Herzlich Willkommen zur Vorlesung Statistik

Thema dieser Vorlesung:
Kovarianz und Korrelation

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Kovarianz und Korrelation: Zusammenhänge zwischen metrischen (bzw. unter bestimmten Bedingungen: ordinalskalierten) Variablen

- Visualisierung
- Kovarianz
- Korrelation
- Korrelation bei ordinalskalierten Merkmalen
- Abschließendes

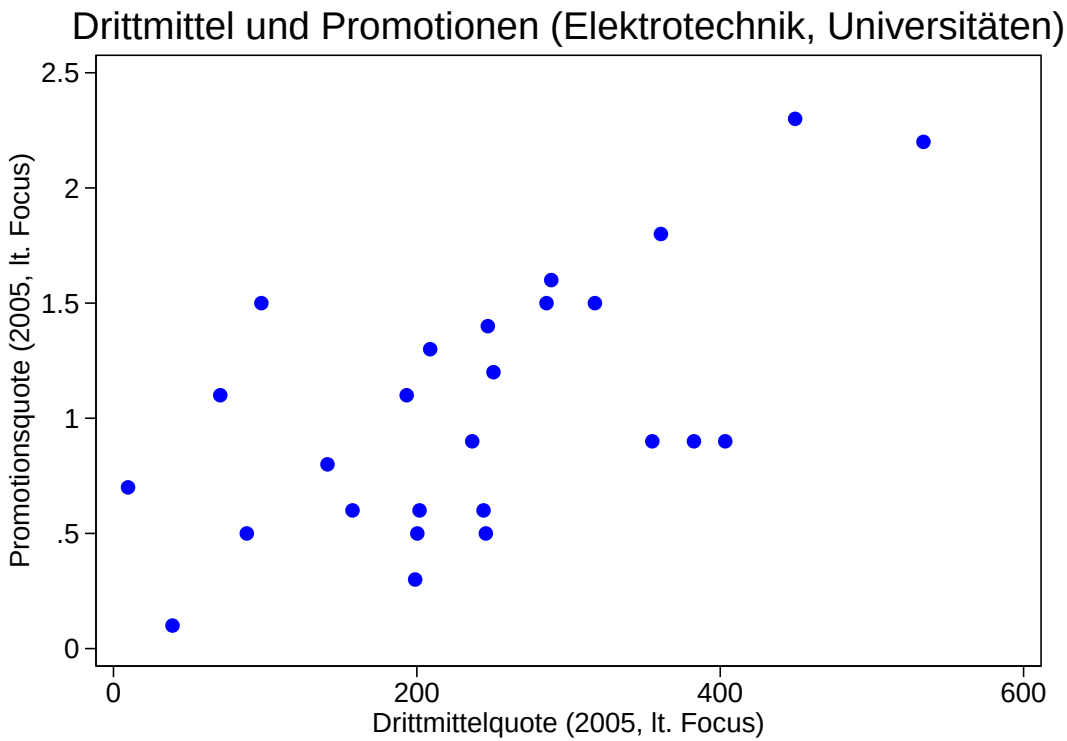
Einführung

- Kovarianz und die daraus abgeleitete Produkt-Moment-Korrelation sind Maßzahlen für den Zusammenhang zwischen zwei metrischen Variablen – oder solchen, die wir mit ein wenig schlechtem Gewissen als metrisch auffassen.
- Wir besprechen am Ende auch zwei Korrelationskoeffizienten für ordinalskalierte Merkmale.

  Der Begriff „Korrelation“ wird nicht ganz einheitlich gebraucht. Manchmal werden auch die in der Vorlesung zu Kreuztabellen erwähnten Assoziationsmaße für ordinalskalierte Variablen als Korrelation bezeichnet, manchal wird der Begriff allgemein im Sinne von „Zusammenhang“ verwendet – bitte jeweils auf den Kontext achten, in dem der Begriff auftaucht.
- Wichtig: Bei der Korrelationsanalyse wird *nicht* zwischen abhängiger und unabhängiger Variable unterschieden (selbst wenn die Forscherin diese Unterscheidung machen kann).

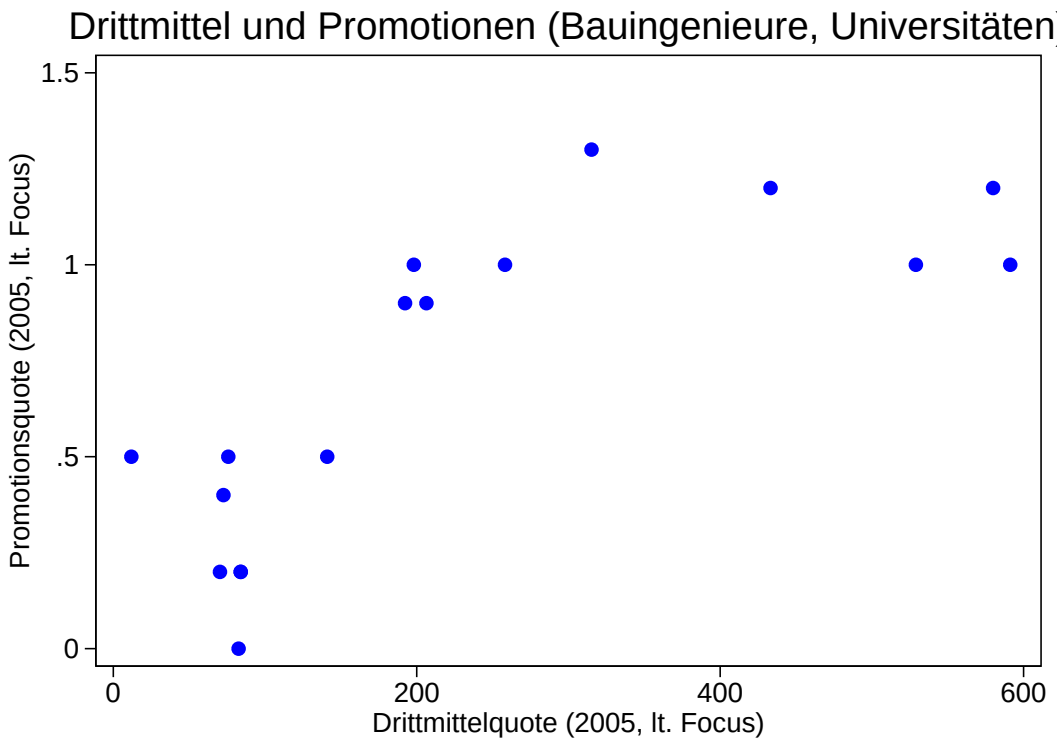
Streudiagramme

Zusammenhänge zwischen metrischen Variablen (bzw. Merkmalen mit vielen zumindest geordneten Ausprägungen) lassen sich manchmal gut anhand von Streudiagrammen visualisieren.



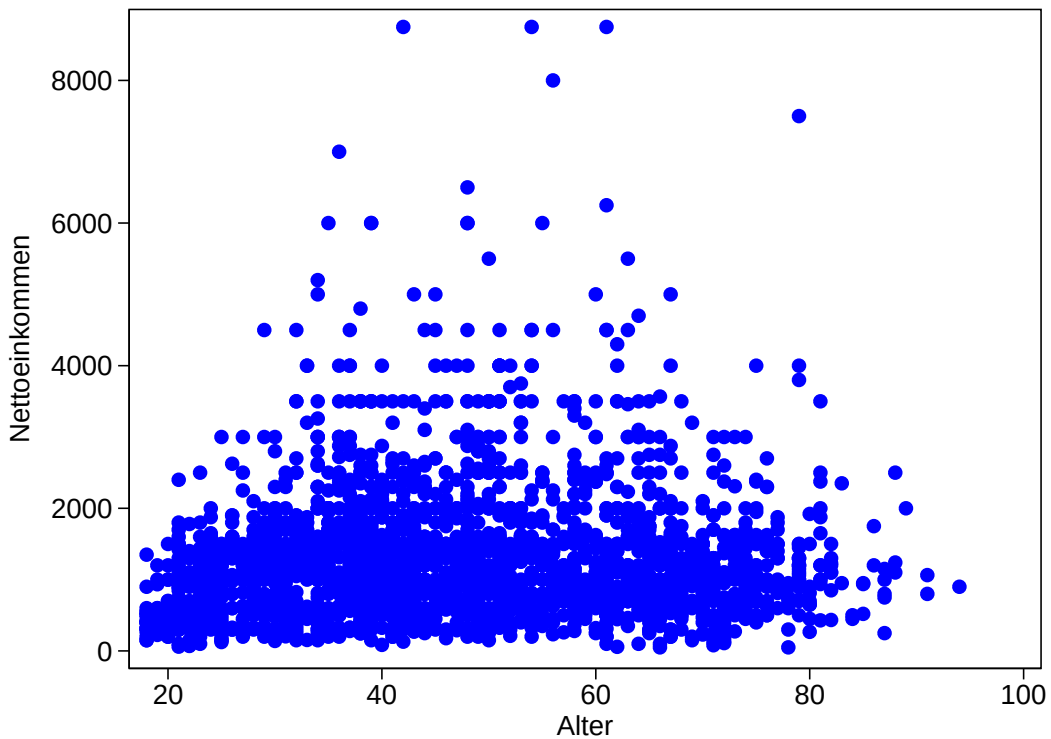
Streudiagramme

Die Visualisierung ist auch wichtig, um bspw. nicht-lineare Zusammenhänge etc. zu entdecken. Übrigens: Die unabhängige Variable (sofern vorhanden) sollte auf der x-Achse abgetragen werden.



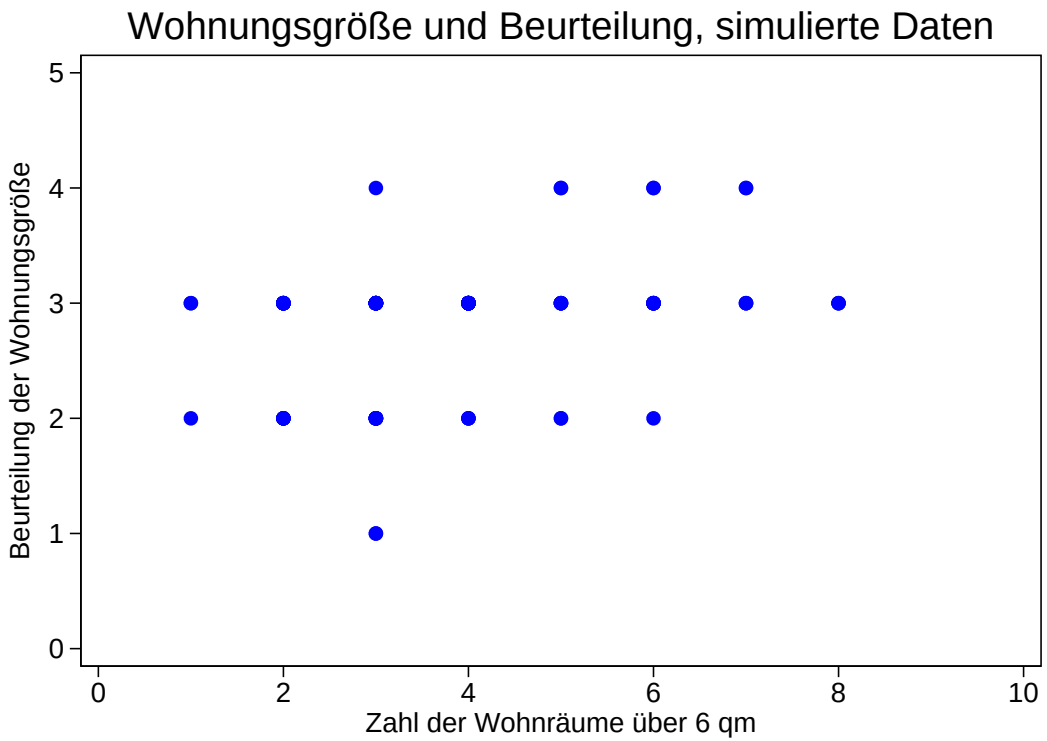
Streudiagramme

In großen Datensätzen sind Streudiagramme oft weniger informativ.



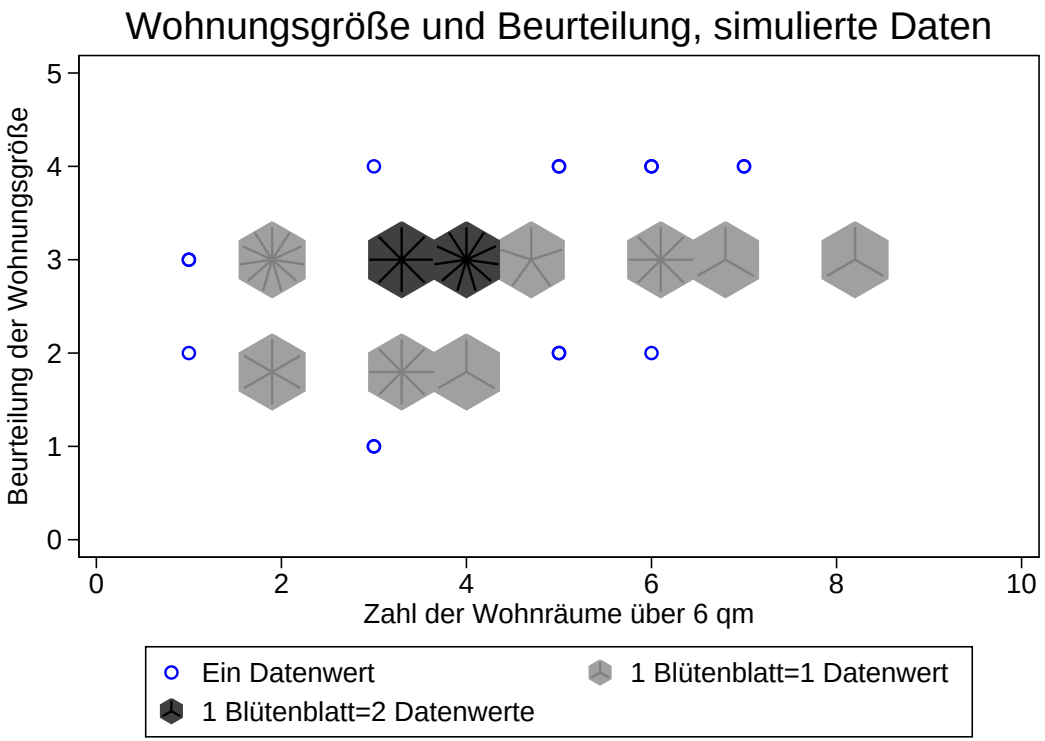
Streudiagramme: Überlappende Werte

Bei Merkmalen mit wenigen Ausprägungen entfallen auf einen Punkt oft mehrere Fälle.



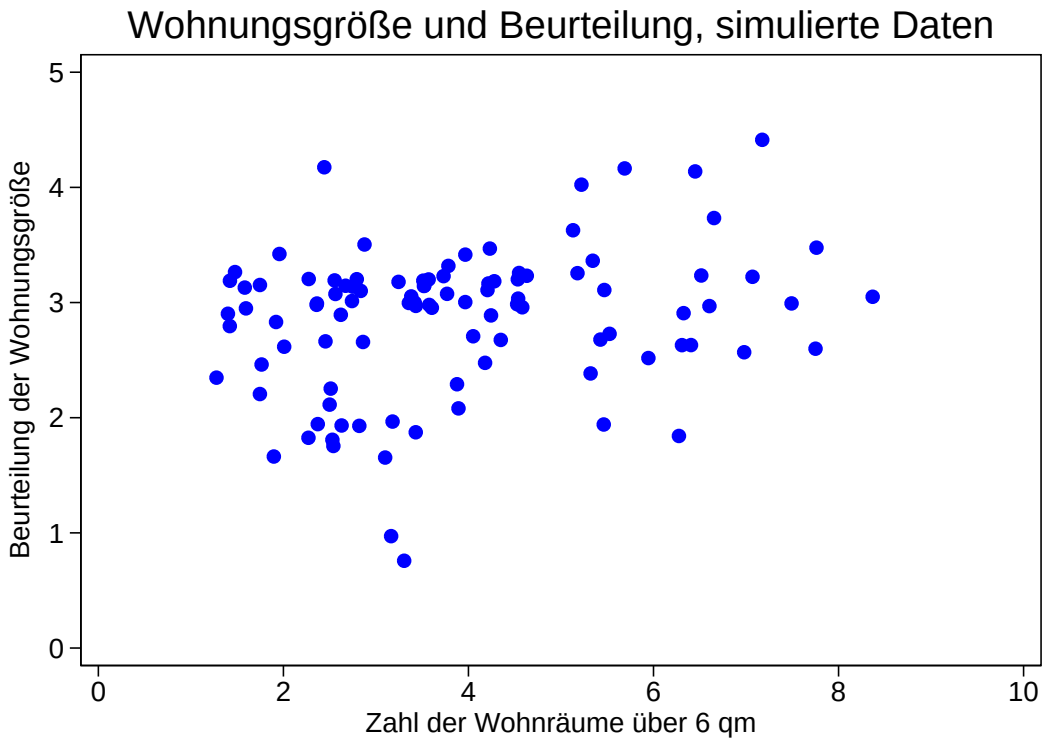
Überlappende Werte: Lösung I

Erste Möglichkeit, überlappende Werte zu visualisieren: das Sonnenblumendiagramm.



Überlappende Werte: Lösung II

Zweite Möglichkeit, überlappende Werte zu visualisieren: das „gejitterte“ Streudiagramm (Ausgangswerte werden mit Zufallszahlen überlagert).



Kovarianz: Einführung

Die Kovarianz ist eine Maßzahl für die „gemeinsame Varianz“ (im Sinne von: „miteinander Variieren“) zweier Variablen.

Grundsätzlich gilt:

- Positiver Zusammenhang: Hohe Werte in der einen Variablen treten tendenziell gemeinsam mit hohen Werten in der anderen Variablen auf, niedrige mit niedrigen $\rightarrow$ positives Vorzeichen.
- Negativer Zusammenhang: Hohe Werte in der einen Variablen treten tendenziell gemeinsam mit niedrigen Werten in der anderen Variablen auf, bzw. niedrige in der einen mit hohen in der anderen $\rightarrow$ negatives Vorzeichen.
- Null: Es besteht kein Zusammenhang zwischen den Werten der einen und denen der anderen Variablen.

Da der Betrag der Kovarianz vom Maßstab der untersuchten Variablen abhängt, informiert die Kovarianz nur über die Richtung des Zusammenhangs, nicht dessen Stärke.

Kovarianz: Die Formeln

Die Kovarianz, die einen gegebenen Datensatz beschreibt:

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

Die Kovarianz als Schätzwert für die Grundgesamtheit:

$$\widehat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Kovarianz: Ein Beispiel

Ein fiktives Beispiel: Bruttogehalt ($\bar{x} = 3000$) und Bildungsjahre ($\bar{y} = 12$)

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})$ $\cdot(y_i - \bar{y})$
	2000	9	-1000	-3	3000
	5000	16	2000	4	8000
	4000	16	1000	4	4000
	1500	9	-1500	-3	4500
	2500	10	-500	-2	1000
Summe	15000	60			20500

Kovarianz für die Daten: $\frac{1}{5} \cdot 20500 = 4100$

Kovarianz als Schätzung für GG: $\frac{1}{4} \cdot 20500 = 5125$

Kovarianz: Ein weiteres Beispiel

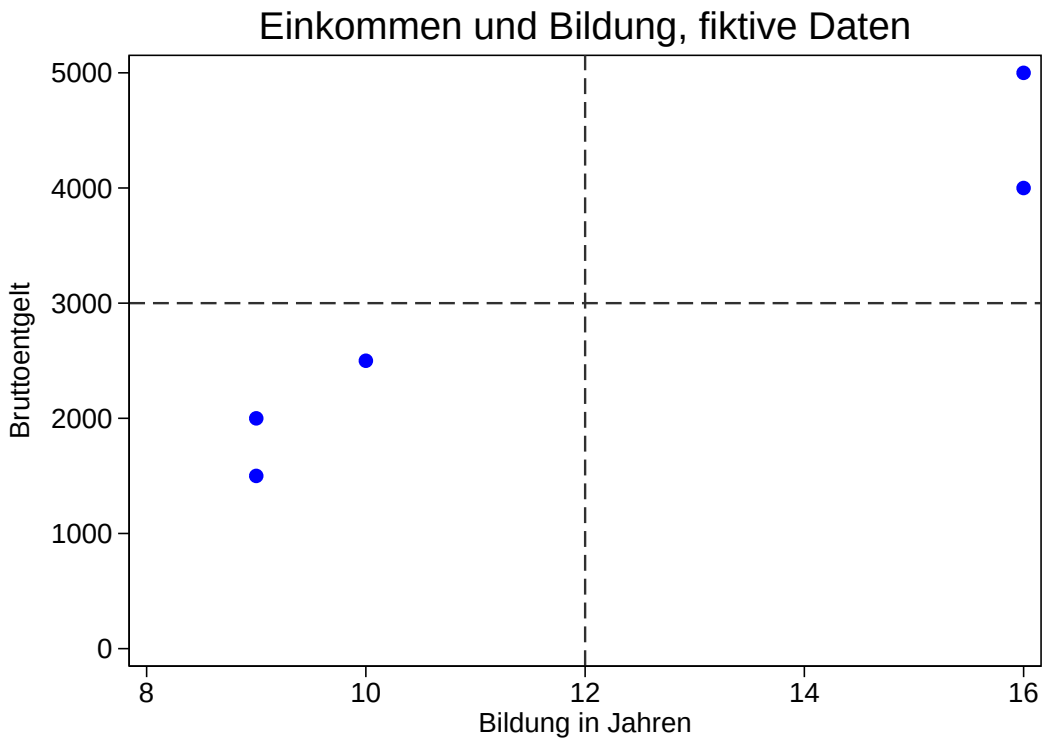
Ein fiktives Beispiel für einen Null-Zusammenhang: Bruttogehalt ($\bar{x} = 3000$) und Körpergröße ($\bar{y} = 1{,}72$)

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})$ $\cdot(y_i - \bar{y})$
	2000	1,55	−1000	−0,17	170
	5000	1,65	2000	−0,07	−140
	4000	1,80	1000	0,08	80
	1500	1,75	−1500	0,03	−45
	2500	1,85	−500	0,13	−65
Summe	15000	8,6			0

Die Kovarianz beträgt mithin 0!

Kovarianz visualisiert I

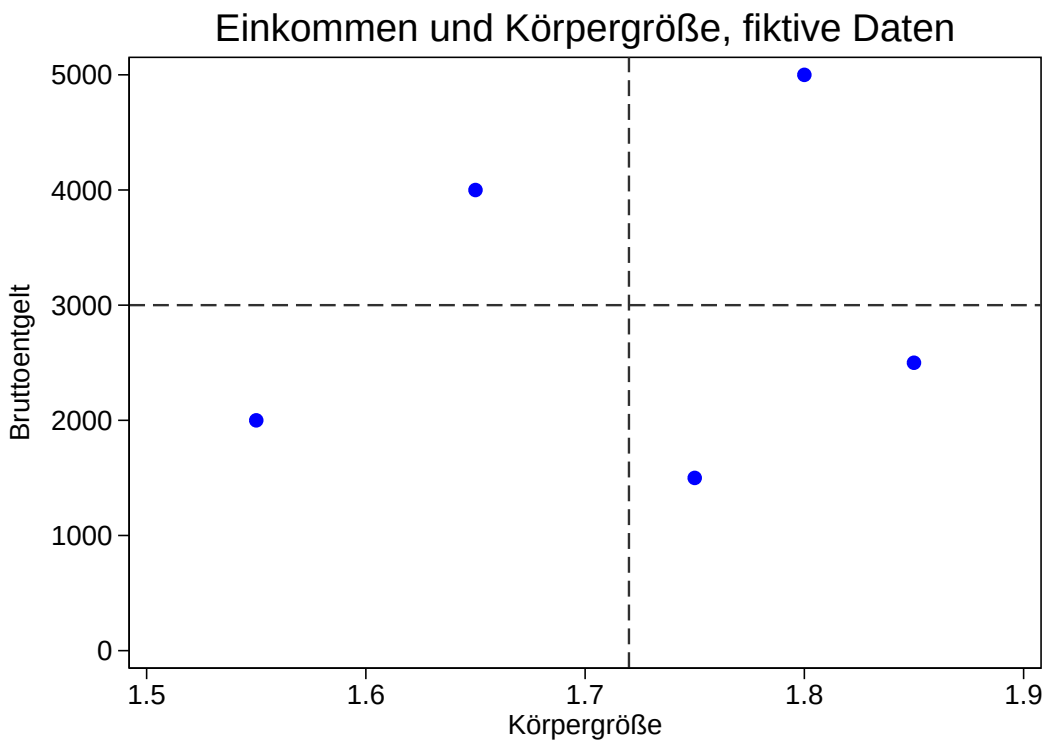
Bei einer positiven Kovarianz überwiegen die Datenwerte in den Quadranten links unten und rechts oben (bei einer negativen Kovarianz ist es genau umgekehrt).



Gestrichelte Linien: Mittelwerte

Kovarianz visualisiert II

Bei einer Kovarianz von Null heben sich die Werte in allen Quadranten wechselseitig auf.



Gestrichelte Linien: Mittelwerte

Korrelation: Die Formel

Der Korrelationskoeffizient (auch: [Bravais-]Pearsonscher Korrelationskoeffizient oder Produkt-Moment-Korrelation) ist die standardisierte Kovarianz:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$ [1]

r kann Werte zwischen -1 (perfekter negativer linearer Zusammenhang) und $+1$ (perfekter positiver linearer Zusammenhang annehmen.

[1] Das gleiche Ergebnis erhält man, wenn man statt s_{xy}, s_x und s_y die Größen $\widehat{\sigma}_{xy}$, $\widehat{\sigma}_x$ und $\widehat{\sigma}_y$ verwendet.

Korrelation im Beispiel

Bei unseren Beispieldaten beträgt die Korrelation

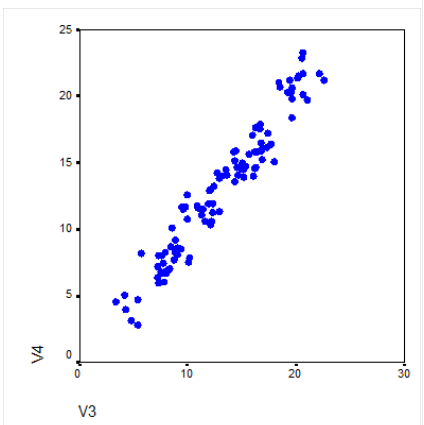
$$r = \frac{4100}{1303{,}84 \cdot 3{,}286} = 0{,}956858 \approx 0{,}96$$

Hinweise:

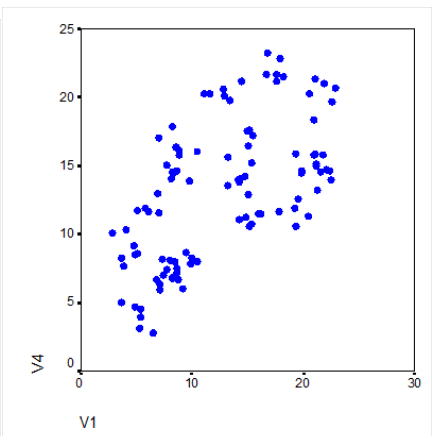
- So hohe Korrelationen treten im echten (sozialwissenschaftlichen) Leben fast nie auf! Man kann bereits ab $|0{,}3|$ von einer mäßigen, ab $|0{,}5|$ von einer starken Korrelation sprechen (Konvention).
- r wird in aller Regel auf nicht mehr als zwei Nachkommastellen genau angeben.
- r ist identisch mit der Kovarianz, wenn man diese aus den standardisierten Werten (z-Werten) der untersuchten Variablen errechnet.

Unterschiedlich starke Korrelationen I

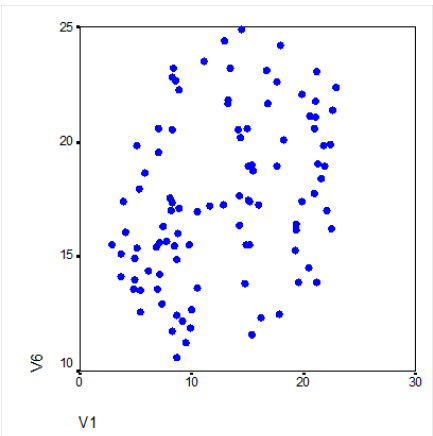
Die folgenden Graphiken zeigen positive Korrelationen unterschiedlicher Stärke:



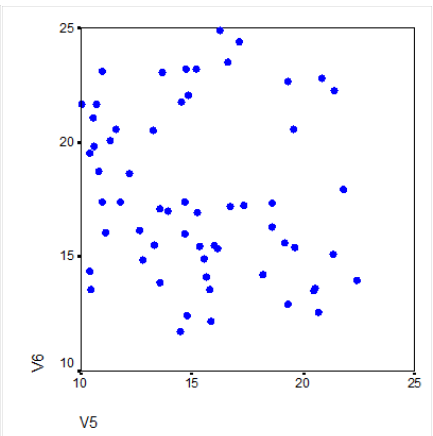
$r = 0{,}97$



$r = 0{,}60$



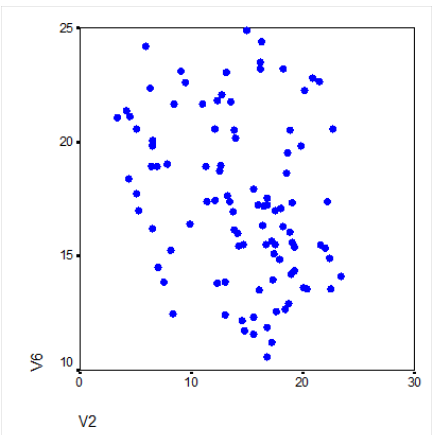
$r = 0{,}33$



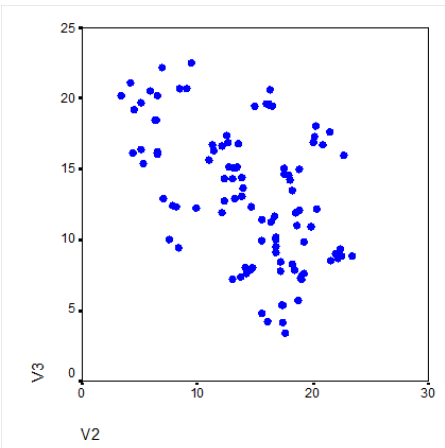
$r = 0{,}04$

Unterschiedlich starke Korrelationen II

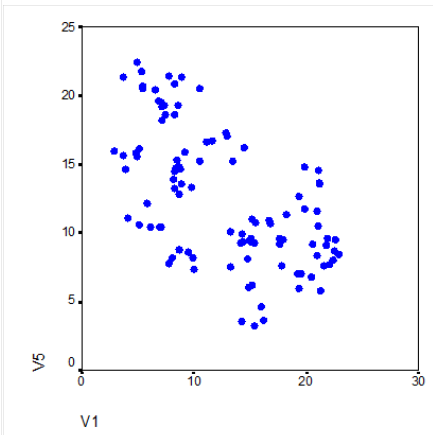
Die folgenden Graphiken zeigen negative Korrelationen unterschiedlicher Stärke:



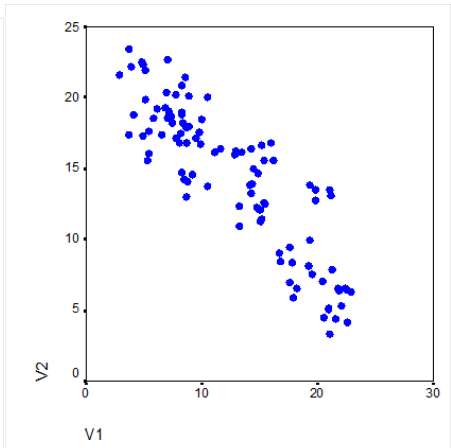
$r = -0{,}26$



$r = -0{,}49$



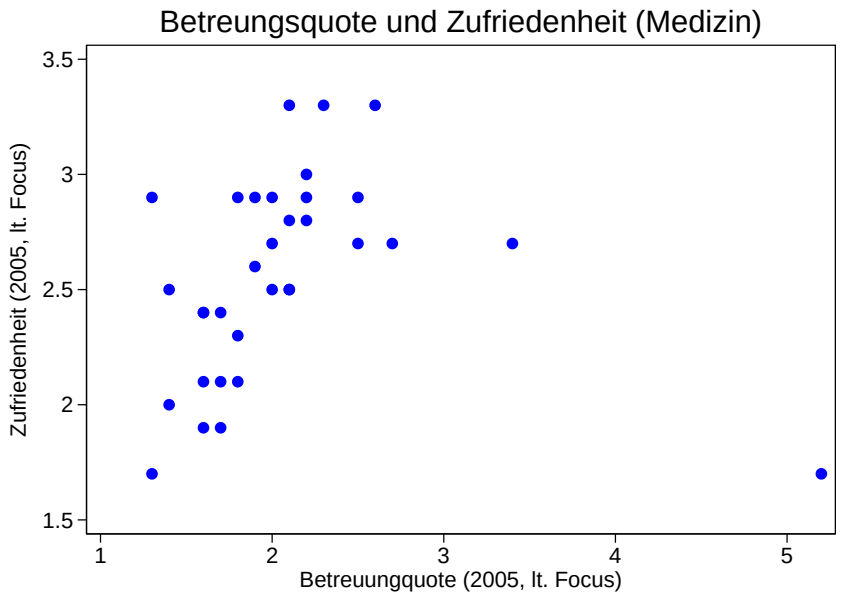
$r = -0{,}60$



$r = -0{,}87$

Probleme bei Korrelation I

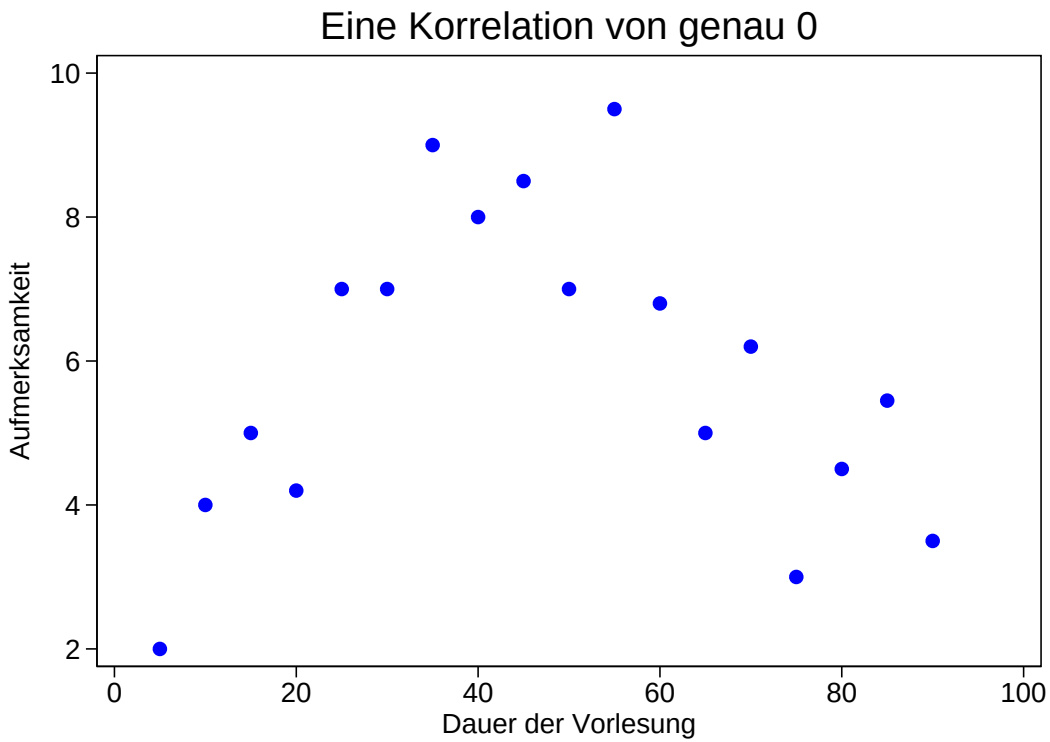
Einzelne Fälle können starken Einfluss ausüben (nicht zuletzt wegen Multiplikation der Abweichungen).



Die Korrelation für die vorliegenden Daten beträgt 0,05, ist also praktisch bedeutungslos. Schließt man jedoch den Wert rechts unten aus (Messfehler?), steigt der Korrelationskoeffizient auf 0,56!

Probleme bei Korrelation II

r berücksichtigt nur lineare Zusammenhänge. Der hier vorliegende starke nicht-lineare Zusammenhang schlägt sich im Korrelationskoeffizienten nicht nieder.



Der Zuammenhang ist rein fiktiv – natürlich sind Sie ständig aufmerksam ☺

Signifikanztest r, Teil I

Die Teststatistik

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

folgt einer t-Verteilung mit $n-2$ Freiheitsgraden. (Die t-Verteilung geht asymptotisch in die Standardnormalverteilung über; ab Fallzahlen von etwa 100 kann man getrost die SNV statt der t-Verteilung verwenden.)

Zum üblichen Signifikanzniveau von $\alpha = 0{,}05$ kann man also bei großen Fallzahlen folgende Ablehnungsbereiche festlegen:

$H_A : r \neq 0$ mit $H_0 : r = 0$: $t < -1{,}96$ und $t > 1{,}96$

$H_A : r > 0$ mit $H_0 : r \leq 0$: $t > 1{,}645$

$H_A : r < 0$ mit $H_0 : r \geq 0$: $t < -1{,}645$

Muss man wegen kleiner Fallzahlen die t-Verteilung heranziehen, werden die Absolutbeträge der kritischen Werte mit abnehmenden Freiheitsgraden immer größer. Beispielsweise lauten bei 20 Freiheitsgraden die kritischen Werte $|2{,}086|$ (statt $|1{,}96|$) und $|1{,}725|$ (statt $|1{,}645|$).

Signifikanztest für r, Teil II

Beispiel aus dem Lehrbuch von Fahrmeir et al. (Achtung: Vor 5. Aufl. dort Fehler in Formel und/oder Ergebnis): $r = 0{,}64$, $n = 20$:

$$t = 0{,}64\sqrt{\frac{20-2}{1-0{,}4096}} = 3{,}5338$$

Der Wert liegt weit außerhalb des 95-%-Intervalls von $-2{,}1$ bis $2{,}1$ (t-Verteilung, 18 Freiheitsgrade), er liegt also im Ablehnungsbereich für die H_0 $r = 0$ bzw. $r \leq 0$.

Rangkorrelation nach Spearman

Als Maß für die Stärke des Zusammenhangs zwischen zwei ordinalskalierten Merkmalen mit vielen Ausprägungen (typischerweise: echte Ränge!) steht der Korrelationskoeffizient nach Spearman (r_{SP}, oft auch als ρ (rho) bezeichnet), zur Verfügung.

Die transparenteste Formel hierfür ist:

$$r_{SP} = \frac{\sum_{i=1}^{n} \left(rg(x_i) - \bar{rg}_x\right) \cdot \left(rg(y_i) - \bar{rg}_y\right)}{\sqrt{\sum_{i=1}^{n} \left(rg(x_i) - \bar{rg}_x\right)^2 \cdot \sum_{i=1}^{n} \left(rg(y_i) - \bar{rg}_y\right)^2}}$$

mit $rg(x_i)$ bzw. $rg(y_i)$ als dem jeweiligen Rangplatz des betreffenden Messwertes in der geordneten Datenreihe und $\bar{rg}_x = \bar{rg}_y = (n+1)/2$

Rangkorrelation nach Spearman: Beispiel

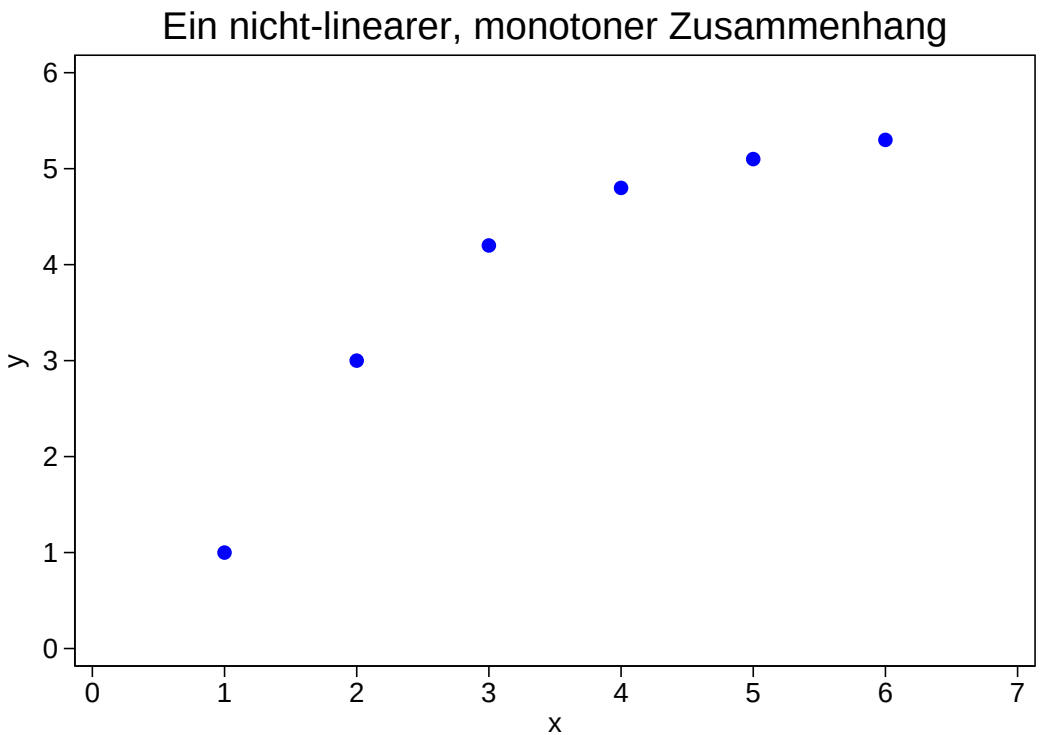
Ein Beispiel (nach Clauß, Günter et al.: Statistik, Frankfurt a. M., 2002, S. 68): Rangplätze nach Leistung und nach Sympathie (n=6, damit $\bar{rg}_x = \bar{rg}_y = 3{,}5$)

$rg(x)$	$rg(y)$	$rg(x)-\bar{rg}_x$	$rg(y)-\bar{rg}_y$	$(rg(x)-\bar{rg}_x)$ $\cdot(rg(y)-\bar{rg}_y)$	$(rg(x)-\bar{rg}_x)^2$	$(rg(y)-\bar{rg}_y)^2$
1	2	-2,5	-1,5	3,75	6,25	2,25
2	3	-1,5	-0,5	0,75	2,25	0,25
3	1	-0,5	-2,5	1,25	0,25	6,25
4	4	0,5	0,5	0,25	0,25	0,25
5	6	1,5	2,5	3,75	2,25	6,25
6	5	2,5	1,5	3,75	6,25	2,25

$$r_{SP} = \frac{13{,}5}{\sqrt{17{,}5 \cdot 17{,}5}} = 0{,}77$$

Mehr zur Rangkorrelation nach Spearman I

Liegen metrische Daten vor, können diese im Prinzip auch in Rangplätze verwandelt werden (wenn nur diese interessieren). In diesem Fall führt bereits ein perfekt monotoner (auch nicht-linearer) Zusammenhang zu $r_{SP} = 1$ bzw. (bei negativem Zusammenhang) $r_{SP} = -1$ (siehe Beispiel in der Graphik).



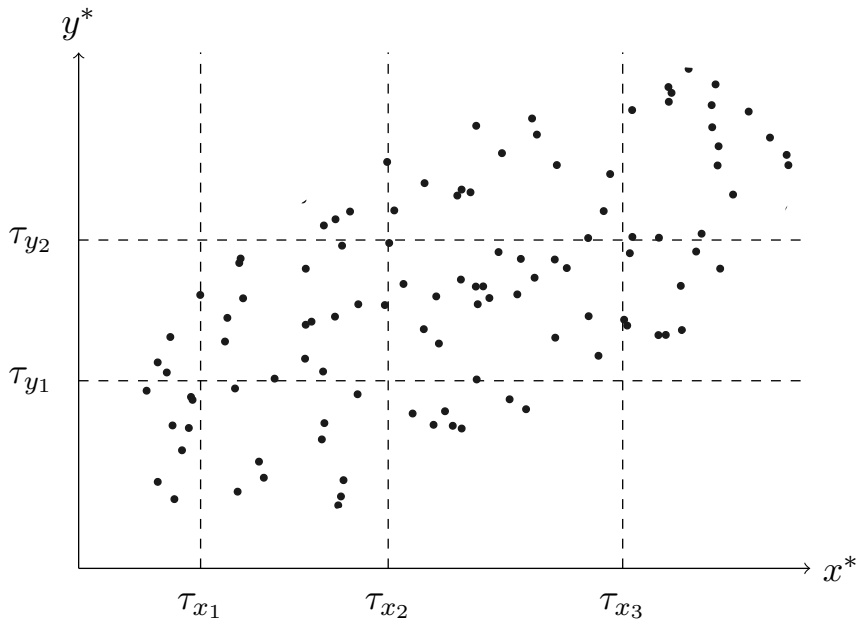
Mehr zur Rangkorrelation nach Spearman II

Was man noch wissen sollte:

- Haben zwei (oder mehr) Fälle den gleichen Messwert, so wird (wie oft beim Sport) allen der gleiche (mittlere) Rangplatz zugewiesen (z. B. wenn der dritte und der vierte Fall den gleichen Messwert haben, erhalten sie beide Rang 3,5).
- Liegen nicht ganz wenige solcher „Bindungen“ (englisch: *ties*) vor, muss eine Formel herangezogen werden, die hierfür korrigiert. Bei sehr vielen Bindungen ist von r_{SP} abzuraten.
- In der Literatur findet man auch andere Formeln für r_{SP}, die oft der Rechenvereinfachung dienen.
- Inferenzstatistik: Für r_{SP} gilt der gleiche Signifikanztest wie für r (wiederum: solange nicht zu viele Bindungen vorliegen)!

Die polychorische Korrelation I

Liegen zwei ordinalskalierte Merkmale mit jeweils wenigen Ausprägungen vor, kann der sog. polychorische Korrelationskoeffizient berechnet werden, sofern angenommen werden kann, dass hinter den manifesten Messwerten (X, Y) normalverteilte metrische Merkmale (X*, Y*) liegen.



Annahme: Der niedrigste gemessene Wert von X* ist 0; er entspricht den niedrigsten Werten der latenten metrischen Variablen. Überschreitet diese einen Schwellenwert τ_{x_1}, wird der Wert 1 angegeben, bei Überschreiten von τ_{x_2} der Wert 2, bei Überschreiten von τ_{x_3} der Wert 3. Analog für Y*.

Die polychorische Korrelation II

Beispiel: Zusammenhang zwischen Einschätzung der persönlichen und der der allgemeinen wirtschaftlichen Lage im ALLBUS (sehr gut/gut; teils-teils; schlecht/sehr schlecht)(absolute Zahlen):

	persönlich			
allgemein	1	2	3	n
1	478	148	59	685
2	561	465	145	1171
3	89	199	219	507
n	1128	812	423	2363

$r_{SP} = 0{,}39$ $r_{polychorisch} = 0{,}49$

Leider beruht die Berechnung von $r_{polychorisch}$ auf iterativen Schätzverfahren; eine einfache Formel kann nicht angegeben werden.

Abschließendes zum Thema Korrelation

In manchen Lehrbüchern kann man noch diverse weitere Korrelationskoeffizienten finden (etwa punktbiseriale, biseriale, biseriale Rang-, tetrachorische ... Korrelation). In der sozialwissenschaftlichen Forschungspraxis werden diese relativ selten benötigt und werden daher an dieser Stelle nicht weiter vertieft.