6 Korrelationsanalyse: Zusammenhangsanalyse stetiger Merkmale

6.1 Korrelationsanalyse

Jetzt betrachten wir bivariate Merkmale (X, Y), wobei sowohl X als auch Y stetig bzw. quasi-stetig und mindestens ordinalskaliert, typischerweise sogar intervallskaliert, sind. Am Rande wird auch der Fall gestreift, dass nur ein Merkmal quasi-stetig und das andere nominalskaliert ist.

6.1.1 Streudiagramm, Kovarianz- und Korrelationskoeffizienten

Beispiele:

- Nettomiete $\longleftrightarrow$ Wohnfläche
- Autoritarismusscore vor/nach einer Informationsveranstaltung
- Monatseinkommen $\longleftrightarrow$ Alter in Jahren
- Wochenarbeitseinkommen $\longleftrightarrow$ Wochenarbeitsstunden
- Wochenarbeitsstunden $\longleftrightarrow$ Hausarbeit in Stunden pro Woche
- Wochenarbeitsstunden (tatsächlich) $\longleftrightarrow$ Wochenarbeit (vertraglich)

6.1.2 Streudiagramme (Scatterplots)

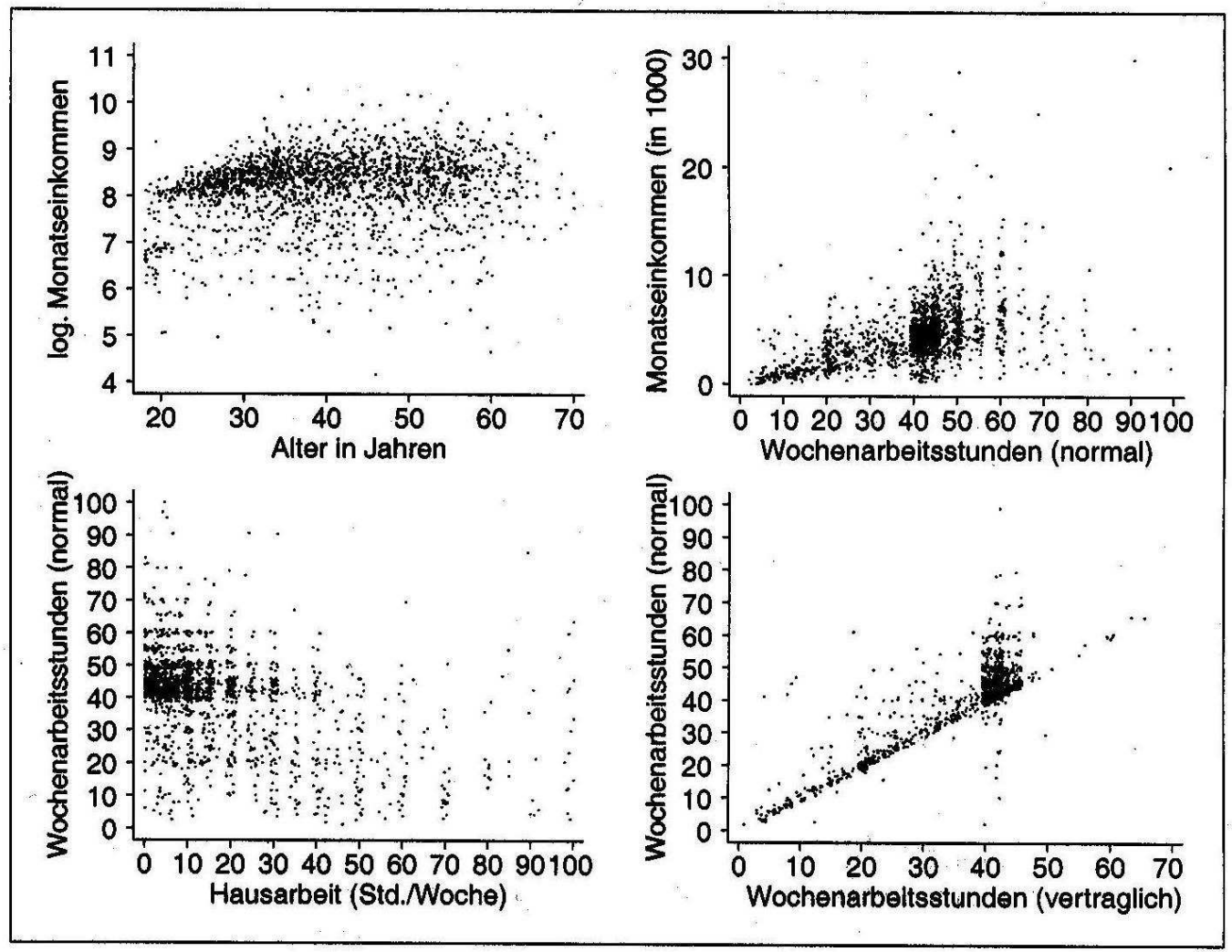
Sind die Merkmale stetig oder zumindestens quasi-stetig (sehr viele verschiedene Ausprägungen), werden Kontingenztabellen sehr unübersichtlich und praktisch aussagelos, da die einzelnen Häufigkeiten in den Zellen der Tabellen natürlicherweise durchwegs sehr klein sind.

Alternative Darstellungsform: $Scatterplot$ / $Streudiagramm$:

Zeichne die Punkte (x_i, y_i), $i = 1, \ldots, n$, in ein X-Y-Koordinatensystem.

$\Longrightarrow$ Guter optischer Eindruck über das Vorliegen, die Richtung und gegebenenfalls die Art eines Zusammenhangs.

$\Longrightarrow$ Ausreißer werden leicht erkannt.



Quelle für Beispiele: Jann (2002), p. 85 ff.

6.1.3 Kovarianz und Korrelation

Wie misst man den Zusammenhang zwischen metrischen Merkmalen?

- Eine Idee die sogenannte Kovarianz (s.u.) zu konstruieren besteht darin nach Konkordanz/Diskordanz zum Schwerpunkt zu fragen und dabei auch die nun interpretierbaren Abstände zur Messung der „individuellen Konkordanzstärke“ heranzuziehen. Negative Werte sprechen für Diskordanz.
- Betrachte den „Mittelpunkt“ der Daten $(\bar{x}, \bar{y})$ und dazu konkordante/diskordante Paare.

- Eine Beobachtung i mit Ausprägung (x_i, y_i) ist
  - *konkordant* zu $(\bar{x}, \bar{y})$, spricht also für einen gleichgerichteten Zusammenhang, wenn

$$(x_i > \bar{x} \text{ und } y_i > \bar{y}) \text{ oder } (x_i < \bar{x} \text{ und } y_i < \bar{y})$$

    also zusammengefasst wenn

$$(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0.$$

  - *diskordant* zu $(\bar{x}, \bar{y})$, spricht also für einen gegengerichteten Zusammenhang, wenn

$$(x_i < \bar{x} \text{ und } y_i > \bar{y}) \text{ oder } (x_i > \bar{x} \text{ und } y_i < \bar{y})$$

    also zusammengefasst wenn

$$(x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0.$$

- Wegen des metrischen Skalenniveaus sind auch die Abstände interpretierbar, das Produkt $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ gibt also sozusagen die Stärke der Konkordanz bzw. Diskordanz an.

- $(x_i - \bar{x})(y_i - \bar{y})$ ist positiv, wenn große (kleine) X-Werte mit großen (kleinen) Y-Werten einhergehen (gleichgerichteter Zusammenhang).

- $(x_i - \bar{x})(y_i - \bar{y})$ ist negativ, wenn große (kleine) X-Werte mit kleinen (großen) Y-Werten einhergehen (gegengerichteter Zusammenhang).

$\Longrightarrow$ Definiere als Zusammenhangsmaß die durchschnittliche individuelle Konkordanzstärke.

Definition: Gegeben sei ein bivariates Merkmal (X, Y) mit metrisch skalierten Variablen X und Y mit $\tilde{s}_X^2 > 0$ und $\tilde{s}_Y^2 > 0$. Dann heißen

$$\mathrm{Cov}(X, Y) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

(empirische) Kovarianz von X und Y,

$$\varrho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\tilde{s}_Y^2 \tilde{s}_X^2}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

(empirischer) Korrelationskoeffizient nach Bravais und Pearson von X und Y, und

$$R_{XY}^2 := (\varrho(X, Y))^2 \tag{6.1}$$

Bestimmtheitsmaß von X und Y.

Bemerkungen:

- Die Kovarianz $\mathrm{Cov}(X, Y)$ ist nicht maßstabsunabhängig.

- Das Teilen durch die Standardabweichungen normiert die Kovarianz und macht sie maßstabsunabhängig.

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sqrt{\tilde{s}_X^2}} \cdot \frac{(y_i - \bar{y})}{\sqrt{\tilde{s}_Y^2}} = \varrho(X, Y)$$

  Also ist - im Sinne obiger Interpretation - der Korrelationskoeffizient die *durchschnittliche standardisierte* Konkordanzstärke.

- Die empirische Kovarianz ist eine Verallgemeinerung der empirischen Varianz. Die Kovarianz eines Merkmals mit sich selbst ist genau die empirische Varianz:

$$\begin{aligned} \mathrm{Cov}(X,X) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,(x_i-\bar{x}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \tilde{s}_x^2 \end{aligned}$$

- Man sieht hier auch, dass die Größe der Kovarianz für sich genommen unanschaulich zu interpretieren ist. Für den Korrelationskoeffizienten hingegen gilt:

$$-1 \leq \varrho(X,Y) \leq 1.$$

  und insbesondere $\varrho(X,X) = 1$.

- Viele der (un)angenehmen Eigenschaften der Varianz (z.B. Ausreißerempfindlichkeit) gelten in analoger Weise.

- Es gilt auch ein Verschiebungssatz:

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \; - \; \bar{x}\bar{y}$$

und damit

$$\varrho(X,Y) = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\displaystyle\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}\cdot\sqrt{\displaystyle\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2}}.$$

Zur Erinnerung:

$$\tilde{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Beispiel: Zunächst inhaltsleere Zahlenbeispiele, zur Interpretation später.

- Gegeben seien die Datenpaare

x_i	37	30	20	28	35
y_i	130	112	108	114	136

Es gilt: $\bar{x} = 30$ und $\bar{y} = 120$, sowie

$$\sum_{i=1}^{n} x_i^2 = 4678 \qquad \sum_{i=1}^{n} y_i^2 = 72600$$

$$\sum_{i=1}^{n} x_i y_i = 18282$$

$$n = 5$$

Tabelle:

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	37	130			
	30	112			
	20	108			
	28	114			
	35	136			
$\sum$					

Basierend auf diesen Hilfsgrößen berechnet sich der Korrelationskoeffizient gemäß Verschiebungssatz als

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

- Gegeben sei ein Merkmal X und das Merkmal $Y = (X - 20)^2$ mit den Datenpaaren.

x_i	10	20	30
y_i	100	0	100
$x_i y_i$	1000	0	3000

Es gilt: $\bar{x} = 20$ und $\bar{y} = \frac{200}{3}$ und damit

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \frac{1}{n} \sum x_i y_i - \bar{x}\bar{y} \\ &= \frac{1}{3}(1000 + 0 + 3000) - 20 \cdot \frac{200}{3} \\ &= \frac{4000}{3} - \frac{4000}{3} = 0 \end{aligned}$$

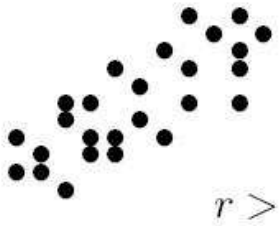
Für den Korrelationskoeffizienten ergibt sich damit ebenfalls $\varrho(X, Y) = 0$!

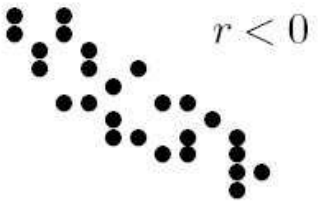
Bemerkungen:

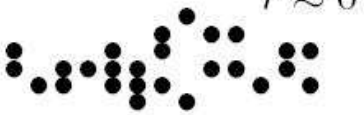
- Es gilt $|\varrho| = 1$ genau dann wenn $Y = aX + b$ mit $a \neq 0$, d.h. X und Y stehen in einem perfekten linearen Zusammenhang.

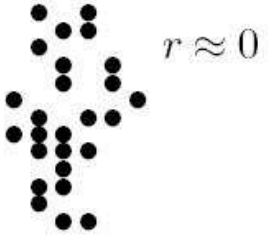
- Ist $\varrho = 0$ (und äquivalent dazu $\mathrm{Cov}(X, Y) = 0$), so nennt man X und Y *unkorreliert*. Es besteht dann keinerlei linearer Zusammenhang.

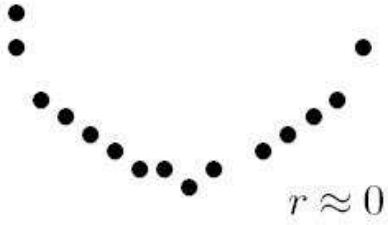
- Die Betonung der *Linearität* des Zusammenhangs ist wesentlich.

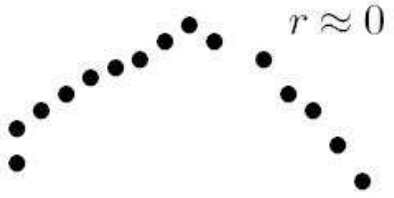

$r > 0$


$r < 0$

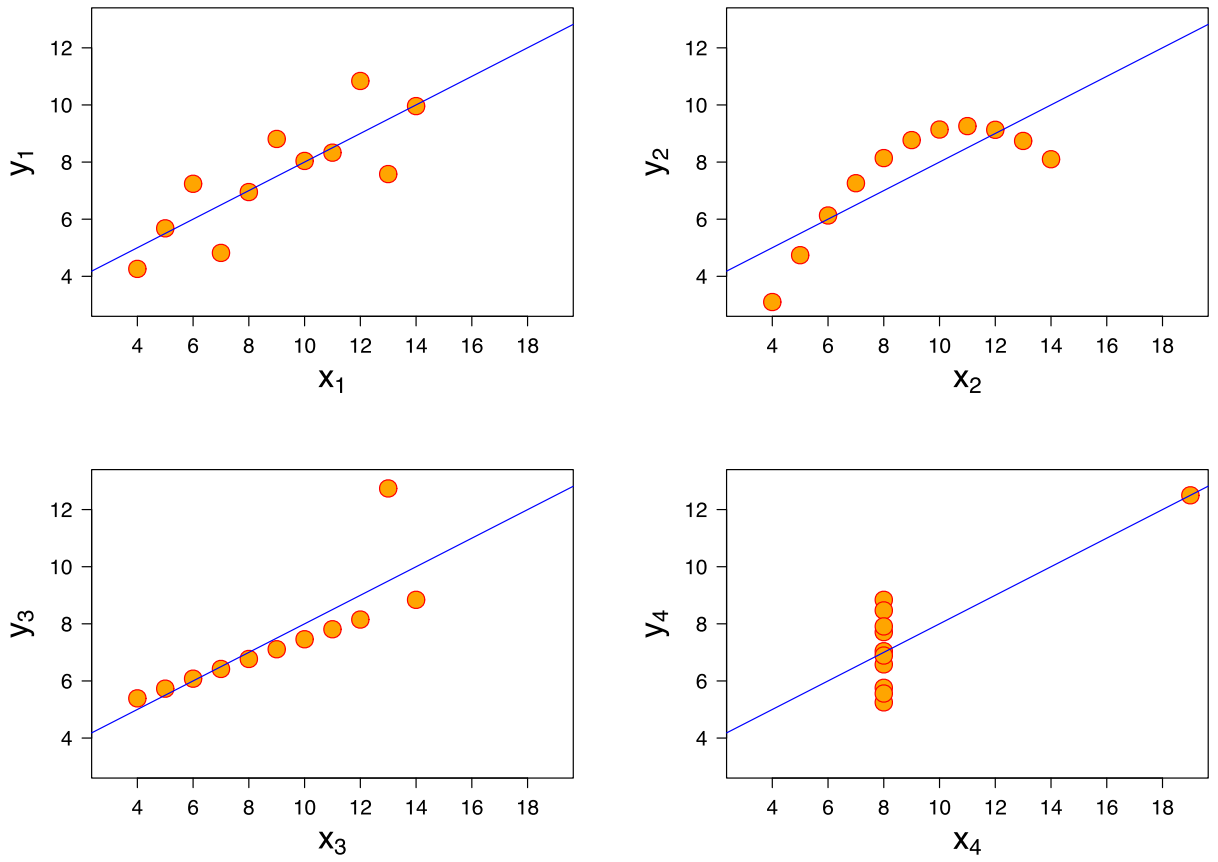

$r \approx 0$


$r \approx 0$


$r \approx 0$


$r \approx 0$

Beispiel: Anscombe's Quartet:



(Quelle: Wikipedia; Anscombe's quartet)

- Allgemein zeigt $|\varrho|$ und R^2 die Stärke eines $linearen$ Zusammenhangs an, also wie gut sich die Datenpaare $(x_1, y_1), \ldots, (x_n, y_n)$ durch eine $Gerade$ beschreiben lassen.
- R^2 ist ein PRE-Maß, das misst, welchen Anteil der gesamten Variation sich durch einen linearen Zusammenhang beschreiben lässt. (Näheres dazu im Abschnitt über die Regression.)
- Gelegentlich wird der Wert des Korrelationskoeffizienten folgendermaßen schematisch interpretiert:
  - $\varrho_{XY} \approx 0$: kein (linearer) Zusammenhang.
  - $\varrho_{XY} > 0$: positive Korrelation, gleichgerichteter (linearer) Zusammenhang.
  - $\varrho_{XY} < 0$: negative Korrelation, gegengerichteter (linearer) Zusammenhang.
  - $|\varrho_{XY}| \leq 0.5$: schwache Korrelation.
  - $0.5 < |\varrho_{XY}| \leq 0.8$: mittlere Korrelation.
  - $|\varrho_{XY}| > 0.8$: starke Korrelation.

- Die Zusammenhangsmaße sind invariant gegenüber Vertauschen von Y und X, unterscheiden also nicht welche Variable als abhängige, wele als unabhängige gilt:

$$\varrho(X,Y) = \varrho(Y,X) \qquad R_{XY} = R_{YX}.$$

- Im Gegensatz zur Kovarianz sind $\varrho(X,Y)$ und R^2_{XY} invariant gegenüber streng monoton steigenden linearen Transformationen. Genauer gilt mit $\tilde{X} := a \cdot X + b$ und $\tilde{Y} := c \cdot Y + d$

$$\varrho(\tilde{X},\tilde{Y}) = \varrho(X,Y)$$

falls $a \cdot c > 0$ und

$$\varrho(\tilde{X},\tilde{Y}) = -\varrho(X,Y)$$

falls $a \cdot c < 0$. Die Korrelation ist also in der Tat maßstabsunabhängig.

Beispiel: Mietspiegel (SPSS-Ausdruck)

Korrelationen

		Nettomiete	Wohnfläche	Baujahr
Nettomiete	Korrelation nach Pearson	1	.600	.223
	Signifikanz (2-seitig)		.000	.006
	N	150	150	150
Wohnfläche	Korrelation nach Pearson	.600	1	-.174
	Signifikanz (2-seitig)	.000		.033
	N	150	150	150
Baujahr	Korrelation nach Pearson	.223	-.174	1
	Signifikanz (2-seitig)	.006	.033	
	N	150	150	150

Zur Interpretation der einzelnen Zellen: In der entsprechenden Zelle stehen Informationen zur Korrelation der Variablen, in der entsprechenden Zeile mit der Variable der jeweiligen entsprechenden Spalte. In der ersten Zeile stehen jeweils die Korrelationskoeffizienten, N ist der Stichprobenumfang, der bei uns mit n bezeichnet wird.
Die zweite Zeile „Signifikanz“ wird erst in Statistik II verständlich. Grob gesprochen gilt: Je kleiner diese Zahl ist, desto sicherer ist man sich, dass der errechnete Korrelationskoeffizient nicht nur zufällig von 0 abweicht.

Beispiele aus Jann (2002) S.87ff

- Arbeitsstunden und Erwerbseinkommen: 0.495
  moderater positiver Zusammenhang.
- Arbeitsstunden und Haushalt: -0.434
  moderater negativer Zusammenhang.
- Vertragliche und geleistete Wochenarbeitsstunden: 0.868
  hoch positiv korreliert (Punkte liegen sehr nahe an „bester Gerade“).

6.1.4 Weitere Korrelationskoeffizienten

Anwendung des Korrelationskoeffizienten nach Bravais-Pearson auf dichotome nominale Merkmale

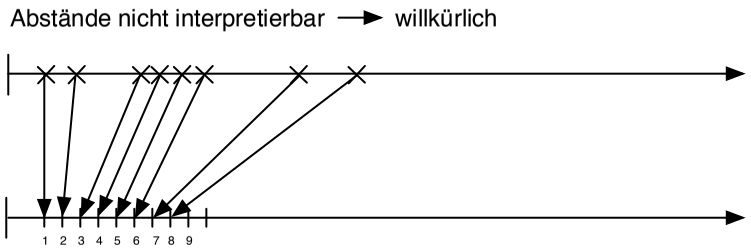
Liegen $dichotome$ nominale Merkmale, d.h. Merkmale mit nur zwei ungeordneten Ausprägungen vor (z.B. ja/nein), und kodiert man die Ausprägung mit 0 und 1, so kann man die Formel des Korrelationskoeffizienten nach Bravais-Pearson sinnvoll anwenden. Man erhält den sogenannten $Punkt\text{-}Korrelationskoeffizienten$, der identisch zu Φ_s aus ($\rightarrow$ Kapitel 5.3) ist.

Im Fall einer dichotomen und einer metrischen Variablen ergibt sich bei Anwendung des Korrelationskoeffizienten nach Bravais-Pearson die sogenannte $Punkt\text{-}biseriale\ Korrelation$. (vgl. etwa Jann (2002, S.90f) oder Wagschal (1999, Kap 10.8).)

Rangkorrelationskoeffizient nach Spearman

- Wir betrachten ein bivariates Merkmal (X, Y), wobei X und Y nur ordinalskaliert sind, aber viele unterschiedliche Ausprägungen besitzen.
- Der Korrelationskoeffizient von Bravais-Pearson darf nicht verwendet werden, da hier die Abstände nicht interpretierbar sind. $(\bar{x}, \bar{y})$ wären willkürliche Zahlen, ebenso $(x_i - \bar{x}), (y_i - \bar{y})$.

  Beispiel



- Liegen keine Bindungen vor, dann rechnet man statt mit $(x_i, y_i)_{i=1,\ldots,n}$ mit $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$ $i = 1, \ldots, n$. Dabei ist

$$\mathrm{rg}(x_i) = j :\Longleftrightarrow x_i = x_{(j)},$$

d.h. der Rang $\mathrm{rg}(x_i)$ ist die Nummer, die x_i in der geordneten Urliste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ einnimmt (analog für $\mathrm{rg}(y_i)$). Der kleinsten Beobachtung wird also der Wert 1 zugeordnet, der zweitkleinsten der Wert 2, usw., der grö§ten der Wert n. Beispiel:

x_i	1	7	2	5.3	16
$\mathrm{rg}(x_i)$					

- Liegen sogenannte Bindungen vor, d.h. haben mehrere Einheiten dieselbe Ausprägung der Variablen X oder der Variablen Y, so nimmt man den Durchschnittswert der in Frage kommenden Ränge (Achtung: etwas anderer Begriff der Bindung als in Kapitel 5).

  Beispiel:

x_i	1	7	7	3	10
Rang					
$\mathrm{rg}(x_i)$					

- Wende nun den Korrelationskoeffizienten nach Bravais-Pearson auf die Rangdaten an. Nach Umformung ergibt sich unter Benutzung von

$$\sum_{i=1}^{n} rg(x_i) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \sum_{i=1}^{n} rg(y_i)$$

folgende Formel:

Definition:

$$\varrho_S(X,Y) := \frac{\sum_{i=1}^{n} \mathrm{rg}(x_i) \cdot \mathrm{rg}(y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n} (\mathrm{rg}(x_i))^2 - n\left(\frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n} (\mathrm{rg}(y_i))^2 - n\left(\frac{n+1}{2}\right)^2}}$$

heißt (empirischer) *Rangkorrelationskoeffizient nach Spearman*.

Bemerkungen:

- Liegen keine Bindungen vor, so gilt

$$\varrho_{S,XY} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

wobei $d_i := \mathrm{rg}(x_i) - \mathrm{rg}(y_i)$.

- Wichtig für Interpretation: Da $\varrho_S(X, Y)$ sich aus der Anwendung von $\varrho(X, Y)$ auf Rangdaten ergibt, behalten die entsprechenden Bemerkungen zum Bravais-Pearson-Korrelationskoeffizienten – auf die Ränge bezogen – ihre Gültigkeit. Insbesondere gilt $-1 \leq \varrho_{S,XY} \leq 1$, und $\varrho_{S,XY}$ ist analog zu interpretieren.

- Im Gegensatz zum Korrelationskoeffizienten von Bravais-Pearson misst der Rangkorrelationskoeffizient nicht nur lineare, sondern allgemeiner monotone Zusammenhänge. Die Anwendung der Rangtransformation bewirkt in gewisser Weise eine Linearisierung monotoner Zusammenhänge.
  Tabelle für $y = x^3$:

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	10	1000	10000	100	1000000
	10	1000	10000	100	1000000
	0	0	0	0	0
	-20	-16000	320000	400	256000000
$\sum$	0	-14000	340000	600	258000000

  Also ist $\varrho(X, Y) = 0.987$, und, da hier in der Tat $rg(x_i) = rg(y_i)$ für alle i, $\varrho_s(X, Y) = 1$.

- Die Bildung von Rängen ist unempfindlich gegenüber Ausreißern, so dass auch der Rangkorrelationskoeffizient ausreißerresistent ist.

Beispiel: (fiktiv, Zahlen aus Jann, 2002/2005)

Zwei Gutachter sollen das autoritäre Verhalten von 5 Gruppenmitgliedern vergleichen, indem sie Scores auf einer Skala zwischen 0 und 100 vergeben. (Dies ist ein typischer Fall einer Ordinalskala; die Abstände sind nicht direkt interpretierbar, sondern nur die Reihenfolge!)

Man berechne den Rangkorrelationskoeffizienten nach Spearman für die Merkmale X und Y mit

X Einstufung durch Gutachter 1
Y Einstufung durch Gutachter 2

Person i	1	2	3	4	5
X: Gutachter 1	10	15	20	20	30
Y: Gutachter 2	20	10	30	40	60
$\mathrm{rg}(x_i)$					
$\mathrm{rg}(y_i)$					

Bemerkung:

- Analog zur punkt-biserialen Korrelation gibt es auch eine $biseriale\ Rangkorrelation$ zur Beschreibung des Zusammenhangs zwischen einer $0-1$-kodierten dichotomen nominalen und einer quasi-stetigen ordinalen Variable (vgl. Wagschal, 1999, Kap 10.7).