2.2 Rangkorrelation nach Spearman

Wir wollen in diesem Kapitel den Rangkorrelationskoeffizienten nach Spearman berechnen. Die erste Datenreihe besteht aus n Realisierungen $x_1, x_2, ..., x_n$ der unabhängig und identisch stetig verteilten Zufallsvariablen $X_1, X_2, ..., X_n$ (die verteilt sind wie X) und die zweite Datenreihe besteht analog aus den Realisierungen $y_1, y_2, ..., y_n$ der unabhängig und identisch stetig verteilten Zufallsvariablen $Y_1, Y_2, ..., Y_n$ (die verteilt sind wie Y). Hier genügt es, wenn das Datenniveau mindestens ordinal ist.

Auf der Basis des Rangkorrelationskoeffizienten nach Spearman kann man einen Test mit den folgenden Hypothesen durchführen:

H_0: Die Zufallsvariablen X und Y sind unabhängig

gegen

H_1: Die Zufallsvariablen X und Y sind abhängig

Wir kommen zu unserem Beispiel:

v1	v2
4	5
4	5
3,7	4,3
2,7	3
5,1	5

Den Output erhalten Sie, wenn Sie auf den Button **→Zusammenhänge untersuchen** klicken**,** wobei Sie zuvor die Variablen v1 und v2 unter diesem Menüpunkt auswählen müssen. Danach können Sie **→Rangkorrelation nach Spearman** wählen.

Der Rangkorrelation nach Spearman

Hier sind die eingegebenen Daten zu sehen:

Spalte 1 (x_i)	Rang(x_i)	Spalte 2 (y_i)	Rang(y_i)
4	3.5	5	4
4	3.5	5	4
3.7	2	4.3	2
2.7	1	3	1
5.1	5	5	4

Stichprobenumfang n	5
Korrelationskoeffizient von Spearman	0.91766293548225
Prüfgröße d (Hotelling-Pabst-Statistik)	1.5
E(D)	17.5
Var(D)	76
p-Wert (approximiert [1])	0.0665

[1] Approximierter p-Wert für n > 20.

Der Korrelationskoeffizient nach Spearman wird wie folgt berechnet: Es werden für beide Datenreihen separat Rangzahlen vergeben. Danach wird mit diesen Rangzahlen der Korrelationskoeffizient nach Pearson berechnet.

Zur Durchführung des Tests kann anstelle des Korrelations-koeffizienten nach Spearman auch die Hotelling-Pabst-Statistik verwendet werden, die etwas einfacher über die Rangzahlen berechnet werden kann, wie wir unten sehen werden.

Die Rangzahlen für die erste Variable sind:

3,5, 3,5, 2, 1, 5

Für die zweite Variable:

4, 4, 2, 1, 4

Wir berechnen den mittleren Rang, der für beide Variablen gleich ist:

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} \text{Rang}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \text{Rang}(y_i) = \frac{n+1}{2}$$

Der Korrelationskoeffizient nach Spearman ergibt sich dann durch:

$$r_S = \frac{\sum_{i=1}^{n} (\text{Rang}(x_i) - \bar{r})(\text{Rang}(y_i) - \bar{r})}{\sqrt{\sum_{i=1}^{n} (\text{Rang}(x_i) - \bar{r})^2} \cdot \sqrt{\sum_{i=1}^{n} (\text{Rang}(y_i) - \bar{r})^2}}$$

Im Beispiel ist $\bar{r}$ = 3. Da der Rangkorrelationskoeffizient nach Spearman im Beispiel mit einem Wert von r_s = 0,9176... recht groß ist (dieser kann Werte zwischen -1 und 1 annehmen), lässt dies eine positive Korrelation vermuten.

Die Hotelling-Pabst-Statistik ist gegeben durch:

$$d = \sum_{i=1}^{n} (\text{Rang}(x_i) - \text{Rang}(y_i))^2$$

Im Beispiel ist d = 1,5. d wird oft als Prüfgröße für den Test auf Korrelation verwendet.

Es gilt:

$$E(D) = n(n-1)(n+1)/6 - \frac{1}{12}\sum_{j=1}^{k_s} s_j(s_j-1)(s_j+1) - \frac{1}{12}\sum_{j=1}^{k_t} t_j(t_j-1)(t_j+1)$$

Im Beispiel gilt: E(D) = 17,5

Die Werte s_j sind die absoluten Häufigkeiten der Wert x_i (siehe Kapitel 1.5). In diesem Beispiel kommt die 2,7 und 3,7 einfach, die 4 doppelt und die 5.1 einfach vor. Somit ist $k_s = 4$, $s_1 = 1$, $s_2 = 1$, $s_3 = 2$ und $s_4 = 1$. Die Werte t_j sind analog die absoluten Häufigkeiten der Wert y_i. In diesem Beispiel kommt die 3 und die 4.3 einfach und die 5 dreifach vor. Somit ist $k_t = 3$, $t_1 = 1$, $t_2 = 1$ und $t_3 = 3$. Es treten somit Bindungen (mehrfach vorkommende Werte bei einer Variablen) auf. Kommen alle Werte nur einfach vor (bei einer Variablen), so entfallen die beiden letzten Summanden und $E(D) = n(n-1)(n+1)/6$.

$$Var(D) = (n^2(n-1)(n+1)^2/36) \cdot \left(1 - \frac{1}{n^3-n}\sum_{j=1}^{k_s} s_j(s_j-1)(s_j+1)\right)\left(1 - \frac{1}{n^3-n}\sum_{j=1}^{k_t} t_j(t_j-1)(t_j+1)\right)$$

Im Beispiel gilt: Var(D) = 76

Nun kann die Prüfgröße z berechnet werden, die Realisierung einer asymptotisch standardnormalverteilten Zufallsvariablen Z ist:

$$z = \frac{d - E(D)}{\sqrt{Var(D)}}$$

Der approximative p-Wert = $2(1-F_{N(0,1)}(|z|))$. Im Output zum Beispiel wird hierfür 0,0665 ausgegeben.

Wie zu sehen ist, könnte die Nullhypothese der Unabhängigkeit auf einem Signifikanzniveau von 5% nicht verworfen werden (denn 0,0665 > 0,05), womit wir keinen signifikanten Zusammenhang zwischen den Messreihen nachweisen können. Hier sollte n aber größer als 20 sein.

Bei diesem Stichprobenumfang sollt aber die exakte Verteilung verwendet werden. In Büchern (wie z.B. in [3], [8] und [9]) findet man hierzu Tabellen (falls keine Bindungen vorhanden sind, wobei man diese auch Näherungsweise verwenden kann).

Wollen wir nun die exakte Verteilung zu diesem Test bestimmen. Dazu müssen für alle möglichen Permutationen der Rangzahlen der Stichproben die Prüfgröße d berechnet werden. Es gibt hier im Beispiel also

$$\frac{n!}{t_1! \cdot \ldots \cdot t_{k_t}!} = \frac{5!}{2!} = 60$$

Möglichkeiten, wenn wir die Rangzahlen der ersten Stichprobe permutieren und

$$\frac{n!}{s_1! \cdot \ldots \cdot s_{k_s}!} = \frac{5!}{3!} = 20$$

Möglichkeiten, wenn wir die Rangzahlen der zweiten Stichprobe permutieren. Wir benötigen also weniger Rechenschritte bei der Permutation der zweiten Stichprobe. Somit gibt es 20 mögliche Rangzahlenkombinationen um r_s oder d zu berechnen.

Es folgen die möglichen Werte für den Spearman-Rangkorrelations-koeffizienten und darunter mögliche Werte für d im Beispiel mit den dazugehörigen (absoluten) Häufigkeiten bei einer Permutation der Rangzahlen der zweiten Teilstichprobe.

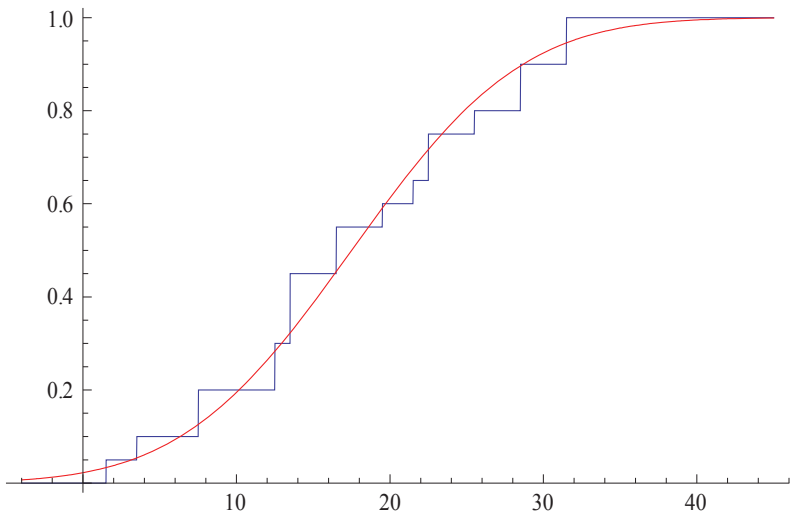
r_S	Häufigkeit
-0,802955	2
-0,630893	2
-0,458831	1
-0,286770	2
-0,229416	1
-0,114708	1
0,057354	2
0,229416	3
0,286770	2
0,573539	2
0,802955	1
0,917663	1

d	Häufigkeit
1,5	1
3,6	1
7,5	2
12,5	2
13,5	3
16,5	2
19,5	1
21,5	1
22,5	2
25,5	1
28,5	2
31,5	2

Wenn wir uns jetzt noch die zugehörige Dichte ausgeben lassen, können wir anhand dieser direkt erkennen, zu welchem α wir die Nullhypothese verwerfen können oder aber auch nicht. Dafür teilen wir jeweils die oberen Häufigkeiten der Prüfgrößen durch die Summe aller Häufigkeiten (also durch 20).

d	P(D = d)	P(D ≤ d)
1,5	0,05	0,05
3,5	0,05	0,1
7,5	0,1	0,2
12,5	0,1	0,3
13,5	0,15	0,45
16,5	0,1	0,55
19,5	0,05	0,6
21,5	0,05	0,65
22,5	0,1	0,75
25,5	0,05	0,8
28,5	0,1	0,9
31,5	0,1	1

Bevor wir allerdings zur Bewertung kommen, wollen wir uns grafisch veranschaulichen, inwieweit sich die Normalverteilung an unsere exakte Verteilung (unten als Treppenfunktion zu sehen, wobei die senkrechten Striche nicht zur Funktion gehören) annähert.



Wir haben weiter oben eine Prüfgröße von d = 1,5 berechnet. Es gilt $P(D \leq 1{,}5) = 1/20$. $P(D \geq 1{,}5) = 1$. Der zweiseitige p-Werte wäre damit $2 \cdot P(D \leq 1{,}5) = 1/10 = 10\%$. Somit könnte man die Nullhypothese erst auf einem Signifikanzniveau α von 10% verwerfen.

Umsetzung mit SAS:

```
data dat1;
input x y;
datalines;
4          5
4          5
3.7        4.3
2.7        3
5.1        5
run;

proc print data=dat1;

proc corr data = dat1 nocorr spearman;
   var x y;
run;
```

SAS-Output zur Prozedur CORR:

Die Prozedure CORR

2 Variablen: x y

Einfache Statistiken						
Variable	**N**	**Mittelwert**	**Std.abweichung**	**Median**	**Minimum**	**Maximum**
x	5	3.90000	0.85732	4.00000	2.70000	5.10000
y	5	4.46000	0.87063	5.00000	3.00000	5.00000

Spearmansche Korrelationskoeffizienten, N = 5 Prob > |r| unter H0: Rho=0		
	x	**y**
x	1.00000	0.91766 0.0280
y	0.91766 0.0280	1.00000

SAS berechnet hier einen anderen p-Wert über:

p-Wert = $2(1-F_{t_{n-2}}(|t|))$ mit

$$t = \sqrt{(n-2)\frac{r_s^2}{1-r_s^2}}$$

und $F_{t_{n-2}}$ als Verteilungsfunktion der t-Verteilung mit n-2 Freiheitsgraden.