

1.7 Wilcoxon Vorzeichenrangtest für eine Stichprobe

Wir gehen wieder davon aus, dass die Stichprobe x_1, x_2, \dots, x_n aus Realisierungen unabhängiger und identisch stetig verteilter Zufallsvariablen X_1, X_2, \dots, X_n besteht. Die Verteilungsfunktion der Zufallsvariablen X_i wird zusätzlich als symmetrisch vorausgesetzt. Das Datenniveau sollte hier ebenfalls sogar metrisch sein. Mit dem Vorzeichenrangtest nach Wilcoxon kann die Hypothese

$$H_0: \text{Median} = \text{Median}_0$$

gegen

$$H_1: \text{Median} \neq \text{Median}_0$$

getestet werden.

Im Beispiel testen wir die Hypothese

$$H_0: \text{Median} = 2$$

gegen

$$H_1: \text{Median} \neq 2$$

und verwenden dabei die Daten aus dem vorangegangenen Kapitel.

Wenn Sie die Daten eingegeben haben, können Sie →**Univariate Statistik** wählen, können Sie neben dem Button Vorzeichentest den Wert für Median_0 eintragen, also hier 2. Dieser Wert gilt auch für den Vorzeichenrangtest. Danach können Sie auf →**Vorzeichenrangtest** klicken und erhalten den folgenden Output:

Vorzeichenrangtest

H0: Median = 2
 gegen
 H1: Median \neq 2

Hier ist u.a. die eingegebene Stichprobe zu sehen:

Stichprobe (x_i)	$y_i = x_i - 2$	Rang($ y_i $)
1	-1	1.5
2		
5	3	5
5	3	5
4	2	3
3	1	1.5
2		
5	3	5

Bemerkung: Bei der Berechnung der Rangzahlen werden nur Differenzen $(x_i - 2) < 0$ berücksichtigt!

Anzahl Werte $x_i - 2 < 0$	6
Rangsumme für Beobachtungen kleiner 2	1.5
t^+ bzw. Rangsumme für Beobachtungen größer 2	19.5
$E(T^+)$	10.5
$\text{Var}(T^+)$	22.125
p-Wert (approximiert ⁽¹⁾)	0.0557

⁽¹⁾ Approximierten p-Wert nur für $n > 20$ beachten.

Nun wollen wir den Output nachvollziehen: Bei diesen Test wird, wie beim vorangegangenen Vorzeichentest, wieder von jeder Beobachtung der hypothetische Median (Median_0) subtrahiert

$$y_i = x_i - \text{Median}_0 = x_i - 2:$$

-1, 0, 3, 3, 2, 1, 0, 3

Dadurch hätten die Zufallsvariablen Y_i unter H_0 den Median 0. Aus dieser transformierten Stichprobe werden alle Nullen entfernt:

-1, 3, 3, 2, 1, 3

Danach werden Rangzahlen für die $|y_i|$ vergeben (siehe Kapitel 1.5):

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

Die Prüfgröße t^+ berechnet sich nun wie folgt: Es wird die Summe über die Rangzahlen gebildet, die zu positiven Werten y_i gehören.

$$t^+ = \sum_{\{i|y_i>0\}} \text{rang}(|y_i|) = 5 + 5 + 3 + 1,5 + 5 = 19,5$$

Im Beispiel ist $t^+ = 5 + 5 + 3 + 1,5 + 5 = 19,5$. Mit dieser Prüfgröße können wir jetzt den p-Wert berechnen und den Test durchführen. Da in unserem Fall Bindungen vorkommen, wollen wir den p-Wert und die Prüfgröße t^+ unter Berücksichtigung dieser Bindungen berechnen.

Es gilt (wie immer unter H_0):

$$E(T^+) = n(n+1)/4$$

$$\text{Var}(T^+) = \frac{n(n+1)(2n+1) - 1/2 \cdot \sum_{j=1}^k t_j(t_j-1)(t_j+1)}{24}$$

Würden keine Bindungen auftreten, dann ergibt sich

$$\text{Var}(T^+) = n(n+1)(2n+1)/24 .$$

Die Werte t_j sind die absoluten Häufigkeiten der Wert $|y_i|$ (siehe Kapitel 1.5). In diesem Beispiel kommt die 1 doppelt, die 2 einfach und die 3 dreifach vor bei den Werten von $|y_i|$. Somit ist $k = 3$, $t_1 = 2$, $t_2 = 1$ und $t_3 = 3$. Im Beispiel gilt also:

$$\sum_{j=1}^k t_j (t_j - 1)(t_j + 1) = 30 .$$

Im Prinzip müssten hier nur die absoluten Häufigkeiten größer als 1 berücksichtigt werden, da für $t_j = 1$ der j -te Summand oben gleich Null ist.

Im Beispiel ist damit $E(T^+) = 10,5$ und $\text{Var}(T^+) = 22,125$.

Ausgegeben wird der approximative p-Wert, der erst für $n > 20$ verwendet werden sollt. Dieser wird wie folgt berechnet:

$$\text{p-Wert} = 2(1 - P_{N(0,1)}(|z|)) \text{ mit } z = \frac{t^+ - E(T^+)}{\sqrt{\text{Var}(T^+)}}$$

Im Beispiel gilt: p-Wert $\approx 0,0557$.

Wie anhand des approximativen p-Werts zu sehen ist, könnte auf einem Signifikanzniveau von 5% die Nullhypothese nicht verworfen werden und demnach kein signifikanter Unterschied des Medians vom Wert 2 nachgewiesen werden.

Für den exakten Test müssten alle theoretisch möglichen Werte der Prüfgröße berücksichtigt werden, d.h. es müssen aus den Rangzahlen

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

alle möglichen Werte für t^+ berechnet werden. Theoretisch gilt: Es könnte kein Wert y_i positiv sein, dann wäre $t^+ = 0$, es könnte ein Wert positiv sein, dann könnte t^+ jeden einzelnen Rangwert annehmen, Dadurch könnte man die exakte Verteilung berechnen und müsste $2^6 = 64$ Summen berücksichtigen.

Wir wollen mal die exakte Verteilung für $n = 6$ bestimmen. Es gilt

$$T^+ = \sum_{i=1}^6 B_i \cdot i,$$

wobei B_i unabhängige $B(1,1/2)$ -verteilte Zufallsvariablen sind, denn unter H_0 tritt jede der n Rangzahlen mit einer Wahrscheinlichkeit von $1/2$ in der Summe für T^+ auf. Der Grund ist: Wenn der Median der Y_i gleich 0 wäre, dann gilt $P(Y_i > 0) = 1/2$, wie beim Vorzeichentest.

t^+	$P(T^+ = t^+)$	$P(T^+ \leq t^+)$
0	0,015625	0,015625
1	0,015625	0,03125
2	0,015625	0,046875
3	0,03125	0,078125
4	0,03125	0,109375
5	0,046875	0,15625
6	0,0625	0,21875
7	0,0625	0,28125
8	0,0625	0,34375

t^+	$P(T^+ = t^+)$	$P(T^+ \leq t^+)$
9	0,078125	0,421875
10	0,078125	0,5
11	0,078125	0,578125
12	0,078125	0,65625
13	0,0625	0,71875
14	0,0625	0,78125
15	0,0625	0,84375
16	0,046875	0,890625
17	0,03125	0,921875
18	0,03125	0,953125
19	0,015625	0,96875
20	0,015625	0,984375
21	0,015625	1

Hier würde gelten: $P(T^+ \geq 19,5) = 0,015625 + 0,015625 = 0,03125$.
 Damit ergibt sich der zweiseitige: p-Wert = $2 \cdot 0,03125 = 0,0625$. Man
 könnte somit H_0 ebenfalls nicht verwerfen (mit $\alpha = 0,05$). Tabellen
 mit Werten für die Grenzen der kritischen Bereiche findet man u.a. in
 [3], [8] oder [9], falls keine Bindungen vorhanden sind (wobei sich die
 Verteilung für $n > 10$ bei Bindungen nicht wesentlich ändert ([3])).

In unserem Fall (mit Bindungen) ergibt sich die im Folgenden
 tabellierte exakte Verteilung:

t^+	$P(T^+ = t^+)$	$P(T^+ \leq t^+)$
0	0,015625	0,015625
1,5	0,03125	0,046875
3	0,03125	0,078125
4,5	0,03125	0,109375
5	0,046875	0,15625
6	0,015625	0,171875
6,5	0,09375	0,265625
8	0,09375	0,359375
9,5	0,09375	0,453125

t^+	$P(T^+ = t^+)$	$P(T^+ \leq t^+)$
10	0,046875	0,5
11	0,046875	0,546875
11,5	0,09375	0,640625
13	0,09375	0,734375
14,5	0,09375	0,828125
15	0,015625	0,84375
16	0,046875	0,890625
16,5	0,03125	0,921875
18	0,03125	0,953125
19,5	0,03125	0,984375
21	0,015625	1

Mit dieser Verteilung, würde gelten:

$$P(T^+ \geq 19,5) = 0,03125 + 0,015625 = 0,046875$$

$$\text{p-Wert} = 2 \cdot 0,046875 = 0,09375$$

Bemerkungen zur Berechnung des p-Wertes mit der exakten Verteilung:

Der p-Wert berechnet sich wie beim Binomialtest

$$\text{p-Wert} = \min \{ 2 \cdot P(T^+ \leq t^+), 2 \cdot P(T^+ \geq t^+), 1 \}$$

nur dass hier nicht, wie bei der Binomialverteilung $P(X \geq k) = 1 - P(X \leq k - 1)$ gilt, denn es kommen nicht nur ganzzahlige Werte von T^+ vor. Hier gilt, wie allgemein für diskrete Verteilungen $P(T^+ \geq t^+) = 1 - P(T^+ < t^+)$. Im Beispiel mit obiger Tabelle gilt $P(T^+ \geq 19,5) = 1 - P(T^+ \leq 18) = 0,046875$ (oder man bildet die Summe wie oben). Da $P(T^+ \leq 19,5) = 0,984375$ größer ist, gilt hier: $\text{p-Wert} = 2 \cdot P(T^+ \geq 19,5) = 0,09375$. Für $n \leq 20$ wird in SAS der „exakte“ p-Wert ausgegeben.

Umsetzung mit SAS:

```
data dat1;
input x;
cards;
1
2
5
5
4
3
2
5
run;

proc univariate data = dat1 mu0=2;
var x;
run;
```

SAS-Output zur Prozedur UNIVARIATE:

Die Prozedur UNIVARIATE
Variable: x

Momente			
N	8	Summe Gewichte	8
Mittelwert	3.375	Summe Beobacht.	27
Std.abweichung	1.59798981	Varianz	2.55357143
Schiefe	-0.2581915	Kurtosis	-1.7422466
Unkorr. Qu.summe	109	Korr. Quad.summe	17.875
Variationskoeff.	47.3478462	Stdfeh. Mittelw.	0.56497471

Grundlegende Statistikmaße			
Lage		Streuung	
Mittelwert	3.375000	Std.abweichung	1.59799
Median	3.500000	Varianz	2.55357
Modalwert	5.000000	Spannweite	4.00000
		Interquartilsabstand	3.00000

Tests auf Lageparameter: $\mu_0=2$				
Test	Statistik		p-Wert	
Studentsches t	t	2.433737	Pr > t 	0.0452
Vorzeichen	M	2	Pr \geq M 	0.2188
Vorzeichen-Rang	S	9	Pr \geq S 	0.0938