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_1	Kai	Export	1.1	14	8.0
e_2	Kai	Verknüpfung	1.2	12	4.9
e_3	Miriam	Export	1.1	12	6.6
e_4	Tina	Verknüpfung	1.2	13	3.2
e_5	Oliver	Export	2.0	17	3.9
e_6	Tina	Export	1.2	11	4.5
e_7	Tina	Verknüpfung	1.2	14	6.1
e_8	Miriam	Export	1.2	10	3.7
e_9	Miriam	Export	1.2	10	4.2
e_{10}	Oliver	Abfrage	1.1	18	8.5
e_{11}	Oliver	Verknüpfung	2.0	16	3.6
e_{12}	Oliver	Abfrage	2.0	15	3.7

X=Aufgabe Y=Anzahl Clicks

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten

Univariate Urlisten

$x_1, \dots, x_N$

$y_1, \dots, y_N$

Univariate Wertebereiche

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

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

$W_Y = \{y(k) \mid k = 1,\dots, K\} = \{y(1), \dots, y(K)\}$

i	x_i	y_i
1	A	D
2	C	E
…	…	
N	B	E

Bivariate Urliste

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

Bivariater Wertebereich

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten

$x_1,\dots,x_N\ ;\ \ y_1,\dots,y_N$ $\qquad (x_1,y_1),\dots,(x_N,y_N)$ $\qquad d_i(j) = I_{x(e_i)=x(j)}$

$x_i \in W_X,\ y_i \in W_Y$ $\qquad (x_i,y_i) \in W_{XY} = W_X \times W_Y$ $\qquad e_i(k) = I_{y(e_i)=y(k)}$

i	**x_i**	**y_i**
1	A	D
2	C	E
…	…	
N	B	E

Dummykodierung →

i	**x_i**	**y_i**	**$d_i(1)$**	**$d_i(2)$**	**$d_i(3)$**	**$e_i(1)$**	**$e_i(2)$**
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			$N_{1\bullet}$	$N_{2\bullet}$	$N_{3\bullet}$	$N_{\bullet 1}$	$N_{\bullet 2}$

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten

Häufigkeitsverteilung eines bivariaten Merkmals

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

$$W_{XY} = \{(x[j],\ y[k]) \mid j = 1,\ldots,J;\ k = 1,\ldots,K\} = \left\{ \begin{matrix} (x[1],y[1]), & \cdots, & (x[1],y[K]), \\ (x[2],y[1]), & \cdots, & (x[2],y[K]), \\ \cdots & \cdots & \cdots, \\ (x[J],y[1]), & \ldots, & (x[J],y[K]) \end{matrix} \right\}$$

Gemeinsame absolute Häufigkeitsverteilung von x und y

$$N_{jk} = N((x[j],\ y[k])),\ j = 1,\ldots,J;\ k = 1,\ldots,K \qquad \begin{matrix} N_{11} & \cdots & N_{1K} \\ N_{21} & \cdots & N_{2K} \\ \cdots & \cdots & \cdots \\ N_{J1} & \ldots & N_{JK} \end{matrix}$$

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten

Häufigkeitsverteilung eines bivariaten Merkmals

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

$$W_{XY} = \{(x[j],\ y[k]) \mid j = 1,\ldots,J;\ k = 1,\ldots,K\} = \begin{Bmatrix} (x[1],y[1]), & \cdots, & (x[1],y[K]), \\ (x[2],y[1]), & \cdots, & (x[2],y[K]), \\ \cdots & \cdots & \cdots, \\ (x[J],y[1]), & \ldots, & (x[J],y[K]) \end{Bmatrix}$$

Gemeinsame relative Häufigkeitsverteilung von x und y

$$f_{jk} = \frac{N_{jk}}{N},\ \ j = 1,\ldots,J;\ k = 1,\ldots,K \qquad \begin{matrix} f_{11} & \cdots & f_{1k} \\ f_{21} & \cdots & f_{2k} \\ \cdots & \cdots & \cdots \\ f_{j1} & \cdots & f_{jk} \end{matrix}$$

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Häufigkeitsverteilung eines bivariaten Merkmals

i	x_i	y_i	$d_i(1)$	$d_i(2)$	$d_i(3)$	$e_i(1)$	$e_i(2)$
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			$N_{1\bullet}$	$N_{2\bullet}$	$N_{3\bullet}$	$N_{\bullet 1}$	$N_{\bullet 2}$

$$N_{jk} = N\big((x[j],\ y[k])\big) = \sum_{i\in\{l \mid e_l(k)=1\}} d_i(j) = \sum_{i\in\{l \mid d_l(j)=1\}} e_i(k) = \sum_{i=1}^{N} d_i(j)\cdot e_i(k)$$

$$N_{j\bullet} = \sum_{i=1}^{N} d_i(j) = \sum_{i\in\{l \mid e_i(1)=1\}} d_i(j) + \sum_{i\in\{l \mid e_i(2)=1\}} d_i(j) + \ldots + \sum_{i\in\{l \mid e_i(K)=1\}} d_i(j)$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{N} d_i(j)\cdot e_k(k) = \sum_{k=1}^{K} N_{jk}$$

i

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

Absolute Häufigkeiten

		Y				
		y(1)	y(2)	…	y(K)	Σ
X	x(1)	N_{11}	N_{12}	…	N_{1K}	$N_{1\bullet}$
	x(2)	N_{21}	N_{22}	…	N_{2K}	$N_{2\bullet}$
	…	…	…	…	…	…
	x(J)	N_{J1}	N_{J2}	…	N_{JK}	$N_{J\bullet}$
	Σ	$N_{\bullet 1}$	$N_{\bullet 2}$	…	$N_{\bullet K}$	N

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

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

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

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)	Σ
X	x(1)	N_{11}	N_{12}	…	N_{1K}	$N_{1\bullet}$
	x(2)	N_{21}	N_{22}	…	N_{2K}	$N_{2\bullet}$
	…	…	…	…	…	…
	x(J)	N_{J1}	N_{J2}	…	N_{JK}	$N_{J\bullet}$
	Σ	$N_{\bullet 1}$	$N_{\bullet 2}$	…	$N_{\bullet K}$	N

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

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

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

		Y				
		y(1)	y(2)	…	y(K)	Σ
X	x(1)	N_{11}	N_{12}	…	N_{1K}	$N_{1\bullet}$
	x(2)	N_{21}	N_{22}	…	N_{2K}	$N_{2\bullet}$
	…	…	…	…	…	…
	x(J)	N_{J1}	N_{J2}	…	N_{JK}	$N_{J\bullet}$
	Σ	$N_{\bullet 1}$	$N_{\bullet 2}$	…	$N_{\bullet K}$	N

Absolute Randhäufigkeitsverteilung von X

Absolute Randhäufigkeitsverteilung von Y

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

Relative Häufigkeiten

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

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

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

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

Gemeinsame relative Häufigkeitsverteilung f_{XY} von X und Y

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

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

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

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

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

$$f_{X\bullet} = \{f_{j\bullet} \mid j = 1,\ldots,J\}$$

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

Relative Randhäufigkeitsverteilung
$f_{X\bullet}$ von X

Relative Randhäufigkeitsverteilung
$f_{\bullet Y}$ von Y

$$f_{\bullet Y} = \{f_{\bullet k} \mid k = 1,\ldots,K\}$$

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)	Σ
X	x(1)					
	x(2)	N_{21}	N_{22}	…	N_{2K}	$N_{2\bullet}$
	…					
	x(J)					

Dieser Datensatz hat Umfang $N_{2\bullet}$

Absolute Häufigkeitsverteilung:

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

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)	Σ
x(1)					
X 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 $N_{2\bullet}$

Relative Häufigkeitsverteilung:

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

		Y				
		y(1)	y(2)	…	y(K)	Σ
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
	Σ					J

Bedingte Verteilung
von Y gegeben X=x(2)

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Nominale Daten: Darstellung einer bivariaten Häufigkeitsverteilung

Kontingenztafel

		Y				
		y(1)	y(2)	…	y(K)	Σ
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
	Σ					J

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

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

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

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

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

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

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

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			
		Abfrage	Export	Verknüpfung	Σ
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
	Σ	2/12	6/12	4/12	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 bedingt auf Bearbeiter(in)

		Aufgabe			
		Abfrage	Export	Verknüpfung	Σ
Bear-bei-ter(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
	Σ				4

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 Bearbeiter(in) bedingt auf Aufgabe

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

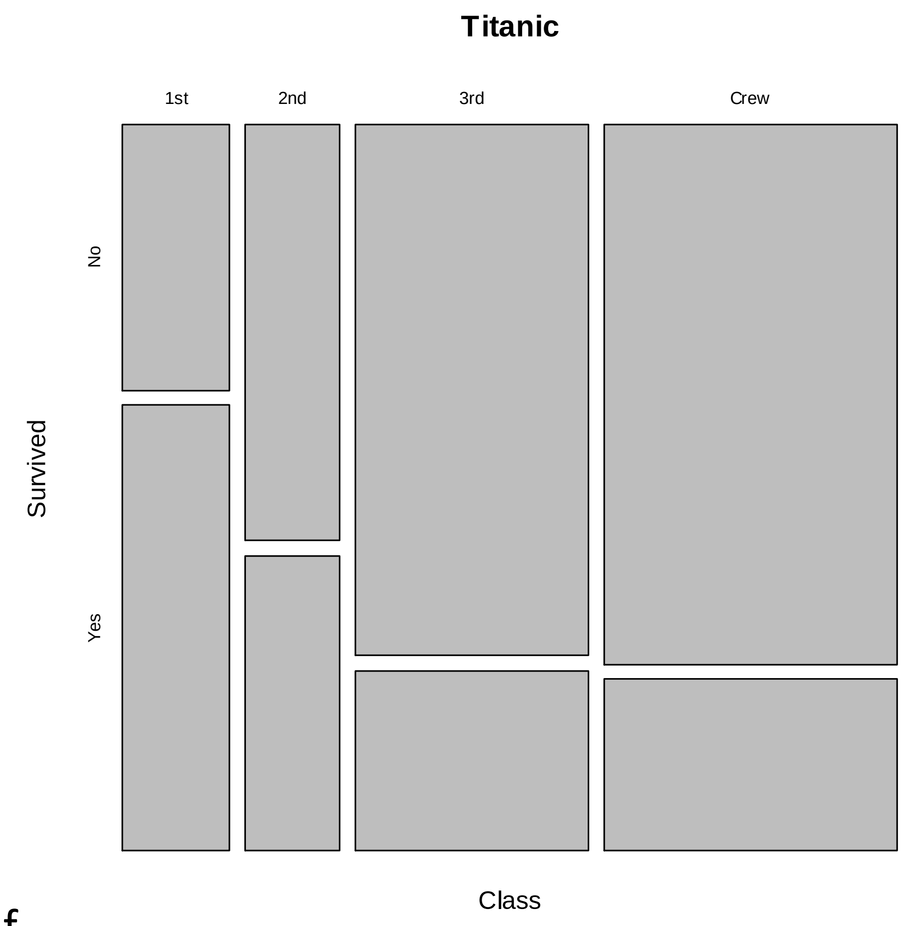
Bivariate Daten: Tabellarische und grafische Darstellungen

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 $f_{SC} = f_{S|C} \cdot f_{\bullet C}$

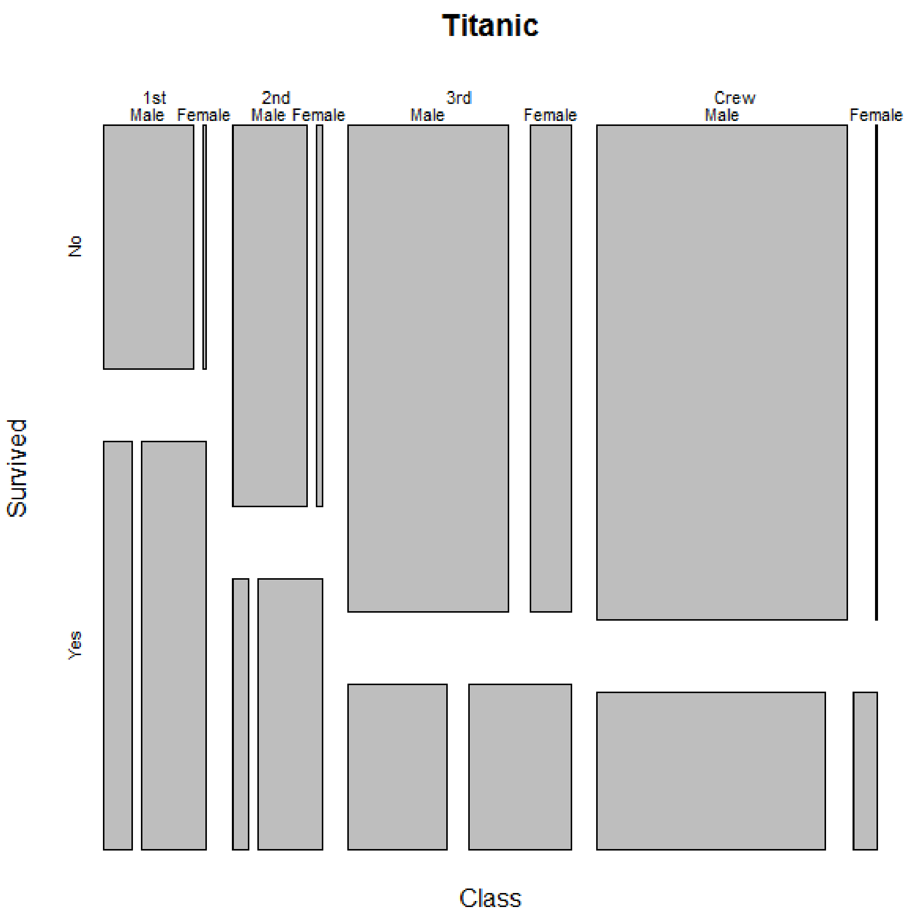
Bivariate Daten: Tabellarische und grafische Darstellungen

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

Bivariate Daten: Tabellarische und grafische Darstellungen

Ordinale Daten

- Kontingenztafeln und Mosaikplots mit geordneten Kategorien

Quantitative Daten

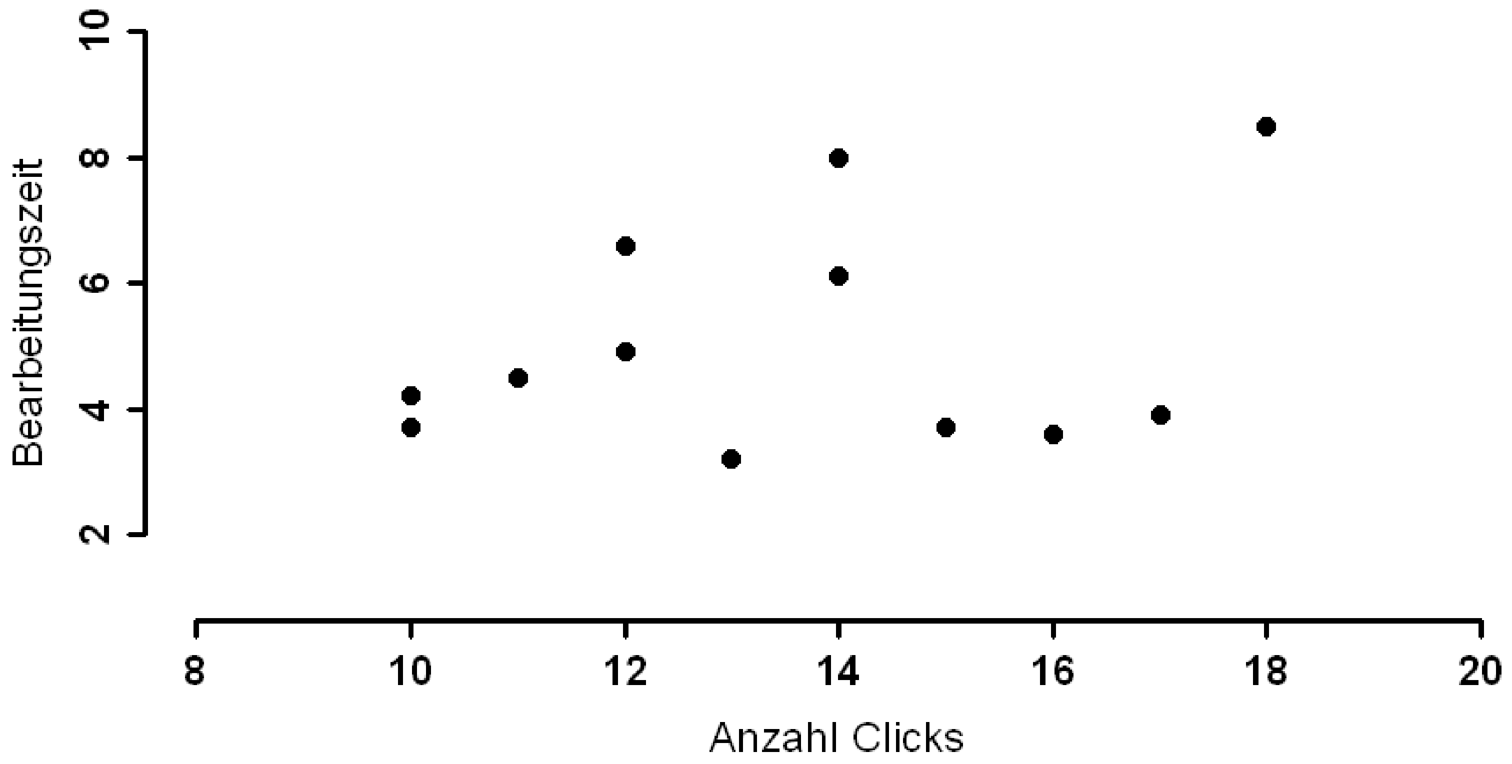
- Kontingenztafeln und Mosaikplots mit klassierten Daten
- Streudiagramme!

Bivariate Daten: Tabellarische und grafische Darstellungen

Quantitative Daten : Beispiel **Bearbeitungen von Softwareaufgaben**

Streudiagramm

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

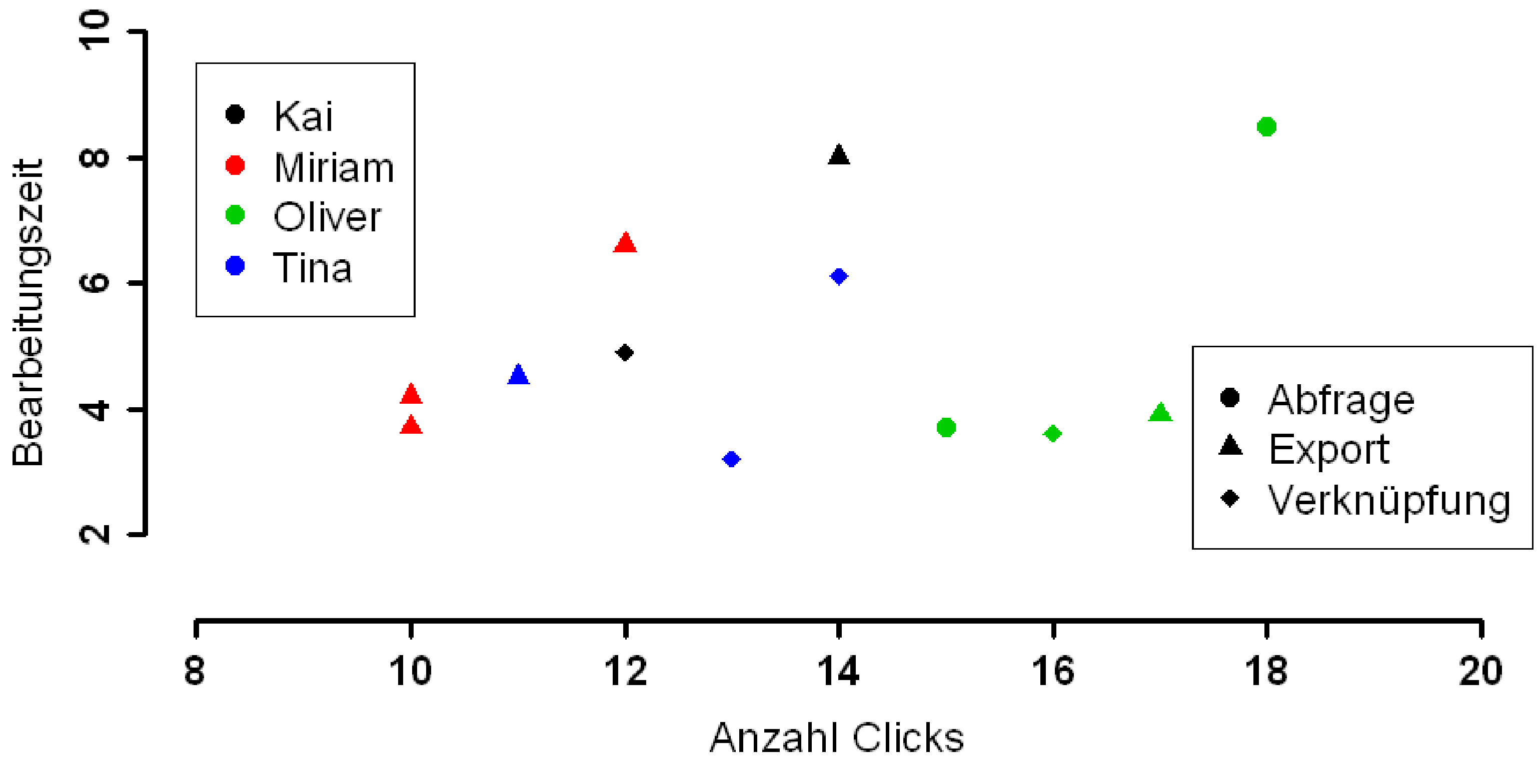


Bivariate Daten: Tabellarische und grafische Darstellungen

Quantitative Daten : Beispiel **Bearbeitungen von Softwareaufgaben**

Streudiagramm

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



Bivariate Daten: Zusammenhangsmaße

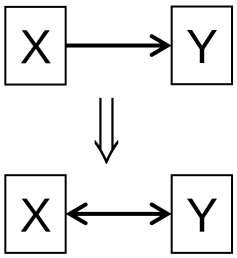
- Erinnerung: Allgemeine Eigenschaft der Streuung univariater Daten: Streuung von X desto höher, je schlechter sich konkrete Werte vorhersagen lassen.
  - Bisher: Vorhersage der Werte von X durch einzelnen Lageparameter.
  - **Jetzt: Vorhersage der Werte von Y unter Verwendung der Werte von X**.
- Allgemein: 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*)

Bivariate Daten: Zusammenhangsmaße

Korrelation und Kausalität

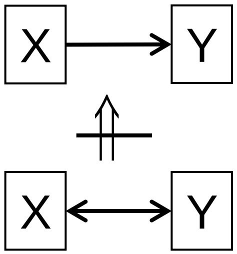
Es gilt:

X ist Ursache von Y   =>   X und Y korrelieren



Aber:

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



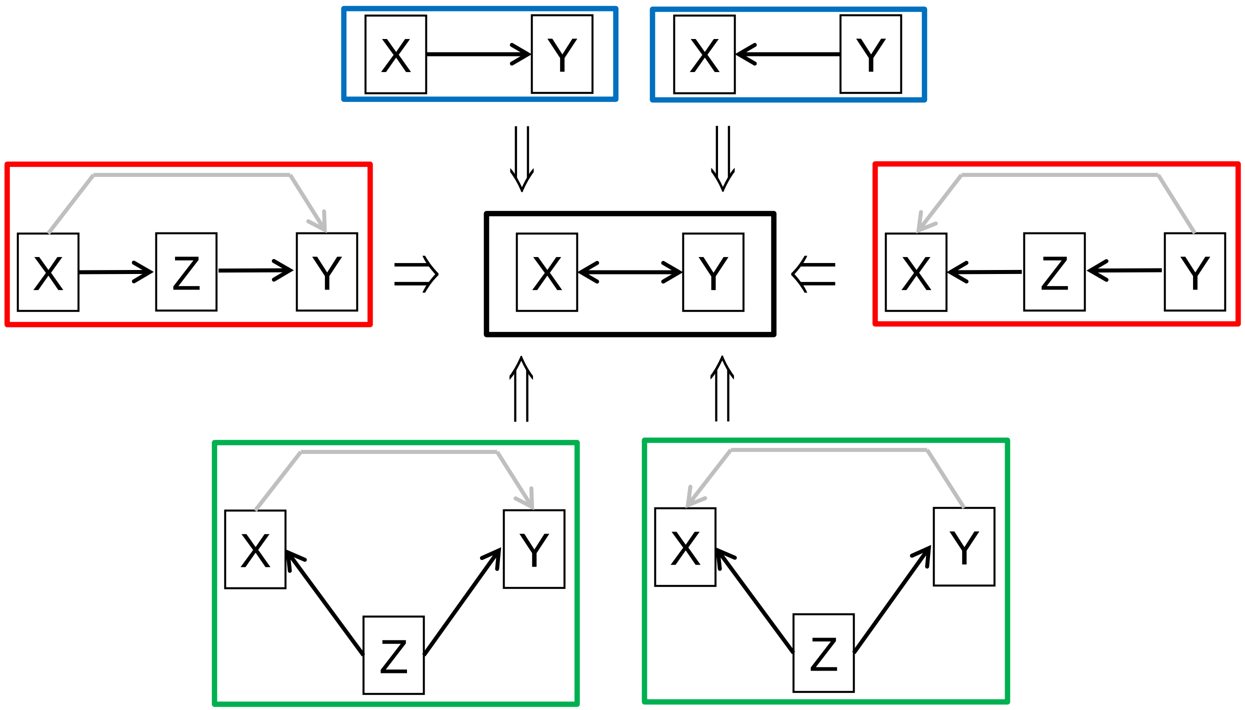
Bivariate Daten: Zusammenhangsmaße

Korrelation und Kausalität

X ist Ursache von Y => X und Y korrelieren

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

Verschiedene Korrelationsquellen möglich



Bivariate Daten: Zusammenhangsmaße

Nominale Daten

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).

		Y				
		y(1)	y(2)	…	y(K)	Σ
X	x(1)	$f_{y;1\|1}$	$f_{y;2\|1}$	…	$f_{y;K\|1}$	1
	x(2)	$f_{y;1\|2}$	$f_{y;2\|2}$	…	$f_{y;K\|2}$	1
	…	…	…	…	…	…
	x(J)	$f_{y;1\|J}$	$f_{y;2\|J}$	…	$f_{y;K\|J}$	1
		$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	

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 $f_{\bullet Y}$ von Y abweicht.

Bivariate Daten: Zusammenhangsmaße

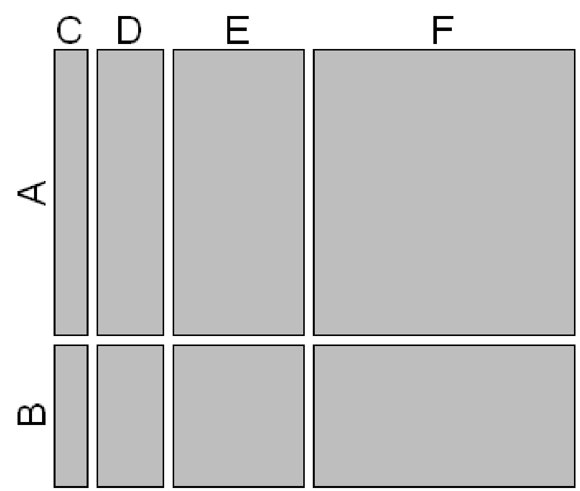
Nominale Daten

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)	…	y(K)	Σ
	x(1)	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	1
X	x(2)	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	1
	…	…	…	…	…	…
	x(J)	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	1
		$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	

Zusammenhang minimal, falls

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



Bivariate Daten: Zusammenhangsmaße

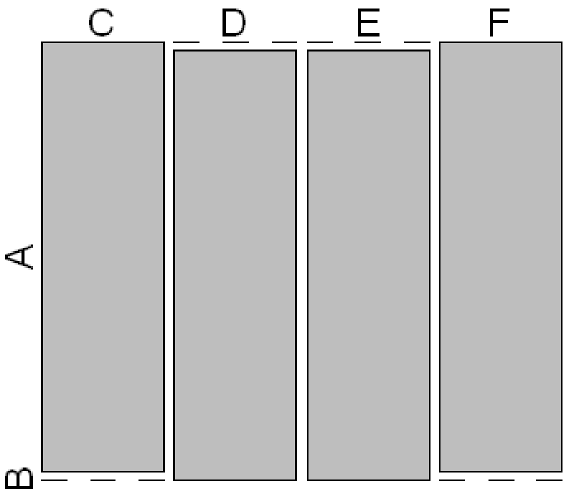
Nominale Daten

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)	…	y(K)	Σ
X	x(1)	0	1	…	0	1
	x(2)	0	0	…	1	1
	…	…	…	…	…	…
	x(J)	1	0	…	0	1
		$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	

Zusammenhang maximal, falls es für alle

$$j \in \{1,\ldots,J\} \text{ ein } k \in \{1,\ldots,K\} \text{ mit } f_{y;k|j} = 1 \text{ gibt}$$



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				Σ
		y(1)	y(2)	…	y(K)	
X	x(1)	$f_{y;1\|1}$	$f_{y;2\|1}$	…	$f_{y;K\|1}$	1
	x(2)	$f_{y;1\|2}$	$f_{y;2\|2}$	…	$f_{y;K\|2}$	1
	…	…	…	…	…	…
	x(J)	$f_{y;1\|J}$	$f_{y;2\|J}$	…	$f_{y;K\|J}$	1
		$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	

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.

Bivariate Daten: Zusammenhangsmaße

Nominale Daten

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.

		Y				
		y(1)	y(2)	…	y(K)	Σ
X	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)	$f_{0;J1}$	$f_{0;J2}$	…	$f_{0;JK}$	$f_{J\bullet}$
	Σ	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet K}$	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.

Bivariate Daten: Zusammenhangsmaße

Nominale Daten

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.

		Y				Σ
		y(1)	y(2)	…	y(K)	
X	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}	$N_{J\bullet}$
	Σ	$N_{\bullet 1}$	$N_{\bullet 2}$	…	$N_{\bullet K}$	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.

Bivariate Daten: Zusammenhangsmaße

Nominale Daten

Je größer die beobachteten Anzahlen N_{jk} von den erwarteten ν_{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 **χ^2-Größe**

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

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

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: die **χ^2-Größe**

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

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

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

$$= \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{Nf_{j\bullet}^2\left(\frac{f_{jk}}{f_{j\bullet}}-f_{\bullet k}\right)^2}{f_{j\bullet}f_{\bullet k}} = \boxed{\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{Nf_{j\bullet}(f_{y;k|j}-f_{\bullet k})^2}{f_{\bullet k}}}$$

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: die **χ^2-Größe**

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

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

Beweis:

$0 \le \chi^2$ klar wegen $N_{j\bullet} > 0$, $N_{\bullet k} > 0$, $(N_{jk} - \nu_{jk})^2 \ge 0$

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

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: die **χ^2-Größe**

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

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

Sei o.B.d.A. $K \le J$.

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

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{N_{jk}^2}{N_{j\bullet}N_{\bullet k}} = K \Leftrightarrow \frac{N_{jk}}{N_{j\bullet}} = 1,$$

d.h. χ^2 wird maximal, wenn es zu jedem j ein k(j) mit $f_{y,k(j)|j} = 1$ gibt.

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: die **χ^2-Größe**

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

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

(Korrigierter) Kontingenzkoeffizient nach Pearson:

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

Eliminiert Abhängigkeit des Koeffizienten vom Stichprobenumfang N und von der Dimension min(J,K)

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

N_{jk}

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

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

v_{jk}

		Aufgabe			
		Abfrage	Export	Verknüpfung	Σ
Bearbei-ter(in)	Kai	0 2·2/12=**1/3**	1 2·6/12=**1**	1 2·4/12=**2/3**	2
	Miriam	0 3·2/12=**1/2**	3 3·6/12=**3/2**	0 3·4/12=**1**	3
	Oliver	2 4·2/12=**2/3**	1 4·6/12=**2**	1 4·4/12=**4/3**	4
	Tina	0 3·2/12=**1/2**	1 3·6/12=**3/2**	2 3·4/12=**1**	3
	Σ	2	6	4	12

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$(N_{jk}-\nu_{jk})^2$

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

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$(N_{jk}-\nu_{jk})^2/\nu_{jk}$

		Aufgabe			
		Abfrage	Export	Verknüpfung	Σ
Bearbei-ter(in)	Kai	0 1·3/(9·1)=**1/3**	1 0/1=**0**	1 1·3/(9·2)=**1/6**	2
	Miriam	0 1·2/(4·1)=**1/2**	3 9·2/(4·3)=**3/2**	0 1/1=**1**	3
	Oliver	2 16·3/(9·2)=**8/3**	1 **1/2**	1 1·3/(9·4)=**1/12**	4
	Tina	0 1·2/(4·1)=**1/2**	1 1·2/(4·3)=**1/6**	2 1/1=**1**	3
	Σ	2	6	4	12

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(N_{jk}-\nu_{jk})^2}{\nu_{jk}} = \frac{1}{12}\begin{pmatrix}4+6+32+6\\+0+18+6+2\\+2+12+1+12\end{pmatrix} = \frac{101}{12} = 8\frac{5}{12} \approx 8.417$$

$(N_{jk}-\nu_{jk})^2/\nu_{jk}$

		Aufgabe			
		Abfrage	Export	Verknüpfung	Σ
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
	Σ	2	6	4	12

Bivariate Daten: Zusammenhangsmaße

Nominale Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

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

$(N_{jk} - \nu_{jk})^2 / \nu_{jk}$

		Aufgabe			
		Abfrage	Export	Verknüpfung	Σ
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
	Σ	2	6	4	12

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.

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 quantit. 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.

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).

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

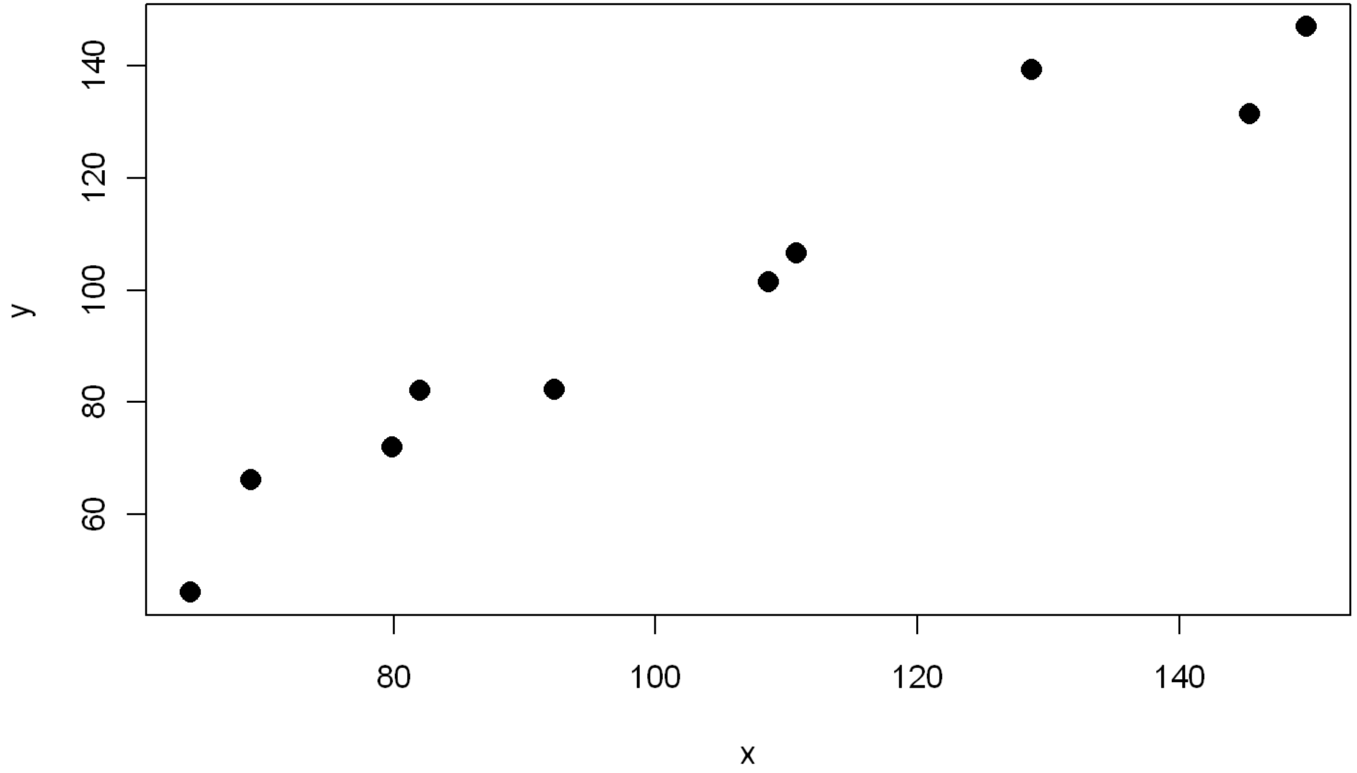
$s_{xy} > 0$, wenn hohe Werte von X in hohem Maße mit hohen Werten von Y einhergehen (Positive Korrelation)

$s_{xy} < 0$, 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)

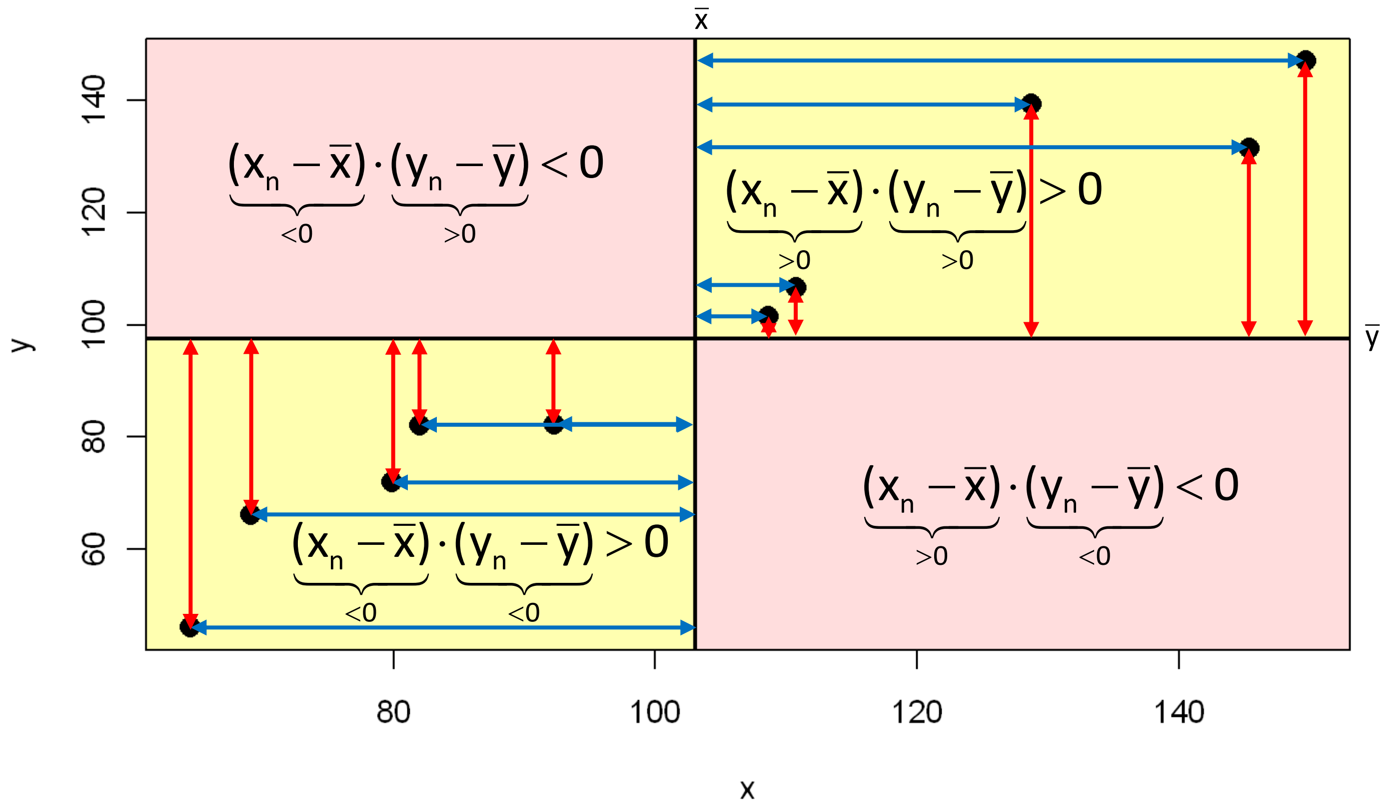
Bivariate Daten: Zusammenhangsmaße

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



Bivariate Daten: Zusammenhangsmaße

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



Bivariate Daten: Zusammenhangsmaße

Quantitative Daten

Kovarianz

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

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

$$\boxed{s_{xy}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \overline{x})(y_n - \overline{y}) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n y_n - x_n\overline{y} - \overline{x}y_n + \overline{x}\cdot\overline{y})$$

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

$$= \frac{N}{N-1}\overline{xy} - \frac{N}{N-1}\overline{x}\cdot\overline{y} - \frac{N}{N-1}\overline{x}\cdot\overline{y} + \frac{N}{N-1}\overline{x}\cdot\overline{y} = \boxed{\frac{N}{N-1}\left(\overline{xy} - \overline{x}\cdot\overline{y}\right)}$$

$\square$

Bivariate Daten: Zusammenhangsmaße

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

Beweis: Spezialfall der Cauchy-Schwarz-Ungleichung:

$$\text{für}(a_n, b_n) \in \Re^2,\ \text{gilt}\left(\sum_{n=1}^{N} a_n b_n\right)^2 \leq \sum_{n=1}^{N} a_n^2 \cdot \sum_{n=1}^{N} b_n^2 \Rightarrow \left(\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})\right)^2 \leq \sum_{n=1}^{N}(x_n - \bar{x})^2 \cdot \sum_{n=1}^{N}(y_n - \bar{y})^2$$

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

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

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

Bivariate Daten: Zusammenhangsmaße

Quantitative Daten: **Korrelationskoeffizient nach Bravais-Pearson**

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad -s_x s_y \leq s_{xy} \leq s_x s_y \Rightarrow -1 \leq r_{xy} \leq 1$$

Gleichheitsbedingung bei der Cauchy-Schwarz-Ungleichung:

Für $(a_n, b_n) \in \Re^2$ gilt

$$\left(\sum_{n=1}^{N} a_n b_n \right)^2 = \sum_{n=1}^{N} a_n^2 \cdot \sum_{n=1}^{N} b_n^2 \Leftrightarrow \text{ es gibt Konstanten } c \text{ und } d \text{ mit } b_n = c + d \cdot a_n \text{ für alle } n$$

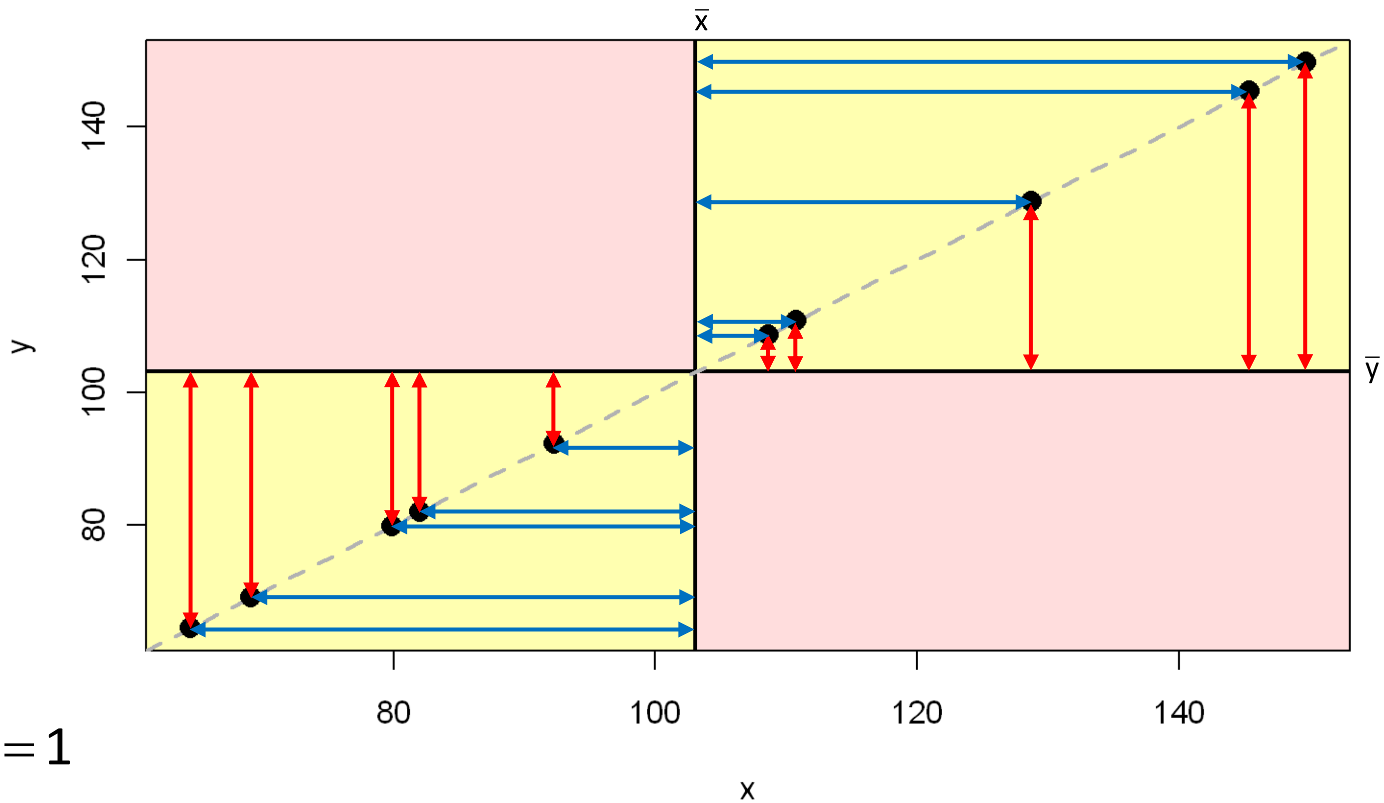
$$\Rightarrow \quad r_{xy} \in \{-1,1\} \Leftrightarrow \left(y_n \text{-} \bar{y} \right) = c + d \cdot \left(x_n \text{-} \bar{x} \right)$$

$$\Leftrightarrow y_n = \tilde{c} + d \cdot x_n \qquad \text{mit } \tilde{c} = c + \bar{y} \text{-} d\bar{x}$$

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

Bivariate Daten: Zusammenhangsmaße

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

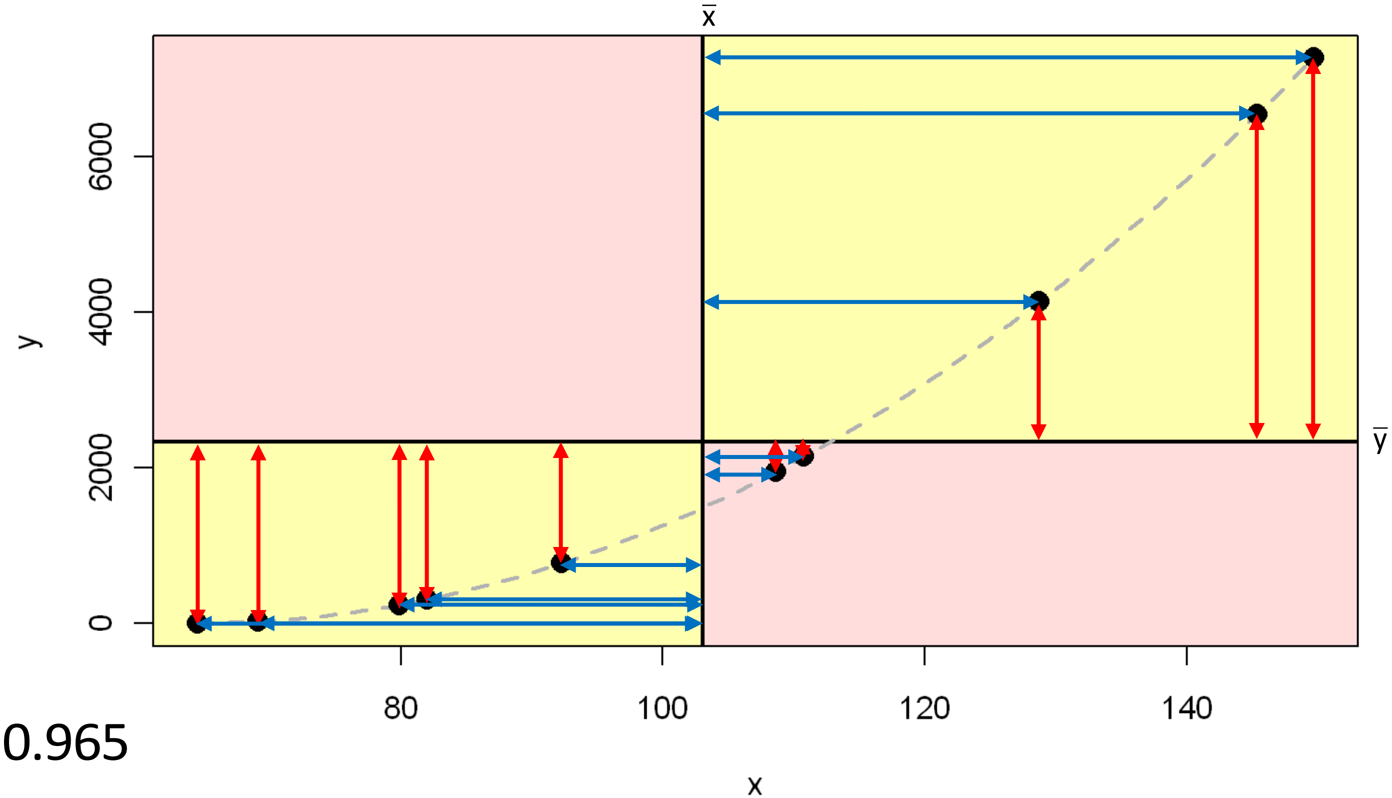


$r_{xy} = 1$

Bivariate Daten: Zusammenhangsmaße

Quantitative Daten: **Korrelationskoeffizient nach Bravais-Pearson**

Nicht-linearer monotoner Zusammenhang

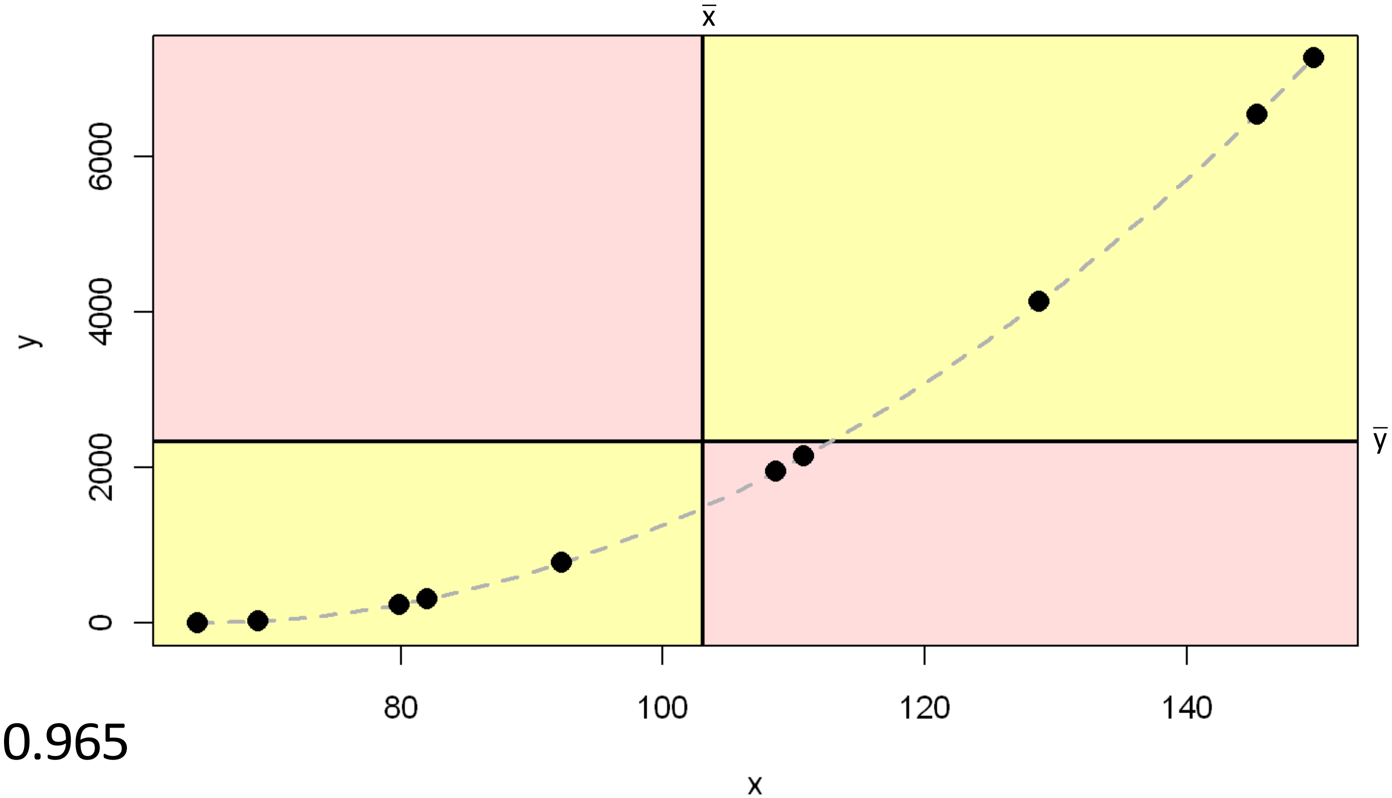


$r_{xy} = 0.965$

Bivariate Daten: Zusammenhangsmaße

Ordinale/Quantitative Daten: Nicht-linearer monotoner Zusammenhang

Übergang zu Rängen

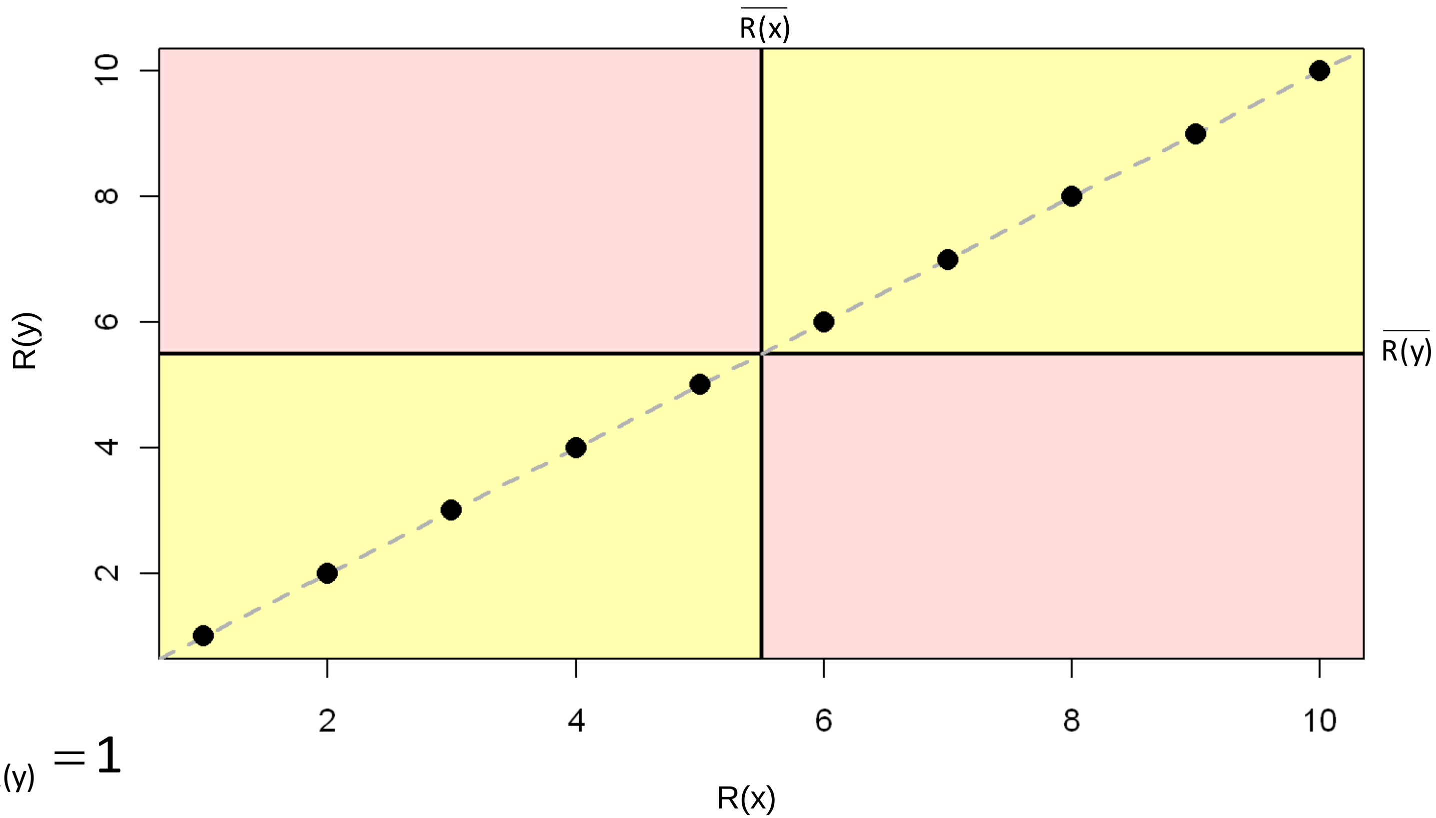


$$r_{xy} = 0.965$$

Bivariate Daten: Zusammenhangsmaße

Ordinale/Quantitative Daten: Nicht-linearer monotoner Zusammenhang

Übergang zu Rängen

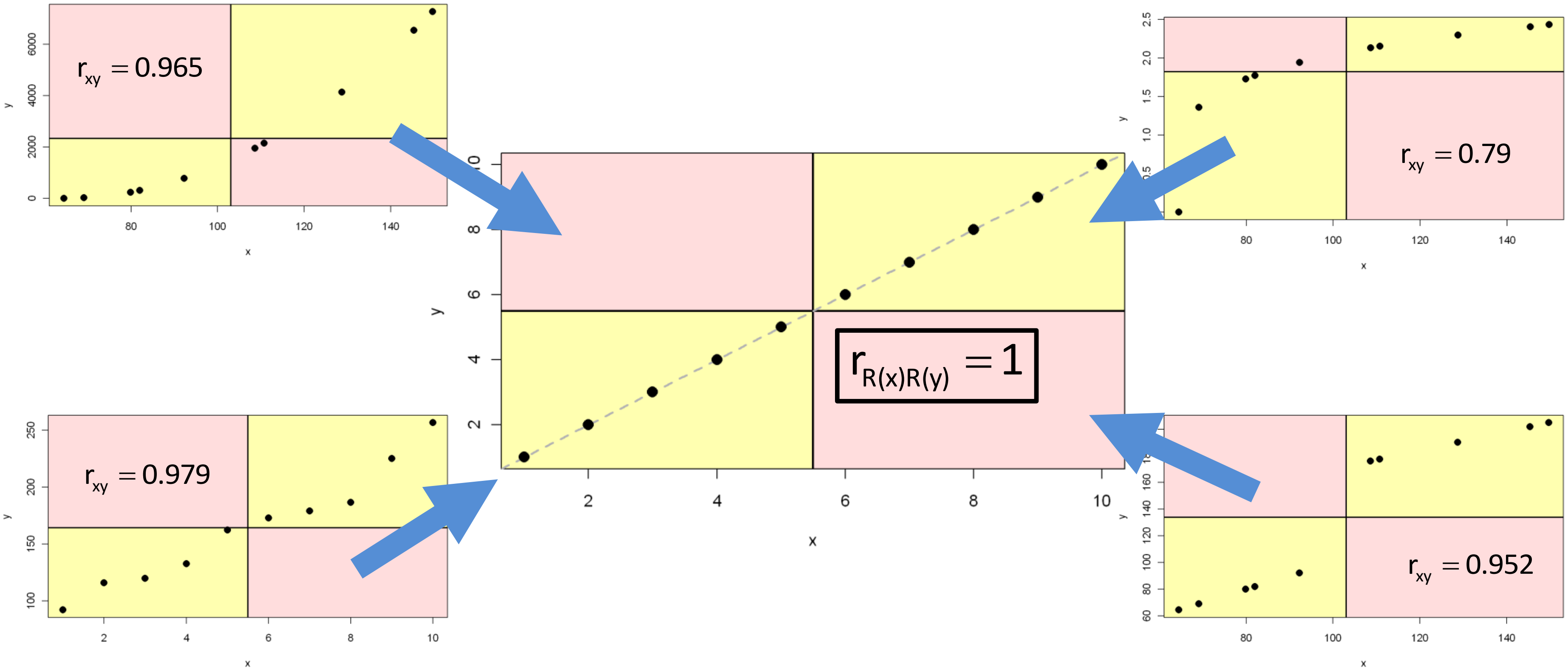


$$r_{R(x)R(y)} = 1$$

Bivariate Daten: Zusammenhangsmaße

Ordinale/Quantitative Daten

Absolute Korrelation von Rängen bei monotonem Zusammenhang immer 1



Bivariate Daten: Zusammenhangsmaße

Ordinale/Quantitative Daten

Falls X und Y mindestens ordinales Skalenniveau haben, so wird der Bravais-Pearson-Korrelationskoeffizient der Ränge R(X) und R(X) von X und Y der **Spearmansche Rangkorrelationskoeffizient** r^{Sp}_{xy} von X und Y genannt:

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

Bivariate Daten: Zusammenhangsmaße

Ordinale/Quantitative Daten

Spearmanscher 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:

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

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

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

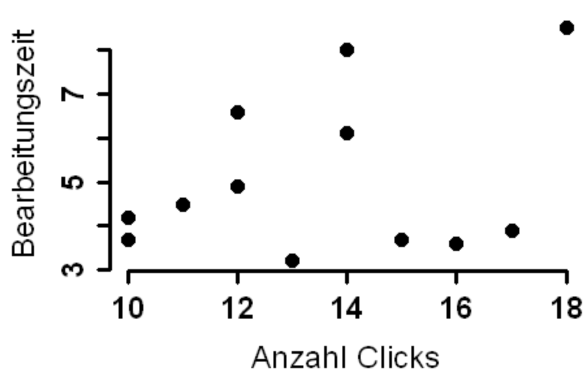
Bivariate Daten: Zusammenhangsmaße

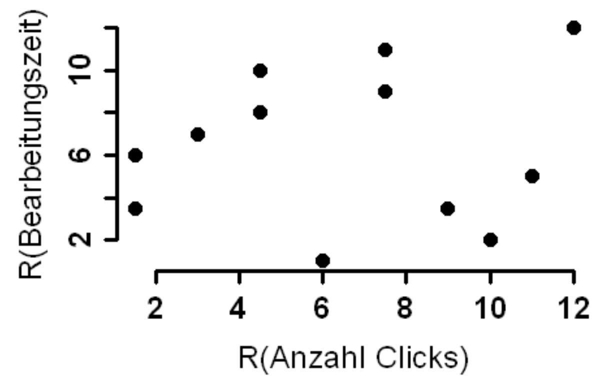
Ordinale/Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

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

$$\boxed{\bar{x}_4 = 13.5 \quad s^2_{x_4} = 7}$$

$$\boxed{\bar{x}_5 = 5.075 \quad s^2_{x_5} = 3.24}$$





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

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

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.

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?

Bivariate Daten: Lineare Regression

Quantitative Daten

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

Für beliebiges (j,k) mit j≠k:

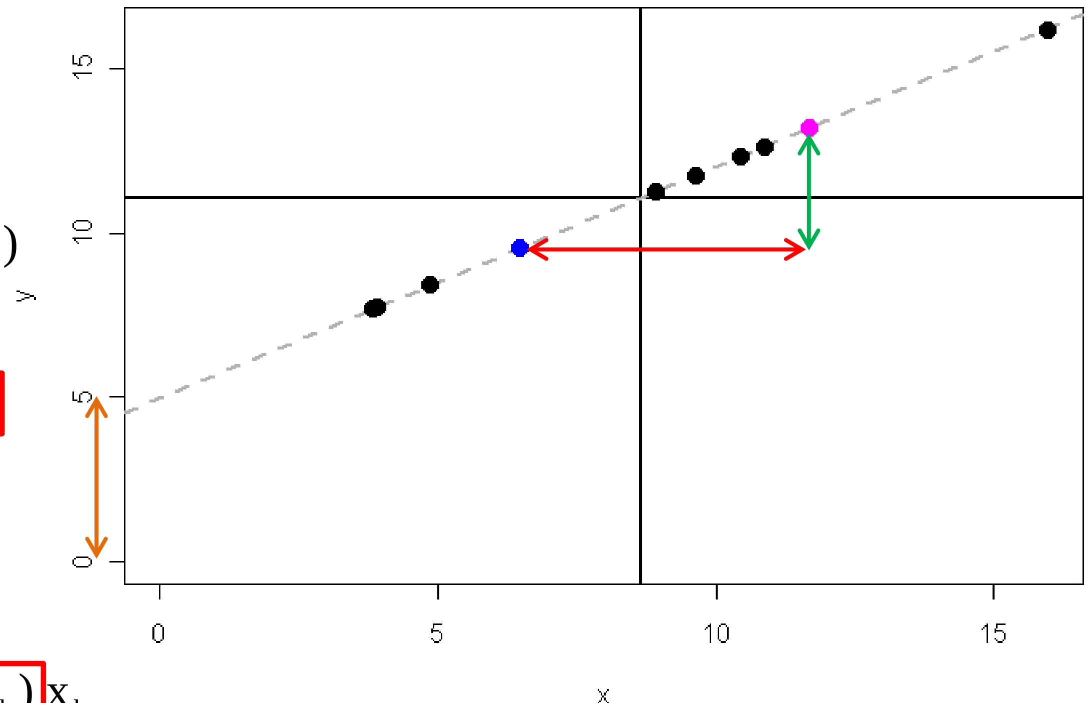
$$\boxed{y_j = c + dx_j}$$

$$\boxed{y_k = c + dx_k}$$

$$\Rightarrow \quad y_j - y_k = (c + dx_j) - (c + dx_k)$$

$$= d(x_j - x_k)$$

$$\Leftrightarrow \quad \boxed{d} = \boxed{(y_j - y_k)} / \boxed{(x_j - x_k)}$$



$$y_k = c + dx_k$$

$$\Leftrightarrow \quad \boxed{c} = y_k - dx_k$$

$$= y_k - \boxed{(y_j - y_k)} / \boxed{(x_j - x_k)} x_k$$

Perfekte Vorhersage durch Einsetzen in die Gleichung.

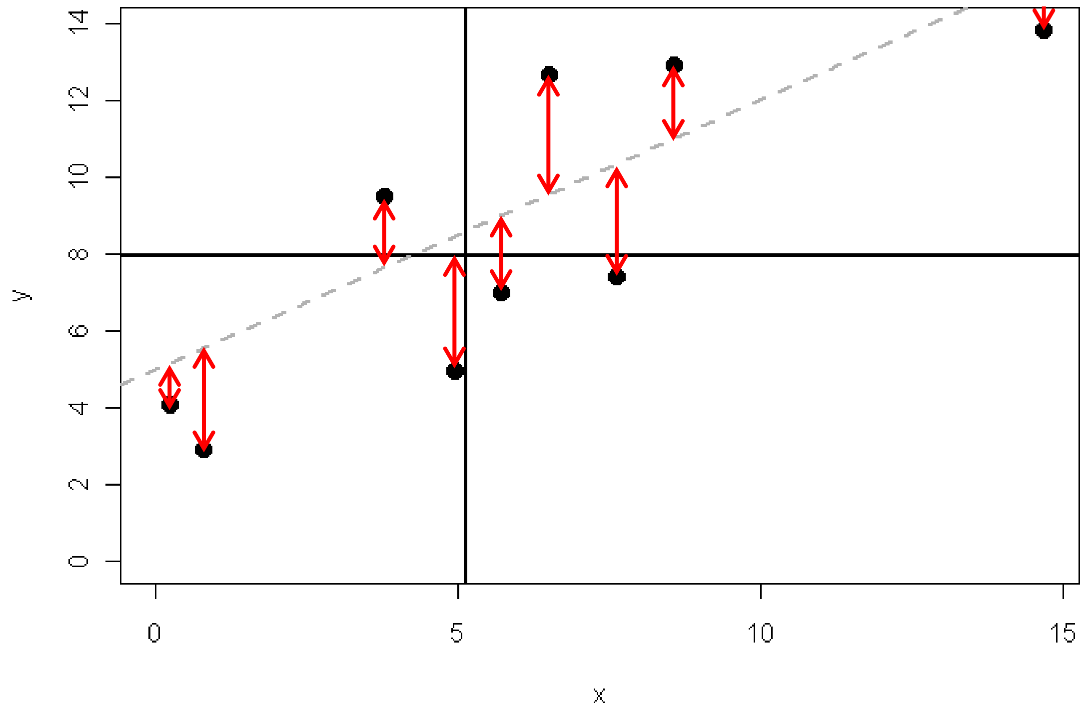
Bivariate Daten: Lineare Regression

Quantitative Daten

$0 < |r_{xy}| < 1 \Leftrightarrow y_n = c + dx_n + \boxed{\varepsilon_n}$ für n=1,...,N

Vorhersagefehler

$\boxed{\varepsilon_n} = y_n - c - dx_n$



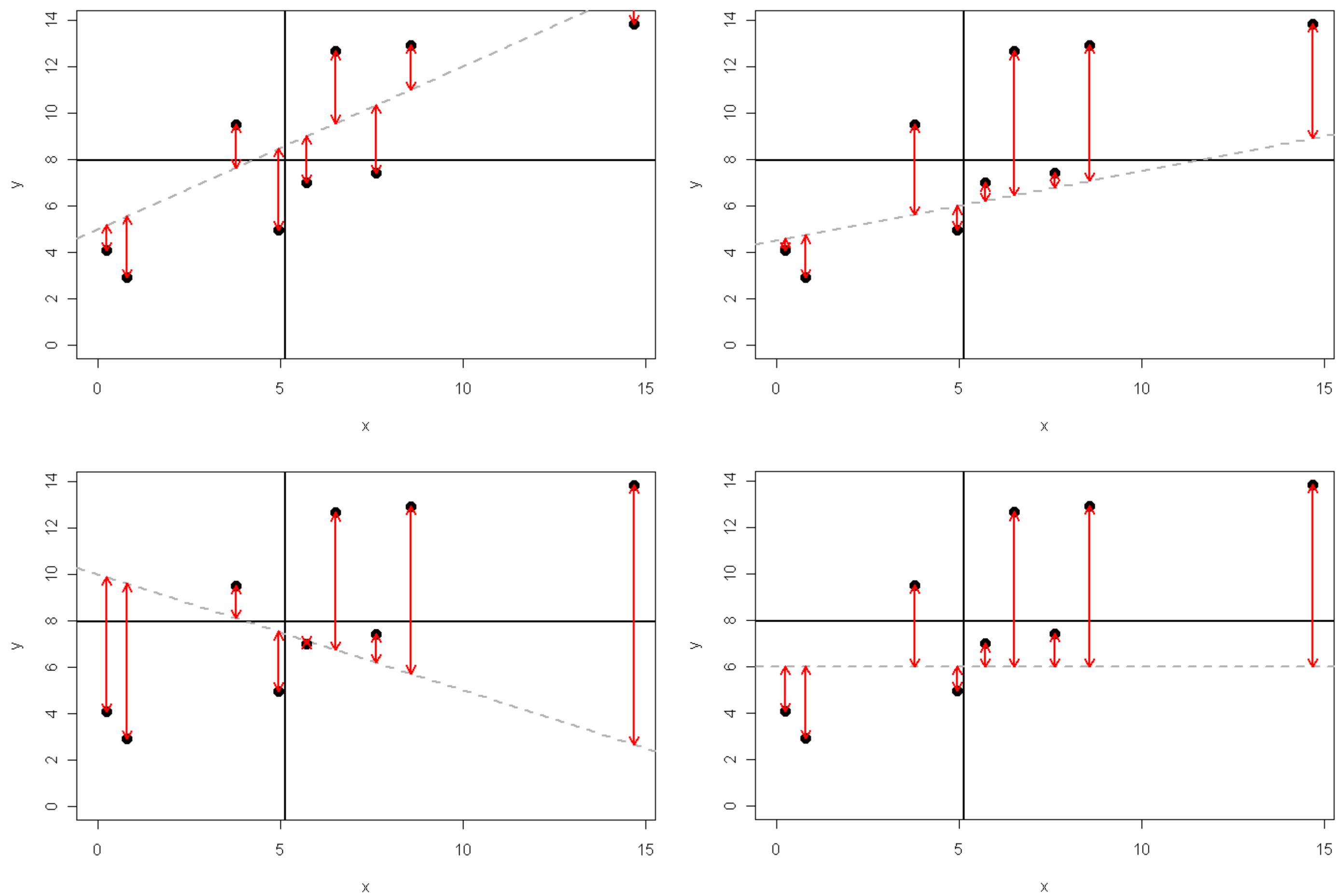
Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

Koeffizienten c und d so bestimmen, dass Fehlerquadrat-summe

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

minimal wird.



Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

Die Fehlerquadratsumme $Q(c,d) = \sum_{n=1}^{N} \left(y_n - c - dx_n\right)^2$ 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

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

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

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

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

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

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

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

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

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

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

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

□

Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme

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

$$\sum_{n=1}^{N} \varepsilon_n^2 = \sum_{n=1}^{N}\left(y_n - (\bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}) - \frac{s_{xy}}{s_x^2}x_n\right)^2 = \sum_{n=1}^{N}\left(y_n - (\bar{y} - r_{xy}\frac{s_y}{s_x}\bar{x}) - r_{xy}\frac{s_y}{s_x}x_n\right)^2$$

$$= \sum_{n=1}^{N}\left((y_n - \bar{y}) - r_{xy}\frac{s_y}{s_x}(x_n - \bar{x})\right)^2 = \sum_{n=1}^{N}\left((y_n - \bar{y})^2 - 2r_{xy}\frac{s_y}{s_x}(y_n - \bar{y})(x_n - \bar{x}) + \left(r_{xy}\frac{s_y}{s_x}\right)^2(x_n - \bar{x})^2\right)$$

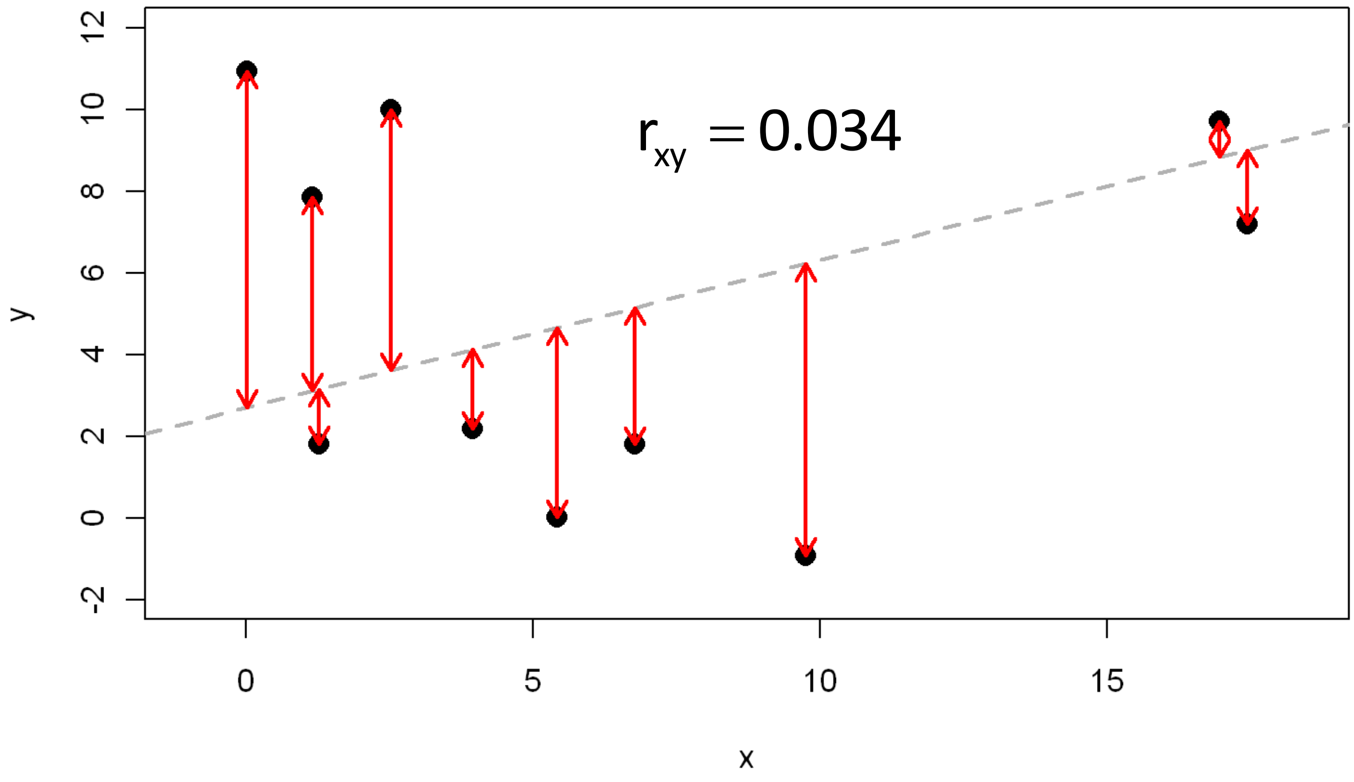
$$= (N-1)\cdot\left(s_y^2 - 2r_{xy}\frac{s_y}{s_x}s_{xy} + \left(r_{xy}\frac{s_y}{s_x}\right)^2 s_x^2\right) = (N-1)\cdot\left(s_y^2 - 2r_{xy}^2 s_y^2 + r_{xy}^2 s_y^2\right)$$

$$= (N-1)\cdot\left(s_y^2 - r_{xy}^2 s_y^2\right) \qquad \square$$

Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

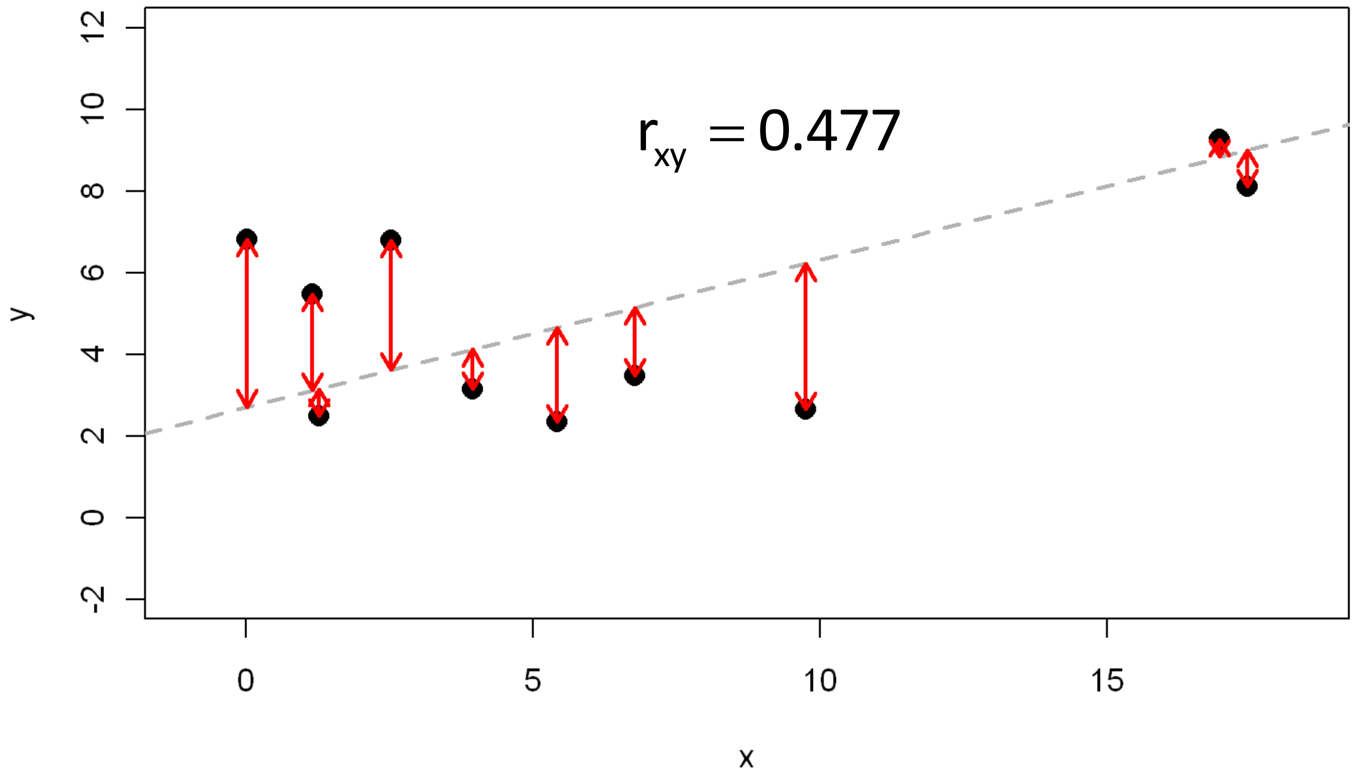
Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

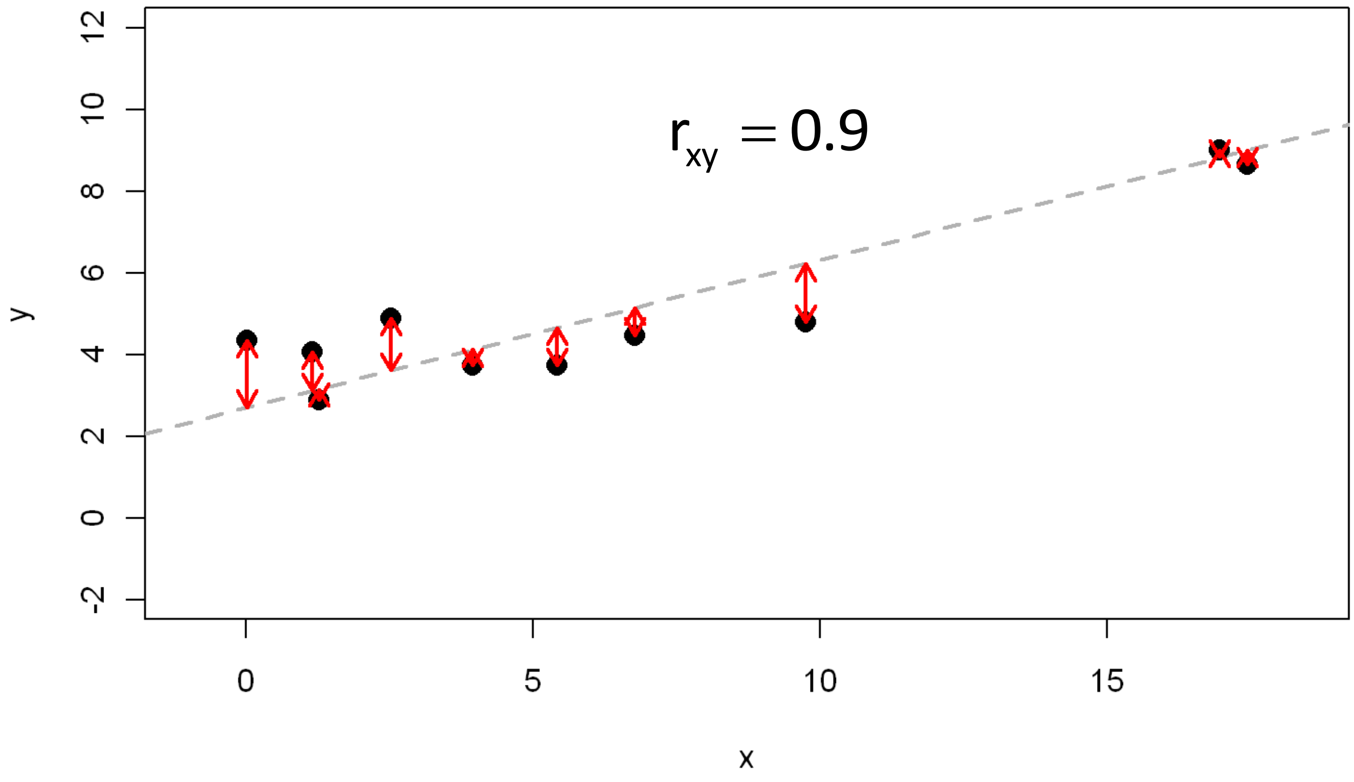
Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

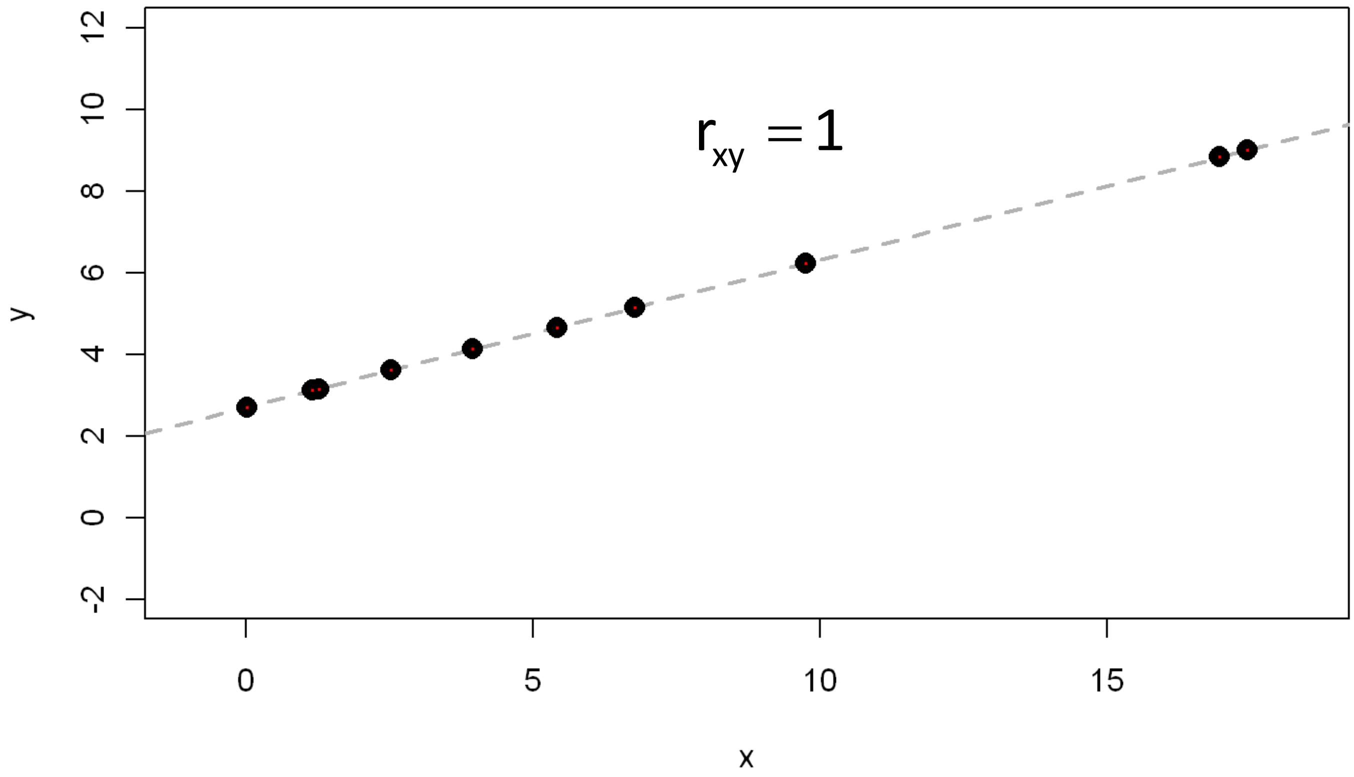
Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



Bivariate Daten: Lineare Regression

Quantitative Daten: **Methode der kleinsten Quadrate**

Je größer die absolute Korrelation, desto kleiner die Fehlerquadratsumme



Bivariate Daten: Lineare Regression

Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Anzahl Clicks	Bear-beitungszeit	$c+dx_4$	ε
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

$$\bar{x}_4 = 13.5 \qquad s^2_{x_4} = 7$$

$$\bar{x}_5 = 5.075 \qquad s^2_{x_5} = 3.24$$

$$r_{x_4 x_5} = 0.301$$

$$\boxed{x_5 = c + dx_4 + \varepsilon}$$

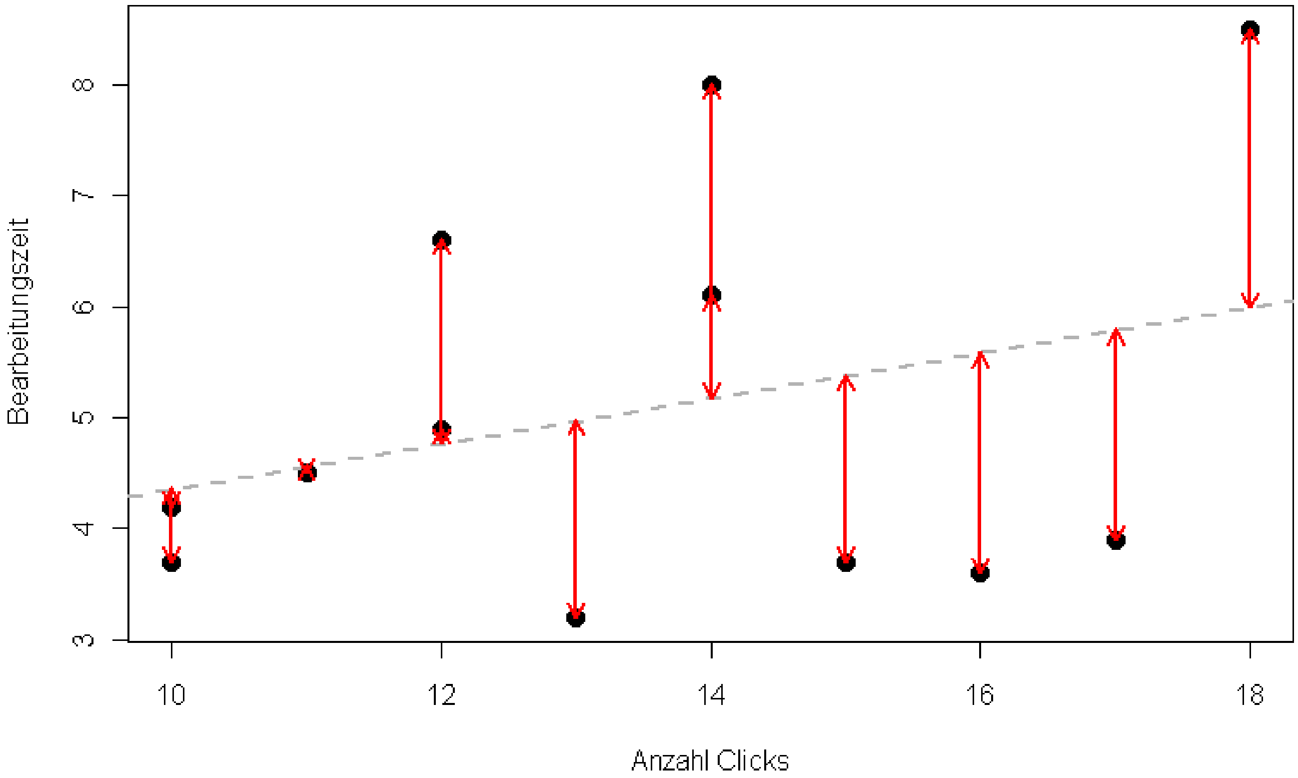
$$\boxed{c} = \bar{x}_5 - r_{x_4x_5}\frac{s_{x_5}}{s_{x_4}}\bar{x}_4 = 5.075 - 0.301\sqrt{\frac{3.24}{7}}\,13.5 = \boxed{2.314}$$

$$\boxed{d} = r_{x_4x_5}\frac{s_{x_5}}{s_{x_4}} = 0.301\sqrt{\frac{3.24}{7}} = \boxed{0.205}$$

Bivariate Daten: Lineare Regression

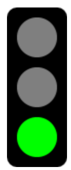
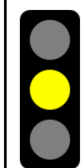
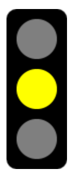
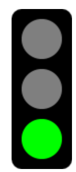
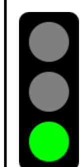
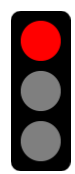
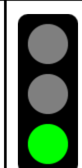
Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$$x_5 = 2.314 + 0.205 x_4 + \boxed{\varepsilon}$$



Bivariate Daten: Lineare Regression

Zusammenfassung

Skalenniveau → **↓Zusammenhangsmaß**	**Nominal**	**Ordinal**	**Quantitativ**
χ^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		 – Ausreißeranfällig – Lin. Zusammenhang + Informationsnutzung