb{R}$

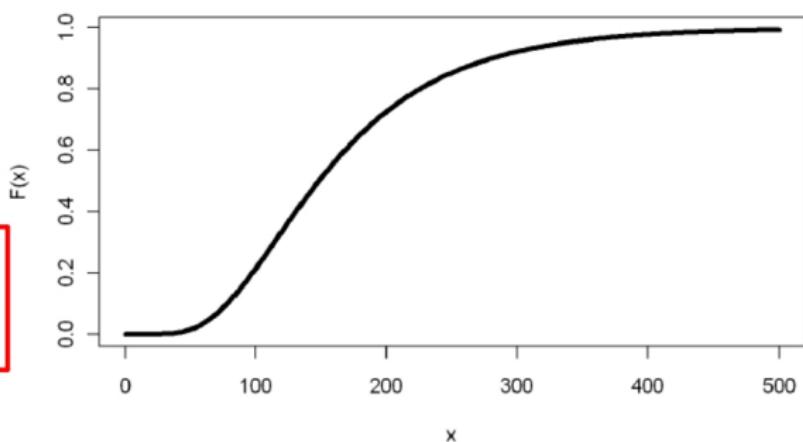
$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit  $F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), x \in \mathbb{R}$

$$(C) \lim_{x \downarrow z} F(x) = F(z) = \lim_{x \uparrow z} F(x)$$

$$(D) P(a < X \leq b) = F(b) - F(a)$$

$$(F) P(X = x) = 0, x \in \mathfrak{N}$$

$$\begin{aligned} (G) P(a < X \leq b) &= P(a \leq X \leq b) \\ &= P(a \leq X < b) = P(a < X < b) \\ &= F(b) - F(a) \end{aligned}$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen

**Spezialfall stetige Verteilungsfunktion** ( $\Omega$  überabzählbar)

$\omega \in \Omega : X(\omega) \in B, B \subseteq \mathbb{R}$

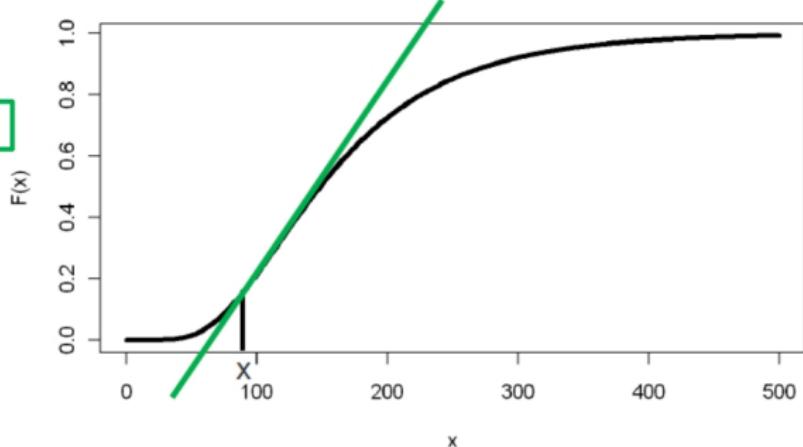
$F = F^X : \mathbb{R} \rightarrow [0, 1]$  mit  $F(x) = P^X((-\infty, x]) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\}), x \in \mathbb{R}$

$$\lim_{a \rightarrow b} P(a < X \leq b) = 0$$

$$\lim_{c \downarrow 0} \frac{F(x+c) - F(x)}{c} = \boxed{F'(x) = f(x)}$$

Die Funktion  $f(x)$  wird **Dichtefunktion** bzw. **Dichte** von  $X$  genannt.

Sie beschreibt die Steigung (Grad der Verdichtung) der Verteilung  $X$



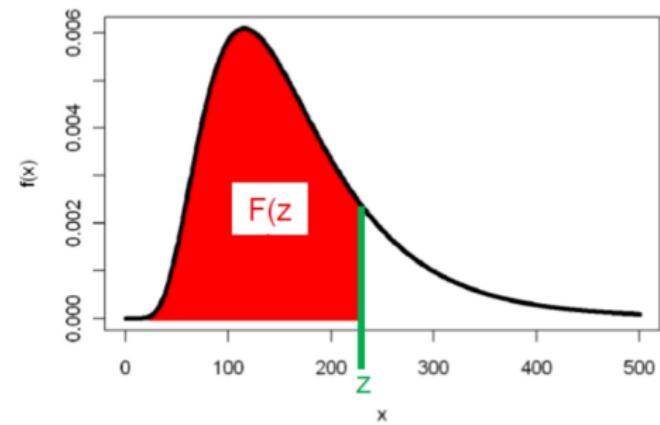
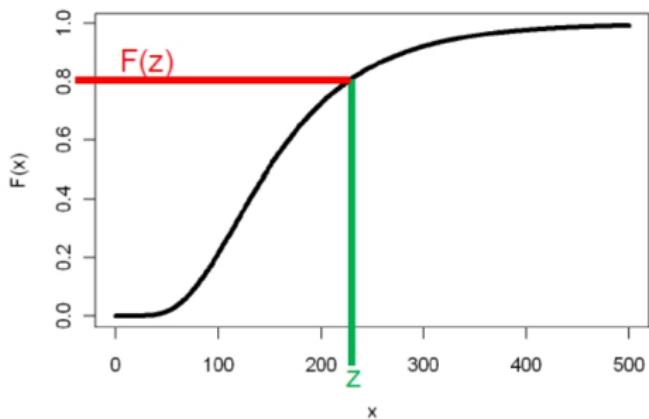
# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung eindimensionaler Zufallsvariablen

**Spezialfall stetige Verteilungsfunktion ( $\Omega$  überabzählbar)**

$$F'(x) = f(x), \quad F(x) = \int_{-\infty}^x f(t) dt, \quad x \in \mathfrak{N}, \quad \int_{-\infty}^{\infty} f(t) dt = 1$$

$$P(X \leq z) = F(z) = \int_{-\infty}^z f(t) dt$$



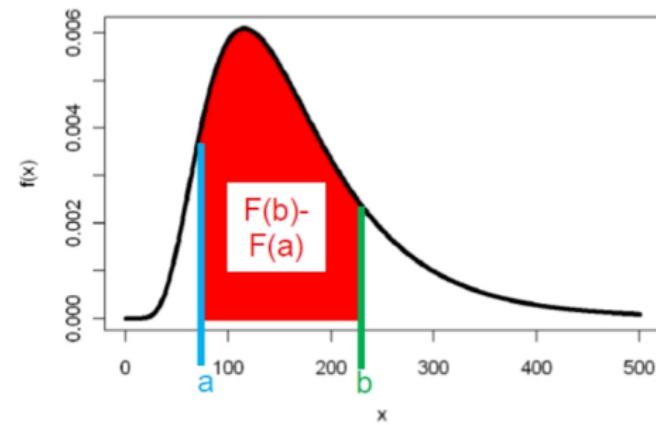
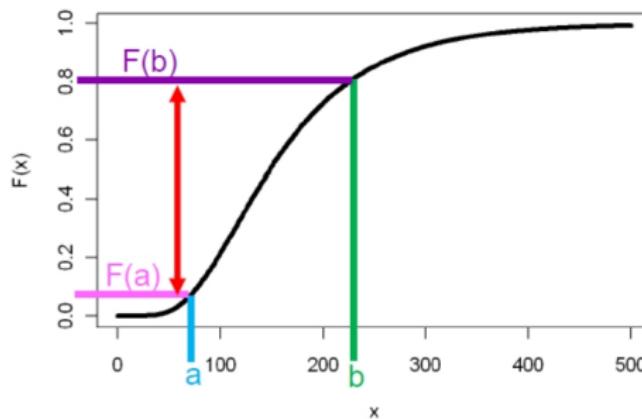
## 6.0 Zufallsvariablen und deren Verteilung

## Verteilung eindimensionaler Zufallsvariablen

### **Spezialfall stetige Verteilungsfunktion ( $\Omega$ überabzählbar)**

$$F'(x) = f(x), \quad F(x) = \int_{-\infty}^x f(t)dt, \quad x \in \mathfrak{N}, \quad \int_{-\infty}^{\infty} f(t)dt = 1$$

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(t) dt$$



## 6.0 Zufallsvariablen und deren Verteilung

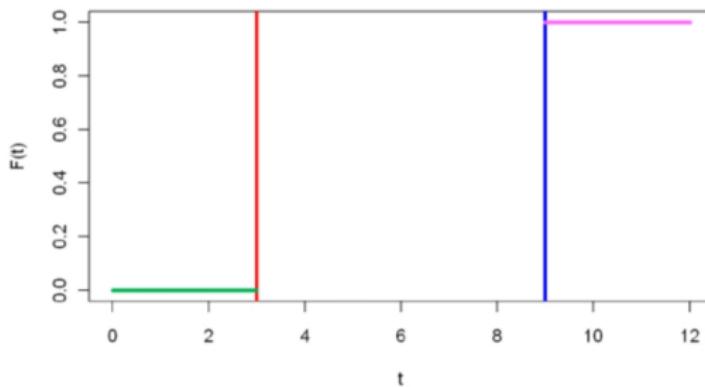
### Verteilung eindimensionaler Zufallsvariablen, stetige Verteilungsfunktion

Beispiel: **Mausaktivität, exakter Zeitpunkt  $T$  des ersten Mausclicks**

Annahme:  $T$  fällt in jedes Intervall gleicher Länge  $c$  zwischen  $t_{\min}$  und  $t_{\max}$  mit derselben Wahrscheinlichkeit

$$P(T < t_{\min}) = 0 = F(t_{\min}) \Rightarrow F(t) = 0, t \leq t_{\min}$$

$$P(T > t_{\max}) = 0 = 1 - F(t_{\max}) \Rightarrow F(t) = 1, t \geq t_{\max}$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen, stetige Verteilungsfunktion

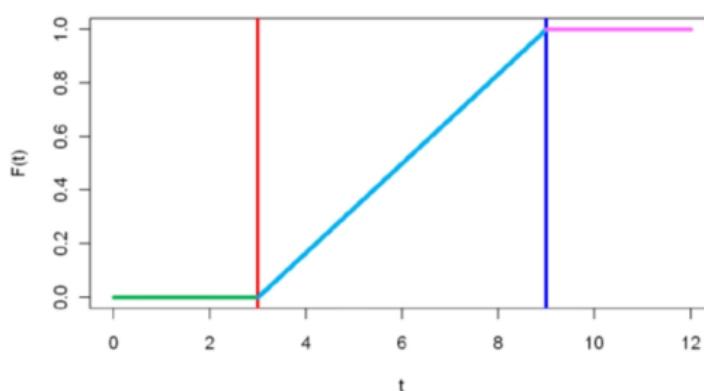
Beispiel: **Mausaktivität, exakter Zeitpunkt  $T$  des ersten Mausclicks**

Annahme:  $T$  fällt in jedes Intervall gleicher Länge  $c$  zwischen  $t_{\min}$  und  $t_{\max}$  mit derselben Wahrscheinlichkeit

$$F(t) = 0, t \leq t_{\min}$$

$$F(t) = \frac{t - t_{\min}}{t_{\max} - t_{\min}}, t_{\min} < t < t_{\max}$$

$$F(t) = 1, t \geq t_{\max}$$



Wahrscheinlichkeitsdichte

$$t \leq t_{\min} : F'(t) = f(t) = \partial 0 / \partial t = 0$$

$$t_{\min} < t < t_{\max} :$$

$$F'(t) = f(t) = \partial \left( \frac{t - t_{\min}}{t_{\max} - t_{\min}} \right) / \partial t = \frac{1}{t_{\max} - t_{\min}}$$

$$t > t_{\max} : F'(t) = f(t) = \partial 1 / \partial t = 0$$

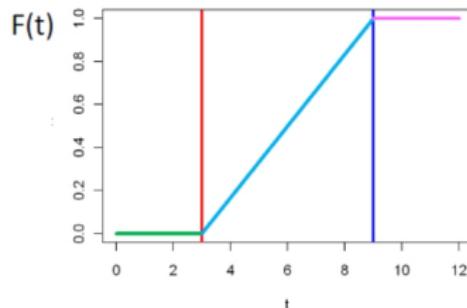
## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung eindimensionaler Zufallsvariablen, stetige Verteilungsfunktion

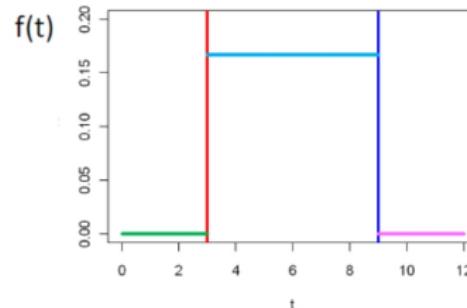
Beispiel: **Mausaktivität, exakter Zeitpunkt  $T$  des ersten Mausclicks**

Annahme:  $T$  fällt in jedes Intervall gleicher Länge  $c$  zwischen  $t_{\min}$  und  $t_{\max}$  mit derselben Wahrscheinlichkeit

$$\begin{aligned} F(t) &= 0, t \leq t_{\min} & F(t) &= 1, t \geq t_{\max} \\ F(t) &= \frac{t - t_{\min}}{t_{\max} - t_{\min}}, t_{\min} < t < t_{\max} \end{aligned}$$



$$\begin{aligned} f(t) &= 0, t \leq t_{\min} & f(t) &= 0, t \geq t_{\max} \\ f(t) &= \frac{1}{t_{\max} - t_{\min}}, t_{\min} < t < t_{\max} \end{aligned}$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

Die **Wahrscheinlichkeitsverteilung** oder kurz **Verteilung** einer zweidimensionalen Zufallsvariablen  $(X, Y)$  ist definiert durch

$$P^{(X,Y)}(B) = P((X, Y) \in B) = P(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in B\}), B \subseteq \mathbb{R}^2$$

Die Funktion  $F = F^{(X,Y)} : \mathbb{R}^2 \rightarrow [0, 1]$  mit

$$\begin{aligned} F(x, y) &= P^{(X,Y)}((-\infty, x] \times (-\infty, y]) = P(X \leq x, Y \leq y) \\ &= P(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}), x, y \in \mathbb{R}, \end{aligned}$$

wird **Verteilungsfunktion** von  $(X, Y)$  genannt.

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

$$P^{(X,Y)}(B) = P((X, Y) \in B) = P(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in B\}), B \subseteq \mathbb{R}^2$$

$$F(x, y) = P(X \leq x, Y \leq y) = P(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}), x, y \in \mathbb{R}$$

#### Eigenschaften

$$1. \lim_{x \rightarrow -\infty} F(x, y) = \lim_{y \rightarrow -\infty} F(x, y) = \lim_{x,y \rightarrow -\infty} F(x, y) = 0$$

Beweis:

$$A = \{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\} = A_x \cap A_y \text{ mit } A_x = \{\omega \in \Omega | X(\omega) \leq x\}$$

$$A_y = \{\omega \in \Omega | Y(\omega) \leq y\}$$

$$F(x, y) = P(A) = P(A_x \cap A_y) = 1 - P(A_x^c \cup A_y^c)$$

$$\boxed{\lim_{x \rightarrow -\infty} F(x, y)} = 1 - (P(A_{-\infty}^c \cup A_y^c)) = 1 - P(\Omega \cup A_y^c)$$

$$= 1 - [P(\Omega) + P(A_y^c) - P(A_y^c)] = 1 - 1 = \boxed{0} \quad \square$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

$$P^{(X,Y)}(B) = P((X, Y) \in B) = P(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in B\}), B \subseteq \mathbb{R}^2$$

$$F(x, y) = P(X \leq x, Y \leq y) = P(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}), x, y \in \mathbb{R}$$

#### Eigenschaften

$$1. \lim_{x \rightarrow -\infty} F(x, y) = \lim_{y \rightarrow -\infty} F(x, y) = \lim_{x, y \rightarrow -\infty} F(x, y) = 0, \quad \boxed{\lim_{x, y \rightarrow \infty} F(x, y) = 1}$$

$$2. \lim_{y \rightarrow \infty} F(x, y) = F^X(x), \quad \lim_{x \rightarrow \infty} F(x, y) = F^Y(y)$$

#### Beweis:

$$A = \{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\} = A_x \cap A_y \text{ mit } A_x = \{\omega \in \Omega | X(\omega) \leq x\} \\ A_y = \{\omega \in \Omega | Y(\omega) \leq y\}$$

$$F(x, y) = P(A) = P(A_x \cap A_y) = 1 - P(A_x^c \cup A_y^c)$$

$$\boxed{\lim_{x \rightarrow \infty} F(x, y)} = 1 - P(A_\infty^c \cup A_y^c) = 1 - P(\emptyset \cup A_y^c) = 1 - P(A_y^c) = P(A_y) = \boxed{F^Y(y)}$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

$$P^{(X,Y)}(B) = P((X, Y) \in B) = P(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in B\}), B \subseteq \mathbb{R}^2$$

$$F(x, y) = P(X \leq x, Y \leq y) = P(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}), x, y \in \mathbb{R}$$

#### Eigenschaften

$$1. \lim_{x \rightarrow -\infty} F(x, y) = \lim_{y \rightarrow -\infty} F(x, y) = \lim_{x,y \rightarrow -\infty} F(x, y) = 0, \quad \boxed{\lim_{x,y \rightarrow \infty} F(x, y) = 1}$$

$$2. \lim_{y \rightarrow \infty} F(x, y) = F^X(x), \quad \lim_{x \rightarrow \infty} F(x, y) = F^Y(y)$$

Beweis (Fortsetzung):

$$\boxed{\lim_{x \rightarrow \infty} F(x, y) = F^Y(y)}$$

Beweis für  $\lim_{y \rightarrow \infty} F(x, y) = F^X(x)$  analog.

$$\boxed{\lim_{x,y \rightarrow \infty} F(x, y) = \lim_{y \rightarrow \infty} F^Y(y) = 1}$$

$F^X(x)$  und  $F^Y(y)$  heißen  
**Randverteilungen** von  $X$  und  $Y$

□

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

$$P^{(X,Y)}(B) = P((X, Y) \in B) = P(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in B\}), B \subseteq \mathbb{R}^2$$

$$F(x, y) = P(X \leq x, Y \leq y) = P(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}), x, y \in \mathbb{R}$$

#### Eigenschaften

$$1. \lim_{x \rightarrow -\infty} F(x, y) = \lim_{y \rightarrow -\infty} F(x, y) = \lim_{x, y \rightarrow -\infty} F(x, y) = 0, \lim_{x, y \rightarrow \infty} F(x, y) = 1$$

$$2. \lim_{y \rightarrow \infty} F(x, y) = F^X(x), \lim_{x \rightarrow \infty} F(x, y) = F^Y(y)$$

$$3. x_1 < x_2 \Rightarrow F(x_1, y) \leq F(x_2, y), y_1 < y_2 \Rightarrow F(x, y_1) \leq F(x, y_2)$$

#### Beweis

$$F(x_i, y) = P(A_i) \text{ mit } A_i = \{\omega \in \Omega | X(\omega) \leq x_i, Y(\omega) \leq y\}$$

Beweis für  
 $F(x, y_1)$  analog

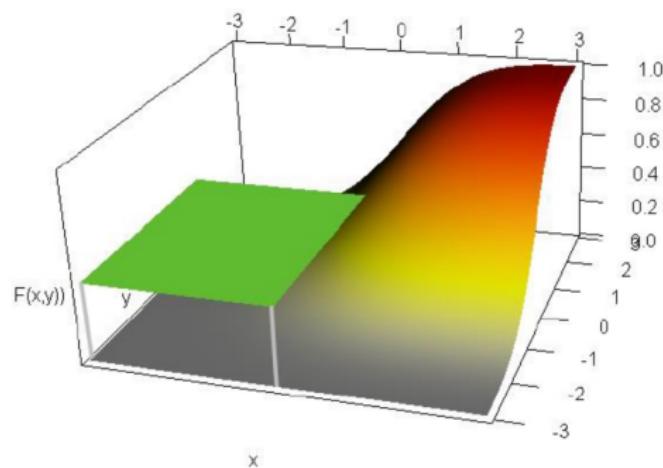
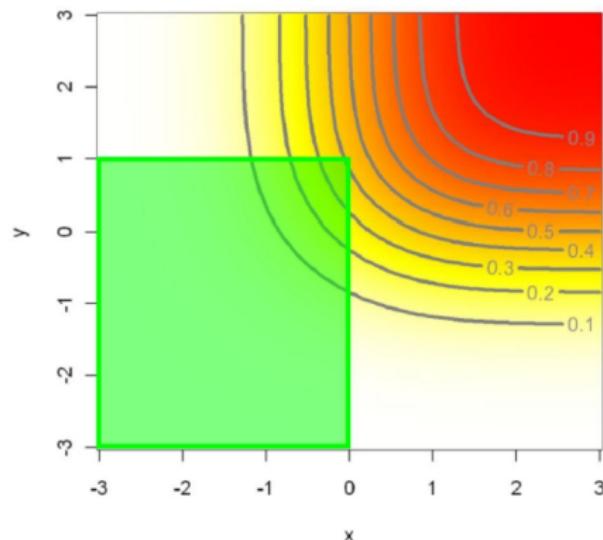
$$x_1 < x_2 \Rightarrow A_1 \subseteq A_2 \Rightarrow P(A_1) \leq P(A_2) \Leftrightarrow F(x_1, y) \leq F(x_2, y)$$

□

# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung mehrdimensionaler Zufallsvariablen

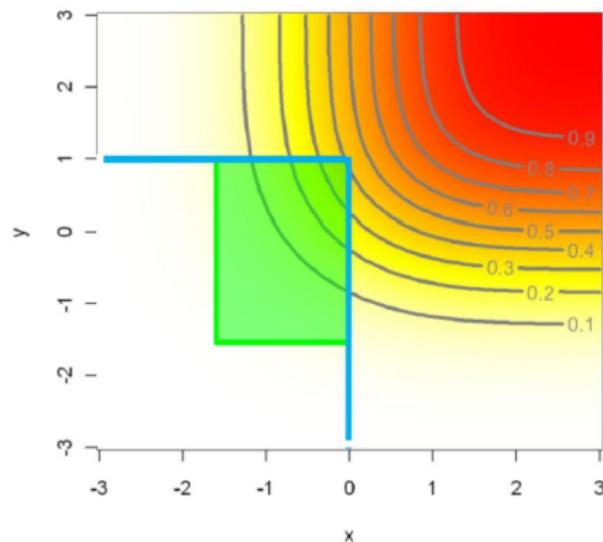
$$F(x,y) = P(X \leq x, Y \leq y), \quad x,y \in \mathbb{R}$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

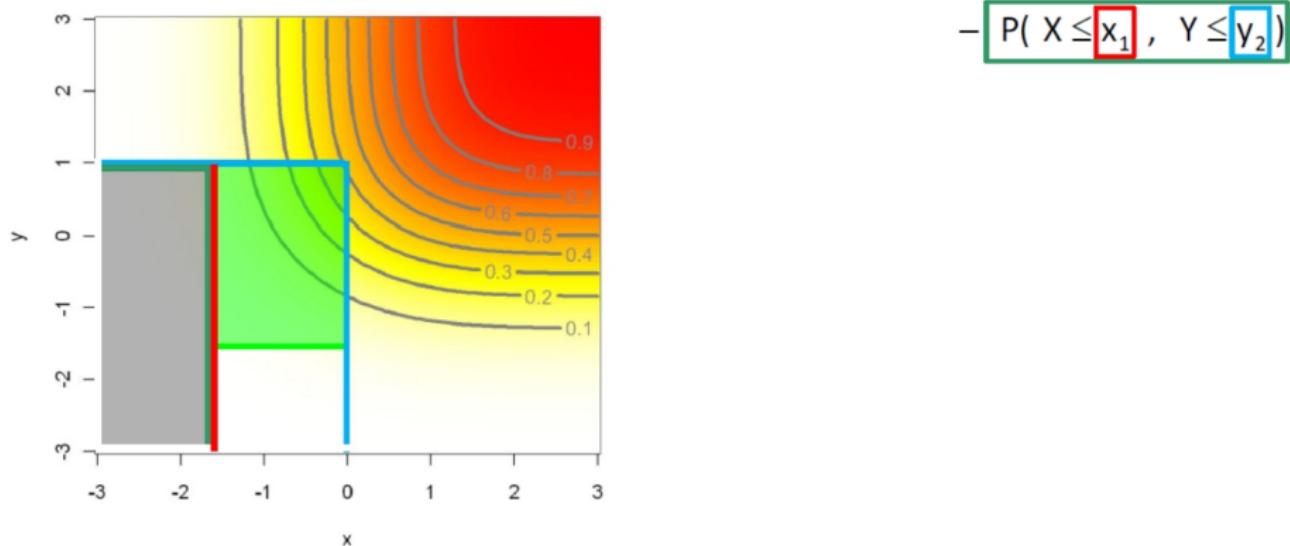
$$P(\boxed{x_1} < X \leq \boxed{x_2}, \boxed{y_1} < Y \leq \boxed{y_2}) = P(X \leq \boxed{x_2}, Y \leq \boxed{y_2})$$



# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung mehrdimensionaler Zufallsvariablen

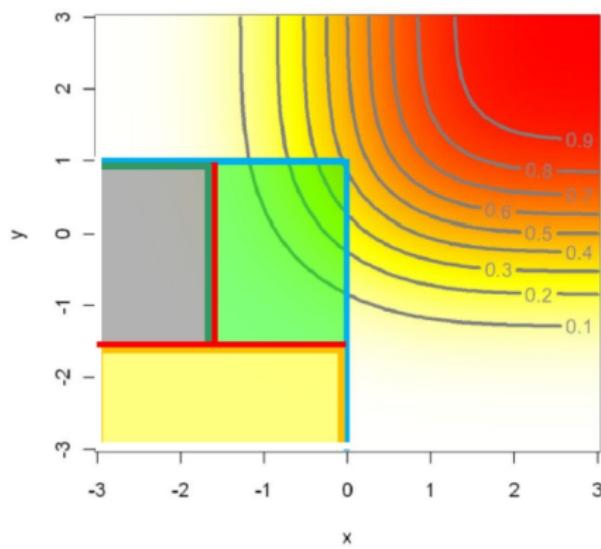
$$P(\boxed{x_1} < X \leq \boxed{x_2}, \boxed{y_1} < Y \leq \boxed{y_2}) = P(X \leq \boxed{x_2}, Y \leq \boxed{y_2}) - P(X \leq \boxed{x_1}, Y \leq \boxed{y_2})$$



# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung mehrdimensionaler Zufallsvariablen

$$P(\boxed{x_1} < X \leq \boxed{x_2}, \boxed{y_1} < Y \leq \boxed{y_2}) = P(X \leq \boxed{x_2}, Y \leq \boxed{y_2}) - P(X \leq \boxed{x_1}, Y \leq \boxed{y_2}) - P(X \leq \boxed{x_2}, Y \leq \boxed{y_1})$$



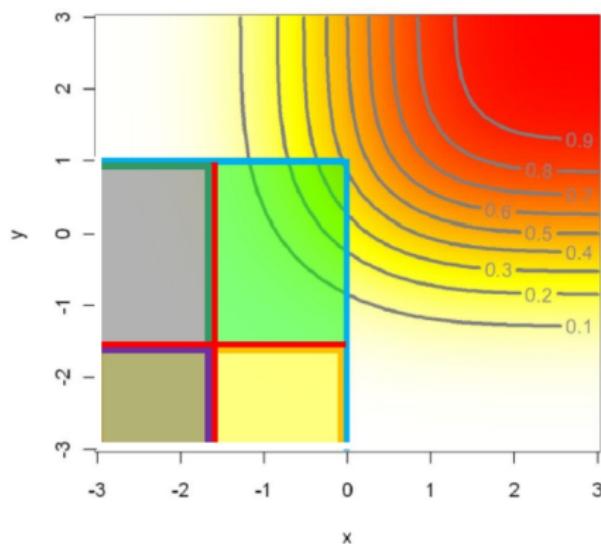
$$- P(X \leq \boxed{x_1}, Y \leq \boxed{y_2})$$

$$- P(X \leq \boxed{x_2}, Y \leq \boxed{y_1})$$

# 6.0 Zufallsvariablen und deren Verteilung

## Verteilung mehrdimensionaler Zufallsvariablen

$$P(\boxed{x_1} < X \leq \boxed{x_2}, \boxed{y_1} < Y \leq \boxed{y_2}) = P(\boxed{X \leq x_2}, \boxed{Y \leq y_2})$$



$$\begin{aligned}
 & - P(\boxed{X \leq x_1}, \boxed{Y \leq y_2}) \\
 & - P(\boxed{X \leq x_2}, \boxed{Y \leq y_1}) \\
 & + P(\boxed{X \leq x_1}, \boxed{Y \leq y_1}) \\
 = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)
 \end{aligned}$$

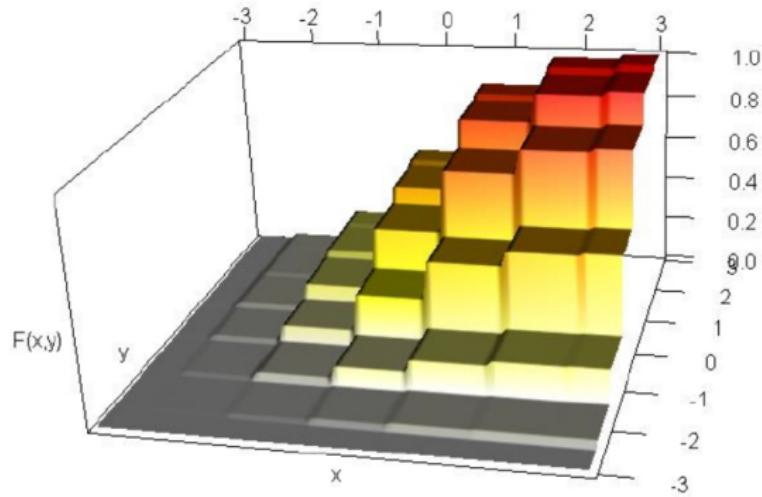
## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

**Spezialfall diskrete Verteilungsfunktion** ( $\Omega$  abzählbar):  $\Omega = \{\omega_1, \dots, \omega_n\}$

$\Rightarrow X \in \{X(\omega_1), \dots, X(\omega_n)\} = \{x_1, \dots, x_n\}$  mit  $-\infty < x_1 \leq \dots \leq x_n < \infty$

$Y \in \{Y(\omega_1), \dots, Y(\omega_n)\} = \{y_1, \dots, y_n\}$  mit  $-\infty < y_1 \leq \dots \leq y_n < \infty$



## 6.0 Zufallsvariablen und deren Verteilung

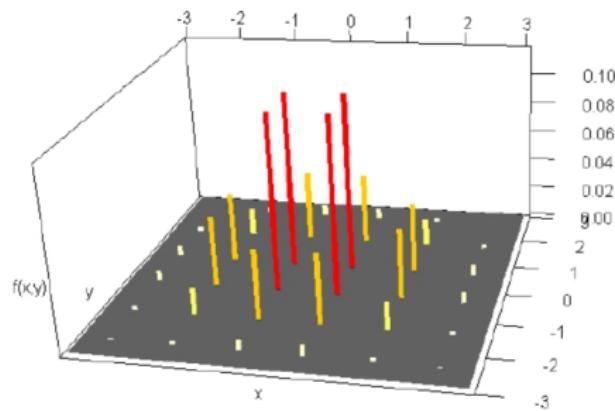
### Verteilung mehrdimensionaler Zufallsvariablen

**Spezialfall diskrete Verteilungsfunktion** ( $\Omega$  abzählbar):  $\Omega = \{\omega_1, \dots, \omega_n\}$

$\Rightarrow X \in \{X(\omega_1), \dots, X(\omega_n)\} = \{x_1, \dots, x_n\}$  mit  $-\infty < x_1 \leq \dots \leq x_n < \infty$

$Y \in \{Y(\omega_1), \dots, Y(\omega_n)\} = \{y_1, \dots, y_n\}$  mit  $-\infty < y_1 \leq \dots \leq y_n < \infty$

Die Funktion  $p : \mathbb{R}^2 \rightarrow [0, 1]$  mit  $p(x, y) = P(X = x, Y = y)$  heißt **Zähldichte von  $(X, Y)$**



$$p(x, y) = \begin{cases} F(x_i, y_i) - F(x_{i-1}, y_i) & , x_i \in \{x_1, \dots, x_n\} \\ -F(x_i, y_{i-1}) + F(x_{i-1}, y_{i-1}) & , y_i \in \{y_1, \dots, y_n\} \\ 0 & , \text{sonst} \end{cases}$$

## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen, diskrete Verteilungsfunktion

Beispiel: **4-facher Münzwurf**,  $X$  = Anzahl Kopf nach 4 Würfen,  $Y$  = Anzahl Kopf nach 2 Würfen

#### Zähldichte

$\downarrow y \quad x \rightarrow$	0	1	2	3	4
0					
1					
2					

## 6.0 Zufallsvariablen und deren Verteilung

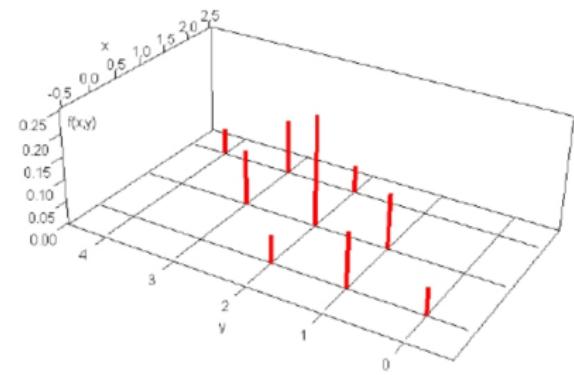
### Verteilung mehrdimensionaler Zufallsvariablen, diskrete Verteilungsfunktion

Beispiel: **4-facher Münzwurf**,  $X$  = Anzahl Kopf nach 4 Würfen,  $Y$  = Anzahl Kopf nach 2 Würfen

#### Zähldichte

$\downarrow y \rightarrow$	0	1	2	3	4
0	1/16	2/16	1/16		
1		2/16	4/16	2/16	
2			1/16	2/16	1/16

$$p(x,y)$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen, diskrete Verteilungsfunktion

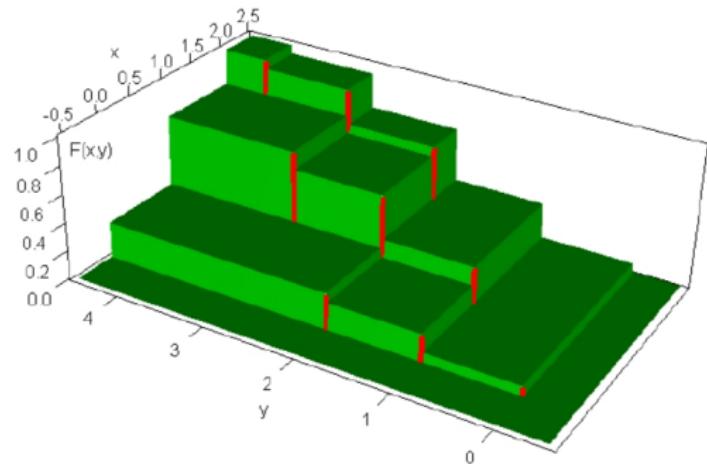
Beispiel: **4-facher Münzwurf**,  $X$  = Anzahl Kopf nach 4 Würfen,  $Y$  = Anzahl Kopf nach 2 Würfen

#### Zähldichte

$\downarrow y$	0	1	2	3	4
$\rightarrow x$					
0	$1/16$	$2/16$	$1/16$		
	$\downarrow =$	$\downarrow =$	$\downarrow =$		
	$1/16$	$3/16$	$4/16$		
1		$2/16$	$4/16$	$2/16$	
	$\downarrow =$	$\downarrow =$	$\downarrow =$		
	$5/16$	$10/16$	$12/16$		
2		$1/16$	$2/16$	$1/16$	
	$\downarrow =$	$\downarrow =$	$\downarrow =$		
	$11/16$	$15/16$	$16/16$		

$$p(x,y)$$

$$F(x,y)$$



## 6.0 Zufallsvariablen und deren Verteilung

### Verteilung mehrdimensionaler Zufallsvariablen

#### Spezialfall stetige Verteilungsfunktion ( $\Omega$ überabzählbar)

$\omega \in \Omega : (X(\omega), Y(\omega)) \in B, B \subseteq \mathbb{R}^2$

$F = F^{XY} : \mathbb{R}^2 \rightarrow [0, 1]$  mit

$F(x, y) = P(X \leq x, Y \leq y) = P(\{\omega \in \Omega | X(\omega) \leq x, Y(\omega) \leq y\}), x, y \in \mathbb{R}$

Die Funktion  $f : \mathbb{R}^2 \rightarrow [0, 1]$  mit  $f(x, y) = \frac{\delta^2 F(x, y)}{\delta x \delta y}$

heißt die **gemeinsame Dichtefunktion** von  $X$  und  $Y$ .

Es gilt:

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(s, t) dt ds, \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(s, t) dt ds = 1$$

Die **Randdichten**  $f^X$  und  $f^Y$  von  $X$  und  $Y$  sind definiert durch

$$f^X(x) = \int_{-\infty}^{\infty} f(x, t) dt \text{ und } f^Y(y) = \int_{-\infty}^{\infty} f(s, y) ds$$

## 6.0 Zufallsvariablen und deren Verteilung

**Verteilung mehrdimensionaler Zufallsvariablen**

**Spezialfall stetige Verteilungsfunktion ( $\Omega$  überabzählbar)**

