Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

10. Vorlesung am 11.01.24: Fortsetzung bivariate Zusammenhänge, metrisch skalierte Merkmale (Pearson‘s r)

Plan für heute

- **Zusammenhänge zwischen (pseudo-)metrischen Merkmalen**
  - Allgemein
  - Hypothesen
  - Grafische Veranschaulichung
  - Zusammenhangsmaß Pearson‘s r

Bivariate Zusammenhänge…

…berechnen

Messniveau	nominal	ordinal	metrisch
nominal	**Chi-Quadrat, Cramer‘s V** Lambda C	**Cramer‘s V** Lambda C	Eta-Koeffizient Mittelwertvergleich (t-test)
ordinal	**Cramer‘s V** Lambda C	**Spearman‘s Rho;** (Kendalls Tau B gamma	**Spearman‘s Rho;** (Kendalls Tau B) gamma
metrisch	Eta-Koeffizient Mittelwertvergleich (t-test)	**Spearman‘s Rho;** (Kendalls Tau B) gamma	**Pearson‘s r**

Bivariate Zusammenhänge…

…darstellen

Messniveau	nominal	ordinal	metrisch
nominal	**Kreuztabelle**	**Kreuztabelle**	(gruppierte) **Boxplots**
ordinal		**Kreuztabelle**	(gruppierte) **Boxplots**
metrisch			**Streudiagramm**

Zusammenhangsmaße

- Maße zur Berechnung, ob und wie stark zwei Merkmale miteinander **statistisch zusammenhängen** oder „**korrelieren**“
- Bei ordinal- und metrisch skalierten Maßen auch Aussagen über die **Richtung des Zusammenhangs** möglich (positiv oder negativ)
- **Pearson‘s r** als Zusammenhangsmaß für zwei metrisch skalierte Merkmale (**Korrelationskoeffizient**)

Folgende Fragen stellen sich bei der statistischen Analyse von Merkmalszusammenhängen:

- 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**?
- Welche **Stärke** hat der Zusammenhang zwischen X und Y?
- (Lässt sich der in der Stichprobe ermittelte Zusammenhang auf die Population übertragen? → Inferenzstatik → nächstes Semester)

Zusammenhangsmaße

- Quantifizierung des Zusammenhangs durch **Korrelationskoeffizienten**
- **Positive Korrelation**:
    - Hohe Werte in der einen Variablen gehen mit hohen Werten in der anderen Variablen einher
    - Niedrige Werte in der einen 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 (Korrelation um 0)**

Zusammenhangsmaße

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

- Vergleich anhand des arithmetischen Mittels der jeweiligen Variablen
  - **Hohe Messwerte** entsprechen Werten über dem arithmetischen Mittel
  - **Niedrige Messwerte** entsprechen Werten unter arithmet. Mittel

Stärke des Zusammenhangs zwischen zwei metrischen Variablen ergibt sich unter Berücksichtigung der **Abweichung der Messwerte vom jeweiligen arithmetischen Mittel**

Hypothesen

In der quantitativen Sozialforschung entwickeln wir Hypothesen, die Zusammenhänge zwischen (mind.) 2 Merkmalen postulieren

- ***Zusammenhangshypothesen***

  Beispiel: Je geringer der soziale Status desto schlechter der Gesundheitszustand im Rentenalter → **linearer Zusammenhang**

- ***Unterschiedshypothesen***

  Beispiel: Rentner*innen mit hohem sozialen Status haben einen besseren Gesundheitszustand als Rentner*innen mit niedrigem sozialen Status

Korrelationen

- Eine **lineare Korrelation** beschreibt einen **linearen** Zusammenhang zwischen zwei Variablen
- D.h. eine Veränderung von X geht mit einer dazu proportionalen Veränderung von Y einher
  - Beispiel: Je größer eine Person ist, desto schwerer ist sie (pro cm mehr Größe, bestimmte Grammzahl mehr)
- Korrelation trifft **keine Aussage** zur Kausalität
- Korrelationsanalyse misst nur, ob sich zwei Merkmale **im Gleichklang** bewegen

Untersuchung von Hypothesen

- Bi- (und multi-)variate Analysen, um Zusammenhänge zwischen Merkmalen zu untersuchen
- Bei **metrisch** skalierten Daten:
  - **Korrelationsanalyse** (Pearson‘s r)
  - zunächst bietet sich eine **grafische Darstellung** an: **Streudiagramm** (auch Punktwolkendiagramm, Scatterplot)

Plan für heute

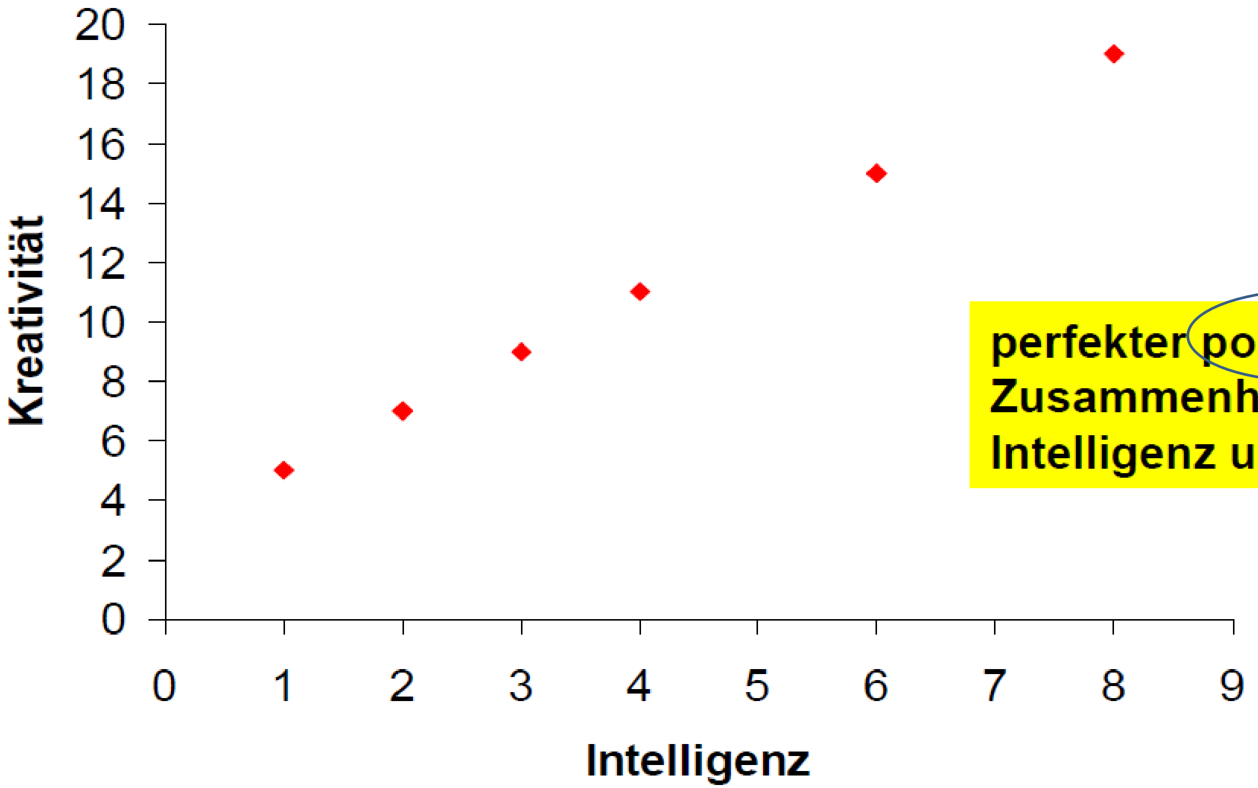
- **Zusammenhänge zwischen (pseudo-)metrischen Merkmalen**
  - Allgemein
  - Hypothesen
  - Grafische Veranschaulichung
  - Zusammenhangsmaß Pearson‘s r

Beispiel: Streudiagramm

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

Beispiel: Streudiagramm

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



(Pearson's)
r=1

perfekter positiver Zusammenhang zwischen Intelligenz und Kreativität

Beispiel 2: Streudiagramm

Zusammenhangshypothese (S. Lehrbrief S. 63/64): „Je intelligenter eine Person ist, desto besser kann sie auch räumlich denken“

Tabelle 34: IQ und Testergebnis beim räumlichen Denken – Urliste

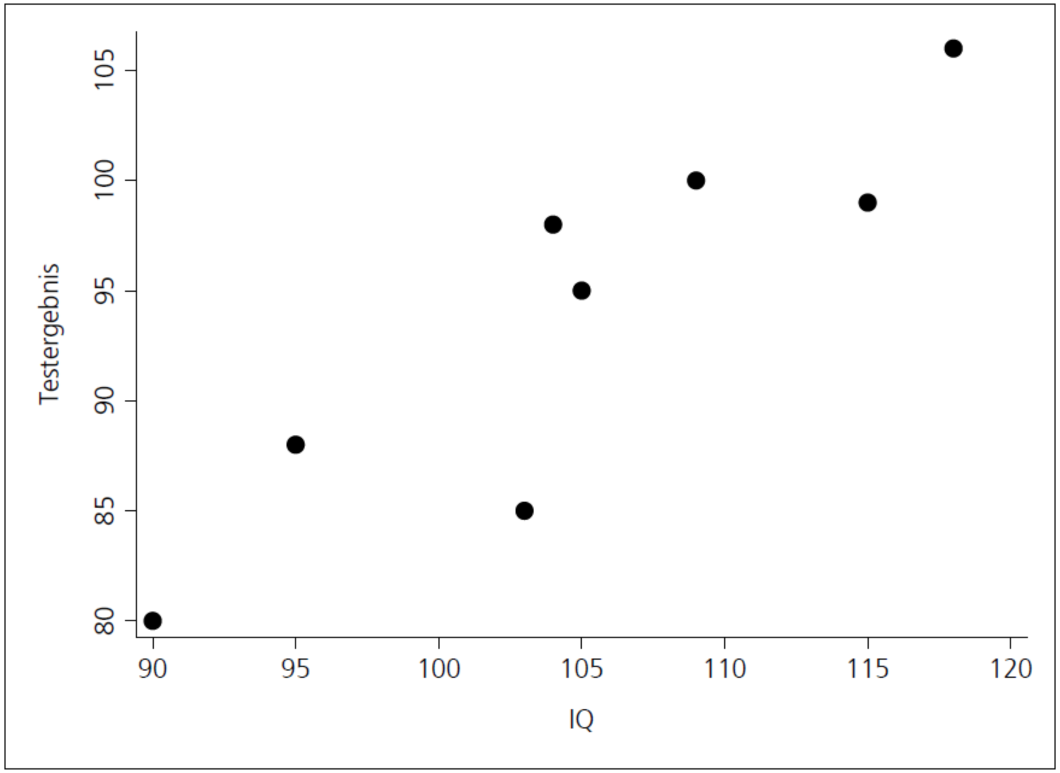
ID	IQ	Testergebnis
1	104	98
2	90	80
3	103	85
4	115	99
5	105	95
6	118	106
7	109	100
8	95	88

Quelle: Eigene Darstellung

Beispiel 2: Streudiagramm

Zusammenhangshypothese (S. Lehrbrief S. 63/64): „Je intelligenter eine Person ist, desto besser kann sie auch räumlich denken.“

Abbildung 13: IQ und Testergebnis beim räumlichen Denken – Streudiagramm

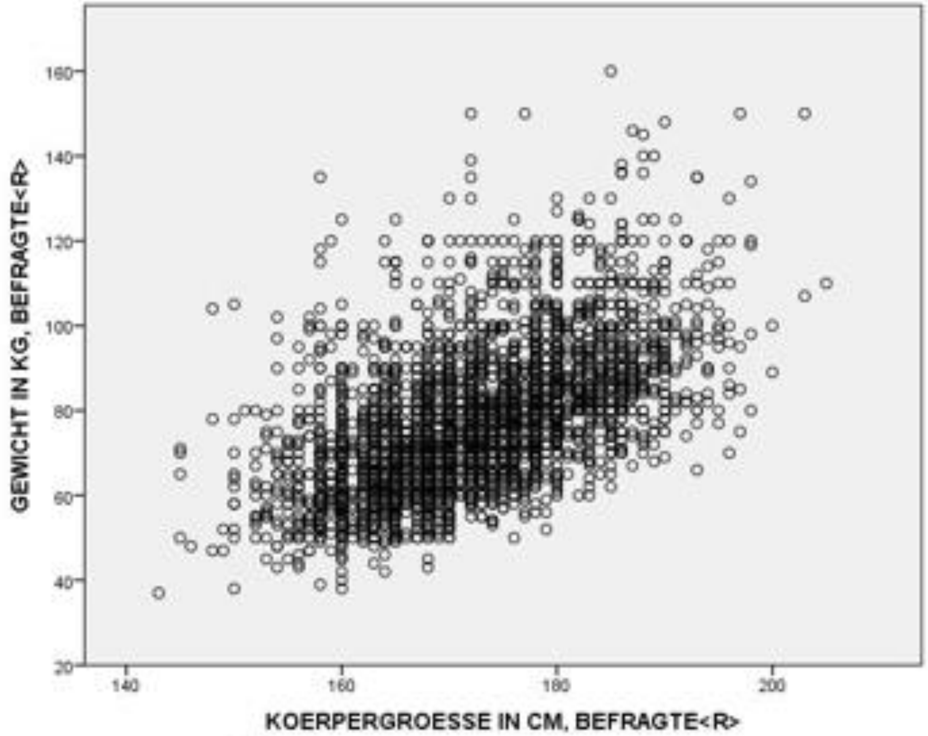


Quelle: Eigene Darstellung

Beispiel 3: Streudiagramm

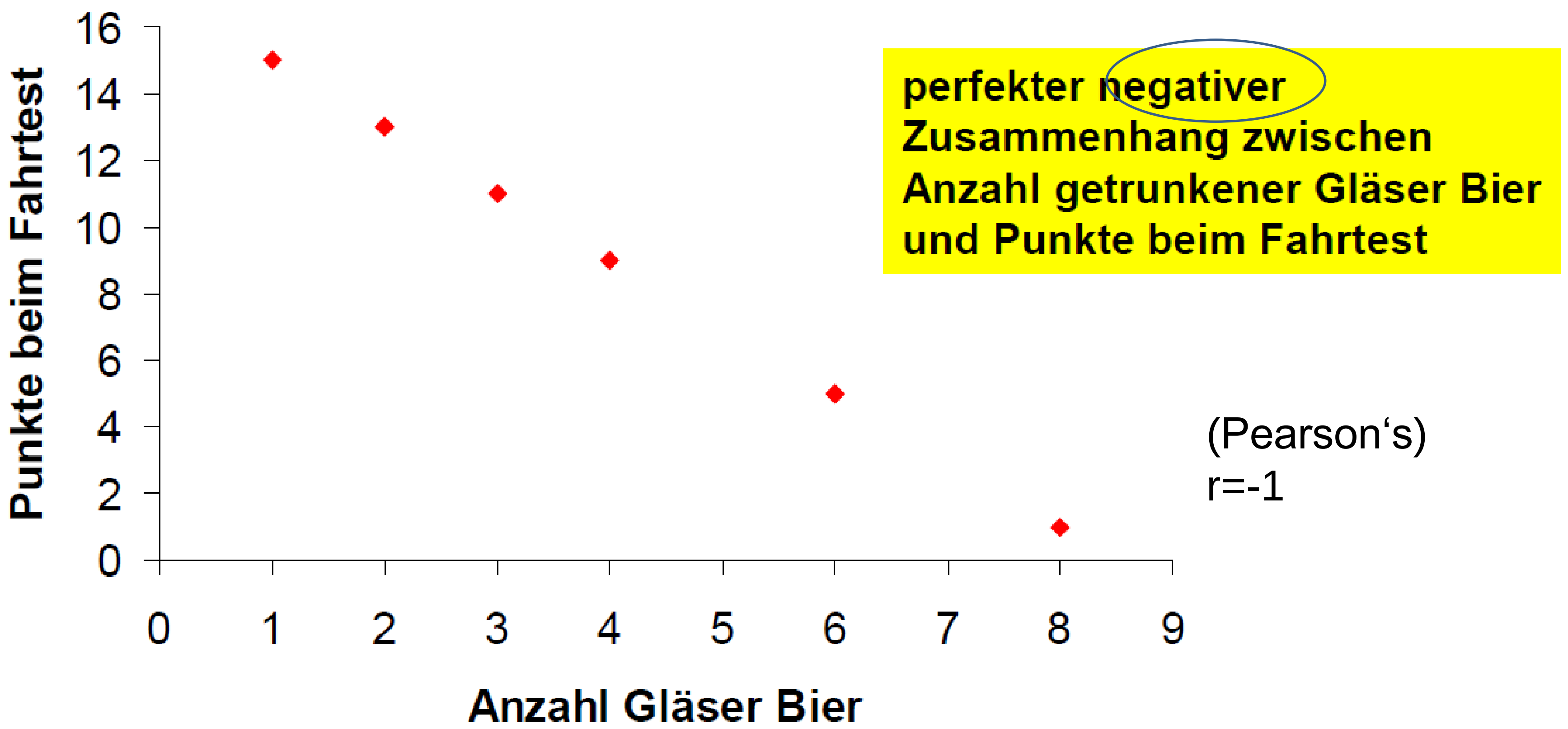
Zusammenhangshypothese: „Je größer man ist, desto schwerer ist man auch.“ (Streudiagramm erstellt auf Basis des Allbus 2016)

r=0,547



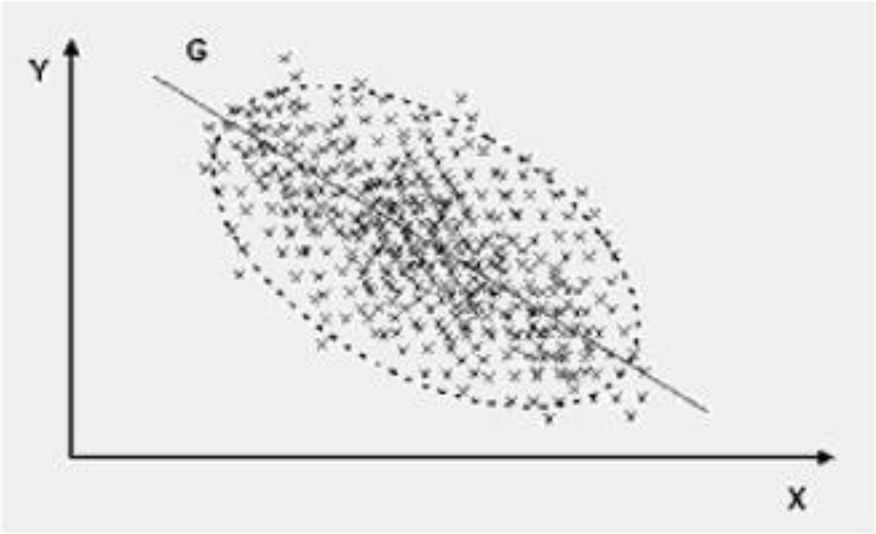
Beispiel 4: Streudiagramm

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



Beispiel 5: Streudiagramm

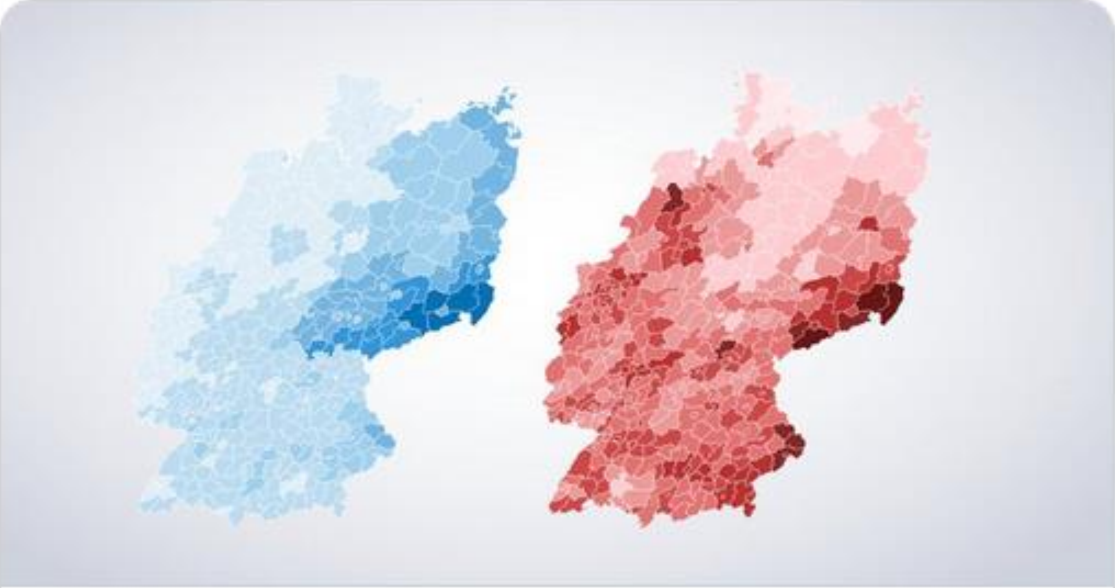
Zusammenhangshypothese: „Je älter ein Auto ist, desto niedriger sein Verkaufswert" – negativer Zusammenhang



Exkurs: grafische Darstellung zweier Merkmale an Beispiel

Matthias Quent @Matthias_Quent · 12. Dez. 2020

Hängen #AfD-Hochburgen und hohe #Coronazahlen zusammen? @plateauton @Tagesspiegel hat unsere Berechnungen @IDZ_Jena überprüft und ausgeweitet. Die Ergebnisse bestätigen die Korrelation. interaktiv.tagesspiegel.de/lab/hotspots-u... via @tagesspiegel

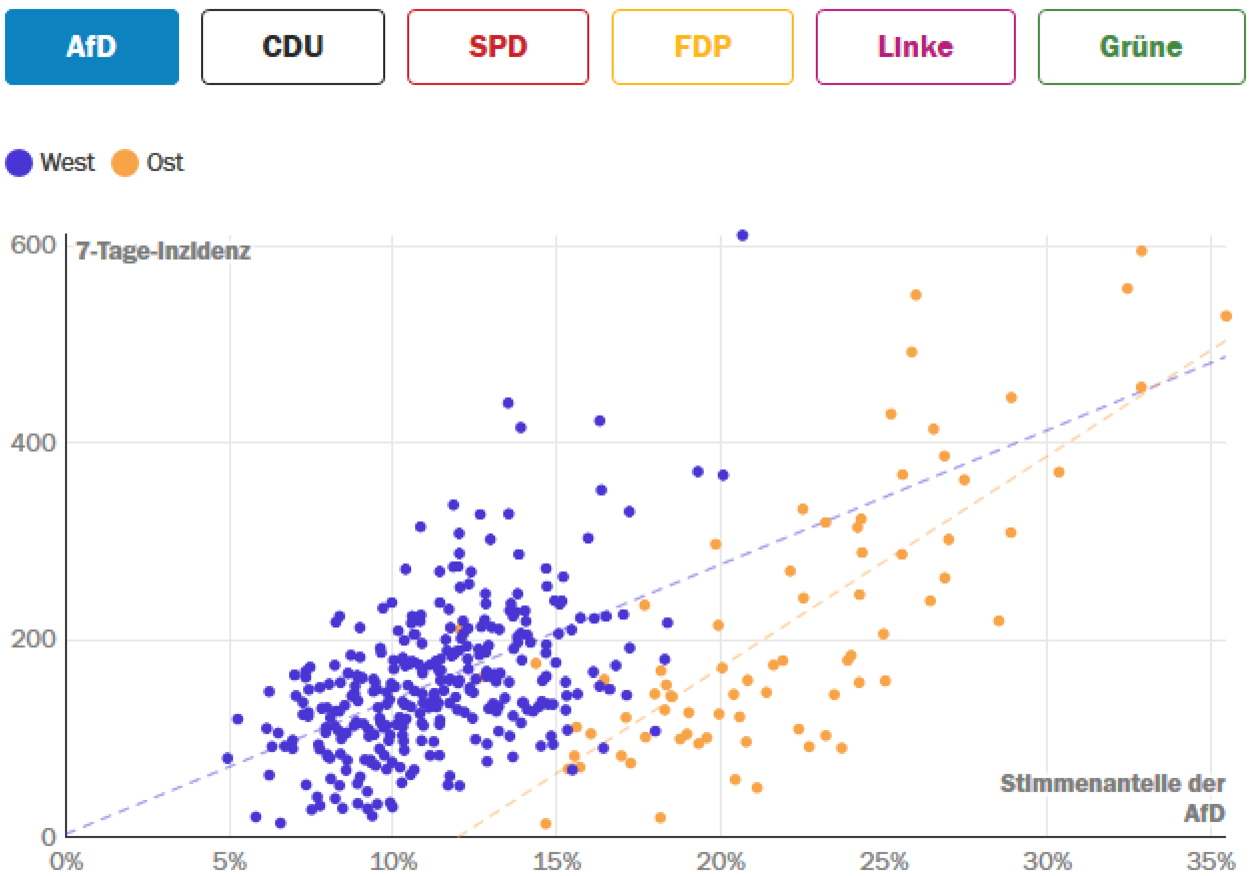


Hängen AfD-Hochburgen und hohe Coronazahlen zusammen?

Eine Analyse legt nahe, dass in Landkreisen mit großer AfD-Wählerschaft auch die Fallzahlen höher sind. Was ist dran? Wir rechnen nach.

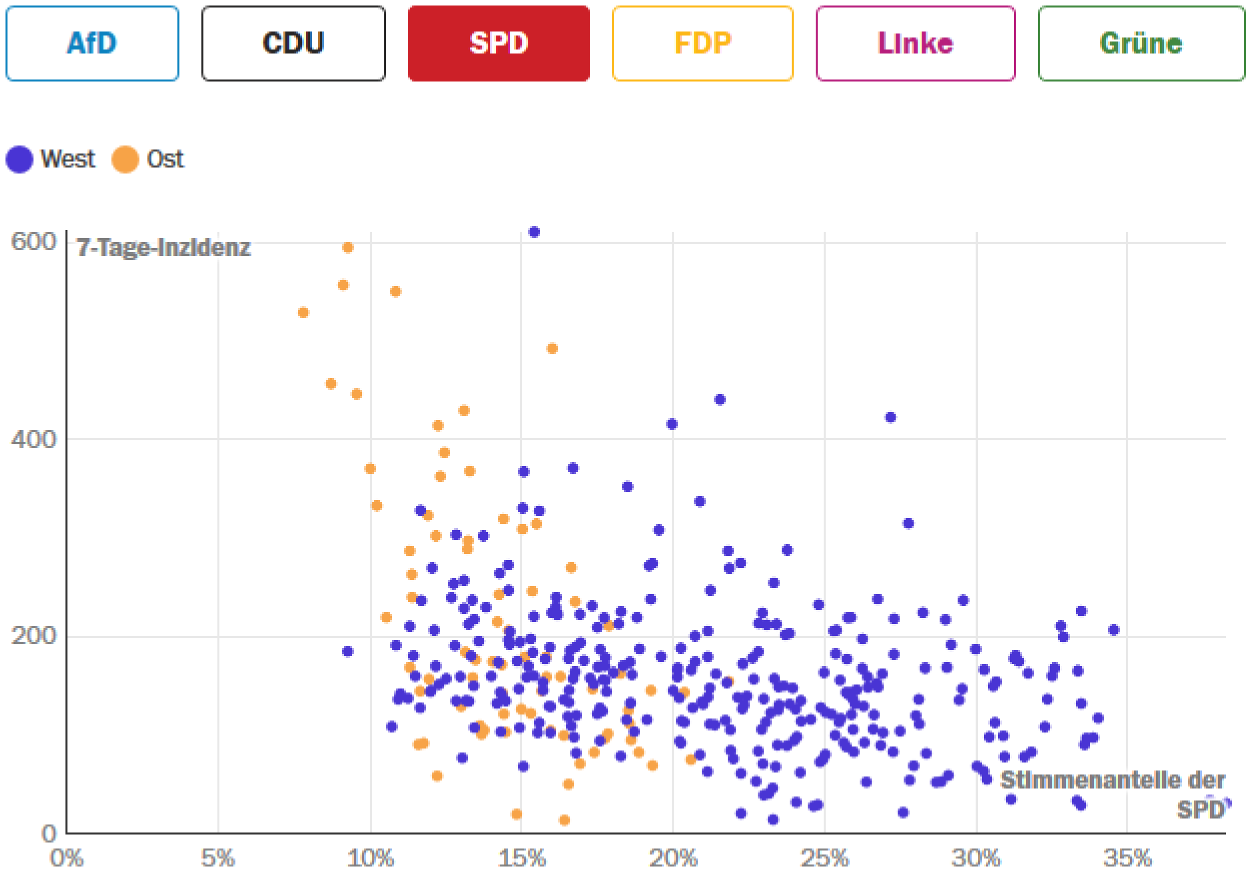
interaktiv.tagesspiegel.de

Exkurs: grafische Darstellung zweier Merkmale an Beispiel



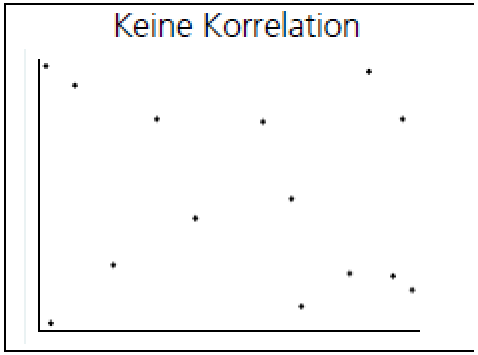
https://interaktiv.tagesspiegel.de/lab/hotspots-und-rechte-haengen-afd-hochburgen-und-corona-hotspots-zusammen/

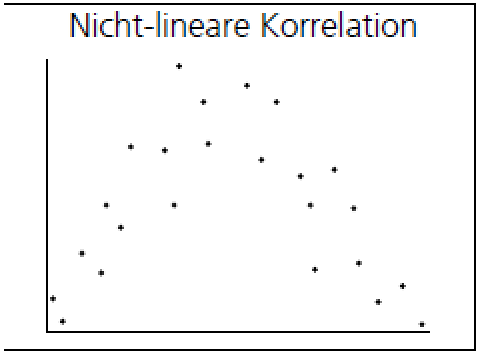
Exkurs: grafische Darstellung zweier Merkmale an Beispiel



https://interaktiv.tagesspiegel.de/lab/hotspots-und-rechte-haengen-afd-hochburgen-und-corona-hotspots-zusammen/

Weitere Streudiagramme und „Zusammenhänge“





Quelle: Eigene Darstellung

Plan für heute

- **Zusammenhänge zwischen (pseudo-)metrischen Merkmalen**
  - Allgemein
  - Hypothesen
  - Grafische Veranschaulichung
  - Zusammenhangsmaß Pearson‘s r

Pearson‘s r

- Auch **Bravais-Pearson Produktkorrelation**
- Berechnet die Stärke des **linearen Zusammenhangs** zwischen zwei (pseudo-)metrisch skalierten Variablen
- Zwischenschritt zur Berechnung: **Kovarianz**

Kovarianz - Berechnung

Kovarianz beschreibt die **gemeinsame Streuung zweier Merkmale:**

1) Abweichung vom Mittelwert für jedes Messwertepaar bestimmen

2) Gemeinsame Abweichung beider Messwerte von ihren Mittelwerten durch Multiplikation berechnen

3) Berechnung der Summe der Abweichungsprodukte

4) Berechnung des durchschnittlichen Abweichungsprodukt (mittels Division durch n)

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

Wiederholung Varianz $$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Kovarianz

- hoch **positiv**, wenn hohe positive Abweichungen mit hohen positiven Abweichungen einhergehen bzw. hohe negative Abweichungen mit hohen negativen Abweichungen
- hoch **negativ**, wenn hohe positive Abweichungen mit hohen negativen Abweichungen einhergehen und umgekehrt.
- gleich **null**, wenn die Richtung der Abweichungen vom Mittelwert in X nicht systematisch mit einer bestimmten Richtung der Abweichungen vom Mittelwert in Y einhergeht.

Kovarianz

- Aber: **unstandardisiertes** Maß, Größe ist abhängig von den jeweiligen Maßeinheiten der beiden Merkmale

- Vergleich zwischen Kovarianzen wird erschwert

→ **Standardisierung durch Korrelationskoeffizienten Pearson‘s r** anhand der Division durch das Produkt der Standardabweichungen beider Merkmale

Pearson‘s r

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

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

Pearson‘s r

- **Wertebereich von -1 bis +1**,
- Vorzeichen zeigt **Richtung** der Korrelation an, Betrag die **Stärke** des Zusammenhanges
    - Negatives Vorzeichen: negativer Zusammenhang
    - Positives Vorzeichen: positiver Zusammenhang

Interpretation Pearson‘s r - Faustregel

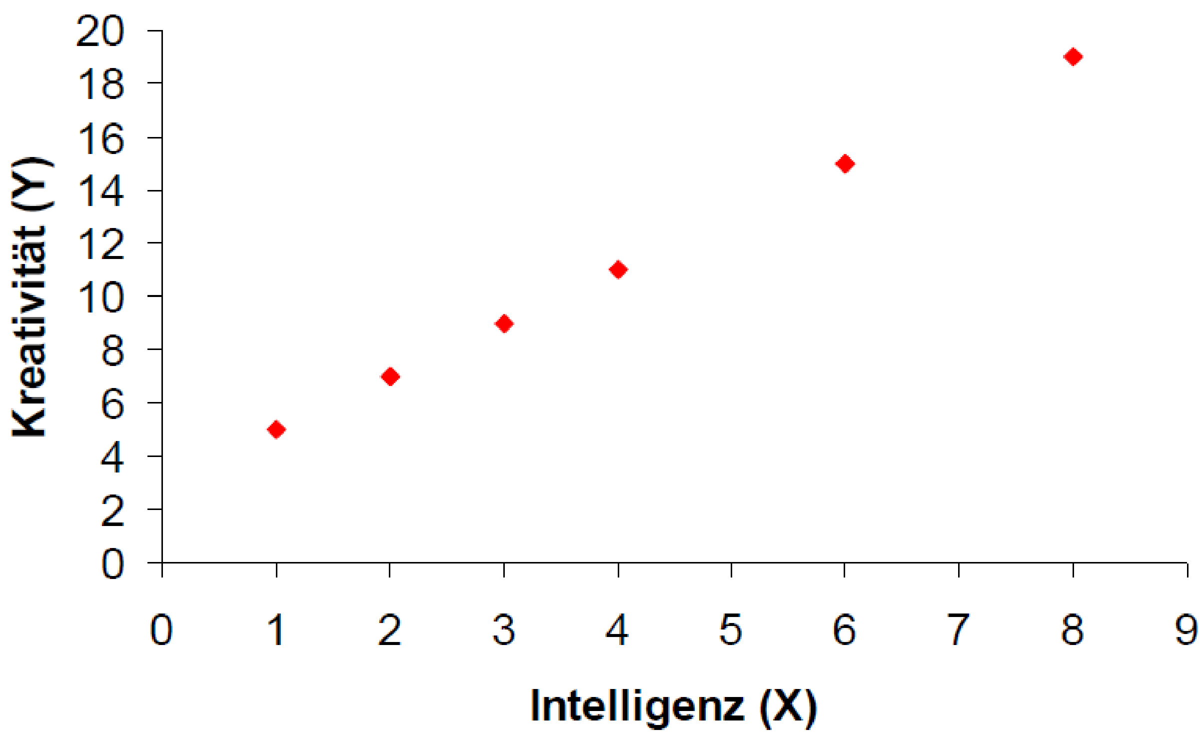
Tabelle 36: Interpretation von Pearson's r

Korrelationskoeffizient ($\|r\|$)	Interpretation
$\leq 0{,}05$	kein Zusammenhang
$> 0{,}05$ bis $\leq 0{,}20$	schwacher Zusammenhang
$> 0{,}20$ bis $\leq 0{,}50$	mittelstarker Zusammenhang
$> 0{,}50$ bis $\leq 0{,}70$	starker Zusammenhang
$> 0{,}70$	sehr starker Zusammenhang

Quelle: Eigene Darstellung

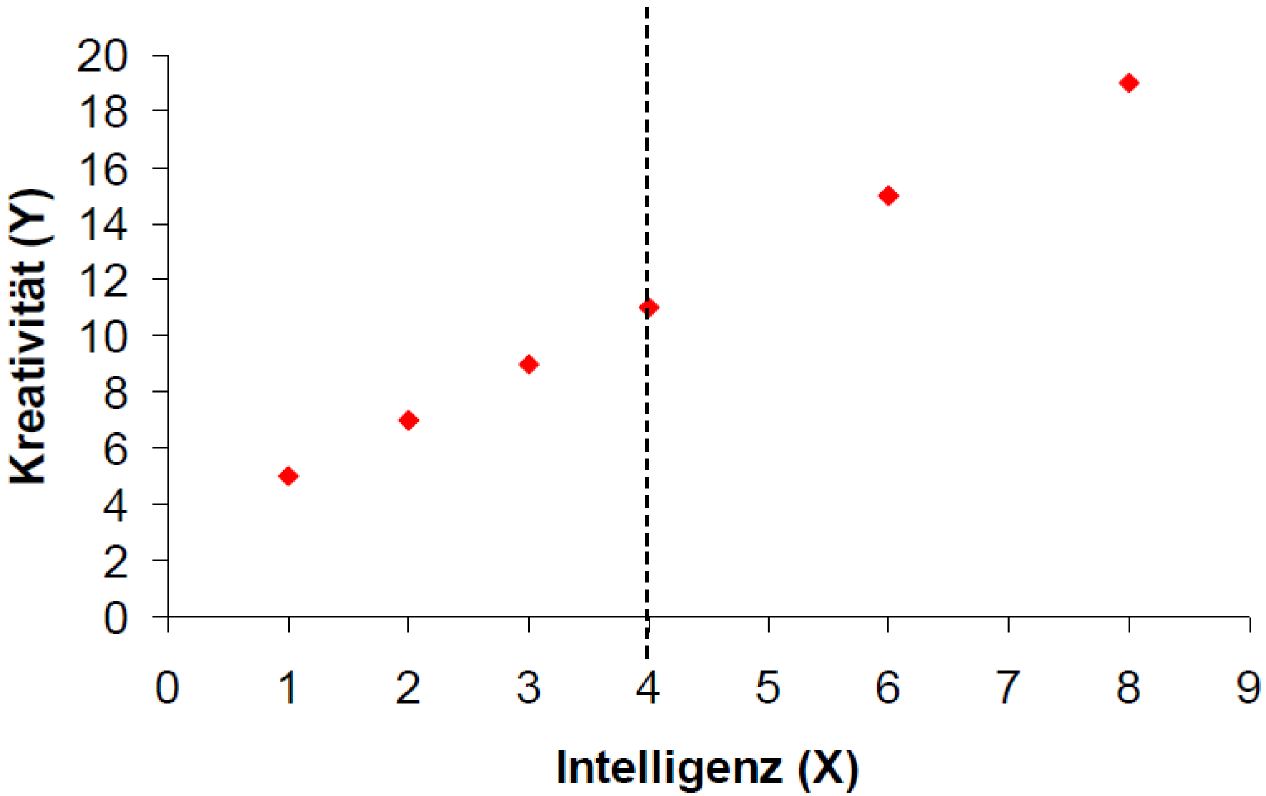
Aber: Die Beurteilung der Höhe einer Korrelation hängt immer von der zugrunde liegenden Fragestellung ab!

Pearson‘s r: Schritt für Schritt



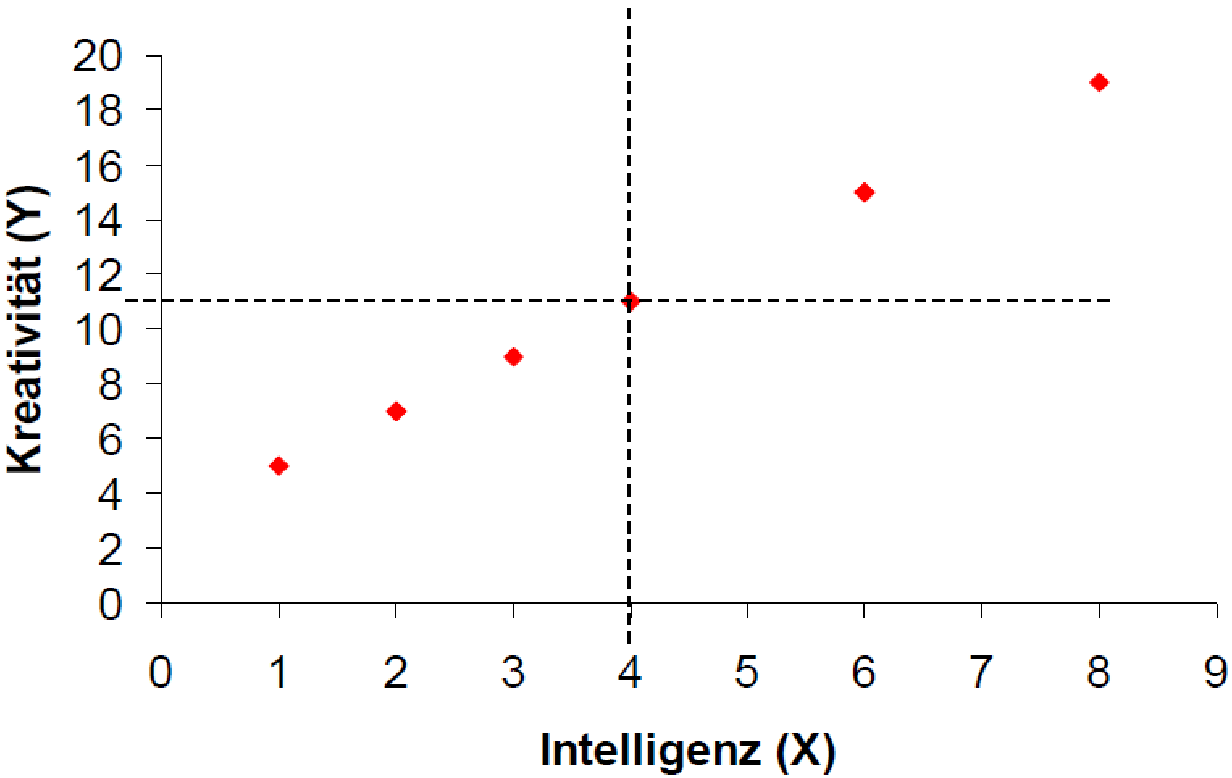
	X	*Y*
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

Pearson‘s r: Schritt für Schritt



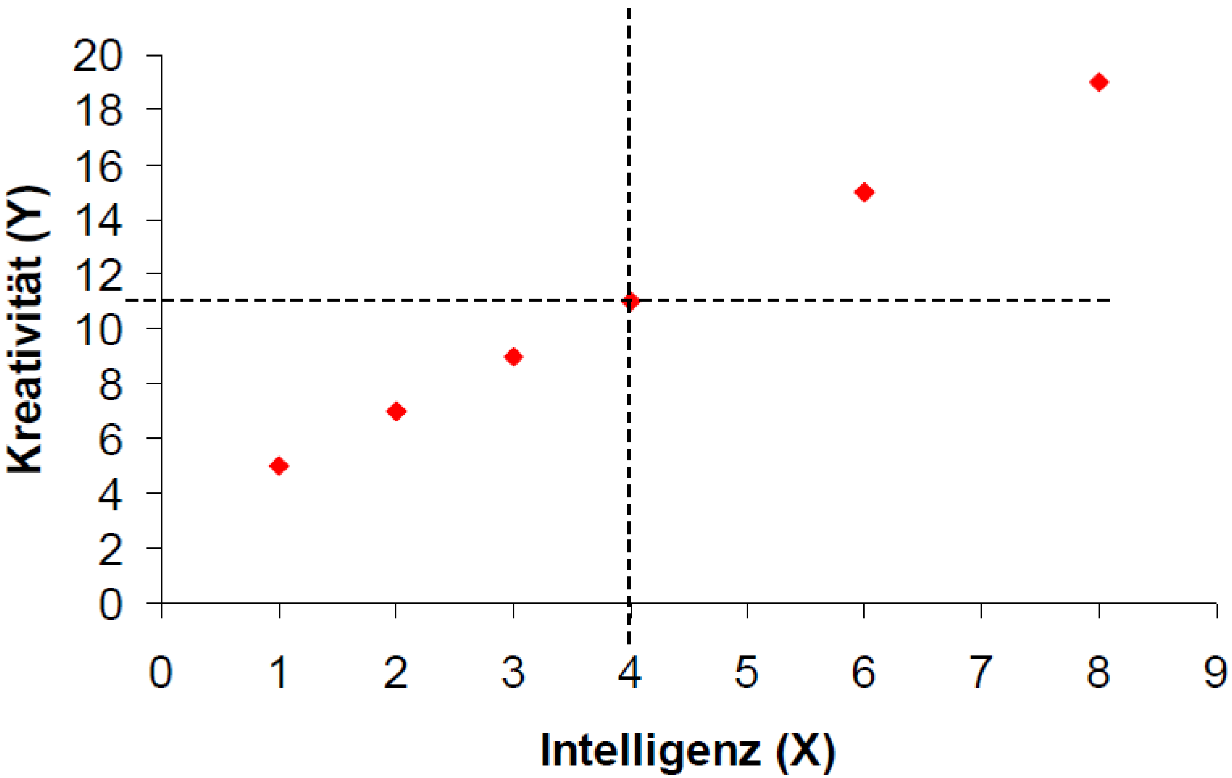
	X	*Y*
	1	5
	2	7
	2	7
	3	9
	4	11
	6	15
	6	15
	8	19
M =	4	11
s^2 =	5,25	21

Pearson‘s r: Schritt für Schritt



	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

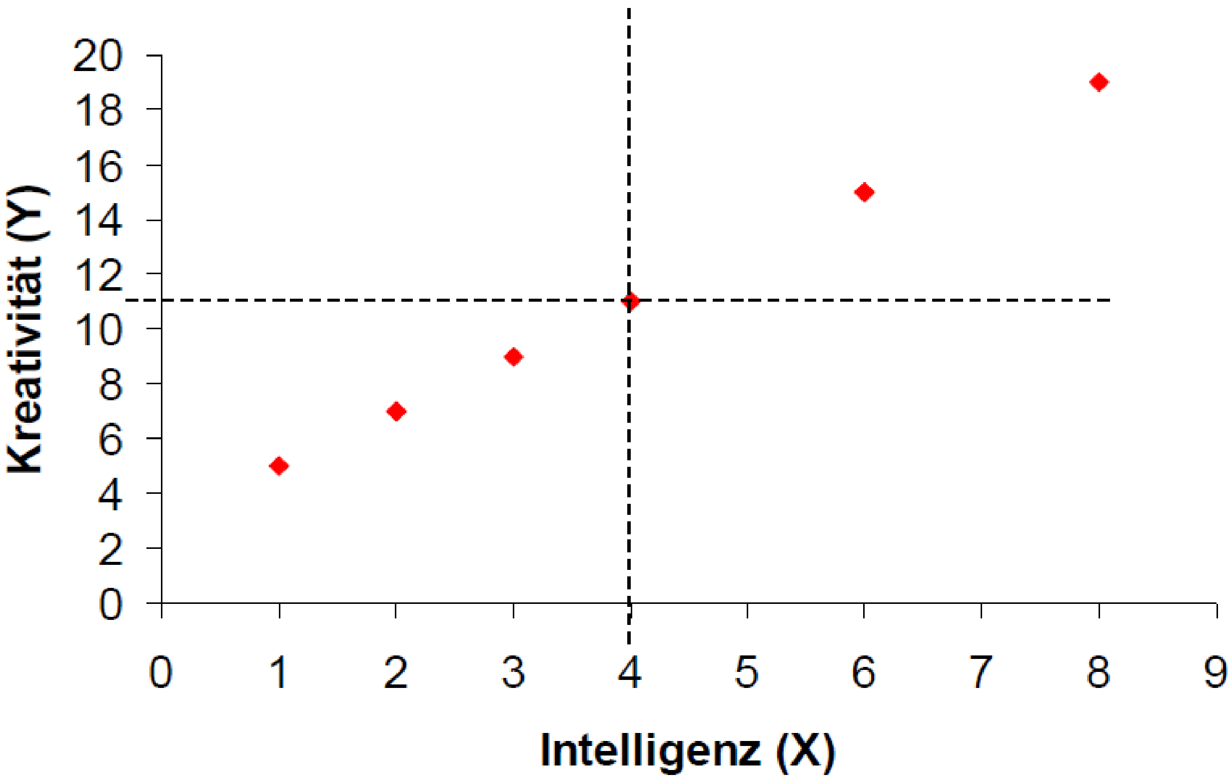
Pearson‘s r: Schritt für Schritt



	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 zwischen zwei Variablen X und Y positiv?

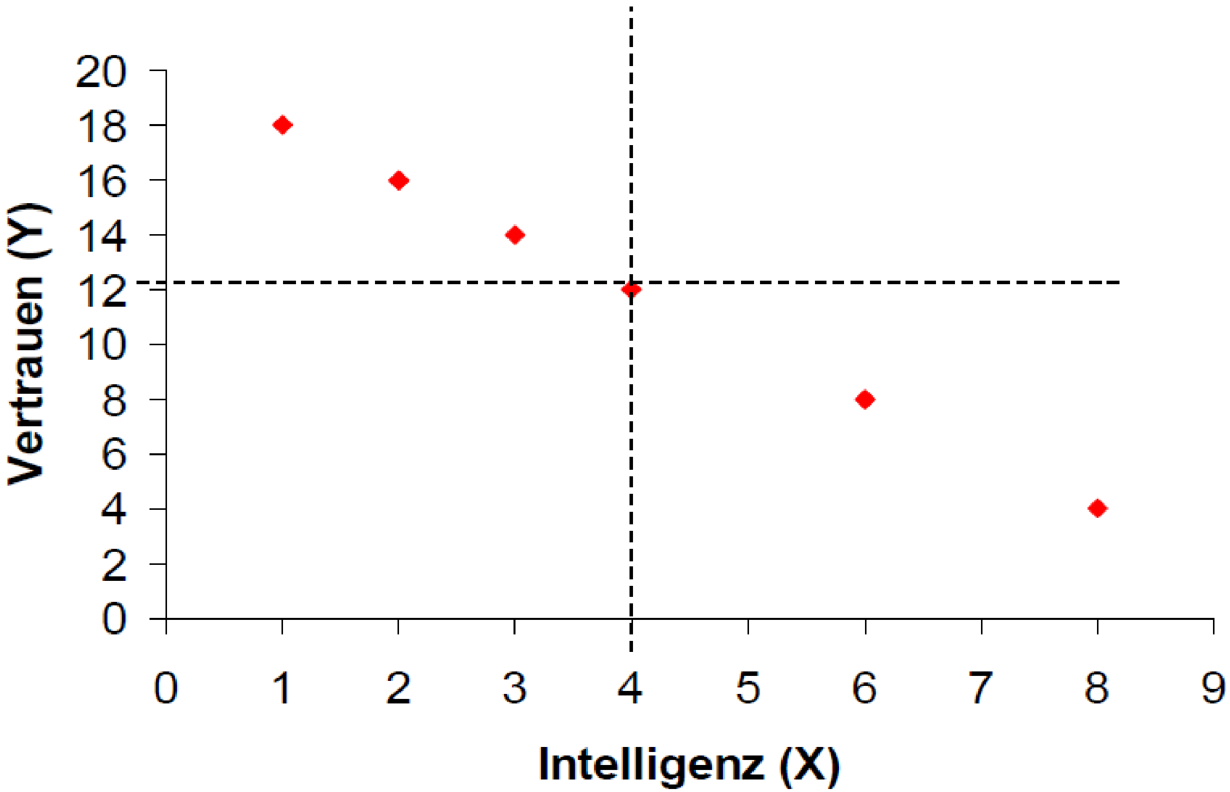
Pearson's r: Schritt für Schritt



	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

Zusammenhang X und Y dann positiv, wenn x-Werte$> \bar{x}$ mit y-Werte $>\bar{y}$ einhergehen (und umgekehrt)

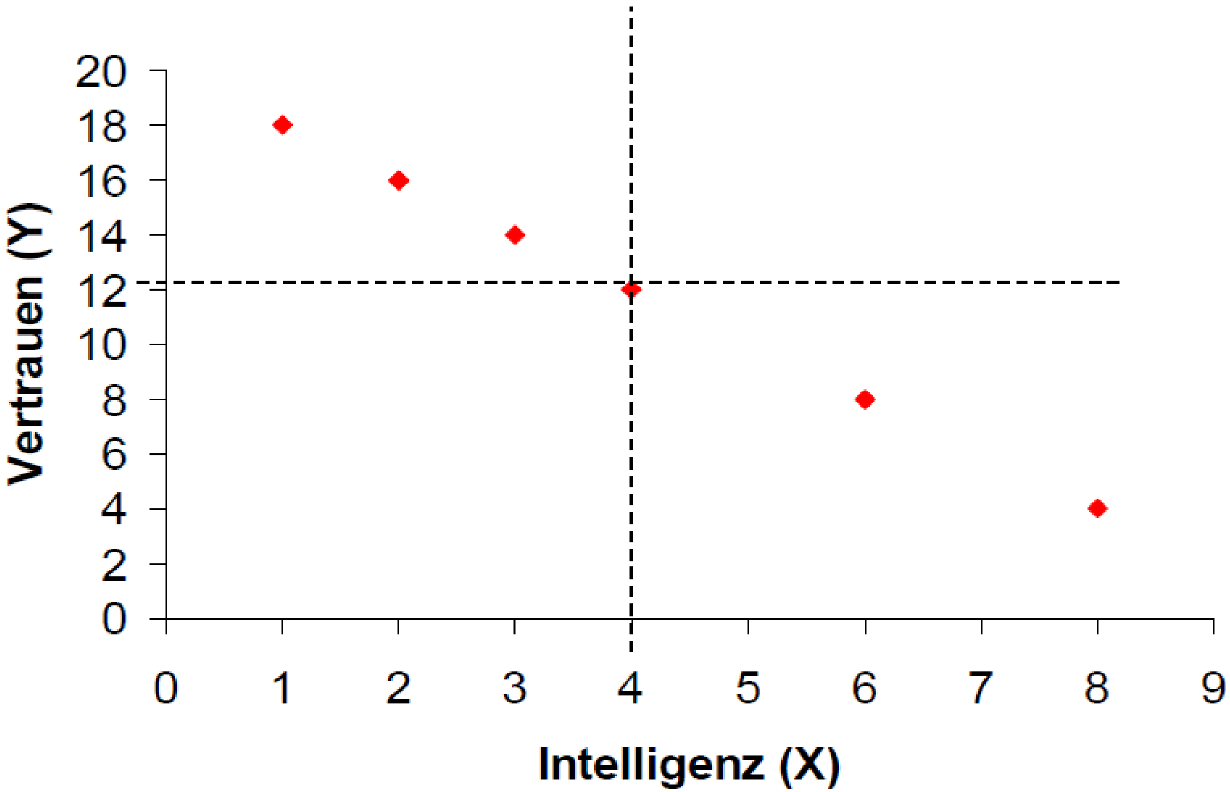
Pearson‘s r: Schritt für Schritt



	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

Negativer Zusammenhang zwischen 2 Variablen X und Y?

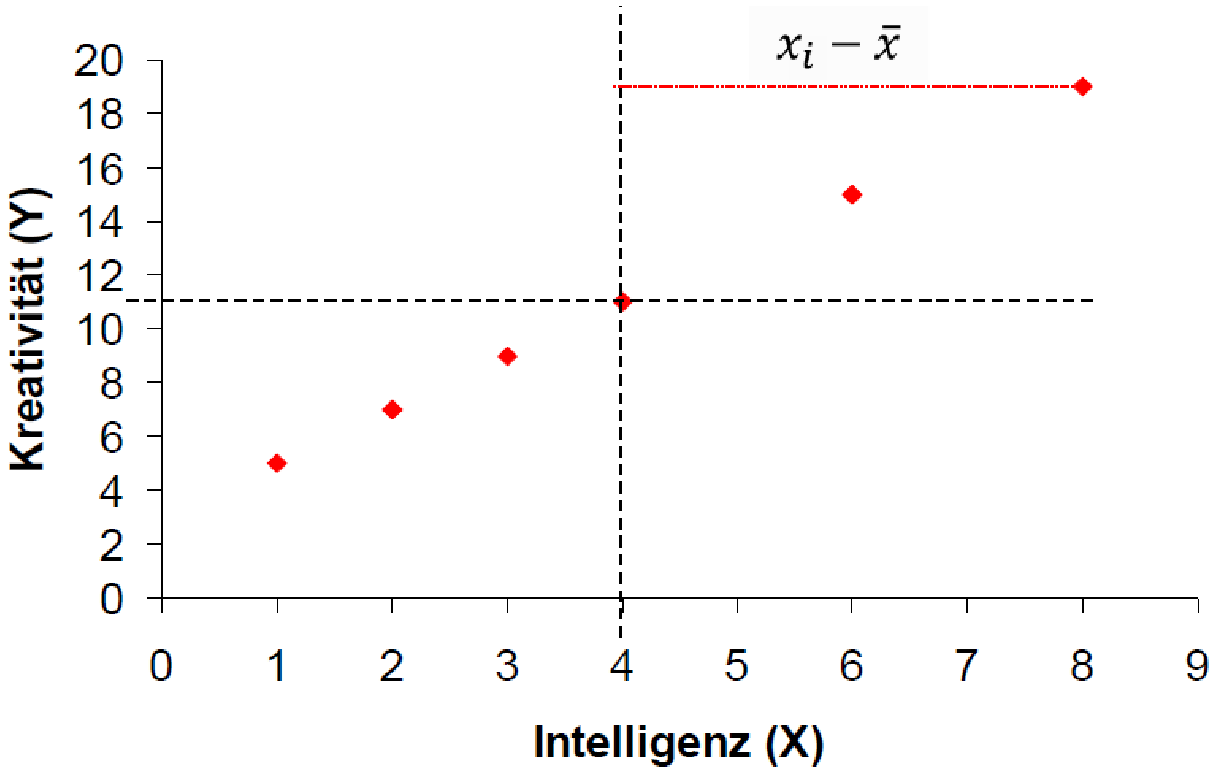
Pearson‘s r: Schritt für Schritt



	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

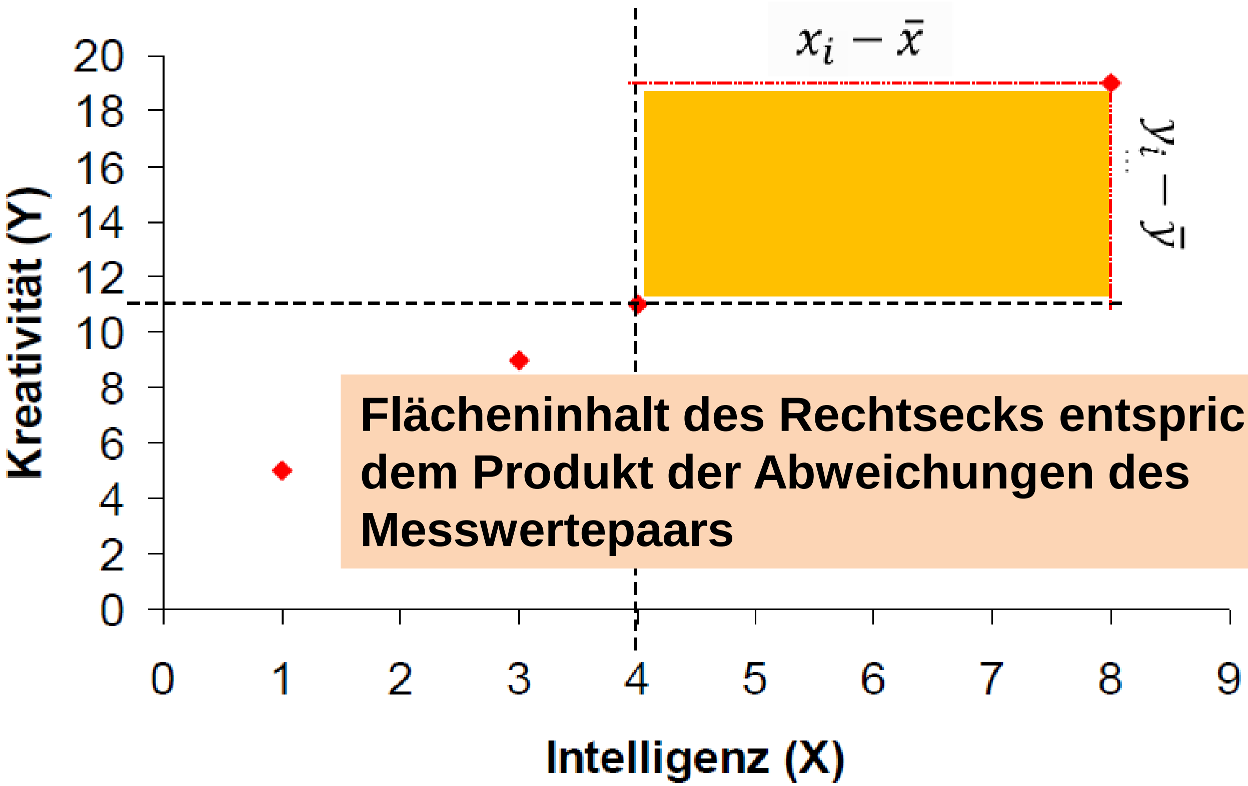
Zusammenhang dann negativ, wenn zwischen 2 Variablen x-Werte >$\bar{x}$ einhergehen mit y-Werte <$\bar{y}$

Pearson‘s r: Schritt für Schritt



	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

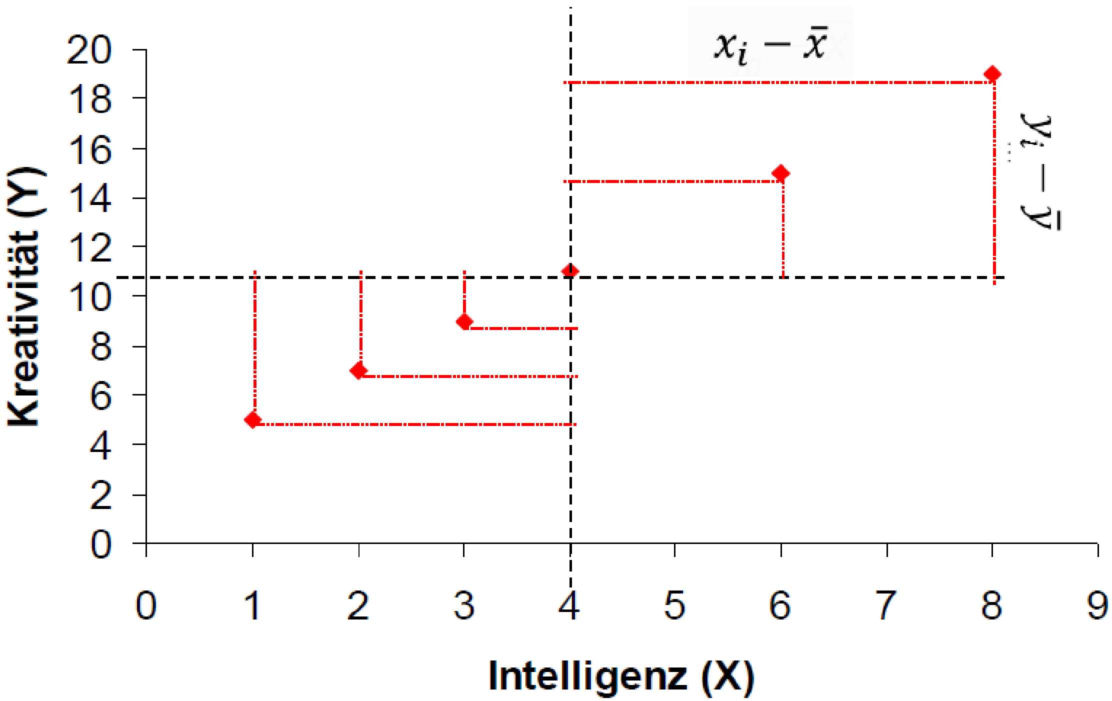
Pearson's r: Schritt für Schritt



Flächeninhalt des Rechtsecks entspricht dem Produkt der Abweichungen des Messwertepaars

	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

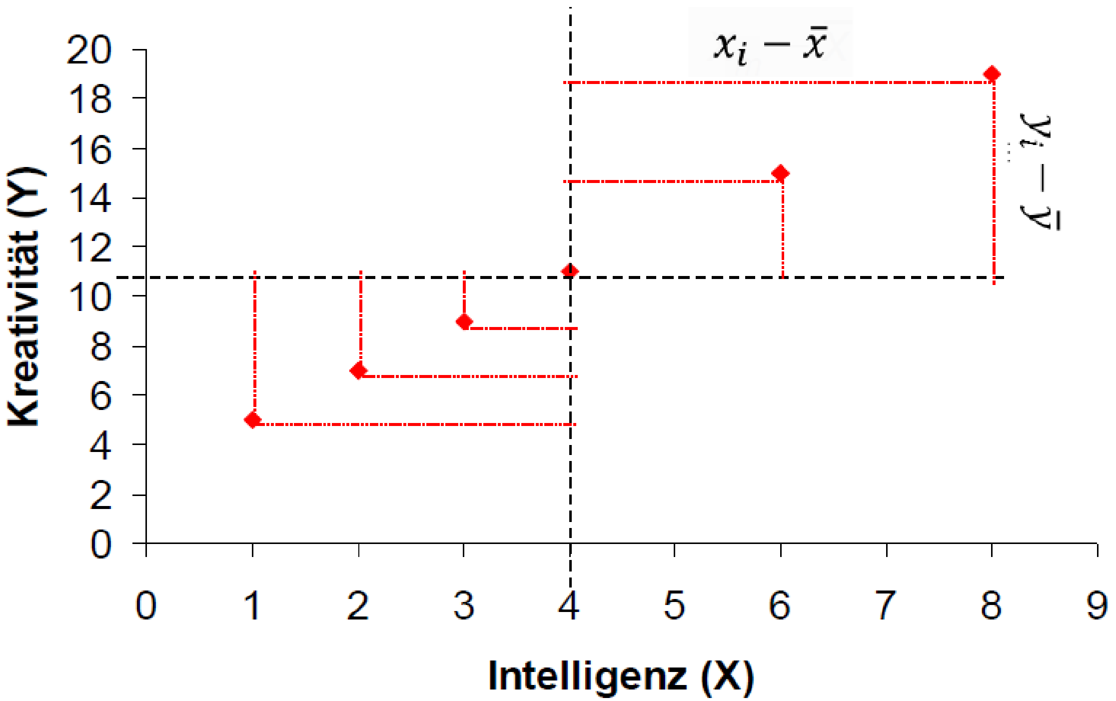
Pearson's r: Schritt für Schritt



$x_i - \bar{x}$	$y_i - \bar{y}$
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: Für jedes x_i sowie y_i wird Differenz vom jeweiligen arithmetischen Mittel berechnet

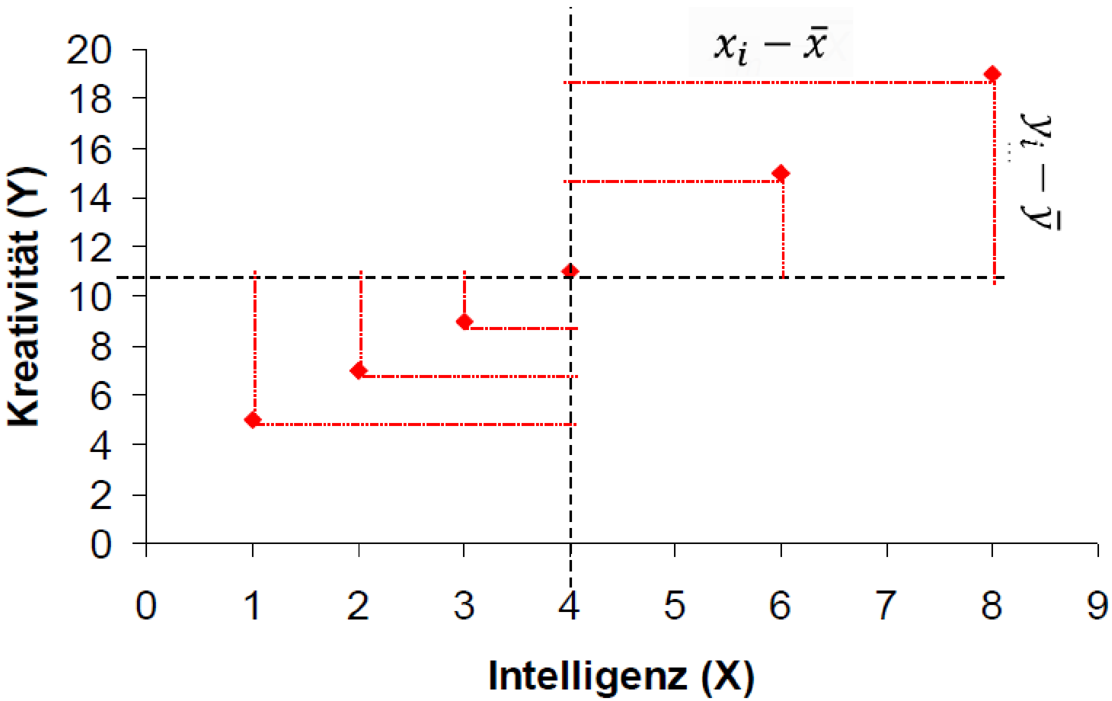
Pearson's r: Schritt für Schritt



$(x_i - \bar{x})(y_i - \bar{y})$		
-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: Für jedes Wertepaar xy_i wird das *Kreuzprodukt*, d.h. das Produkt der Mittelwertsabweichung berechnet

Pearson's r: Schritt für Schritt

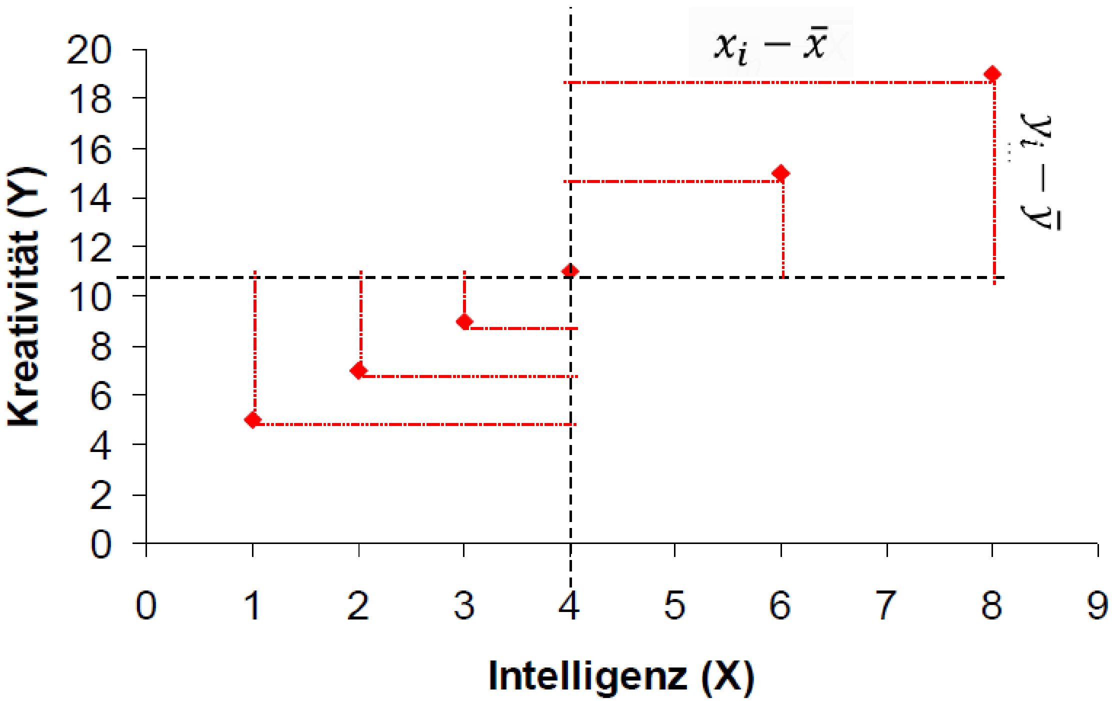


$(x_i - \bar{x})(y_i - \bar{y})$		
-3	-6	18
-2	-4	8
-2	-4	8
-1	-2	2
0	0	0
2	4	8
2	4	8
4	8	32
Summe:		84

Schritt 3: Berechnung der 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: Schritt für Schritt



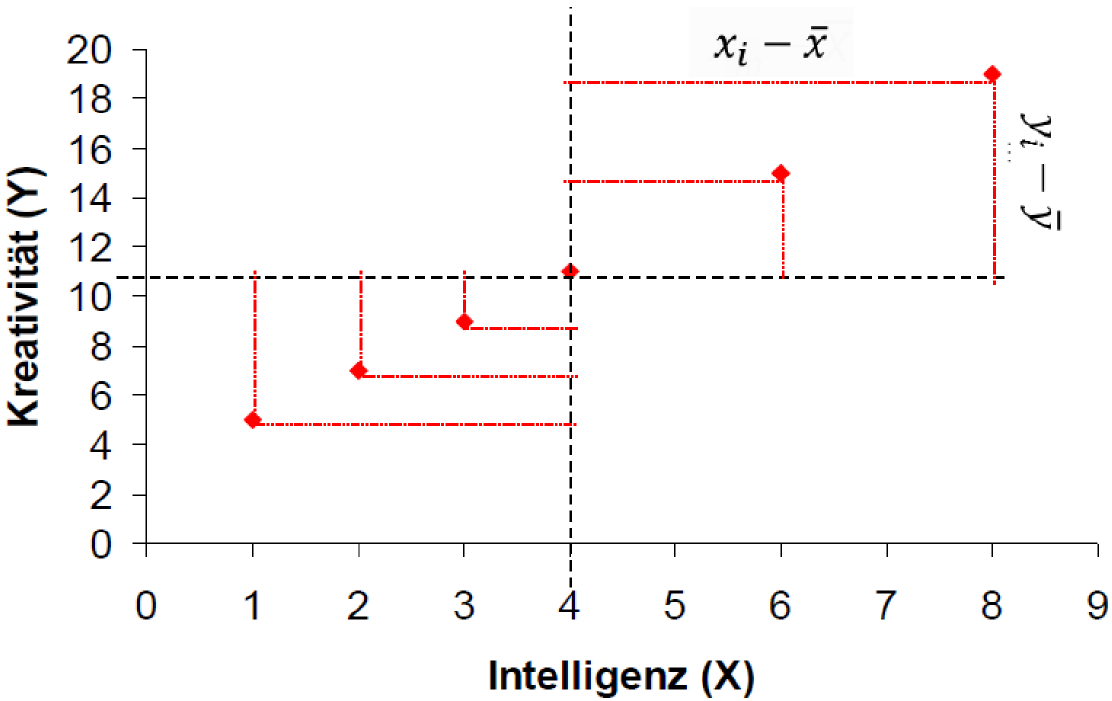
$(x_i - \bar{x})(y_i - \bar{y})$		
-3	-6	18
-2	-4	8
-2	-4	8
-1	-2	2
0	0	0
2	4	8
2	4	8
4	8	32

Summe:	84
Kovarianz:	10,5

Schritt 4: Berechnung Kovarianz, indem durch n geteilt wird („mittleres Kreuzprodukt“)

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

Pearson‘s r: Schritt für Schritt



$s_X^2 = 5{,}25 \quad s_X = 2{,}29$

$s_Y^2 = 21 \quad s_Y = 4{,}58$

$s_X \cdot s_Y = 2{,}29 \cdot 4{,}58 = 10{,}5$

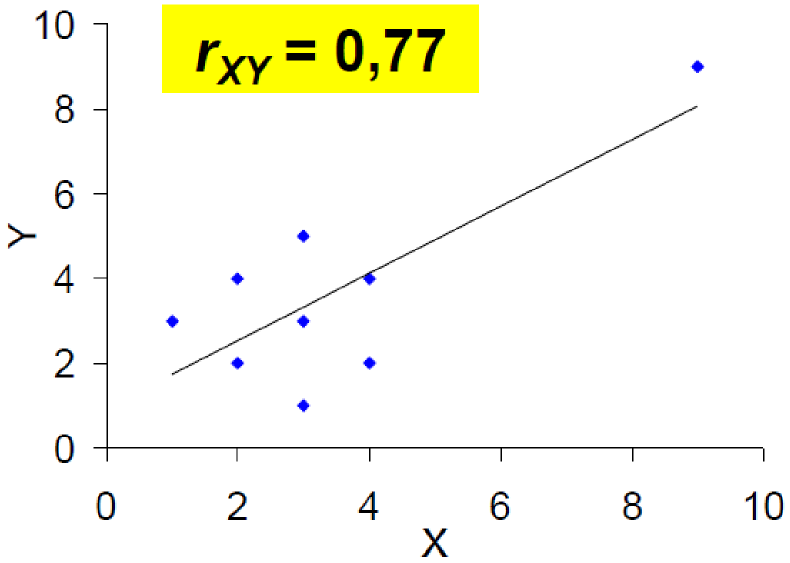
$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$\mathrm{r} = \frac{10{,}5}{10{,}5} = 1$$

Schritt 5: Berechnung Pearson‘s r durch Relativierung der empirischen Kovarianz am Produkt der Standardabweichungen (Maximale Kovarianz)

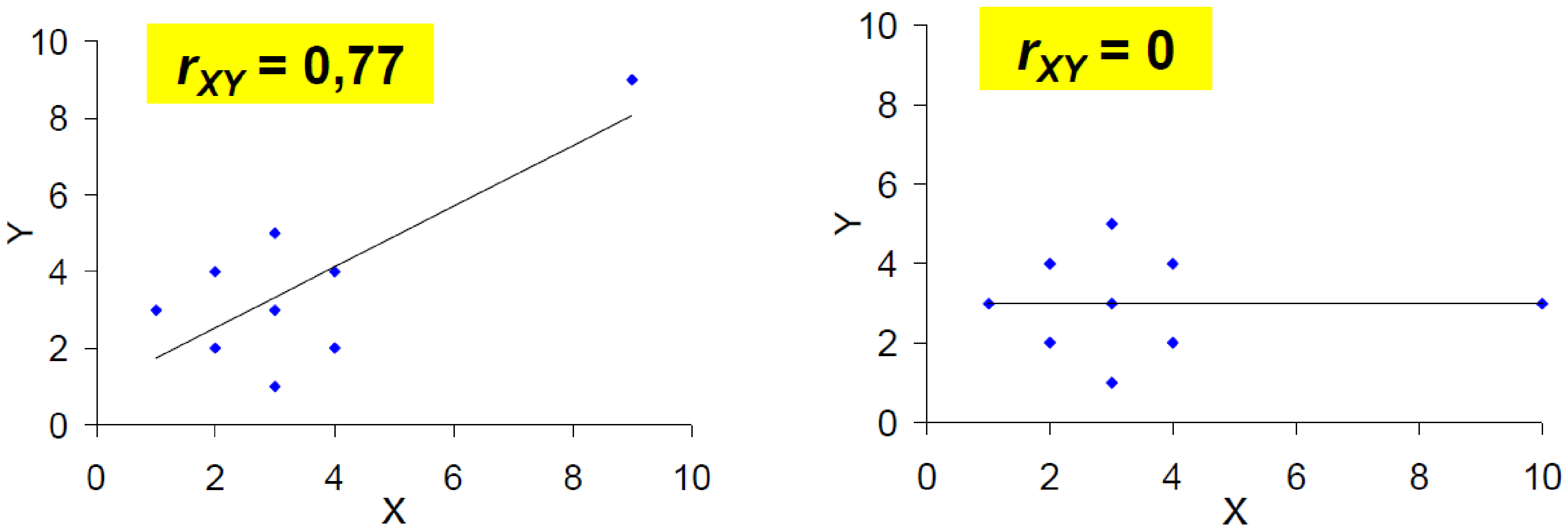
Pearson's r: Ergänzung

Korrelationskoeffizienten sind sensitiv gegenüber Ausreißern und Extremwerten, v.a. bei kleinen Stichproben

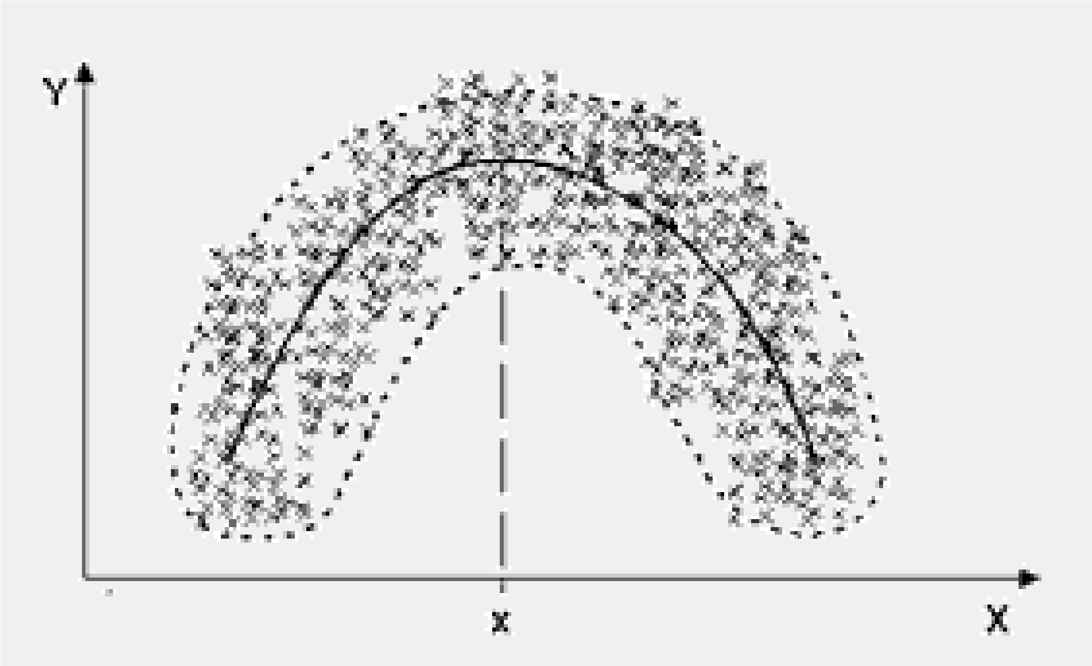


Pearson‘s r: Ergänzung

Korrelationskoeffizienten sind sensitiv gegenüber Ausreißern und Extremwerten, v.a. bei kleinen Stichproben



Keine lineare Korrelation! r=0 (Bsp. Leistungsfähigkeit und Anspannung während Klausur)



1) **Zeichnen Sie ein Streudiagramm**
2) **Berechnen Sie Pearson‘s r**
3) **Interpretieren Sie das Ergebnis**

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