Themen: Zusammenhangsmaße für intervall-skalierte (metrische) Variablen – Kovarianz & Pearson's *r*

Prof. Dr. Elmar Schlüter
Justus-Liebig-Universität Giessen
Fachbereich Sozial- und Kulturwissenschaften
Institut für Soziologie
Wintersemester 2014/2015

Inhalt

- Grafische Veranschaulichung bivariater Zusammenhänge
  - Scatterplot
- Kreuzproduktsumme
- Kovarianz
- Pearson‘s *r*

Materialien in Anlehnung an Christ (2012) & Pöge (2010)

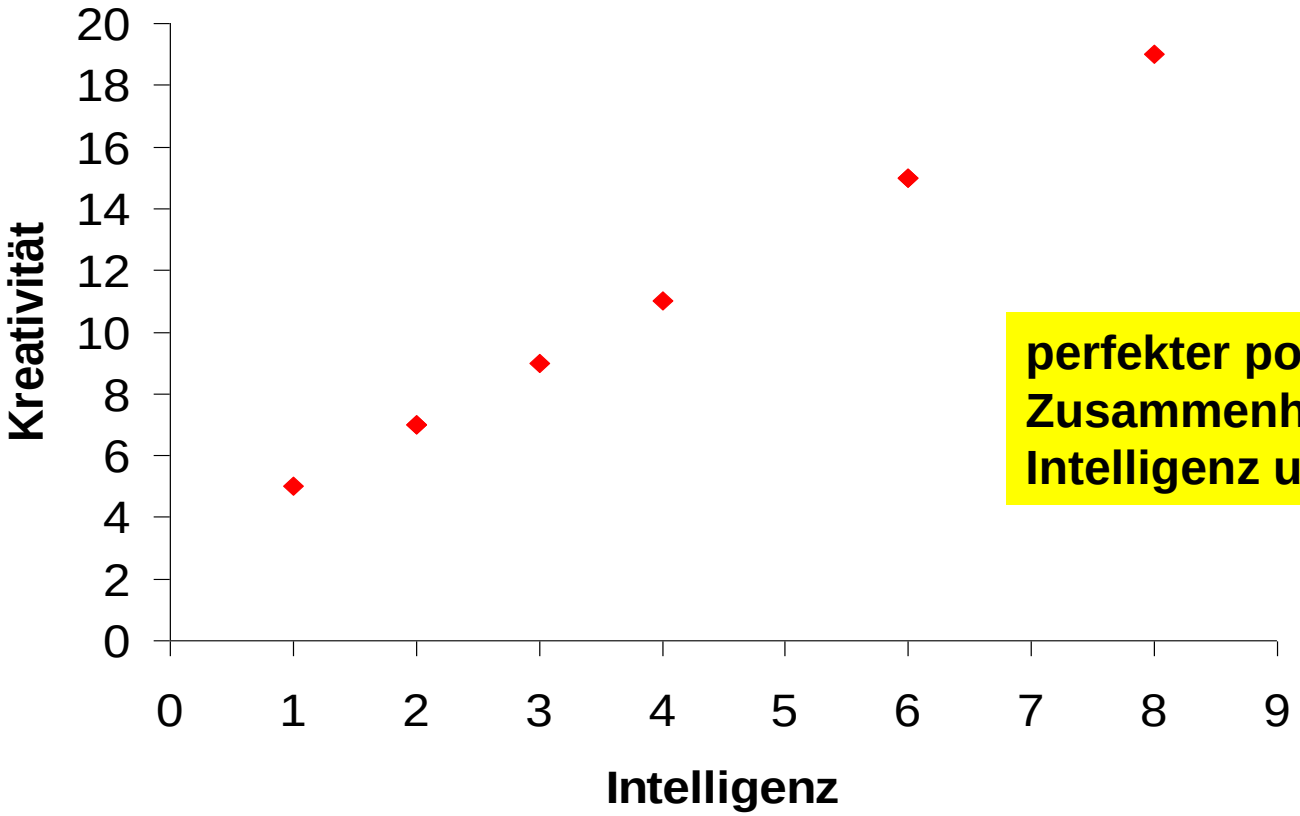
Zusammenhangshypothesen

- Beispiele:
- Je intelligenter eine Person ist, desto kreativer ist sie auch.
- Je mehr Alkohol man trinkt, desto schlechter fährt man Auto.
- Je mehr Vertrauen man in andere Menschen hat, desto zufriedener lebt man.
- …

Grafische Veranschaulichung bivariater Zusammenhänge

Scatterplots

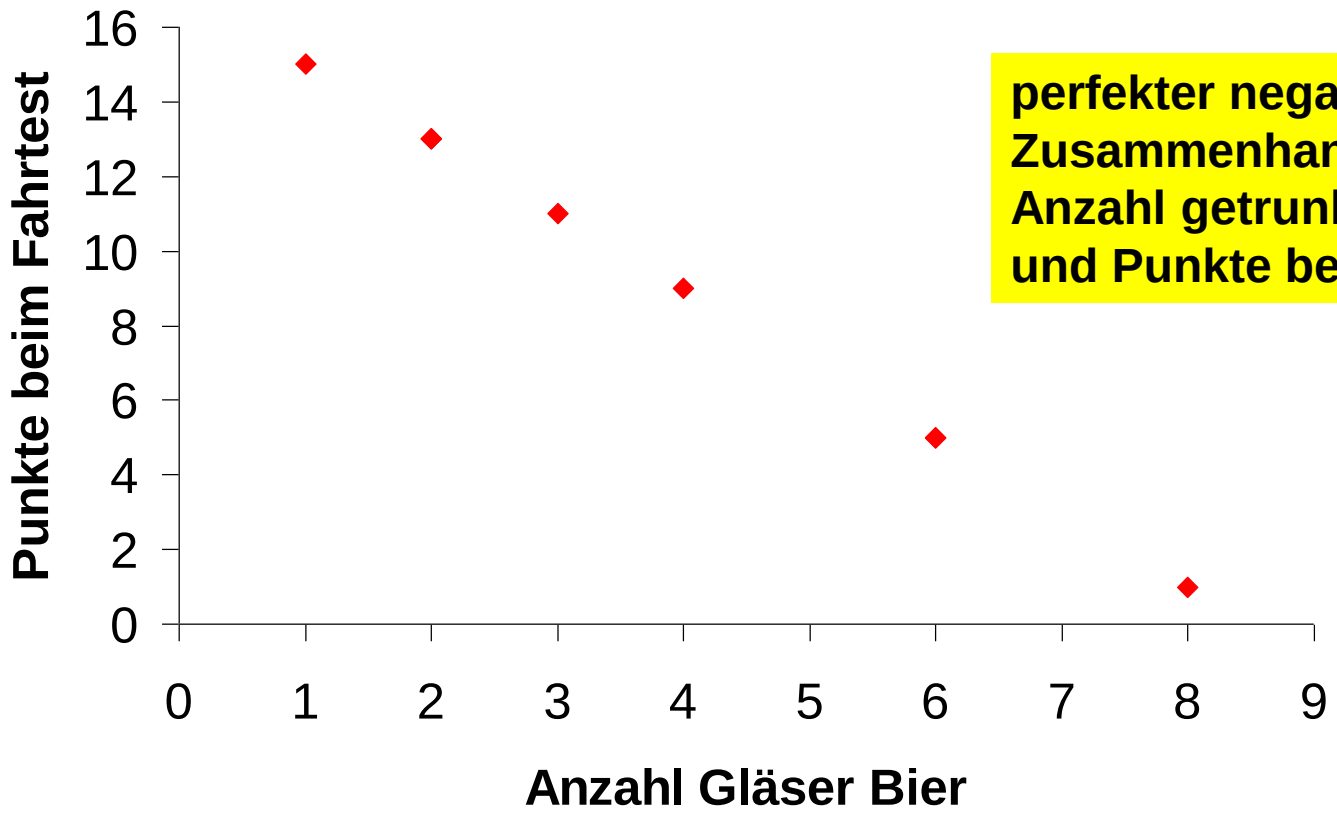
Hypothese: „Je intelligenter eine Person ist, desto kreativer ist sie auch."



perfekter positiver Zusammenhang zwischen Intelligenz und Kreativität

Scatterplots

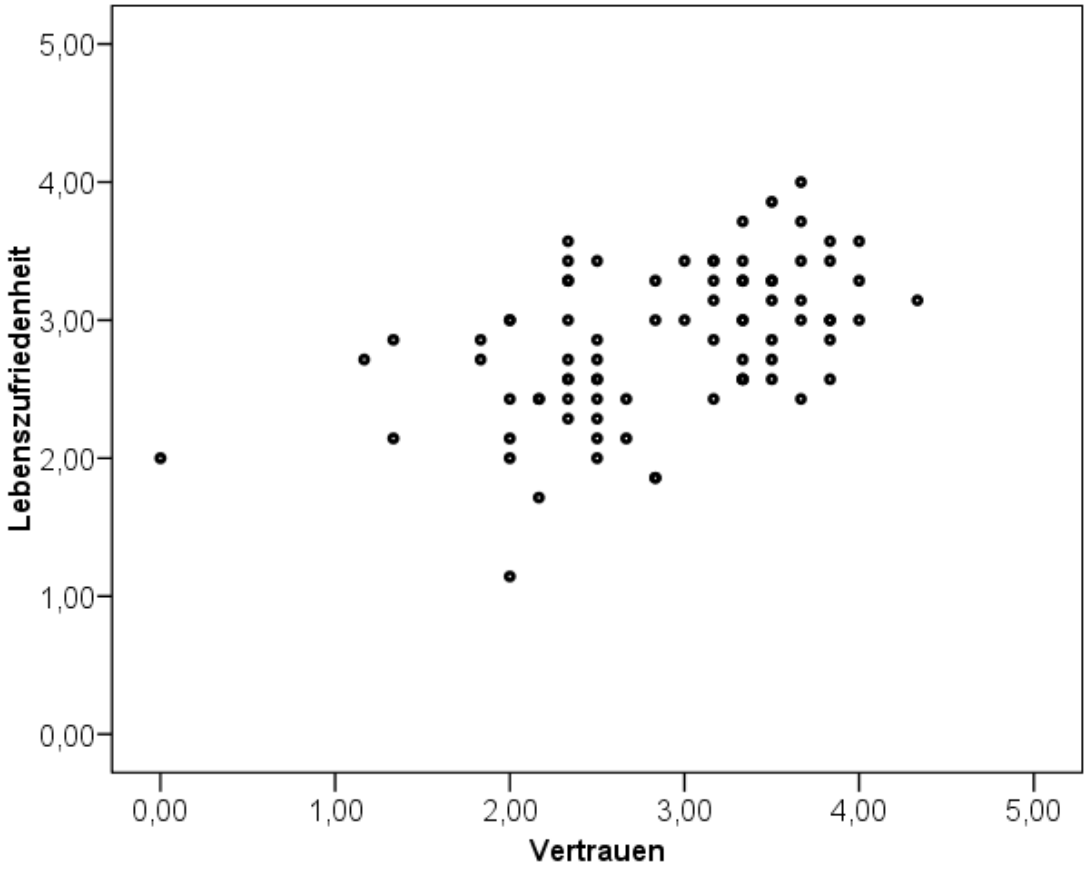
Hypothese: „Je mehr Alkohol man trinkt, desto schlechter fährt man Auto.“



perfekter negativer Zusammenhang zwischen Anzahl getrunkener Gläser Bier und Punkte beim Fahrtest

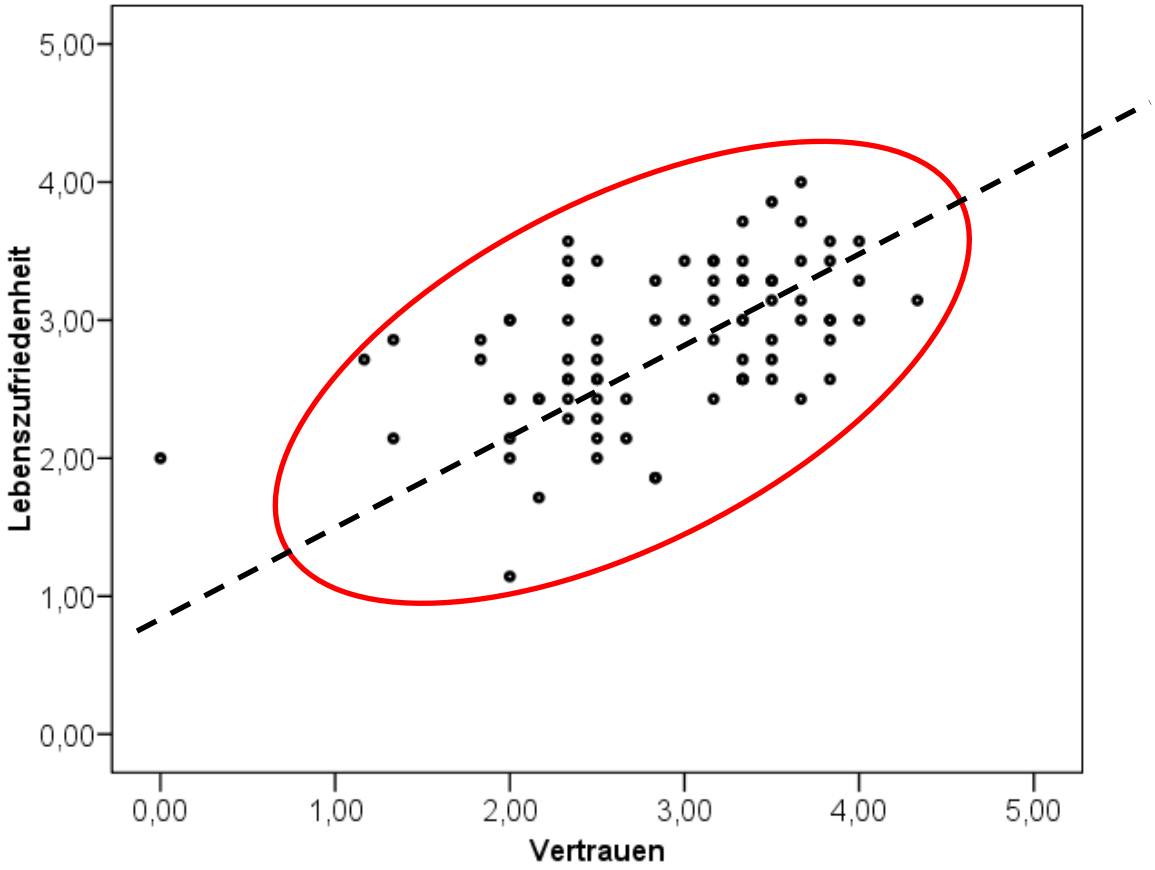
Scatterplots

Hypothese: „Wer Vertrauen in andere Menschen hat, lebt glücklicher und zufriedener.“



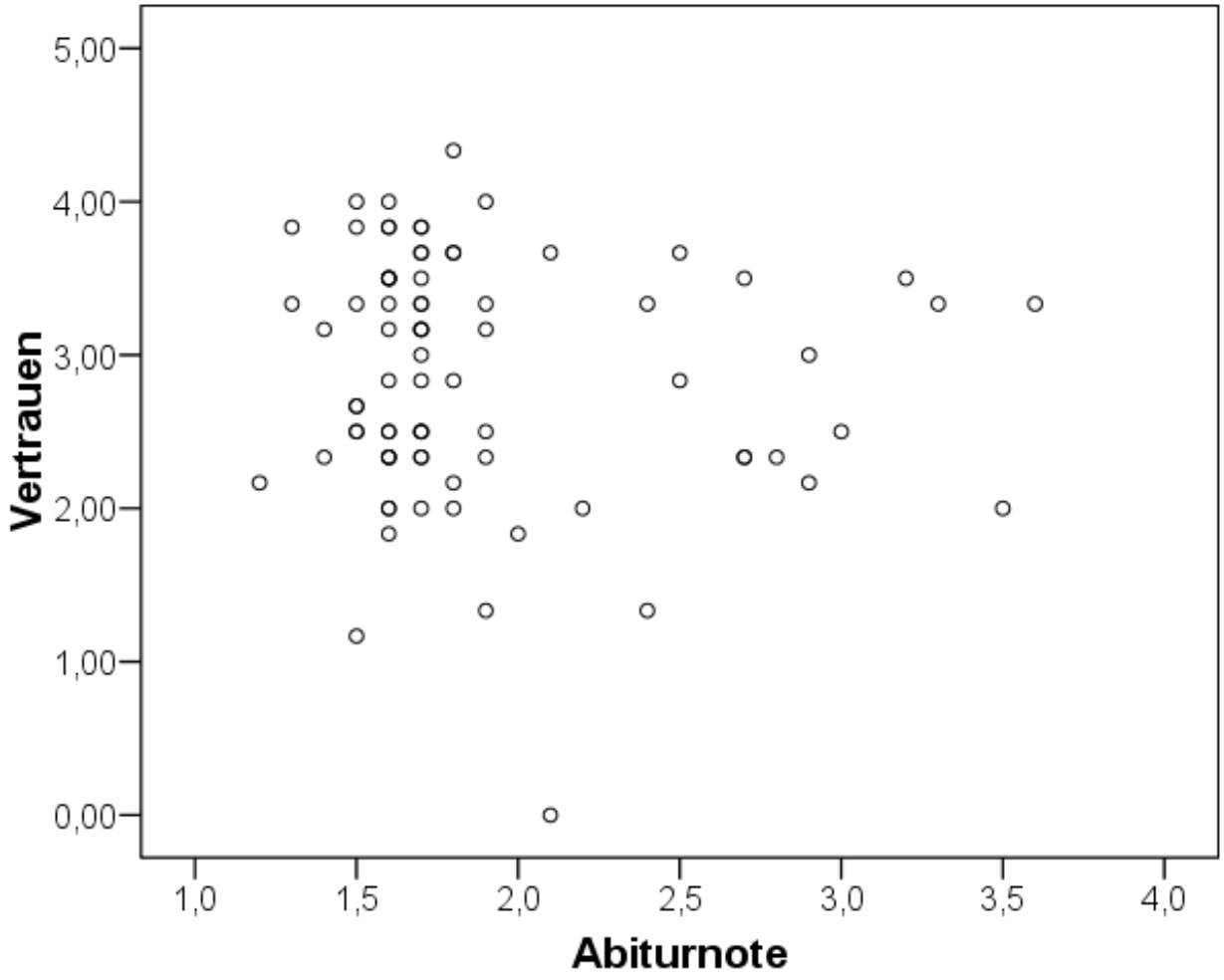
Scatterplots

Hypothese: „Wer Vertrauen in andere Menschen hat, lebt glücklicher und zufriedener."

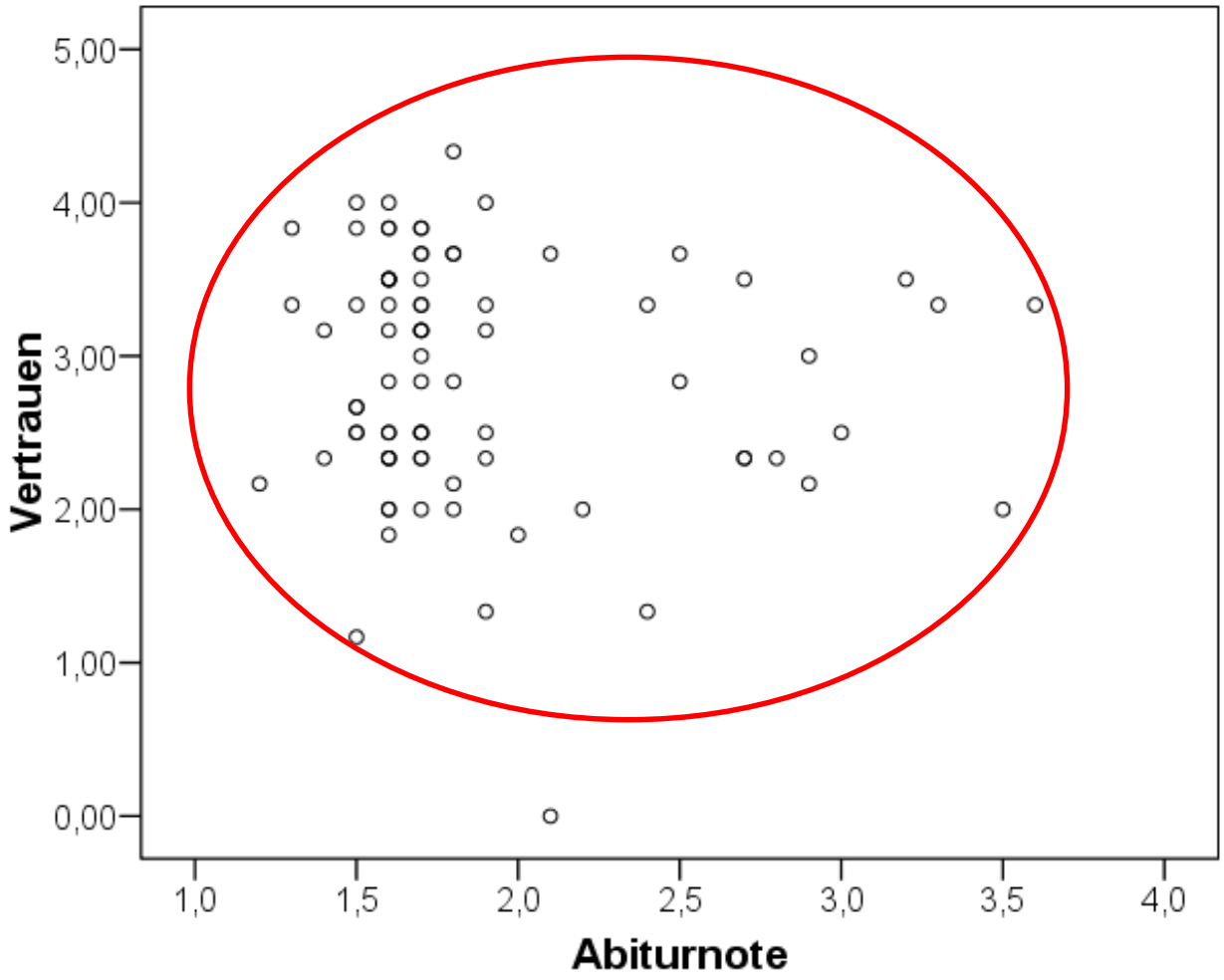


nicht so ganz perfekter, aber positiver Zusammenhang zwischen Vertrauen und Lebenszufriedenheit

Scatterplots



Scatterplots



Zusammenhangsmaße für metrische Variablen

Zusammenhangsmaße

- Für eine quantitative Analyse von Merkmalszusammenhängen sind folgende Fragen von Bedeutung:

  - Wie lässt sich die Form des Zusammenhangs zwischen *X* und *Y* beschreiben?
  - Welche Richtung hat der Zusammenhang zwischen *X* und *Y*, d.h. ist er negativ oder positiv?
  - Wie stark ist der Zusammenhang zwischen *X* und *Y*?
  - *Lässt sich der in der Stichprobe ermittelte Zusammenhang auf die Population übertragen? (Inferenzstatistik)*

Zusammenhangsmaße

- Quantifizierung durch Zusammenhangsmaße → Korrelationskoeffizienten
- Positive Korrelation:
  - Hohe Werte in der einen Variablen gehen mit hohen Werten in der anderen Variablen einher
  - Niedrige Werte in der einen Variable gehen mit niedrigen Werten in der anderen Variablen einher
- Negative Korrelation:
  - Hohe (niedrige) Werte in der einen Variablen gehen mit niedrigen (hohen) Werten in der anderen Variablen einher
- Ggfs. auch kein Zusammenhang

Zusammenhangsmaße

- Wann ist ein Messwert „hoch“? Wann ist ein Messwert „niedrig“?

- Vergleich anhand des Mittelwertes der jeweiligen Variablen
  - Hohe Messwerte entsprechen Werten über dem Durchschnitt
  - Niedrige Messwerte entsprechen Werten unter dem Durchschnitt

- Stärke des Zusammenhangs zwischen zwei metrischen Variablen ergibt sich durch die Abweichung der Messwerte vom jeweiligen Mittelwert

Zusammenhangsmaße

- Vorgehensweise:
  - Bestimme für jedes Messwertepaar die Abweichung vom Mittelwert
  - Berechne die gemeinsame Abweichung beider Messwerte von Ihren Mittelwerten durch Multiplikation
  - Berechne die Summe der Abweichungsprodukte
  - Berechne das durchschnittliche Abweichungsprodukt mittels Division durch die Anzahl der Fälle (n)

Zusammenhangsmaße

- Vorgehensweise:

– Bestimme für jedes Messwertepaar die Abweichung vom Mittelwert

$x_i - \bar{x}$ $y_i - \bar{y}$

– Berechne die gemeinsame Abweichung beider Messwerte von Ihren Mittelwerten durch Multiplikation

$(x_i - \bar{x})\ (y_i - \bar{y})$

– Berechne die Summe der Abweichungsprodukte (SAP)

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

Zusammenhangsmaße

- Vorgehensweise:
  - Berechne das durchschnittliche Abweichungsprodukt - die Kovarianz - mittels Division durch die Anzahl der Fälle (n):

$$\text{cov (x,y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{n}$$

  - Kovarianz beschreibt die gemeinsame Streuung zweier Merkmale

  - Zur Erinnerung: Varianz

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Zusammenhangsmaße

- Die Kovarianz ist definiert als:

$$\text{cov}\,(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\;(y_i-\bar{y})}{n}$$

- Die Kovarianz ist dann hoch positiv, wenn hohe positive Abweichungen auch mit hohen positiven Abweichungen einhergehen und hohe negative Abweichungen auch mit hohen negativen Abweichungen einhergehen.

- Die Kovarianz ist dann hoch negativ, wenn hohe positive Abweichungen mit hohen negativen Abweichungen einhergehen und umgekehrt.

- Die Kovarianz ist dann Null, wenn die Richtung der Abweichungen vom Mittelwert in X nicht systematisch mit einer bestimmten Richtung der Abweichungen vom Mittelwert in Y einhergeht.

Zusammenhangsmaße

- Die Kovarianz ist definiert als:

$$\text{cov}\,(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\ (y_i-\bar{y})}{n}$$

- Die Kovarianz ist ein unstandardisiertes Maß
- Dies bedeutet, ihre Größe ist abhängig von den gewählten Maßeinheiten
- Unschön, erschwert den Vergleich zwischen unterschiedlichen Kovarianzen
- Lösung: Standardisierung anhand der Division durch das Produkt der Standardabweichungen beider Merkmale → Pearson's *r*

Pearson‘s *r*

- Formal:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson‘s r entspricht der anhand des Produkts der Standardabweichungen standardisierten Kovarianz

- Vereinfacht:

$$r = \frac{Cov_{xy}}{S_x S_y}$$

Pearson‘s *r*

- Pearson‘s *r*
- Wertebereich von -1 bis +1
  - Negatives Vorzeichen = negativer Zusammenhang
  - Positives Vorzeichen = positiver Zusammenhang

- Vorzeichen indiziert die Richtung, Betrag die Stärke des Zusammenhanges

- Für Pearson‘s *r* hat Cohen (1988) folgenden Taxonomievorschlag gemacht:

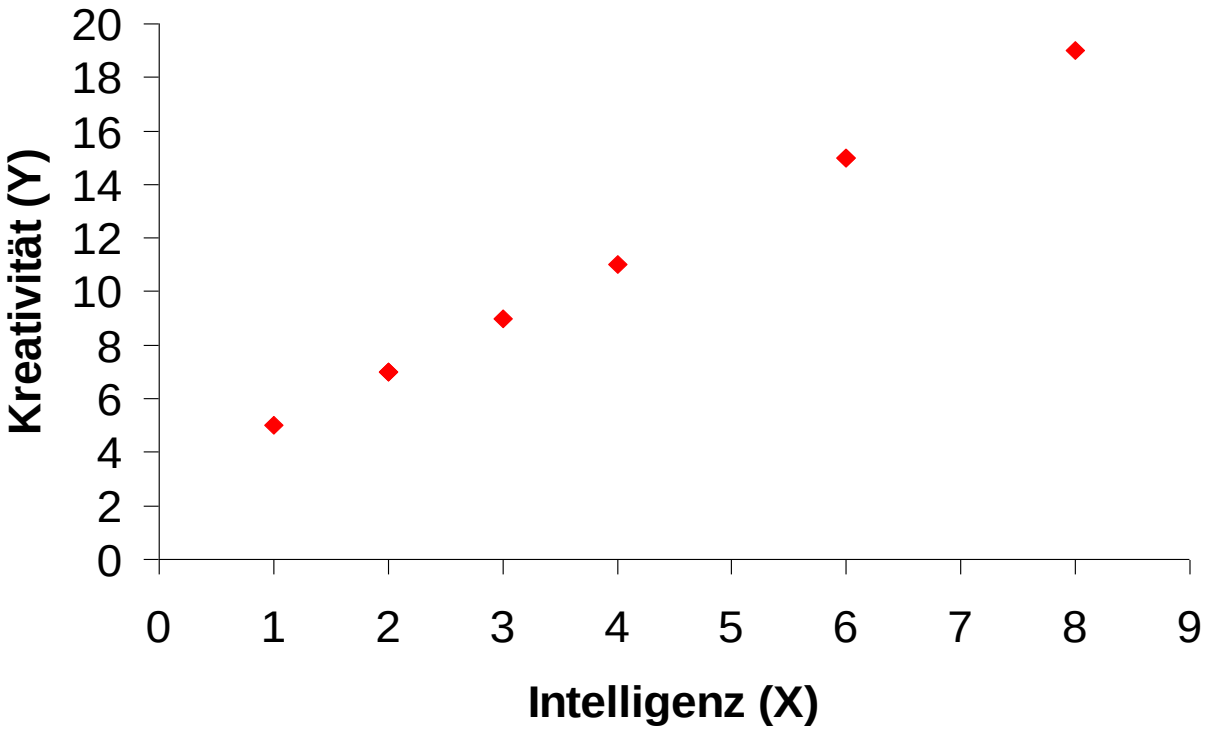
  $|r| \approx 0{,}10$ ⇨ „schwacher“ Zusammenhang

  $|r| \approx 0{,}30$ ⇨ „mittlerer“ Zusammenhang

  $|r| \approx 0{,}50$ ⇨ „starker“ Zusammenhang

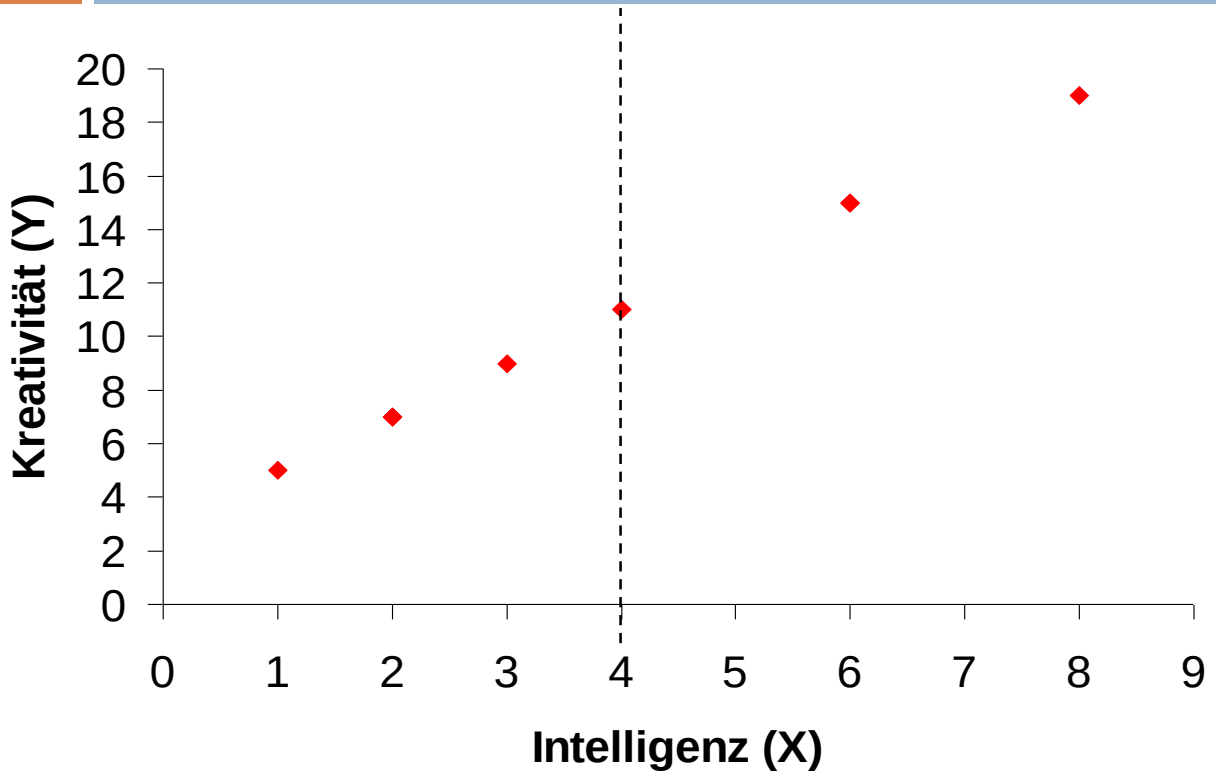
- Aber: Die Beurteilung der Höhe einer Korrelation hängt immer von der zugrunde liegenden Fragestellung ab!

Pearson‘s *r*



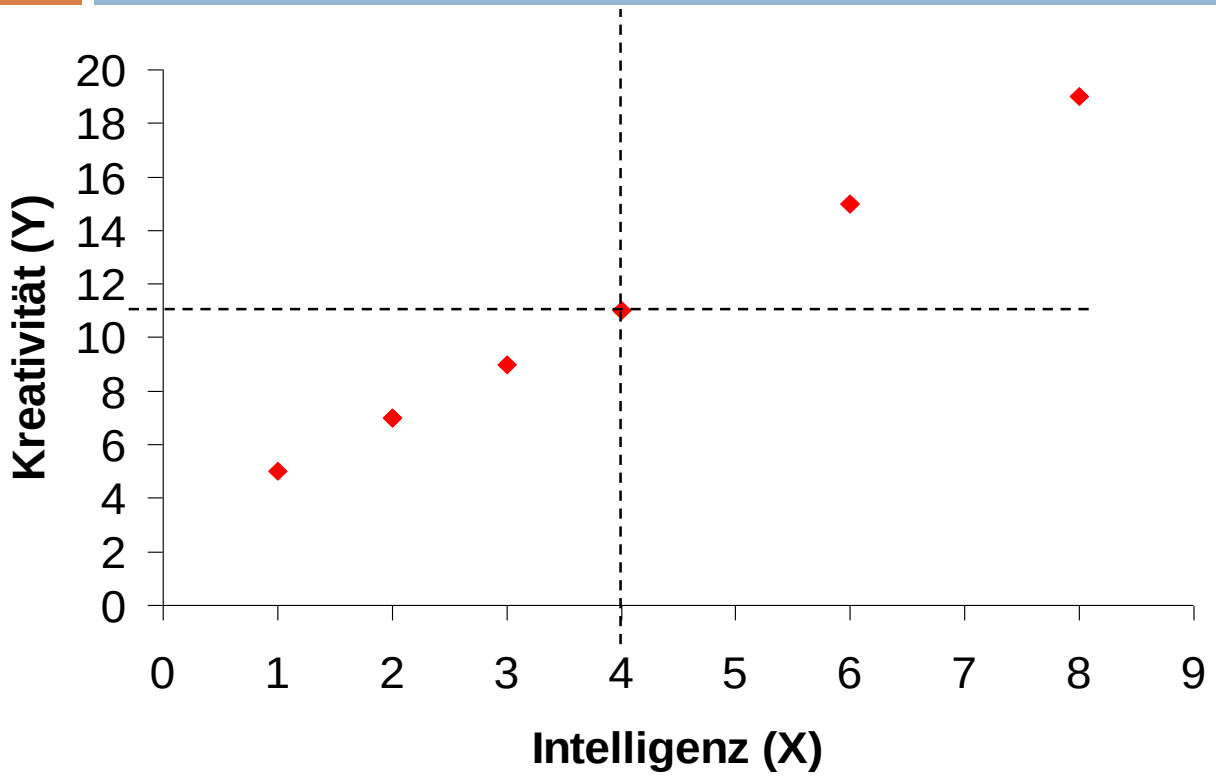
	X	*Y*
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

Pearson‘s *r*



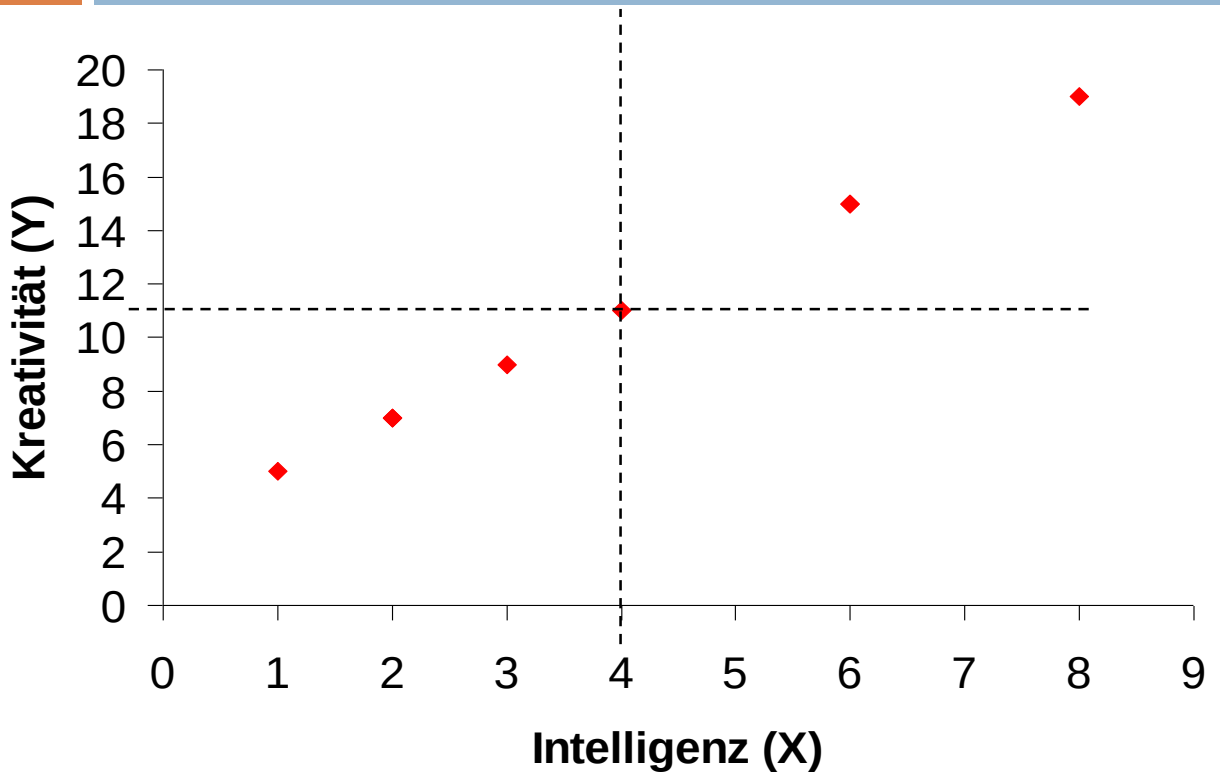
	X	*Y*
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

Pearson‘s *r*



	X	*Y*
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

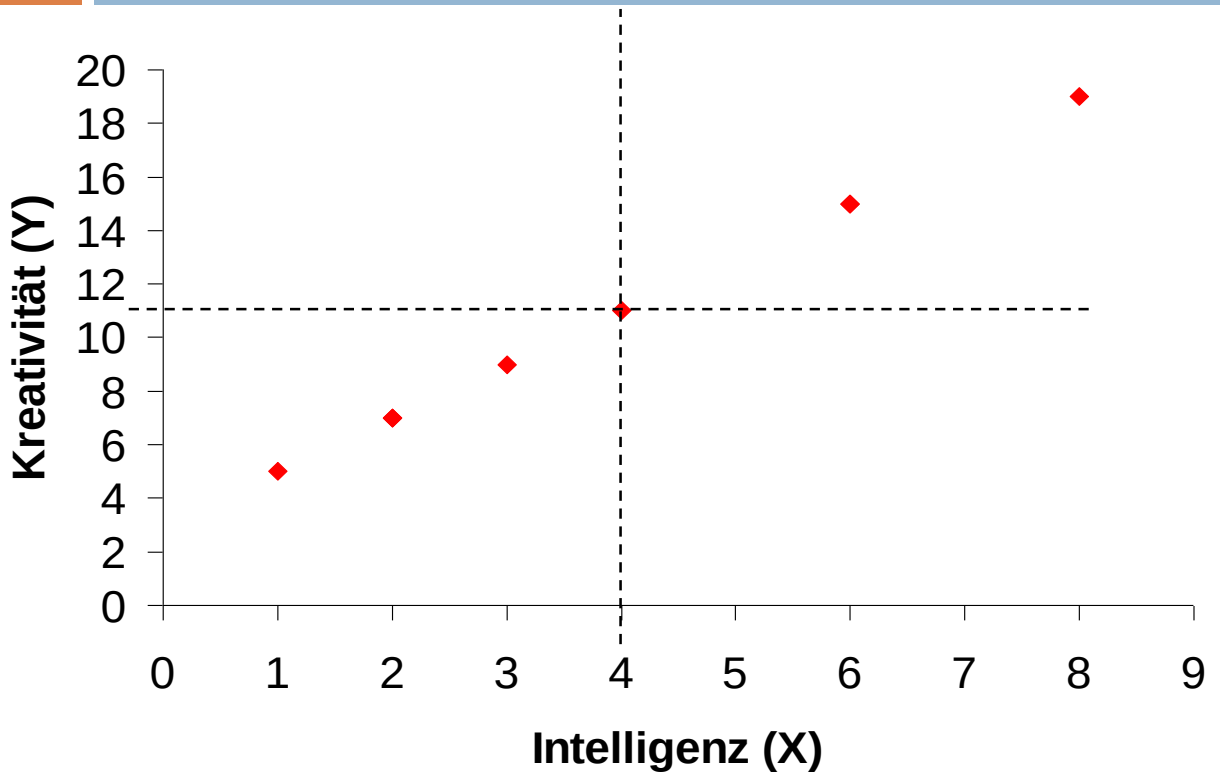
Pearson‘s *r*



	X	*Y*
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

Wann ist der Zusammenhang zweier Variablen *X* und *Y* positiv?

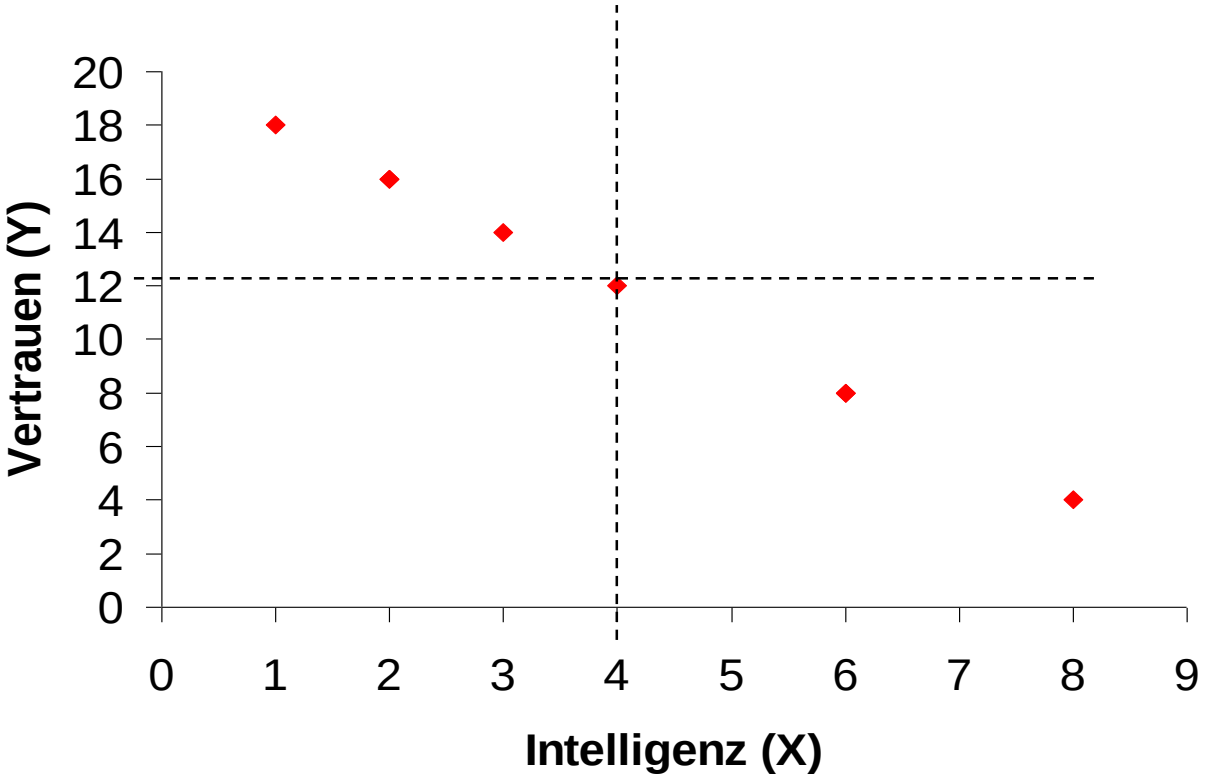
Pearson‘s *r*



	X	*Y*
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

Der Zusammenhang zweier Variablen *X* und *Y* ist dann positiv, wenn *x*-Werte, die oberhalb ihres Mittelwerts $\bar{x}$ liegen, mit *y*-Werten einhergehen, die ebenfalls oberhalb ihres Mittelwerts $\bar{y}$ liegen (und umgekehrt).

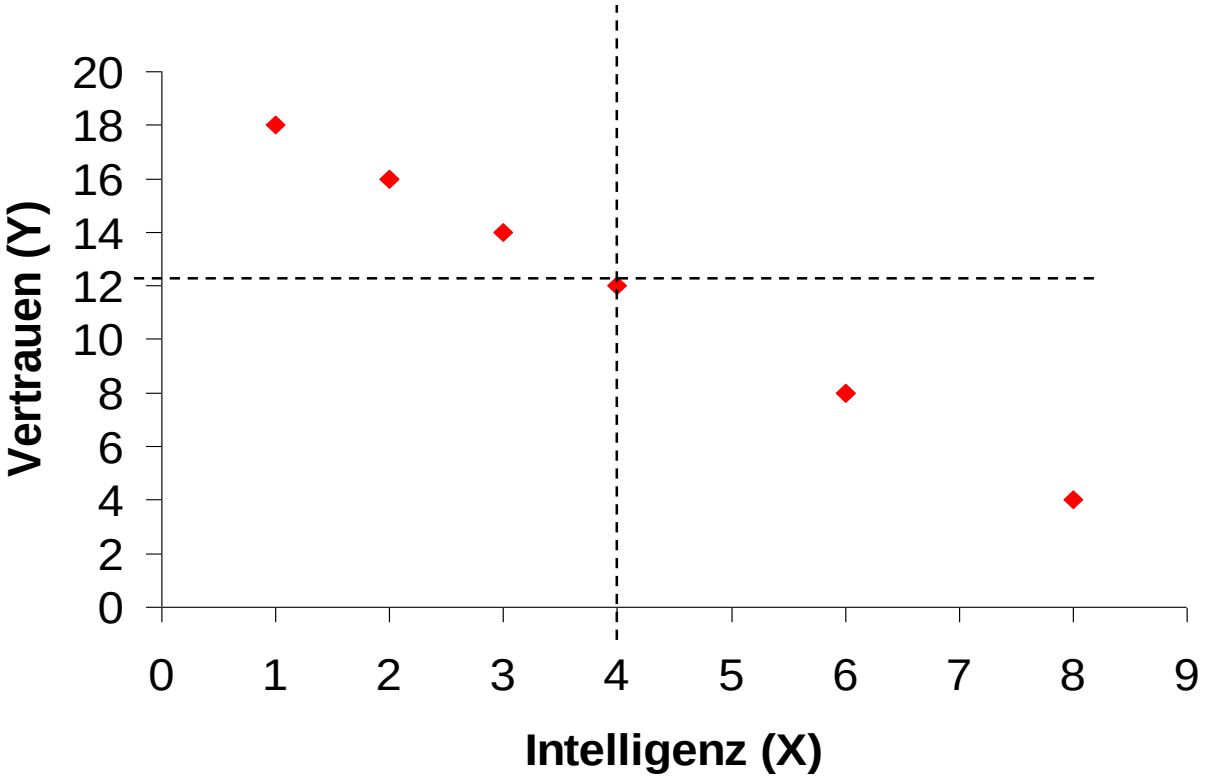
Pearson‘s *r*



	X	*Y*
	1	18
	2	16
	2	16
	3	14
	4	12
	6	8
	6	8
	8	4
M =	4	12
s^2 =	5,25	21

Wann ist der Zusammenhang zweier Variablen *X* und *Y* negativ?

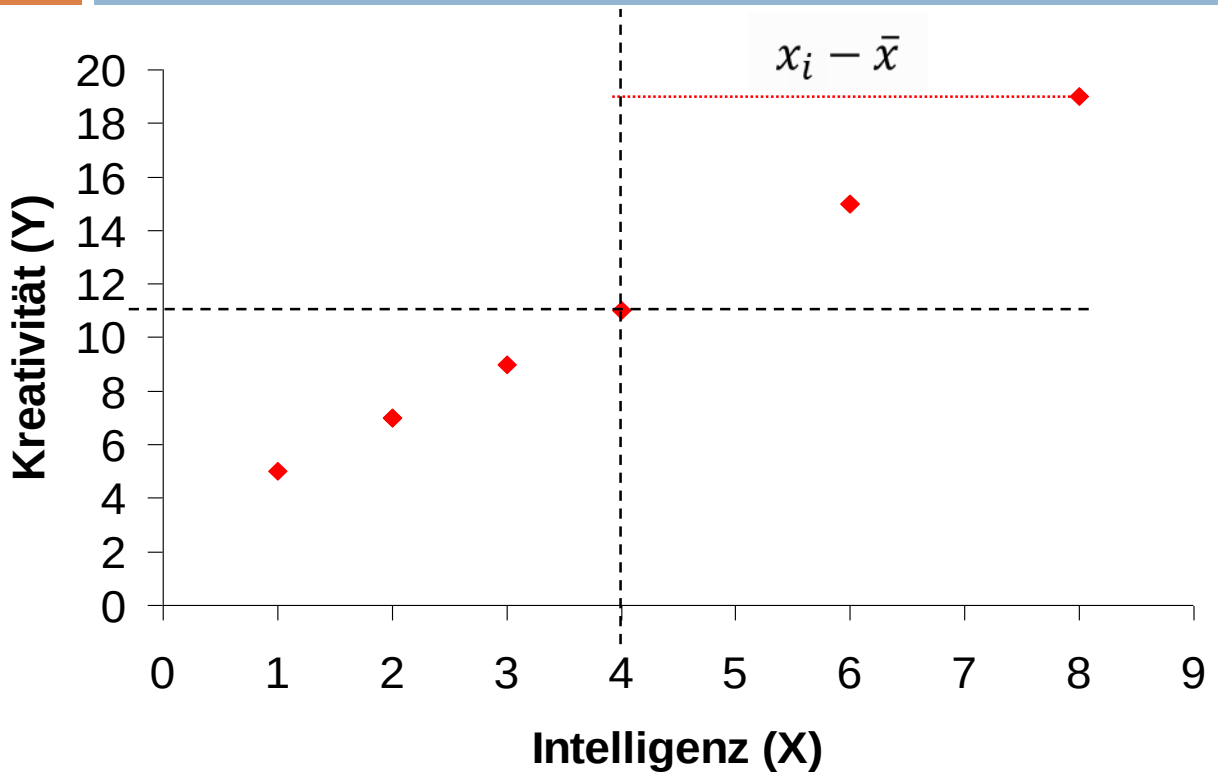
Pearson‘s *r*



	X	*Y*
	1	18
	2	16
	2	16
	3	14
	4	12
	6	8
	6	8
	8	4
M =	4	12
s^2 =	5,25	21

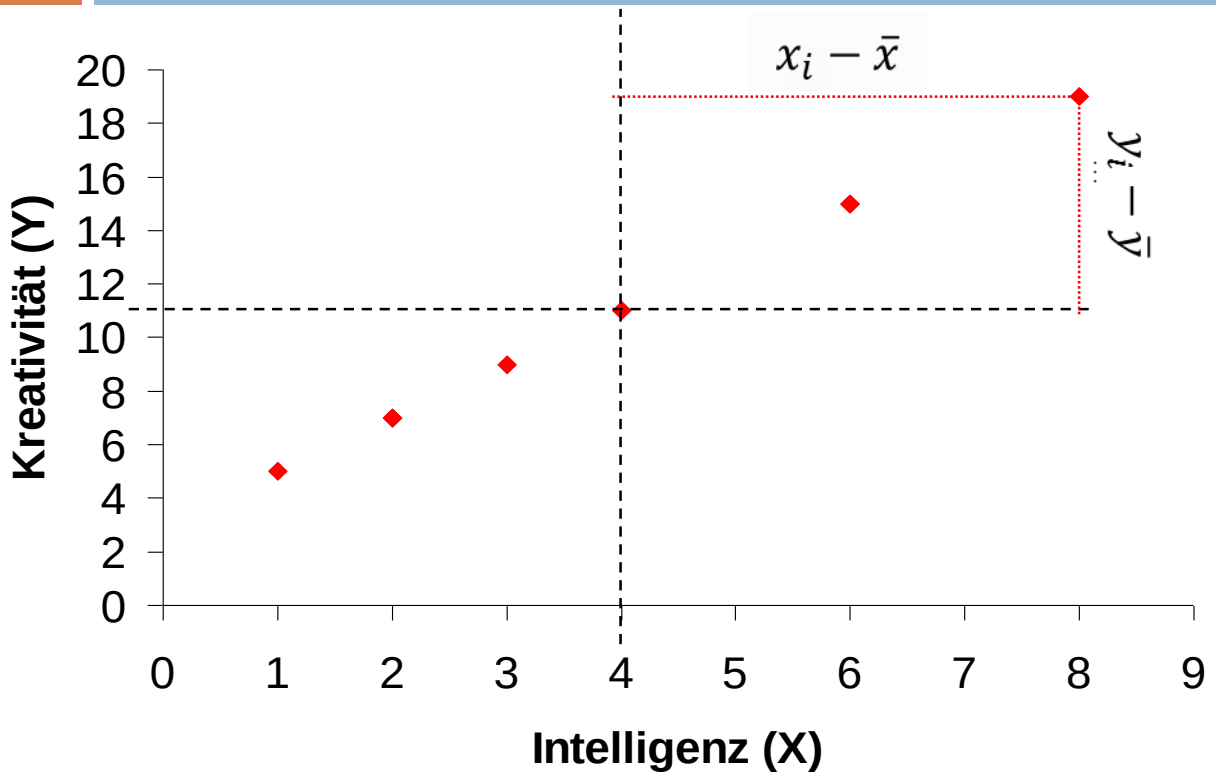
Der Zusammenhang zweier Variablen *X* und *Y* ist dann negativ, wenn *x*-Werte, die oberhalb ihres Mittelwerts $\overline{x}$ liegen, mit *y*-Werten einhergehen, die unterhalb ihres Mittelwerts $\overline{y}$ liegen (und umgekehrt).

Pearson‘s *r*



	X	Y
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

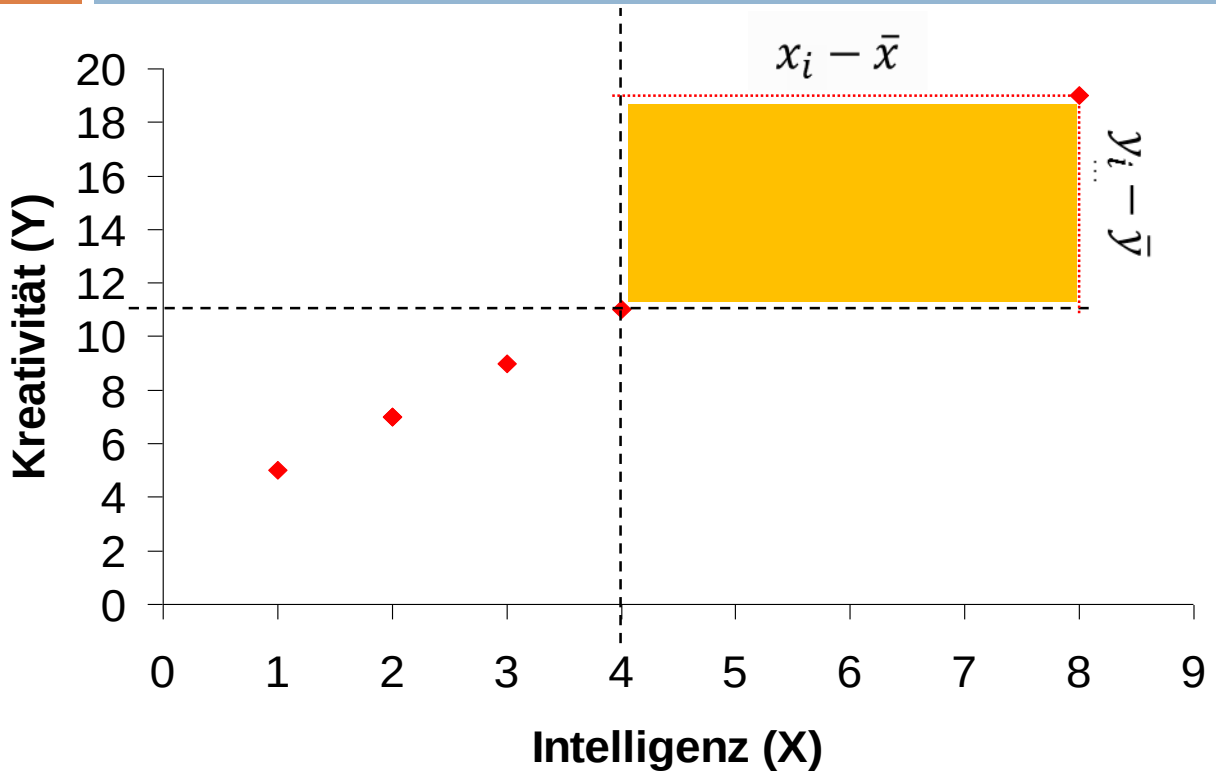
Pearson‘s *r*



	X	Y
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

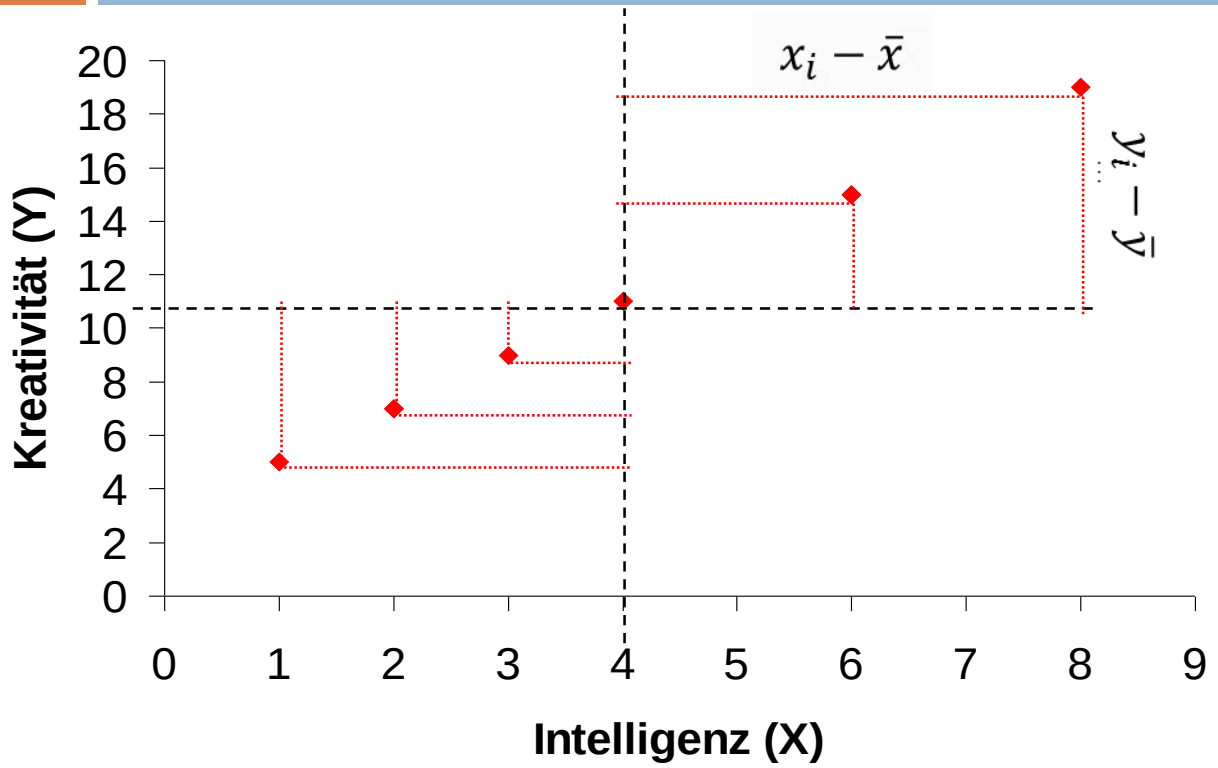
Pearson‘s *r*

Flächeninhalt des Rechtecks entspricht dem Produkt der Abweichungen des Messwertepaares



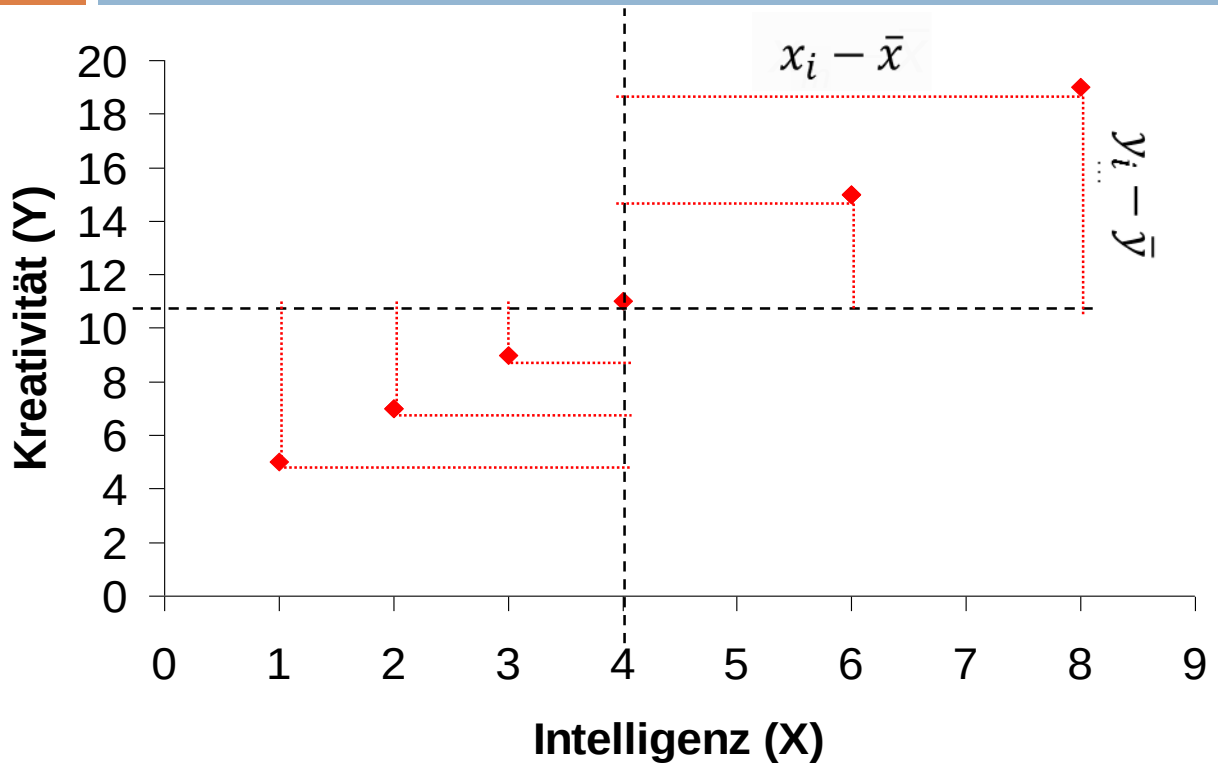
	X	Y
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

Pearson‘s *r*



$x_i - \bar{x}$	$y_i - \bar{y}$
1 – 4 = -3	5 – 11 = -6
2 – 4 = -2	7 – 11 = -4
2 – 4 = -2	7 – 11 = -4
3 – 4 = -1	9 – 11 = -2
4 – 4 = 0	11 – 11 = 0
6 – 4 = 2	15 – 11 = 4
6 – 4 = 2	15 – 11 = 4
8 – 4 = 4	19 – 11 = 8

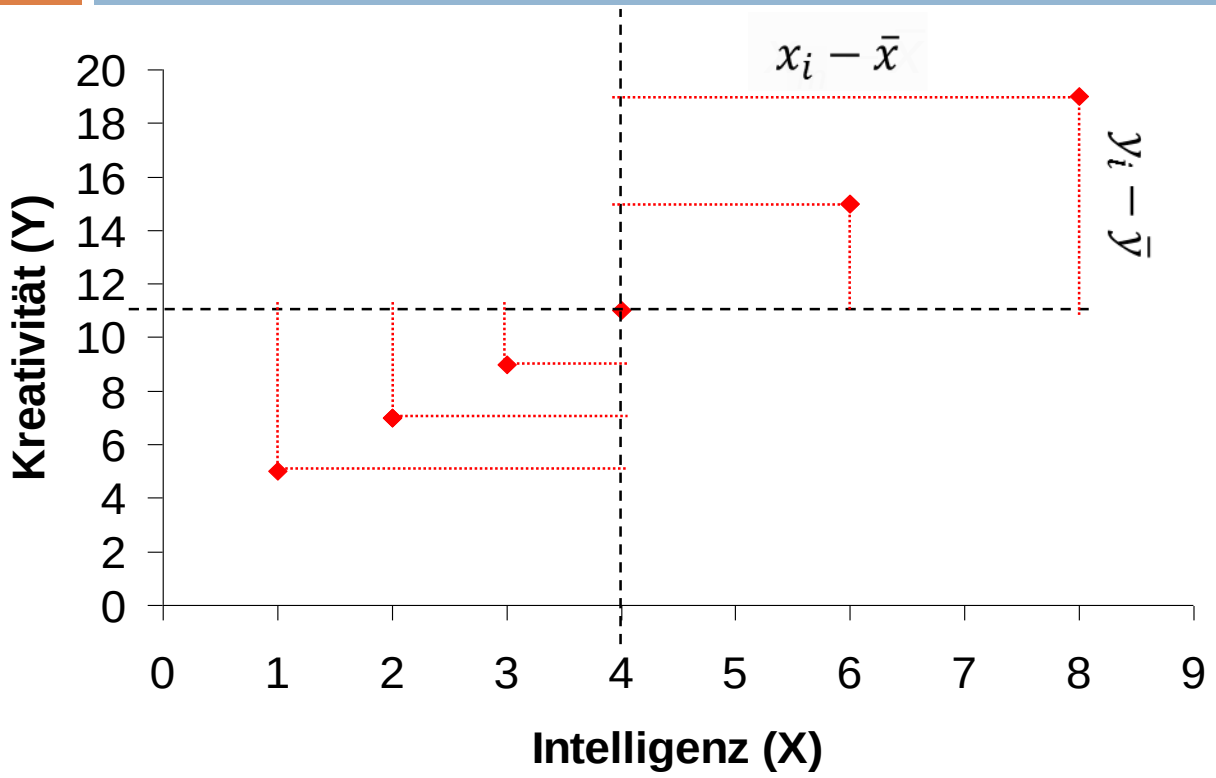
Pearson‘s *r*



$x_i - \bar{x}$	$y_i - \bar{y}$
1 – 4 = -3	5 – 11 = -6
2 – 4 = -2	7 – 11 = -4
2 – 4 = -2	7 – 11 = -4
3 – 4 = -1	9 – 11 = -2
4 – 4 = 0	11 – 11 = 0
6 – 4 = 2	15 – 11 = 4
6 – 4 = 2	15 – 11 = 4
8 – 4 = 4	19 – 11 = 8

Schritt 1: Wir berechnen für jeden Wert x_i sowie für jeden Wert y_i die Differenz vom jeweiligen Mittelwert.

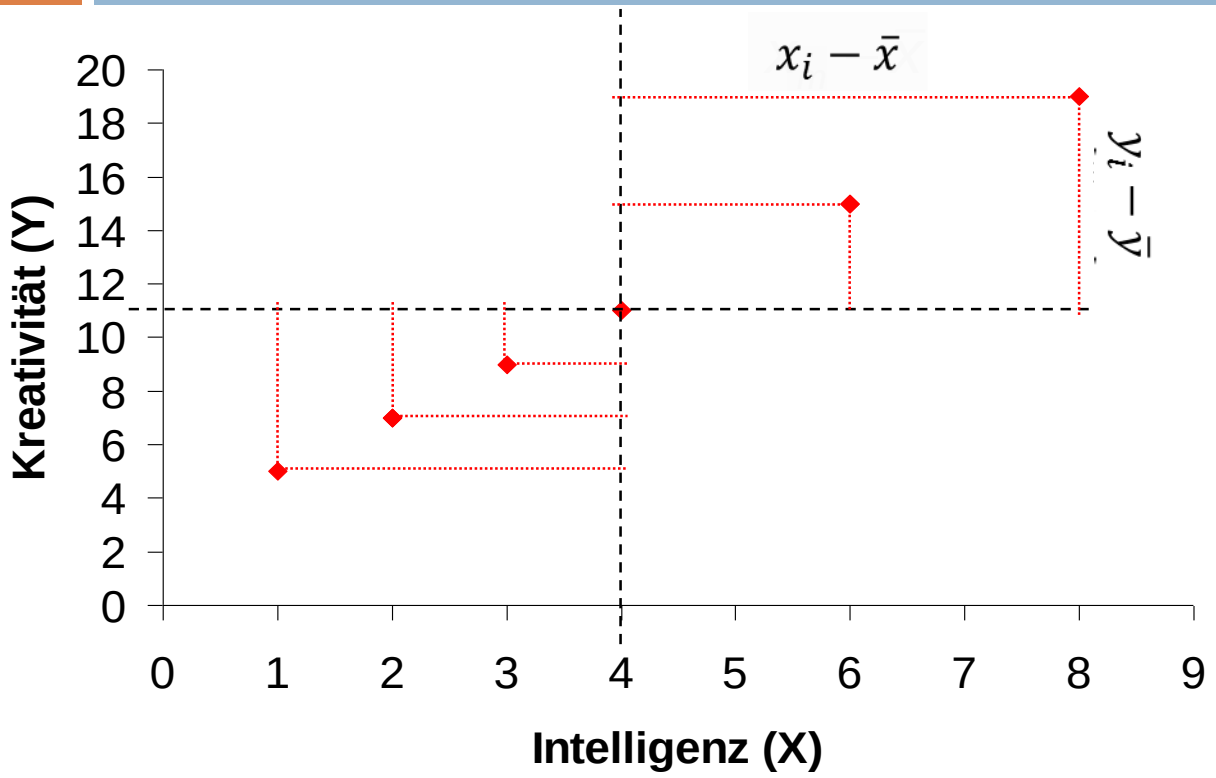
Pearson‘s *r*



$(x_i - \bar{x})(y_i - \bar{y})$ ↓		
-3	-6	18
-2	-4	8
-2	-4	8
-1	-2	2
0	0	0
2	4	8
2	4	8
4	8	32

Schritt 2: Wir berechnen für jedes Wertepaar xy_i das ***Kreuzprodukt***, d.h. das Produkt der Mittelwertsabweichung.

Pearson‘s *r*

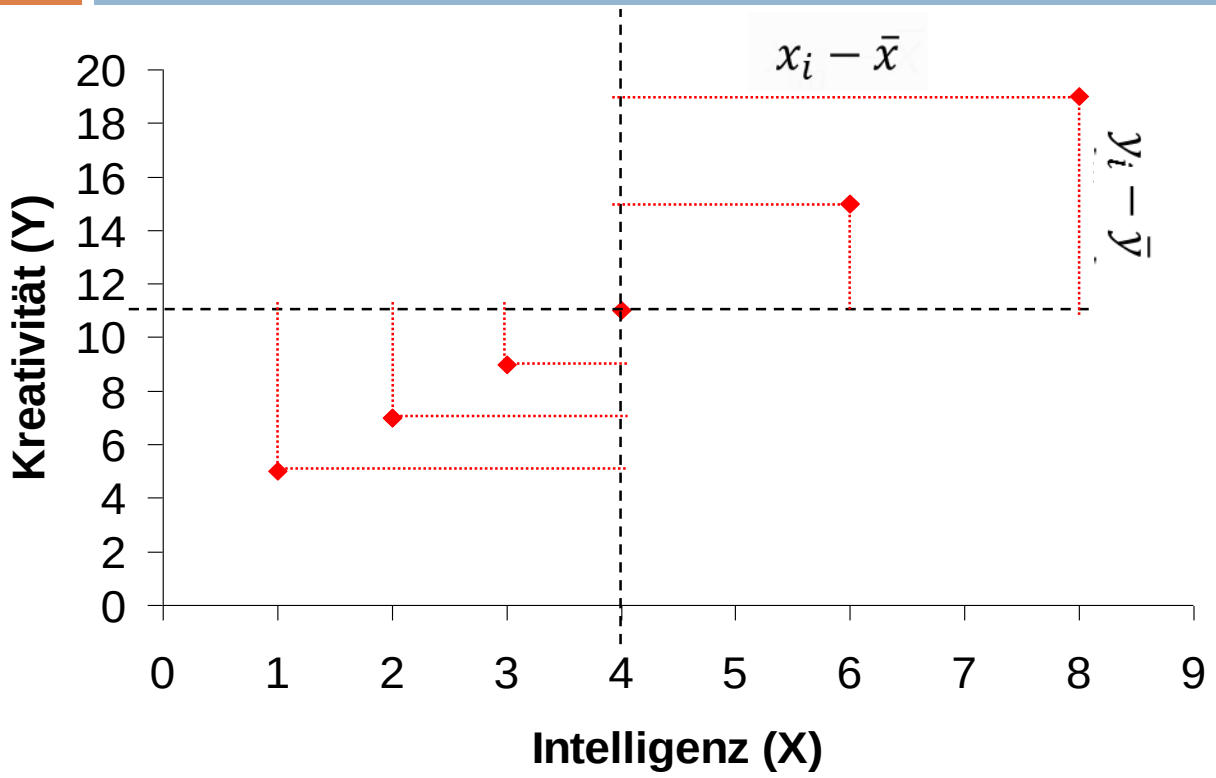


$(x_i - \bar{x})(y_i - \bar{y})$		
-3	-6	18
-2	-4	8
-2	-4	8
-1	-2	2
0	0	0
2	4	8
2	4	8
4	8	32
Summe:		84

Schritt 3: Wir berechnen die ***Kreuzproduktsumme***, d.h. die Summe aller Kreuzprodukte von *i* = 1 bis *n*.

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

Pearson‘s *r*

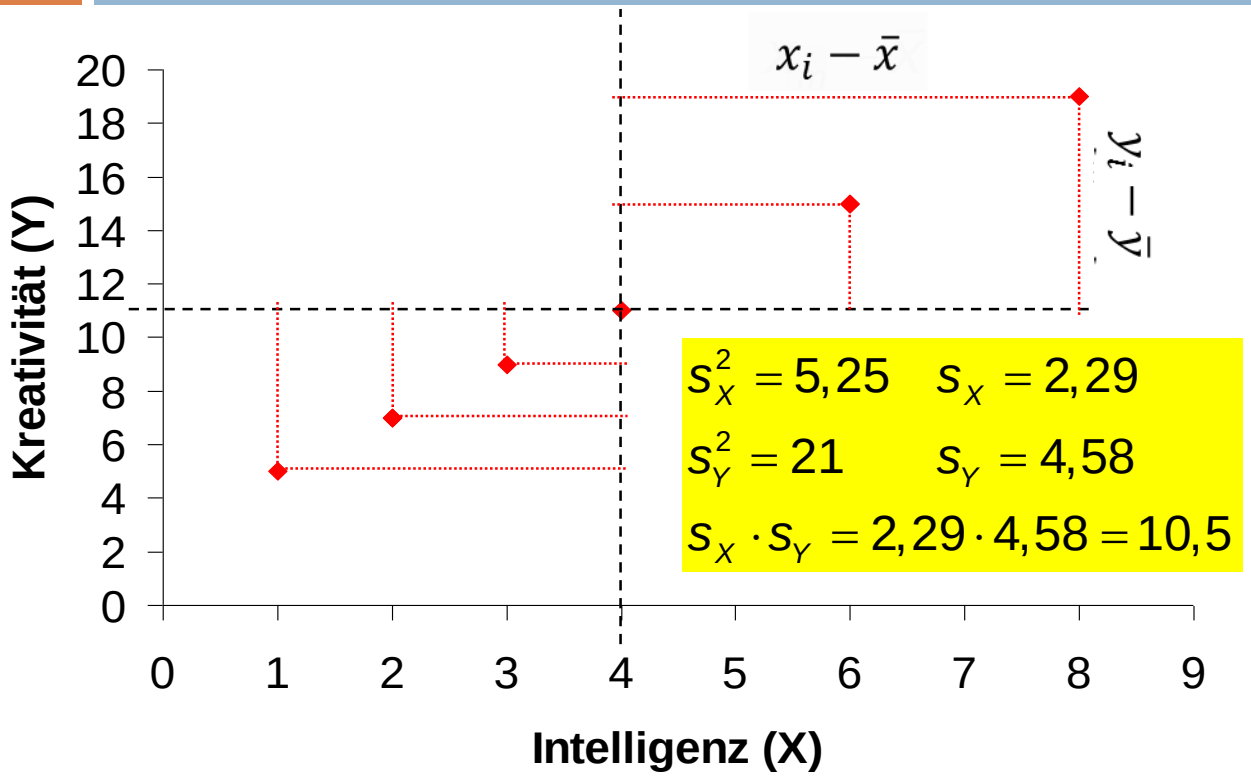


$(x_i - \bar{x})(y_i - \bar{y})$		
-3	-6	18
-2	-4	8
-2	-4	8
-1	-2	2
0	0	0
2	4	8
2	4	8
4	8	32
Summe:		84
Kovarianz:		10,5

Schritt 4: Wir berechnen das mittlere Kreuzprodukt oder die ***Kovarianz*** (Cov), indem wir durch *n* teilen.

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

Pearson‘s *r*

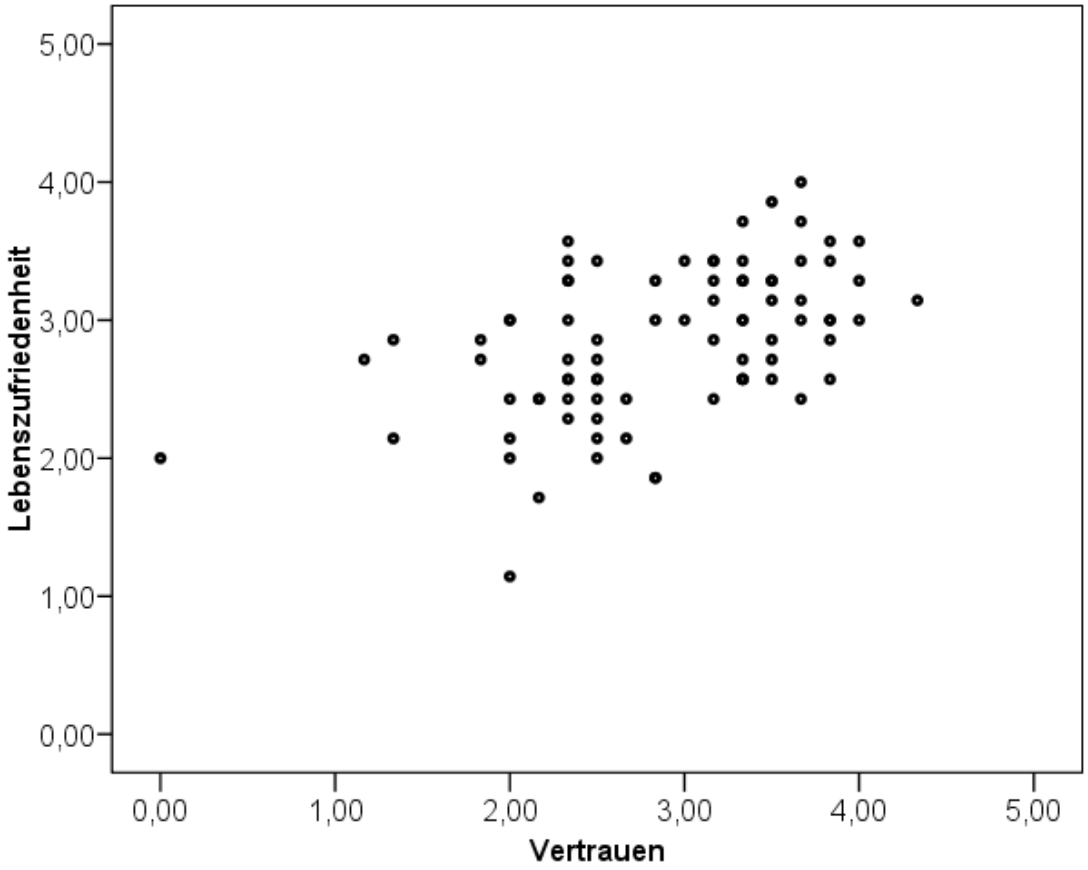


$(x_i - \bar{x})(y_i - \bar{y})$		
-3	-6	18
-2	-4	8
-2	-4	8
-1	-2	2
0	0	0
2	4	8
2	4	8
4	8	32
Summe:		84
Kovarianz:		10,5
Korrelation:		1

Schritt 5: Wir relativieren die empirische Kovarianz an der maximalen Kovarianz und erhalten Pearson‘s *r*.

$$r = \frac{Cov_{xy}}{s_x s_y}$$

Pearson‘s *r*



$$r = \frac{Cov_{xy}}{s_x s_y}$$

Mini-Übung Pearson‘s *r*

1. Zeichnen Sie ein Punktediagramm für die nachfolgend dargestellten Werte

2. Berechnen Sie Pearson‘s r.

3. Interpretieren Sie Ihr Ergebnis.

Mini-Übung Pearson's *r*

	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
A	0	2					
B	10	6					
C	4	2					
D	8	4					
E	8	6					

$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Mini-Übung Pearson's *r*

	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
A	0	2	-6	-2			
B	10	6	4	2			
C	4	2	-2	-2			
D	8	4	2	0			
E	8	6	2	2			
	6	4					

$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Mini-Übung Pearson's *r*

	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
A	0	2	-6	-2	12	36	4
B	10	6	4	2	8	16	4
C	4	2	-2	-2	4	4	4
D	8	4	2	0	0	4	0
E	8	6	2	2	4	4	4
	6	4			28	64	16

$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Mini-Übung Pearson's *r*

	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
A	0	2	-6	-2	12	36	4
B	10	6	4	2	8	16	4
C	4	2	-2	-2	4	4	4
D	8	4	2	0	0	4	0
E	8	6	2	2	4	4	4
	6	4			28	64	16

$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$r = \frac{28}{\sqrt{64(16)}}$$

Pearson's *r*

	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
A	0	2	-6	-2	12	36	4
B	10	6	4	2	8	16	4
C	4	2	-2	-2	4	4	4
D	8	4	2	0	0	4	0
E	8	6	2	2	4	4	4
	6	4			28	64	16

$$r = \frac{Cov_{xy}}{s_x s_y}$$

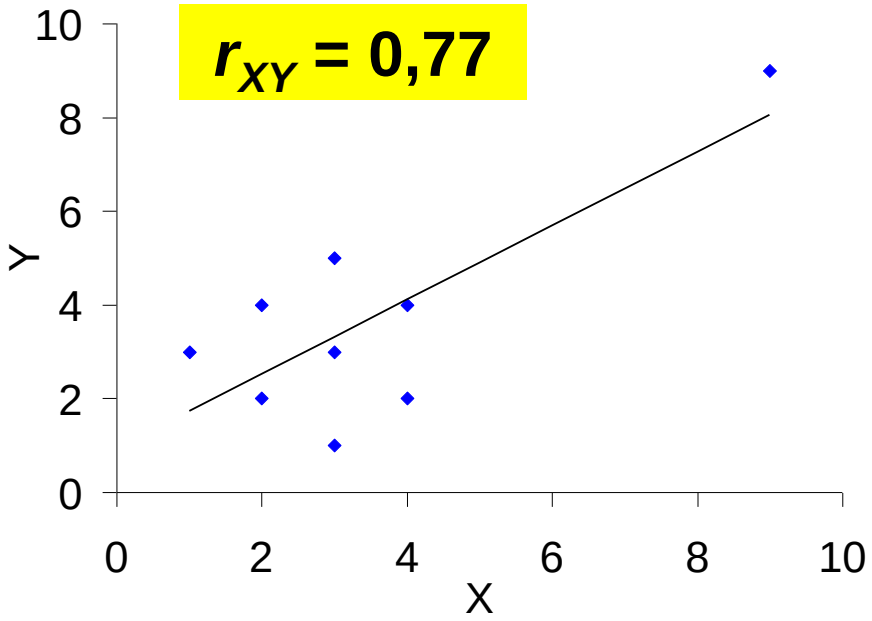
$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$r = \frac{28}{\sqrt{64(16)}} = \frac{28}{32} = 0.875$$

Zusammenhangsmaße für metrische Variablen: Zugabe

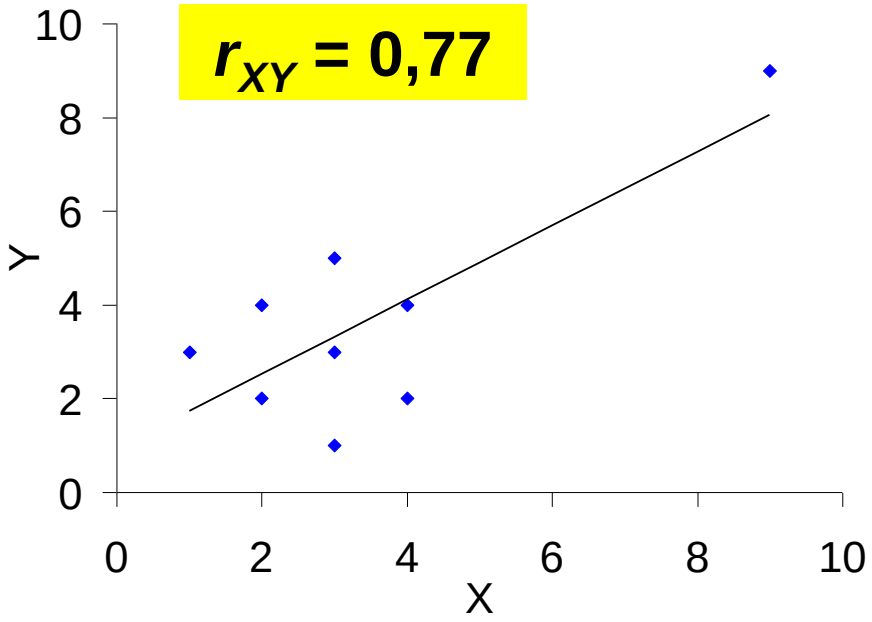
Pearson‘s *r*

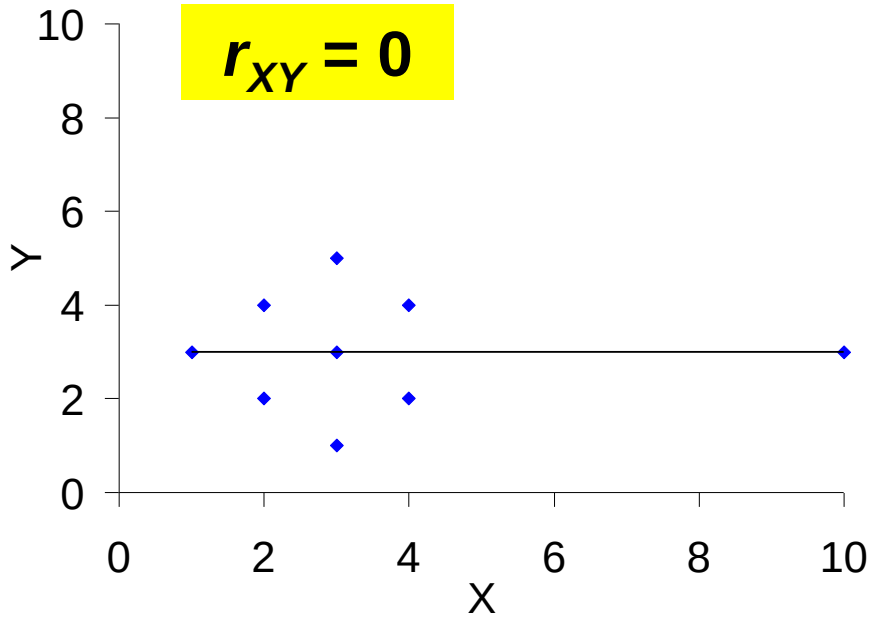
- Korrelationskoeffizienten sind sensitiv gegenüber Ausreißern und Extremwerten; insbesondere bei kleinen Stichproben



Pearson‘s *r*

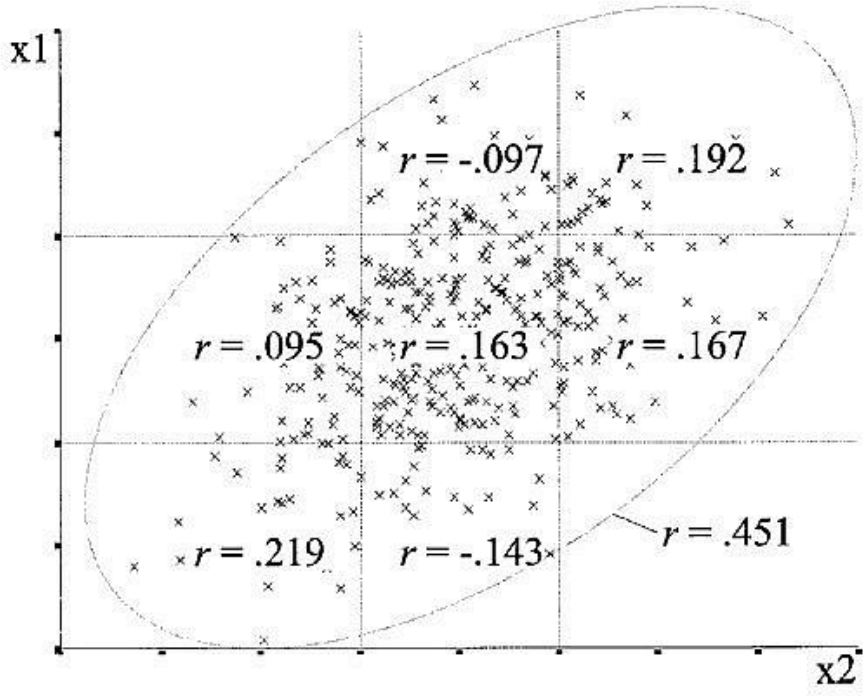
- Korrelationskoeffizienten sind sensitiv gegenüber Ausreißern und Extremwerten; insbesondere bei kleinen Stichproben





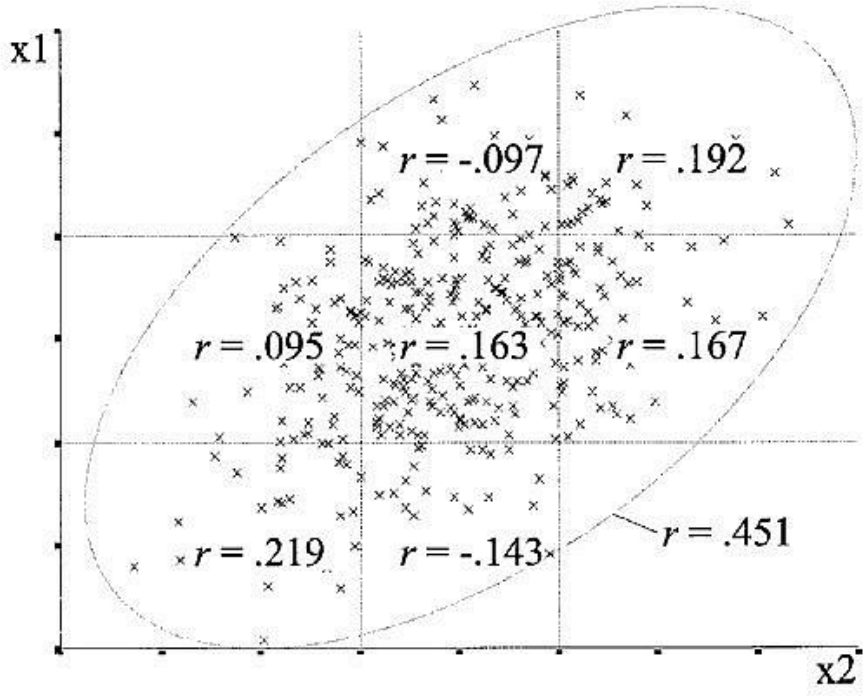
Pearson‘s *r*

- Korrelationskoeffizienten können stark variieren, je nachdem, welcher „Ausschnitt“ auf der Dimension möglicher Werteausprägungen betrachtet wird!
- Das Problem stellt sich insbesondere häufig bei nicht-repräsentativen (z.B. selektiven) Stichproben



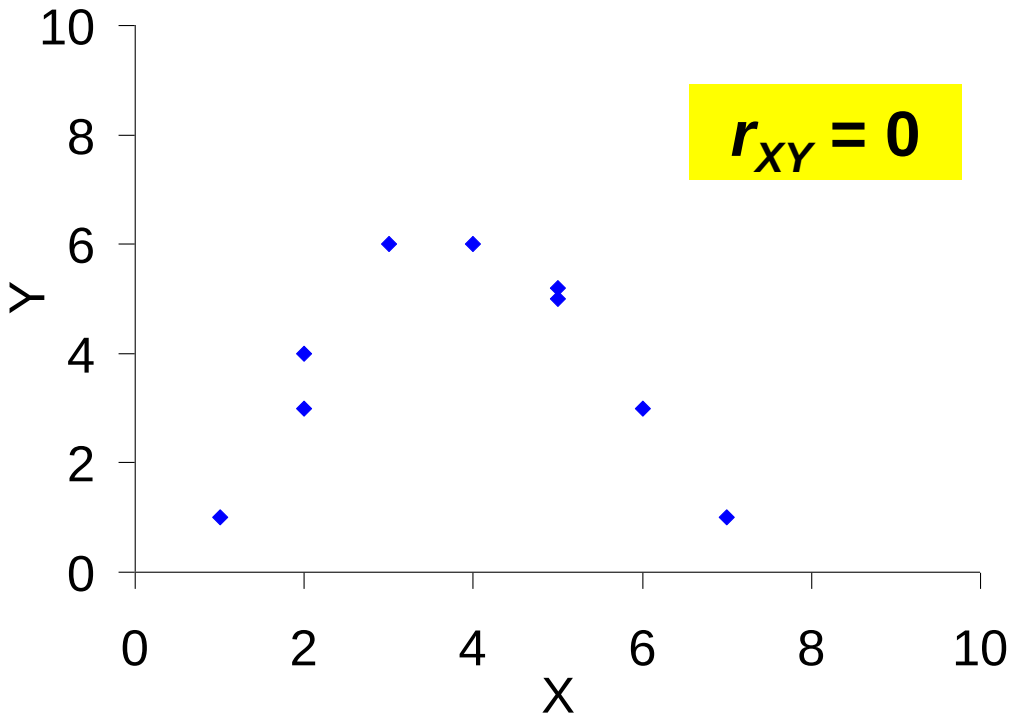
Pearson‘s *r*

- Korrelationskoeffizienten können stark variieren, je nachdem, welcher „Ausschnitt“ auf der Dimension möglicher Werteausprägungen betrachtet wird!
- Das Problem stellt sich insbesondere häufig bei nicht-repräsentativen (z.B. selektiven) Stichproben



Pearson‘s *r*

- Wenn die Form des Zusammenhangs *nicht-linear* ist, kann die Produkt-Moment-Korrelation nicht interpretiert werden!



Pearson‘s *r*

- Wenn die Form des Zusammenhangs *nicht-linear* ist, kann die Produkt-Moment-Korrelation nicht interpretiert werden!