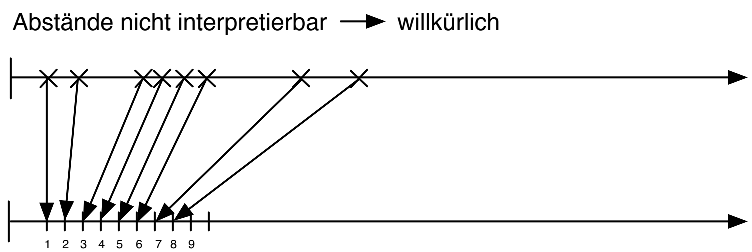
Rangkorrelationskoeffizient nach Spearman

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

Andererseits macht eine Darstellung und Analyse über Kontingenztafeln wenig Sinn, da wegen der vielen unterschiedlichen Ausprägungen die meisten Zellen bestenfalls spärlich besetzt sind.

Bsp. 6.7.



Idee: Nicht willkürlich sind aber die Ränge
Natürliche Skala für ordinale Daten soll die Ordnung widerspiegeln $\rightarrow$ Ränge betrachten und diese als intervallskaliert (d.h. insbes. als gleichabständig) betrachten (nicht unumstritten!).

- Liegen keine *$Bindungen$* vor, dann rechnet man bei ordinalskalierten Merkmalen statt mit $(x_i, y_i)_{i=1,\dots,n}$ mit den zugehörigen Rängen $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$ $i = 1, \dots, n$. 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 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ einnimmt (analog für $\mathrm{rg}(y_i)$). Der kleinsten Beobachtung wird also der Wert 1 zugeordnet, der zweitkleinsten der Wert 2, usw., der größten der Wert n.

Bsp. 6.8.

x_i	1	7	2	5.3	16
$\mathrm{rg}(x_i)$	1	4	2	3	5

- Liegen sogenannte Bindungen vor, d.h. hier: 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).

Bsp. 6.9.

x_i	1	7	7	3	10
Rang	1	3 oder 4	3 oder 4	2	5
$\mathrm{rg}(x_i)$	1	3.5	3.5	2	5

- 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*.

Bem. 6.10.

- Liegen keine Bindungen vor, so gilt (Beweis durch Nachrechnen: z.B. Ferschl (1985^3). Deskriptive Statistik, S. 285)

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

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

$d_i \approx 0$ für alle i zwischen den Rängen von X und Y besteht kaum ein Unterschied. In der Tat dann $\rightarrow$

$$\frac{\sum_{i=1} d_i^2}{n(n^2-1)}$$

klein und damit

$$1 - \frac{6 \sum_{i=1} d_i^2}{n(n^2-1)} \quad \text{fast} \quad 1$$

maximaler Unterschied: d_i^*
Gegenläufige Ränge

X	1	2	3	$\cdots$
Y	n	n-1	n-2	$\cdots$

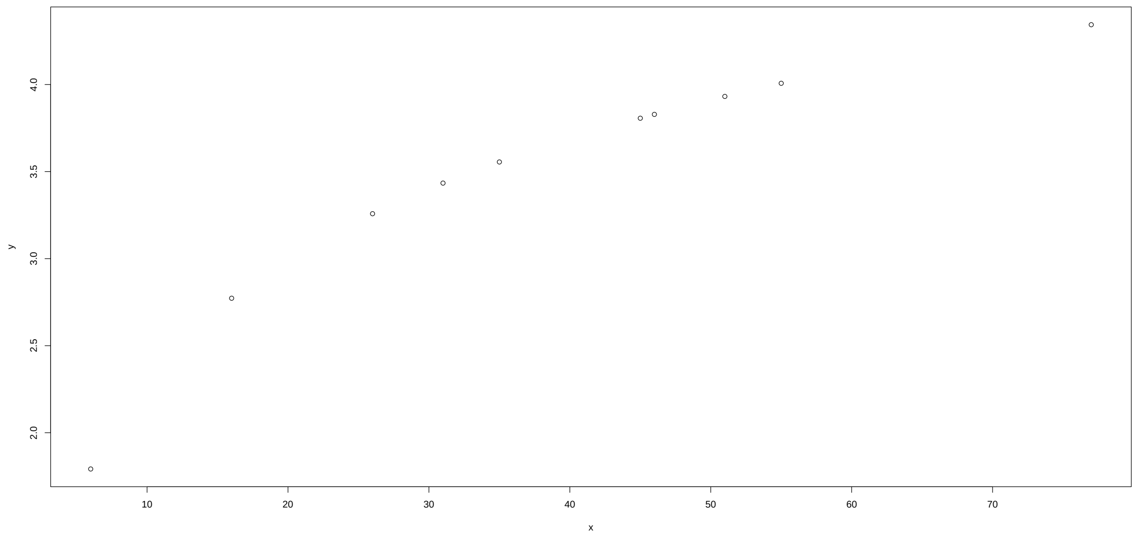
$d_1 = (n-1) = n-1$
$d_2 = (n-1-2) = n-3$
$d_3 = (n-2-3) = n-5$
$d_i = 2 * (\frac{n}{2} - i + 1)$

- 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	-8000	160000	400	64000000
$\sum$	0	-6000	180000	600	66000000

  Also ist $\varrho(X,Y) = 0.973$, und, da hier in der Tat $rg(x_i) = rg(y_i)$ für alle i, $\varrho_s(X,Y) = 1$.



y
x

- Die Bildung von Rängen ist deutlich unempfindlicher gegenüber Ausreißern, so dass auch der Rangkorrelationskoeffizient ausreißerresistenter ist.

Multipliziert man beispielsweise die größte Beobachtung mit 100, so beeinflusst dies stark das arithmetische Mittel $\bar{x}$ und damit den Bravis-Pearson-Korrelationskoeffizienten, aber nicht den Rangplatz.

Bsp. 6.11.

(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
$\mathrm{rg}(x_i)$	1	2	3.5	3.5	5
$\mathrm{rg}(y_i)$	2	1	3	4	5

$$
\begin{aligned}
\varrho_s(X,Y) &= \frac{\sum_{i=1}^{n} \mathrm{rg}(x_i)\,\mathrm{rg}(y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n}(\mathrm{rg}(x_i))^2 - n\left(\frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}(\mathrm{rg}(y_i))^2 - n\left(\frac{n+1}{2}\right)^2}} \\
&= \frac{(1\cdot 2 + 2\cdot 1 + 3.5\cdot 3 + 3.5\cdot 4 + 5\cdot 5) - 5\cdot(\frac{5+1}{2})^2}{\sqrt{1^2 + 2^2 + 3.5^2 + 3.5^2 + 5^2 - 5\cdot(\frac{5+1}{2})^2}\sqrt{2^2 + 1^1 + 3^2 + 4^2 + 5^2 - 5\cdot(\frac{5+1}{2}}} \\
&= 0.872
\end{aligned}
$$