Tests von Wilcoxon

Tobias Steinherr

28. Mai 2014

Inhaltsverzeichnis

1 Einleitung

Der amerikanische Chemiker Frank Wilcoxon (*1892 in Irland, † 1965 in Florida) war sehr interessiert darin, statistische Methoden zur Untersuchung wissenschaftlicher Daten zu gewinnen. Unter seinen etwa 70 Veröffentlichungen gilt der Artikel 'Individual Comparisons by Ranking Methods', veröffentlicht 1945 in der Zeitschrift 'Biometrics Bulletin', als sein bedeutendster. In diesem stellte er zwei von ihm entwickelte Tests, den Wilcoxon-Rangsummentest und den Wilcoxon-Vorzeichen-Rang-Test, vor, die bis heute eine enorme Bedeutung in der Statistik haben und vor allem angewandt werden, wenn unter gewissen Voraussetzungen andere Tests nicht funktionieren. Beide Tests vergleichen zwei Stichproben miteinander, einmal für den Fall unabhängiger (Rangsummentest), einmal für den Fall abhängiger Stichproben (Vorzeichen-Rang-Test). Zweitgenannter Test eignet sich ebenso für den Ein-Stichprobenfall.

Für diese Fragestellungen kommen meist t-Tests in Frage, die jedoch an gewisse Voraussetzungen gebunden sind. So verlangen diese Stichproben aus einer normalverteilten Grundgesamtheit respektive einen bzw. zwei hinreichend große Stichprobenumfänge, sodass der zentrale Grenzwertsatz erfüllt ist und eine Normalverteilung zumindest approximativ angenommen werden kann. Da dies nicht bei allen Daten der Fall sein muss, suchte Frank Wilcoxon nach einer Alternative dazu. Grundlegendes seiner Methodik ist es, dass er exakte metrische Werte und Differenzen zwischen einzelnen Werten mehr oder weniger außen vor lässt. Vielmehr sortiert Wilcoxon die einzelnen Werte der Daten der Größe nach und vergibt entsprechende Ränge. Aus einem metrischen Skalenniveau wird so ein ordinales.

Wilcoxons Tests wird teilweise als der Anfang der nonparametrischen Statistik gesehen, was jedoch nicht der Wahrheit entspricht. [1, S. 191] Sicher ist aber, dass sie unter den nonparametrischen Verfahren zu den bedeutendsten zählen. Im Folgenden werden diese beiden Tests genauer besprochen.

2 Wilcoxon-Rangsummentest

2.1 Erklärung

Der Wilcoxon-Rangsummentest untersucht die Fragestellung, ob sich zwei voneinander unabhängige Stichproben hinsichtlich eines bestimmten Merkmals voneinander unterscheiden. Es handelt sich also um einen Zwei-Stichproben-Test, der im Gegensatz zu einem in etwa vergleichbaren Zwei-Stichproben-t-Test ohne Parameter auskommt. Grundsätzliche Annahme dieses Tests ist es, dass zwei voneinander unabhängige Stichproben einer Verteilungsfunktion der gleichen Form folgen, die lediglich eventuell um einen Betrag δ links- oder rechtsseitig verschoben ist. Sei F_X die Verteilungsfunktion der Zufallsvariablen X (Größe: n) und F_Y diejenige der Zufallsvariablen Y (Größe: m), so soll gelten:

$$F_X(x) = F_Y(x - \delta)$$

Dabei ist δ ein fester, unbekannter Parameter. Der Wilcoxon-Rangsummentest soll nun zeigen, ob und inwiefern sich δ von 0 unterscheidet. Dabei gibt es drei Möglichkeiten, in jeder davon lautet die Nullhypothese

$$H_0 : F_X(x) = F_Y(x)$$

und die Alternativhypothese

$$H_1 : F_X(x) = F_Y(x - \delta).$$

Die drei Fälle unterscheiden sich lediglich hinsichtlich δ.

- Im Fall a) ist $\delta > 0$, es wird getestet, ob F_X einseitig nach links gegenüber F_Y verschoben ist
- Der Fall b) ist genau umgekehrt, hier wird überprüft, ob $\delta < 0$, ob also F_X gegenüber F_Y einseitig nach rechts verschoben ist
- Fall c) überprüft sozusagen beides gleichzeitig, nämlich ob $\delta \neq 0$, also zweiseitig, ob es überhaupt eine Verschiebung unter den beiden Verteilungsfunktionen gibt

Was hier womöglich noch kompliziert klingen mag, erweist sich gerade in der Praxis als einfacher und äußerst nützlich. Denn gleiche Verteilungsfunktionen implizieren unter anderem gleiche Mediane und daher wird mit den Fällen a), b) und c) gleichbedeutend überprüft, ob und wie sich die Mediane der beiden Gruppen voneinander unterscheiden. Und so lauten die Hypothesen auch:

- Fall a): $H_0 : x_{med} \geq y_{med}$ $\quad$ $H_1 : x_{med} < y_{med}$
- Fall b): $H_0 : x_{med} \leq y_{med}$ $\quad$ $H_1 : x_{med} > y_{med}$
- Fall c): $H_0 : x_{med} = y_{med}$ $\quad$ $H_1 : x_{med} \neq y_{med}$

Um einen dieser Fälle zu überprüfen, bedarf es noch der Teststatistik, die in jedem der Fälle identisch ist. Wie bereits erwähnt, werden aus den Daten Ränge gebildet. Dazu werden zunächst alle Beobachtungen $x_1, x_2, ..., x_n, y_1, y_2, ..., y_m$ der Größe nach sortiert und entsprechend Ränge $rg(x_1)$, $rg(x_2)$,...,$rg(x_n)$, $rg(y_1)$, $rg(y_2)$,..., $rg(y_m)$ vergeben. Rangplatz 1 erhält die kleinste Beobachtung, statt des metrischen Werts wird hier also im Folgenden mit der natürlichen Zahl 1 gerechnet. Man spricht von Bindungen, wenn mehrere Beobachtungen den gleichen Wert besitzen. Ist dies der Fall, so wird für jede der Beobachtungen der Mittelwert der eigentlichen Ränge gebildet. Der Testwert T_W ergibt sich nun als die Summe der Ränge von X, also

$$T_W = \sum_{i=1}^{n} rg(X_i).$$

Dieser Testwert muss nun mit entsprechenden Grenzwerten verglichen werden, damit festgestellt werden kann, ob die ermittelten eventuellen Abweichungen zwischen den Stichproben auch signifikant groß sind. Für jeden der drei Fälle sind die Grenzwerte unterschiedlich. Es wird dabei immer berücksichtigt,

wie wahrscheinlich das Auftreten eines gewissen Testwerts bzw. eines Wertes der mindestens so groß oder klein wie dieser ist. Beispielsweise wird im ersten Fall bedacht, wie groß die Wahrscheinlichkeit ist, einen Wert zu erhalten, der maximal so groß ist wie der Testwert. Um bei einem Test eine gewisse Signifikanz zu gewährleisten, wird berechnet, ab oder bis zu welchem Wert die Wahrscheinlichkeit, diesen zu erreichen, so groß ist wie das Signifikanzniveau. Dies kann von Hand berechnet werden (siehe Kapitel 4), ist allerdings auch in Tabellen [2, S.591] festgehalten, da sich die Komplexität der Berechnung mit zunehmendem Stichprobenumfang ebenso erhöht. Die Ablehnungsbereiche lauten:

- Fall a): $T_W < w_\alpha(n, m)$
- Fall b): $T_W > w_{1-\alpha}(n, m)$
- Fall c): $T_W > w_{1-\alpha/2}(n, m)$ oder $T_W < w_{\alpha/2}(n, m)$

$w_{\tilde{\alpha}}$ bedeutet das $\tilde{\alpha}$-Quantil der Verteilung. Über- oder unterschreitet die Teststatistik entsprechend eines der Quantile, so wird die Nullhypothese verworfen und man geht von der Richtigkeit der Alternativhypothese aus.

2.2 Anwendung

Ein fiktives Beispiel: Bei einem Sprachkurs werden vor Beginn des Unterrichts die bisherigen Vorkenntnisse aller Teilnehmer geprüft. Hierbei sollen Wörter und Sätze übersetzt, korrekt verstanden und selbst ausgesprochen werden. Maximal sind dabei 100 Punkte zu erreichen bzw. damit gleichbedeutend 100 Fehlerpunkte. Bei der Korrektur vermutet der Kursleiter, dass unter den Teilnehmern die $n = 11$ Frauen besser abschneiden als die $m = 10$ Männer. Im Folgenden ist das ungeordnete Ergebnis des Tests zu sehen. Zu sehen ist das Geschlecht (w = weiblich, m = männlich) und die Fehlerpunkte der Teilnehmer.

m	w	w	m	w	m	m	m	w	w	m
86	54	37	89	76	63	45	83	43	36	49

w	w	w	m	w	w	m	w	m	m
76	53	46	81	83	43	58	54	93	78

Tabelle 1: Ungeordnetes Ergebnis des ersten Tests (Fehlerpunkte)

Nun werden aus diesen Fehlerpunkten nach Gruppen (hier: Geschlechter) getrennt die Ränge verteilt. Rangplatz 1 erhält derjenige Teilnehmer, dessen Test die wenigsten Fehlerpunkte aufweist. Theoretisch wäre es umgekehrt genauso möglich, hierbei würde sich nur die Interpretation ändern. Beim Ordnen wird schnell deutlich, dass mehrere Personen die selbe Anzahl an Fehlerpunkten erzielten. So tritt beispielsweise der Wert 43 zweimal auf, auf den potentiellen Rängen 3 und 4. Hier wird das Mittel der Ränge berechnet, in diesem Fall

$(3+4)/2 = 3,5$, und an jedem entsprechenden Platz als Rang verwendet. Selbiges gilt für die Werte 54 und 83, die ebenso doppelt auftauchen.

Frauen		**Männer**	
Rang	(Fehler)	**Rang**	(Fehler)
1	36	5	45
2	37	7	49
3,5	43	11	58
3,5	43	12	63
6	46	15	78
8	53	16	81
9,5	54	17,5	83
9,5	54	19	86
13,5	76	20	89
13,5	76	21	93
17,5	83		

Tabelle 2: Rangplätze unterteilt nach Gruppen

Nun werden die Rangsummen gebildet. Dazu werden aus den beiden Gruppen einfach alle Ränge addiert. Für die Frauen ergibt sich so eine Rangsumme von $\sum_{i=1}^{n} rg(X_i) = 1 + 2 + 3,5 + ... + 17,5 = 87,5$ und für die Männer $\sum_{i=1}^{m} rg(Y_i) = 143,5$. Wie zu vermuten war, ist die Rangsumme bei den Männern trotz minimal kleinerem Stichprobenumfang weitaus höher als die der Frauen. Die Teststatistik wird nun zeigen, ob dieser Unterschied auch signifikant ist. Getestet wird im Folgenden die einseitige Hypothese

$$H_0 : x_{med} \geq y_{med}$$

gegen die Alternativ-Hypothese

$$H_1 : x_{med} < y_{med}.$$

Der Testwert ist gegeben durch die Rangsumme der Frauen, also $T_W = 87,5$ Der Ablehnungsbereich ist in diesem Fall $T_W < w_\alpha(n,m)$. Der Wert $w_\alpha(n,m)$ mit $\alpha = 0,05$ ergibt sich als 98 und ist somit deutlich größer als der Testwert von $87,5$. Somit kann der Kursleiter seine Vermutung, dass die Frauen seines Kurses bessere Vorkenntnisse haben als die männlichen Teilnehmer durch den Wilcoxon-Rangsummentest auf einem Signifikanzniveau von 5% als bestätigt ansehen.

3 Wilcoxon-Vorzeichen-Rang-Test

3.1 Erklärung

Auch für den Einstichprobenfall und den Fall für gepaarte Stichproben entwickelte Wilcoxon einen parameterfreien Test, also beispielsweise wieder Alternativen für die entsprechenden t-Tests. Dieser dient einerseits der Fragestellung, ob sich Werte innerhalb gepaarter Stichproben unterscheiden, andererseits derer, ob ein Merkmal innerhalb einer Stichprobe durchschnittlich eine bestimmte feste Größe hat oder nicht.

Mit gepaarten Stichproben ist gemeint, dass zwei Zufallsvariablen in einem direkten Zusammenhang zueinander stehen und voneinander völlig abhängig sind. Ein Standardbeispiel hierfür ist, wenn das gleiche Merkmal der gleichen Personen zu zwei verschiedenen Zeitpunkten gemessen wird, die Elemente der Stichprobe sind also die selben.

Ähnlich wie im Fall des Wilcoxon-Rangsummentests wird hier erneut mit Rängen gearbeitet, allerdings unterscheiden sich die Vorgehensweisen dennoch voneinander. Kaum unterschiedlich ist der Wilcoxon-Vorzeichen-Rang-Test allerdings, ob nun gepaarte Stichproben verglichen werden oder ob überprüft wird, ob und inwiefern sich ein Merkmal einer Stichprobe von einem hypothetischen Wert unterscheidet. Seien $X_{(1)}$ und $X_{(2)}$ abhängige, gepaarte Stichproben der Größe n und φ ein fester Wert, so kann der Wilcoxon-Vorzeichen-Rang-Test folgende Hypothesen gegeneinander abwägen:

Gepaarte Stichproben

- Fall a): $H_1 : x_{(1)med} \geq x_{(2)med}$ $H1 : x_{(1)med} > x_{(2)med}$
- Fall b): $H_1 : x_{(1)med} \leq x_{(2)med}$ $H1 : x_{(1)med} < x_{(2)med}$
- Fall c): $H_1 : x_{(1)med} = x_{(2)med}$ $H1 : x_{(1)med} \neq x_{(2)med}$

Prüfen auf einen hypothetischen Wert

- Fall a): $H_1 : x_{(1)med} \geq \varphi$ $H1 : x_{(1)med} < \varphi$
- Fall b): $H_1 : x_{(1)med} \leq \varphi$ $H1 : x_{(1)med} > \varphi$
- Fall c): $H_1 : x_{(1)med} = \varphi$ $H1 : x_{(1)med} \neq \varphi$

Während der Wilcoxon-Rangsummentest ohne andere Vorarbeit aus den Daten sofort Ränge bildet, müssen beim Wilcoxon-Vorzeichen-Rang-Test zunächst Differenzen berechnet werden. Bei gepaarten Stichproben werden die Werte $x_{(2)i}$ von $x_{(1)i} (i = 1, 2, 3, ..., n)$, bei einem Test auf einen bestimmten Wert wird ebendieser Wert φ von jedem $x_{(1)i}$ subtrahiert, also

$$D_i = x_{(1)i} - x_{(2)i}$$
bzw.
$$D_i = x_{(1)i} - \varphi$$

Aus diesen Differenzen werden nun Ränge gebildet. Rang 1 erhält die betragsmäßig kleinste Differenz, den größten Rang die Differenz mit dem größten Betrag. Zu beachten ist hier allerdings noch, dass das Vorzeichen der Differenz ebenso mit in den Rang eingeht; wenn die Differenz negativ ist, wird der dazugehörige Rang ebenso mit einem Minus versehen. Auch hier können Bindungen entstehen, diese werden äquivalent behandelt wie im Wilcoxon-Rangsummentest. Entsteht die Differenz 0, wo wird der entsprechende Rang halbiert und die eine Hälfte mit einem positiven, die andere mit einem negativen Rang versehen. Die Teststatistik W^+ berechnet sich nun als die Summe der positiven Ränge, also

$$W^+ = \sum_{i=1}^{n} rg(|D_i|)Z_i$$

mit

$$Z_i = \begin{cases} 1, & D_i > 0 \\ 0, & \text{sonst} \end{cases}$$

Nun muss berechnet werden, wie hoch die Wahrscheinlichkeit für den resultierenden Testwert oder einen extremeren Wert in Richtung der Alternativhypothese ist. Da dies bereits bei recht geringen Stichprobenumfängen zu sehr hohem Arbeitsaufwand führt (vgl. Kapitel 4), empfiehlt sich auch hier, einen Blick in eine Tabelle zu werfen, in der einige wichtige Quantile der Verteilung von W^+ mit entsprechendem Stichprobenumfang aufgelistet sind [2, S.591]. Die Ablehnbereiche für die drei Fälle a), b) und c) lauten:

- Fall a): $W^+ < w_{\alpha}^+$
- Fall b): $W^+ > w_{1-\alpha}^+$
- Fall c): $W^+ < w_{\alpha/2}^+$ oder $W^+ > w_{1-\alpha/2}^+$

$w_{\tilde{\alpha}}^+$ bezeichnet hierbei das $\tilde{\alpha}$-Quantil der Verteilung von w^+.

3.2 Anwendung

3.2.1 Gepaarte Stichproben

Nachdem der Sprachkurs nach einigen wöchentlichen Sitzungen beendet ist, will der Kursleiter überprüfen, wie viel seine Schüler im Rahmen der Veranstaltung gelernt haben und stellt sie vor eine weitere Prüfung, äquivalent zu der Einstiegsprüfung. Der Kursleiter ist guter Dinge und erwartet wenig Fehlerpunkte bei den Teilnehmern. Spätestens nach der Korrektur fällt er allerdings zunächst aus allen Wolken, da er bessere Ergebnisse erwartete und von mehreren Leistungen geradezu schockiert ist. Voller Pessimismus stellt er die Vermutung auf, dass sich die Sprachkenntnisse der Frauen durch seinen Kurs insgesamt nicht einmal verbessert haben.

Um diese These zu überprüfen, wird der Wilcoxon-Vorzeichen-Rang-Test für gepaarte Stichproben benutzt. Die dazugehörigen Hypothesen, die der Kursleiter dazu aufstellt, lauten:

$$H_0 : x_{(n)med} \leq x_{(v)med} \qquad H_1 : x_{(n)med} > x_{(v)med}$$

Mit X_v sind die Ergebnisse der Frauen vor dem Sprachkurs bezeichnet, mit X_n diejenigen danach. Nun wird zunächst für jede Teilnehmerin i die Differenz D_i zwischen dem Ergebnis vor und nach dem Sprachkurs gebildet und diese werden nach dem Betrag geordnet. Es werden also wieder Ränge gebildet, Rangplatz 1 erhält hierbei die Differenz mit dem kleinsten Betrag. Zu beachten ist, dass jeder Rangplatz mit dem selben Vorzeichen versehen wird wie die dazugehörige Differenz. Ist die Fehleranzahl nach dem Sprachkurs geringer geworden, so erhält auch der dazugehörige Rangplatz ein negatives Vorzeichen.

Fehler davor	**Fehler danach**	**Differenz**	**Rangplatz**
43	45	2	1
37	31	-6	-2,5
54	49	-6	-2,5
36	27	-9	-4,5
46	37	-9	-4,5
83	94	11	6
76	63	-13	-7
53	39	-14	-8
43	12	-31	-9
54	19	-35	-10
76	4	-72	-11

Tabelle 3: Ergebnisse der Frauen mit Rangplätzen

Wie zu sehen ist, verbesserten 9 der 11 Teilnehmerinnen ihre Prüfungsergebnisse. Unter den Teilnehmerinnen befinden sich jedoch auch 2, die schlechtere Resultate erzielten als vor dem Sprachkurs. Eine Teilnehmerin um 2 Punkte, was die kleinste Vorher-Nachher-Differenz bedeutet und eine Teilnehmerin um 11 Punkte, der Rang des Differenzbetrags hiervon ist Nummer 6. Nun wird die Teststatistik W^+ berechnet. Dazu werden die positiven Differenzränge addiert, also

$$W^+ = \sum_{i=1}^{n} rg(|D_i|)Z_i$$

mit

$$Z_i = \begin{cases} 1, & D_i > 0 \\ 0, & \text{sonst} \end{cases}$$

Hier:

$$W^+ = 1 \cdot 1 + 2,5 \cdot 0 + 2,5 \cdot 0 + ... + 11 \cdot 0 = 7$$

Signifikant größer als $x_{(v)med}$ wäre $x_{(n)med}$ nach diesem Test erst dann, wenn die Teststatistik größer wäre als $w^+_{1-\alpha}(n)$, was in diesem Fall 52 ist (Signifikanzniveau: 5%), also weit entfernt von 7. Das Ergebnis ist also alles andere

als signifikant. Ganz im Gegenteil: Die Fehleranzahl in der Prüfung nach dem Kurs ist sogar auf dem 1%-Signifikanzniveau kleiner als die vor dem Kurs, da $7 < w^+_{0,01}(11) = 8$. Der Kursleiter kann also beruhigt sein: Zumindest eine Verschlechterung der Sprachkenntnisse durch seinen Kurs hat laut dem Wilcoxon-Vorzeichen-Rang-Test definitiv nicht stattgefunden und sein Kurs war nicht völlig nutzlos.

3.2.2 Prüfen auf einen konkreten Wert

Doch wie sieht es mit den Männern aus? Auch von deren Prüfungsergebnissen ist der Kursleiter nicht gerade angetan. Er ist sogar davon überzeugt, dass die Männer unter den 100 zu erreichenden Punkten nur etwa die Hälfte erreichten. Gegeneinander getestet werden also folgende Hypothesen:

$$H_0 : y_{(n)med} = 50 \qquad H_1 : y_{(n)med} \neq 50$$

Auch diese Problematik lässt sich mit dem Wilcoxon-Vorzeichen-Rang-Test überprüfen, sehr ähnlich zu dem vorherigen Test. Hier gilt: Erweist sich das Testergebnis als nicht signifikant, dann kann nicht davon ausgegangen werden, dass sich die Prüfungsergebnisse der Männer überzufällig von 50 Fehlern unterscheiden. Nun werden Differenzen zwischen allen $m = 10$ Prüfungsergebnissen und 50 gebildet und daraus wieder positive und negative Ränge gebildet.

Fehler	**Differenz zu 50**	**Rang**
50	± 0	± 1
47	-3	-2
55	5	3
42	-8	-4
39	-11	-5
34	-16	-6,5
66	16	6,5
31	-19	-8
12	38	-9
1	-49	-10

Tabelle 4: Ergebnisse der zweiten Prüfung der Männer

Da ein Teilnehmer exakt 50 Fehlerpunkte hatte, ist hier die Differenz zu 50 0, was weder ein positiver noch ein negativer Wert ist. Der dazugehörige Rang 1 wird zur Hälfte als positiv gesehen und zur Hälfte als negativ, geht also jeweils in die positiven und negativen Rangsummen als 0,5 ein. Nun werden wieder die positiven Ränge addiert und man erhält $W_+ = 0,5 + 3 + 6,5 = 10$.

Nun kann man die Nullhypothese nicht ablehnen, falls sich der Testwert W_+ im Bereich $[w^+_{\alpha/2}; w^+_{1-\alpha/2}]$ befindet. Mit $\alpha = 0,05$ ergibt sich ein Bereich von $[9; 46]$, wählt man für α den Wert $0,1$, so beschränkt sich der Bereich auf $[11; 44]$. Im ersten Fall ist der Testwert 10 nicht enthalten ($\rightarrow H_0$ wird beibehalten), im

zweiten Fall jedoch schon ($\rightarrow H_0$ wird verworfen). Dies bedeutet konkret: Die Wahrscheinlichkeit, dass sich die erhaltene durchschnittliche Fehleranzahl der Prüfungsergebnisse der Männer nur zufällig von 50 unterscheiden, liegt nach dem Wilcoxon-Vorzeichen-Rang-Test zwischen 5 und 10%.

4 Exakte Berechnung der Verteilung der Teststatistik

4.1 Erklärung

Bisher wurde bei der Überprüfung auf Signifikanzen lediglich auf tabellierte Quantile der Verteilungen der Wilcoxon-Teststatistiken hingewiesen und das Prinzip lediglich angeschnitten, wie die Berechnung der exakten Wahrscheinlichkeiten für bestimmte Testwerte funktioniert. Nun soll zweiteres für beide Tests genauer ausgeführt werden. Grundsätzliche Überlegungen dabei sind immer mitunter folgende Fragestellungen:

- Wie viele mögliche Rangsummen sind insgesamt zu erreichen?
- Auf wie viele verschiedene Arten sind bestimmte Rangsummen zu erreichen?

4.1.1 Wilcoxon-Rangsummentest

Bei einem Wilcoxon-Rangsummentest mit den Gruppen X vom Umfang n und Y vom Umfang m gilt immer Folgendes:

1. Die niedrigste mögliche Rangsumme für X ist $r_{min(X)} = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ und für Y gleich $r_{min(Y)} = \sum_{i=1}^{m} i = \frac{m(m+1)}{2}$
2. Die höchste mögliche Rangsumme für X ist $r_{max(X)} = \sum_{i=m+1}^{n+m} i = \frac{n(3n+1)}{2}$ und für Y gleich $r_{max(X)} = \sum_{i=n+1}^{n+m} i = \frac{m(3m+1)}{2}$
3. Damit gibt es insgesamt $s = r_{max(X)} - r_{min(X)} + 1 = r_{max(Y)} - r_{min(Y)} + 1$ unterschiedliche mögliche Rangsummen für X und Y
4. Die Summe beider Rangsummen ist immer $\frac{(n+m)(n+m+1)}{2}$ $\rightarrow$ Hat X die niedrigste mögliche Rangsumme, so hat Y die höchste mögliche Rangsumme, hat X die zweitniedrigste mögliche Rangsumme, so hat Y die zweithöchste mögliche Rangsumme, usw.
5. Insgesamt gibt es $l = \binom{n+m}{n} = \binom{n+m}{m}$ verschiedene Möglichkeiten, die Ränge auf die beiden Gruppen aufzuteilen

Diese Informationen reichen jedoch noch nicht komplett aus, um exakte Wahrscheinlichkeiten zu berechnen. Da $l \geq s$ bedeutet dies, dass es meist Rangsummen gibt, die auf mehrere Weisen zu erreichen sind. Wie viele Möglichkeiten

es für welche Rangsummen gibt, ist die Information, die zur Berechnung noch fehlt. Wenn die Gruppe X eine Rangsumme r hat, muss überlegt werden, wie viele Möglichkeiten es gibt, diese Zahl als Summe aus n verschiedenen Summanden (jeweils $\in \mathbb{N}$ und kleiner gleich $(n+m)$) zu erhalten. Bei kleinen Werten für r stellt dies noch kein großes Problem dar, doch mit zunehmender Größe nimmt der Aufwand für das Ausmachen der Anzahl der Möglichkeiten immer schneller zu. Zum gleichen Ergebnis wie zu überlegen, wie viele Möglichkeiten es gibt, r als Summe aus n ungleichen Summanden $\leq (n+m)$ darzustellen, führt folgender Gedanke:
Wie viele Möglichkeiten gibt es, die Zahl $(r - \binom{n}{2})$ als Summe aus $n \in \mathbb{N}$ Summanden (unbedeutend, ob gleich oder ungleich), die höchstens den Wert $(m+1)$ haben, zu erhalten? [3, S. 82]
Sei $F_P(a, b, c)$ die Funktion, die ausgibt, wie viele Möglichkeiten (Partitionen) es gibt, die Zahl a als Summe von b Summanden $\leq c$ darzustellen, so berechnet sich die Wahrscheinlichkeit für eine bestimmte Rangsumme r durch die Population X der Größe n, der gegenüber eine Population Y der Größe m steht, zu

$$P(T_W = r) = \frac{F_P(r-\binom{n}{2},n,m+1)}{\binom{n+m}{n}}$$

Da allerdings meist nach der Wahrscheinlichkeit, höchstens oder mindestens eine gewisse Rangsumme zu erreichen, gefragt wird, müssen in diesen Fällen die Wahrscheinlichkeiten für alle Rangsummen größer oder kleiner gleich dieser gewissen Rangsumme kumuliert werden, also:

$$P(T_W \leq r) = \frac{\sum_{i=r_{min(X)}}^{r}(F_P(i-\binom{n}{2},n,m+1))}{\binom{n+m}{n}}$$

bzw.

$$P(T_W \geq r) = \frac{\sum_{i=r}^{r_{max(X)}}(F_P(i-\binom{n}{2},n,m+1))}{\binom{n+m}{n}}$$

Außerdem gilt:

$$P(T_W \leq r) = 1 - P(T_W > r)$$

Hier wird nun der Zusammenhang zu den Grenzwerten deutlich. Ist $P(T_W \leq r)$ maximal so groß wie ein Wert α, so kann auf einem α-Signifikantniveau davon ausgegangen werden, dass der Median von X kleiner ist als der von Y (siehe Fall a)). Umgekehrt kann davon auf dem gleichen Signifikanzniveau davon ausgegangen werden, dass der Median von X größer ist als der von Y, wenn $P(T_W \geq r)$ höchstens den Wert α annimmt (siehe Fall b)). Wird überprüft, ob sich die Mediane überhaupt unterscheiden, unabhängig davon, welcher nun größer ist, so wird auf dem Signifikanzniveau α davon ausgegangen, dass dem so ist, wenn $P(T_W \leq r) + P(T_W \geq r) \leq \alpha$ (siehe Fall c)).

Anmerkung:
In Wilcoxons Paper wurde zunächst nur der Spezialfall mit zwei gleich großen

Teilstichproben (also $n = m$) betrachtet. Für diesen Fall nutzte Wilcoxon die Äquivalenz der Anzahl an Partition eines bestimmten Werts aus einer bestimmten Anzahl (hier: n) an ungleichen Summanden mit einer maximalen Größe (hier: $2n$) zu einer anderen Partition aus. Seien die möglichen Rangsummen aufsteigend mit den Zahlen $t = 0, 1, 2, 3, ...$ bezeichnet, so gilt bis zur Rangsumme mit der Nummerierung 1 bis n für die Anzahl an Partitionen, dass sie äquivalent zur grundsätzlichen Anzahl an Partitionen von t sind. Für die niedrigste mögliche Rangsumme ($t = 0$) gilt, dass sie nur auf eine mögliche Art zu erreichen ist.

Sei als Beispiel $n = m = 4$, so ist die niedrigste mögliche Rangsumme gleich $1 + 2 + 3 + 4 = 10$.

- 11 als Partition mit 4 Teilen ≤ 8 zu erhalten, ist nur auf eine Weise möglich, da es für das dazugehörige $t = 11 - 10 = 1$ nur eine Partition gibt: 1 selbst
- Für 12 gibt es 2 Möglichkeiten, da 2 aus 2 Partitionen gebildet werden kann: 2 selbst und $(1 + 1)$
- Für 13 gibt es 3 Möglichkeiten, da 3 aus 3 Partitionen gebildet werden kann: 3 selbst, $(1 + 2)$ und $(1 + 1 + 1)$
- Für 14 gibt es 5 Möglichkeiten, da 4 aus 5 Partitionen gebildet werden kann: 4 selbst, $(2 + 2)$, $(1 + 3)$, $(1 + 1 + 2)$ und $(1 + 1 + 1 + 1)$
- Für größere Rangsummen ist $t > n$ und die Äquivalenz der beiden Partitionen ist nicht mehr gegeben

Für die weitere Ausführung der exakten Berechnung der Teststatistik siehe [3, S.81 und 82]

4.1.2 Wilcoxon-Vorzeichen-Rang-Test

Für den Wilcoxon-Vorzeichen-Rang-Test, angewandt auf die gepaarten Stichproben $X_{(1)}$ und $X_{(2)}$ der Größe n oder die Stichprobe $X_{(1)}$ und den hypothetischen Wert φ, gilt immer Folgendes, was zur Berechnung exakter Wahrscheinlichkeiten für gewisse Rangsummen ausgenutzt werden kann bzw. benötigt wird:

1. Der niedrigste mögliche mögliche Testwert für W^+ ist 0. Dieser kann nur erreicht werden, wenn alle Differenzen (und somit Ränge) negativ sind
2. Der höchste mögliche Testwert für W^+ ist $W^+_{max} = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$. Dieser kann nur erreicht werden, wenn alle Differenzen (und somit Ränge) positiv sind
3. Damit gibt es $W^+_{max} + 1$ verschiedene mögliche Rangsummen
4. Insgesamt gibt es 2^n unterschiedliche Möglichkeiten, die $W^+_{max} + 1$ verschiedenen Rangsummen zu erhalten

Wie beim Wilcoxon-Rangsummentest fehlt auch hier zur Berechnung der exakten Wahrscheinlichkeiten nur noch die Information, auf wie viele unterschiedliche Möglichkeiten die jeweiligen Rangsummen zu erreichen sind. Während beim Wilcoxon-Rangsummentest allerdings nur von Interesse war, wie diese Rangsummen als Summe einer einzigen bestimmten Anzahl an Summanden zu erhalten sind, so muss beim Wilcoxon-Vorzeichen-Rang-Test beachtet werden, dass die Rangsummen hier als Summe von bis zu n (nur für den Fall, dass $W^+ = W^+_{max}$) Summanden entstehen können. Das heißt, dass für jede mögliche Rangsumme überprüft werden muss, wie viele Möglichkeiten es gibt, diese Summe aus bis zu n unterschiedlichen Summanden darzustellen.
Auch hier kann der selbe 'Trick' angewendet werden wie im Rangsummentest von Wilcoxon.
Sei w ein bestimmter Testwert W^+, $j \in \mathbb{N}$ die Anzahl an ungleichen Summanden und n die maximale Wert für einen Summanden, so ist die Anzahl an gleichen oder ungleichen Partitionen von $w - \binom{j}{2}$ aus j Teilen $(< n - j + 1)$ gleich der Anzahl an an ungleichen Partitionen von w aus j Teilen $(< n)$.

Somit ist die Wahrscheinlichkeit für einen bestimmten Testwert w gleich

$$P(W^+ = w) = \left(\frac{\sum_{j=2}^{n} F_P\left(w-\binom{j}{2},j,n-j+1\right)+q}{2^n}\right)$$

Da $\binom{j}{2}$ für $j = 1$ nicht definiert ist, beginnt die Summe erst bei 2. Dies hat zur Folge, dass die Summe Partitionen mit nur einem Element nicht berücksichtigt. Falls $w \leq n$, gibt es diese Partition jedoch und somit auch eine Möglichkeit mehr, die Rangsumme w zu erhalten. Damit gilt für q:

$$q = \begin{cases} 1, & w \leq n \\ 0, & \text{sonst} \end{cases}$$

Für die Testwerte $w = \{0; 1; 2\}$ verkürzt sich die Formel zu $P(W^+ = w) = \frac{1}{2^n}$.

Für den Fall $P(W^+ \leq w)$ muss diese Formel noch ein wenig erweitert werden, da hier die eben erwähnte Wahrscheinlichtkeit für jeden Wert kleiner oder gleich w berechnet und aufsummiert wird:

$$P(W^+ \leq w) = \left(\frac{\sum_{i=3}^{w}\sum_{j=2}^{n} F_P\left(i-\binom{j}{2},j,n-j+1\right)+Q}{2^n}\right)$$

Die äußere Summe beginnt erst bei 3, da Partitionen mit ungleichen Summanden für die Werte $w \leq 2$ ohnehin nur aus einem Teil j bestehen können, nämlich $w = j$. Besagte Partitionen sind in der Formel mit Q abgedeckt: Q ist hier ähnlich zu q im Fall $P(W^+ = w)$. Es ist nichts weiter als alle Partitionen mit nur einem Element addiert und damit gleich:

$$Q = \begin{cases} n + 1, & w > n \\ w + 1, & \text{sonst} \end{cases}$$

Für die Testwerte $w = \{0; 1; 2\}$ verkürzt sich auch diese Formel, nämlich zu $P(W^+ \leq w) = \frac{w+1}{2^n}$.

$P(W^+ > w)$ berechnet sich als Gegenwahrscheinlichkeit wieder als $1 - P(W^+ \leq w)$ und $P(W^+ \geq w) = P(W^+ > w-1)$ entsprechend als $1 - P(W^+ \leq w-1)$.

4.2 Anwendung

4.2.1 Wilcoxon-Rangsummentest

Voller Freude über seinen grandiosen Erfolg von nur einem Fehlerpunkt will ein Teilnehmer nach Beendigung der letzten Sitzung des Sprachkurses noch in einer Kneipe mit anderen Mitstreitern anstoßen. Zu fünft besuchen sie die nächste Gaststätte und sitzen sich gegenüber, $n = 3$ Teilnehmer auf einer Sitzbank, $m = 2$ auf Stühlen. Der Spitzenreiter, auf der Bank sitzend, wird nach einigen Drinks übermütig und gibt an: "Wir auf der Bank haben alle bessere Prüfungsergebnisse als ihr auf den Stühlen, das kann doch kein Zufall sein!"

Tatsächlich sind die drei besten Ergebnisse unter den fünf von denjenigen, die auf der Bank sitzen, was eine Rangsumme von $T_W = 1 + 2 + 3 = 6$ ausmacht, die niedrigste, die möglich ist. Da es insgesamt $\binom{5}{3} = 10$ unterschiedliche Möglichkeiten gibt, die verschiedenen Ränge auf die Teilnehmer zu verteilen, ist die Wahrscheinlichkeit dafür, genau diese Rangsumme zu erreichen gleich $P(T_w = 6) = \frac{1}{10} = 0,1$. Und da es in diesem Fall keine niedrigere Rangsumme gibt, ist diese Wahrscheinlichkeit auch gleich der Wahrscheinlichkeit $P(T_w \leq 6)$. Hier sieht man: Aufgrund des niedrigen Stichprobenumfangs ist bei diesem Test kein Signifikanzniveau von unter 10% möglich. Eine Übersicht über die Wahrscheinlichkeiten für sämtliche mögliche Rangsummen r ist in der folgenden Tabelle zu finden:

r	**Möglichkeiten für r**	$\#(r)$	$P(T_W = r)$	$P(T_W \leq r)$
6	{1;2;3}	1	0,1	0,1
7	{1;2;4}	1	0,1	0,2
8	{1;2;5},{1;3;4}	2	0,2	0,4
9	{1;3;5},{2;3;4}	2	0,2	0,6
10	{1;4;5},{2;3;5}	2	0,2	0,8
11	{2;4;5}	1	0,1	0,9
12	{3;4;5}	1	0,1	1

Tabelle 5: Wahrscheinlichkeiten für sämtliche Realisationen von r

Die dritte Spalte gibt die Anzahl der Möglichkeiten für eine Realisation r an, die vierte die Wahrscheinlichkeit, ebendiese Realisation der Rangsumme zu erreichen und die fünfte und letzte Spalte die Wahrscheinlichkeit, höchstens den

Wert r zu erlangen. In diesem Beispiel ist es natürlich noch trivial, die einzelnen Möglichkeiten für die unterschiedlichen Testwerte r zu ermitteln. Es sei trotzdem kurz veranschaulicht, wie die Anzahl der unterschiedlichen Möglichkeiten für einen Testwert w alternativ berechnet werden kann. Als Beispiel sei die Rangsumme $r = 9$ herangezogen. Unterschiedliche Möglichkeiten, diesen Wert $r = 9$ als Summe von $n = 3$ ungleichen Summanden $\in \mathbb{N}$ und $\leq (n+m) = 5$ gibt es ebenso viele wie es Möglichkeiten gibt, die Zahl $(r - \binom{n}{2}) = 6$ als Summe aus $n = 3$ Summanden, ungleich oder gleich, darzustellen, wobei die Summenaden höchstens den Wert $m + 1 = 3$ annehmen dürfen. Dies ist nur auf zwei verschiedene Arten möglich, nämlich durch 1,2 und 3 und durch 2,2 und 2. Sicher: in diesem Fall erschwerte diese Vorgehensweise die Arbeit sogar noch, bei höheren Werten für r ändert sich dies jedoch.

4.2.2 Wilcoxon-Vorzeichen-Rang-Test

Da die Sprachkursteilnehmer auf den Stühlen ihrem angeberischen Mitstreiter erklärten, dass die Wahrscheinlichkeit für die angeblich überzufällige Überlegenheit der Teilnehmer auf der Sitzbank mit 10% doch nicht allzu niedrig ist, sucht er verzweifelt nach anderen Argumenten: "Dafür haben wir alle drei unsere Fehlerpunkte verringert, das will doch was heißen!"Was der Teilnehmer meint, bedeutet für den Wilcoxon-Vorzeichen-Rang-Test nichts anderes, als dass die 3 Differenzen zwischen den Fehlerpunkten nachher und vorher immer negativ ausfallen und somit auch deren Ränge. Somit ergibt sich für die Teststatistik W^+ (zur Erinnerung: die Summe der positiven Ränge) ein Wert von $w = 0$. Dieser Wert ist nur auf eine Weise zu erreichen und insgesamt gibt es $2^3 = 8$ Möglichkeiten für die Rangverteilung. $P(W^+ = 0)$ ist somit $\frac{1}{8} = 0,125$, ebenso wie $P(W^+ \leq 0)$, da 0 der kleinstmögliche Testwert ist. Durch diese doch recht hohe Wahrscheinlichkeit lassen sich die Teilnehmer von den Sprüchen des Angebers nicht beeindrucken. Dieser ist eingeschnappt und gibt zurück: "Dafür bin ich besser als ihr alle!"

Die Wahrscheinlichkeiten für sämtliche mögliche Rangsummen w sind in folgender Tabelle aufgelistet:

$\boldsymbol{w}$	**Möglichkeiten für** $\boldsymbol{w}$	$\#(\boldsymbol{w})$	$\boldsymbol{P(W^+ = w)}$	$\boldsymbol{P(W^+ \leq w)}$
0	kein Rang > 0	1	0,125	0,125
1	{1}	1	0,125	0,25
2	{2}	1	0,125	0,375
3	{2;1},{3}	2	0,25	0,625
4	{1;3}	1	0,125	0,75
5	{2;3}	1	0,125	0,875
6	{1;2;3}	1	0,125	1

Tabelle 6: Wahrscheinlichkeiten für sämtliche Realisationen von w

Für größere Werte ist es wieder einfacher, den 'Trick' anzuwenden. Hier ist

dies prinzipiell absolut nicht notwendig, da das Beispiel sehr einfach gewählt wurde. Die Frage, die sich für jeden Wert w stellt, ist hierbei: Wie viele unterschiedliche Arten gibt es, jeden Wert w als Summe aus $j = 1, 2, 3$ ungleichen Summanden $\in \mathbb{N} \leq 3$ darzustellen? Dies ist gleichbedeutend mit der Frage, wie viele Möglichkeiten es gibt, den Wert $w - \binom{j}{2}$ als Summe aus j gleichen oder ungleichen Summanden $\in \mathbb{N}$ und ≤ 3 zu erhalten? Da j in diesem Fall größer als 2 sein muss, wurde in der Formel $P(W^+ = w) = (\frac{\sum_{j=2}^{n} F_P(w-\binom{j}{2},j,n-j+1)+q}{2^n})$ der Wert q eingeführt, der hier zu 1 wird, wenn $w \leq 3$, es also die Möglichkeit gibt, w als Summe aus nur einem Summanden $\in \mathbb{N}$ und ≤ 3 darzustellen. Was für jeden Summanden ausgemacht werden muss, sei hier am Beispiel des Wertes $w = 6$ demonstriert:

- Ist 6 als Summe aus einem Summanden $\in \mathbb{N}$ und ≤ 3? $\rightarrow$ Nein, da $6 > 3$
- Wie viele Möglichkeiten gibt es, 6 als Summe aus zwei ungleichen Summanden $\in \mathbb{N}$ und ≤ 3 darzustellen? $\rightarrow$ Keine, da $6 - \binom{2}{2} = 5$ nicht als Summe aus 2 gleichen oder ungleichen Summanden $\in \mathbb{N}$ und $\leq 3 - 2 + 1 = 2$ darzustellen ist
- Wie viele Möglichkeiten gibt es, 6 als Summe aus drei ungleichen Summanden $\in \mathbb{N}$ und ≤ 3 darzustellen? $\rightarrow$ Eine, da sich $6 - \binom{3}{2} = 3$ ebenso auf eine Weise als Summe aus 3 gleichen oder ungleichen Summanden $\in \mathbb{N}$ und $\leq 3 - 3 + 1 = 1$ darstellen lässt, nämlich als $3 = 1 + 1 + 1$

Literatur

[1] S. Kotz and N.L. Johnson. *Breakthroughs in Statistics II.: Methodology and Distribution.* Springer series in statistics: Perspectives in statistics. Springer-Verlag GmbH, 1992.

[2] Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, and Gerhard Tutz, editors. *Statistik*. Springer-Lehrbuch. Springer, Berlin [u.a.], 6., überarb. aufl. edition, 2007.

[3] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 12 1945.