26 Nichtparametrische Tests

Das Menü und auch die Ergebnisausgabe von „Nichtparametrische Tests“ ist seit der Version 18 neu gestaltet. Die Darstellung in der 8. Auflage des Buches bezieht sich ausschließlich darauf. Jedoch können Nutzer, die mit dem alten Menü arbeiten wollen, dies weiterhin über die Option „Alte Dialogfelder“ tun. Für diese Nutzer ist der vorliegende Text gedacht. Dieser entspricht weitgehend dem in der 7. Auflage. Er behandelt die nichtparametrischen Tests, wie Sie im Untermenü „Alte Dialogfelder“ in der SPSS-Version 18 bis 21 dargestellt sind bzw. wie sie in der SPSS-Version 17 und früher im Menü „Nichtparametrische Tests“ gestaltet waren.

26.1 Einführung und Überblick

Überblick über die Tests in SPSS. Aus der Übersicht in Abb. 26.1 kann man entnehmen, welche nichtparametrischen Tests von SPSS bereitgestellt werden. Die Reihenfolge orientiert sich an der im Menü „Nichtparametrische Tests“ in SPSS. Es wird im Überblick kurz angeführt, welchen Testzweck die einzelnen Tests haben, welches Messniveau für die Variablen erforderlich ist, um wie viel Stichproben es sich handelt und ob es sich um ein Design von unabhängigen oder verbundenen Stichproben handelt.

Exakte Tests. Für die Anwendungen mit SPSS Base werden bei den einzelnen Tests Prüfgrößen berechnet und theoretische Verteilungen dienen zur Signifikanzprüfung. Aber nicht immer sind die Bedingungen dafür gegeben, dass die Verteilung der Prüfgrößen hinreichend durch die theoretischen Verteilungen approximiert werden dürfen. SPSS für Windows bietet daher in Ergänzung zum Basismodul das Modul „Exakte Tests“ an. Nach Installation dieses Moduls steht in den Dialogboxen zur Durchführung nichtparametrischer Tests zusätzlich eine Schaltfläche „Exakt...“ zur Verfügung. Durch Klicken auf die Schaltfläche kann man die Dialogbox „Exakte Tests“ öffnen und zwischen zwei Verfahren zur Durchführung exakter Tests wählen (ausführlicher ⇨ Kap. 31).

26.2 Tests für eine Stichprobe

26.2.1 Chi-Quadrat-Test (Anpassungstest)

Der Chi-Quadrat-Test ist schon im Zusammenhang mit der Kreuztabellierung behandelt worden (⇨ Kap. 10.3). Dort geht es um die Frage, ob zwei nominalskalierte Variable voneinander unabhängig sind oder nicht (Chi-Quadrat-Unabhängigkeitstest).

Hier geht es um die Frage, ob sich für eine Zufallsstichprobe eine (nominal- oder ordinalskalierte) Variable in ihrer Häufigkeitsverteilung signifikant von erwarteten Häufigkeiten der Grundgesamtheit unterscheidet (Anpassungs- bzw. „Goodness of Fit“-Testtyp). Die erwarteten Häufigkeiten können z.B. gleichverteilt sein oder einer anderen Verteilung folgen.

Das folgende Beispiel bezieht sich auf Befragungsdaten der Arbeitsgruppe Wahlforschung an der Hamburger Hochschule für Wirtschaft und Politik zur Vorhersage der Wahlergebnisse für die Bürgerschaft der Freien und Hansestadt Hamburg im Herbst 1993 (Datei WAHLEN2.SAV). Unter anderem wurde gefragt, welche Partei zur Bürgerschaftswahl 1991 gewählt worden ist.

Tabelle 26.1. Übersicht über nichtparametrische Tests von SPSS

Testname	**Mess-niveau***	**Testzweck**	**Anzahl der Stichproben**	**Stichpro-bendesign#**
1. Chi-Quadrat	n	Empirische gleich erwartete Häufigkeit?	1	-
2. Binomial	d	Empirische Häufigkeit binomialverteilt?	1	-
3. Sequenzanalyse	d	Reihenfolge der Variablenwerte zufällig?	1	-
4. Kolmogorov-Smirnov	o	Empirische Verteilung gleich theoretischer?	1	-
5. Mann-Whitney U	o	2 Stichproben aus gleicher Verteilung?	2	u
6. Moses	o	2 Stichproben aus gleicher Verteilung?	2	u
7. Kolmogorov-Smirnov Z	o	2 Stichproben aus gleicher Verteilung?	2	u
8. Wald-Wolfowitz	o	2 Stichproben aus gleicher Verteilung?	2	u
9. Kruskal-Wallis H	o	k Stichproben aus gleicher Verteilung?	k	u
10. Median	o	2 oder k Stichproben aus Verteilung mit gleichem Median?	2 bzw. k	u
11. Jonckheere-Terpstra	o	k Stichproben aus gleicher Verteilung. Für geordnete Verteilungen	k	u
12. Wilcoxon	o	2 verbundene Stichproben aus gleicher Verteilung?	2	v
13. Vorzeichen	o	2 verbundene Stichproben aus gleicher Verteilung?	2	v
14. McNemar	d	2 Stichpr. verändert im Vorher/Nachher-Design	2	v
15. Marginale Homogenität	n	2 Stichpr. verändert im Vorher/Nachher-Design	2	v
16. Friedman	o	k verbundene Stichpr. aus gleicher Verteilung?	k	v
17. Kendall´s W	o	k verbundene Stichpr. aus gleicher Verteilung?	k	v
18. Cochran Q	d	k verbundene Stichpr. mit gleichem Mittelwert?	k	v

* n = nominal, o = ordinal, d = dichotom

\# u = unabhängig, v = verbunden

Die Verteilung dieser Variable (mit PART_91 bezeichnet) mit den Werten 1 bis 7 (für die Parteien SPD, CDU, Grüne/GAL, F.D.P., Republikaner und Sonstige; der Wert 6 kommt nicht vor) soll mit den tatsächlichen Wahlergebnissen in 1991 für diese Parteien verglichen und getestet werden, ob sich ein signifikanter Unterschied in den Verteilungen ergibt. Ergibt sich ein signifikanter Unterschied, so könnte das als ein Indikator dafür gesehen werden, dass die Stichprobe nicht hinreichend repräsentativ ist. Die Hypothese H_0 lautet also, die Stimmenverteilung auf die Parteien in der Stichprobe entspricht dem Ergebnis der Bürgerschaftswahl. Entsprechend lautet die H_1-Hypothese, dass die Verteilungen signifikant unterschiedlich sind. Nach Öffnen der Datei WAHLEN2.SAV gehen Sie wie folgt vor:

- ▹ Wählen Sie die Befehlsfolge „Analysieren“, „Nichtparametrische Tests ▹“; „Chi-Quadrat···“. Es öffnet sich dann die in Abb. 26.1 wiedergegebene Dialogbox.
- ▹ Aus der Quellvariablenliste wird die Testvariable PART_91 in das Eingabefeld „Testvariablen:“ übertragen. Sollen für weitere Variablen Tests durchgeführt werden, so sind auch diese zu übertragen.
- ▹ Die gewählte Option „Aus den Daten“ in der Auswahlgruppe „Erwarteter Bereich“ bedeutet, dass der gesamte Wertebereich der Variablen (hier: 1 bis 7) für den Test benutzt wird. Soll nur ein Teilwertebereich für den Test ausgewertet werden, so kann dieses mit der Option „Angegebener Bereich verwenden“ geschehen, indem man in das Eingabefeld „Minimum“ den kleinsten (z.B. 1) und in das Eingabefeld „Maximum“ den größten Wert (z.B. 4) eingibt.
- ▹ In „Erwartete Werte“ kann man aus zwei Optionen auswählen. „Alle Kategorien gleich“ wird man wählen, wenn die gemäß der Hypothese H_0 erwarteten Häufigkeiten der Kategorien der Variablen (hier die Parteien) gleich sind (Gleichverteilung). Für unser Beispiel ist die Option „Werte“ relevant. In das Eingabefeld von Werte gibt man die gemäß der H_0-Hypothese erwarteten Häufigkeiten für die Kategorien (Parteien) ein. Wichtig ist, dass sie in der Reihenfolge entsprechend der Kodierung der Variable, beginnend mit dem kleinsten Wert (hier: 1 für SPD), eingegeben werden. Mit „Hinzufügen“ werden die jeweils in das Werte-Eingabefeld eingetragenen Häufigkeiten nacheinander in das darunter liegende Textfeld übertragen. Die erwarteten Werte können sowohl als prozentuale als auch absolute Häufigkeiten eingegeben werden. Die in der Abb. 23.1 sichtbaren Eintragungen ergeben sich daraus, dass bei der Bürgerschaftswahl 1991 die SPD 48,0 %, die CDU 35,1 %, die Grünen/GAL 7,2 %, die FDP 5,4 % und Sonstige 3,1 % Stimmenanteile erhalten haben [da in der Datei für den codierten Wert 5 (für Republikaner) keine Fälle enthalten sind, darf man den Stimmenanteil der Republikaner nicht angeben, weil sonst von SPSS der Test mit einer Fehlermeldung abgebrochen wird]. Hat man sich bei schon eingegebenen Werten vertan, so kann man sie markieren und mittels „Entfernen“ aus dem Textfeld entfernen.

.

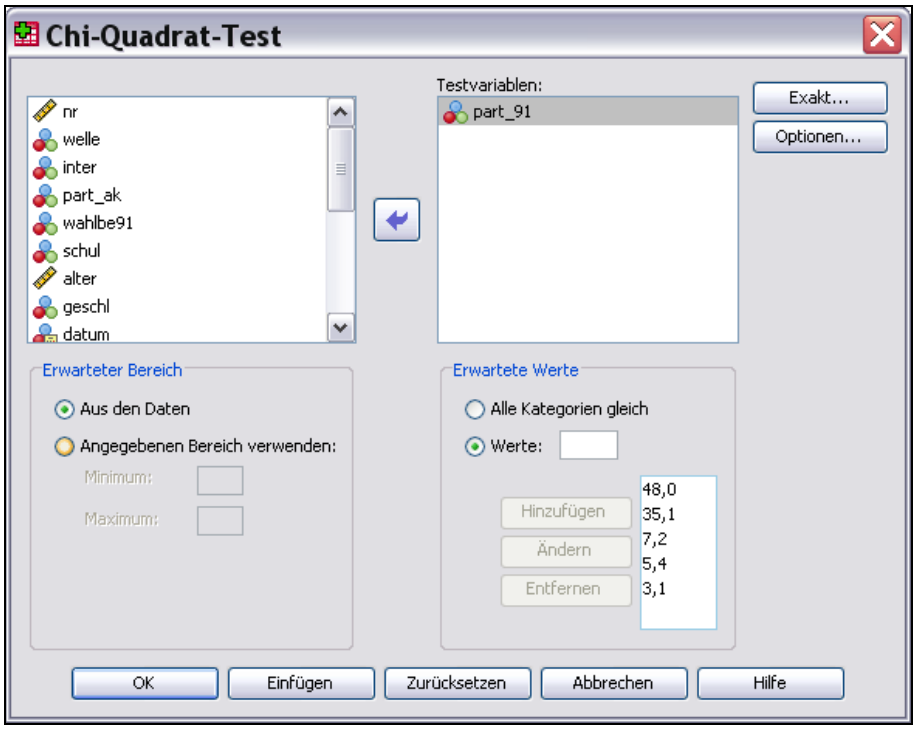

Abb. 26.1. Dialogbox „Chi-Quadrat-Test“

In Tabelle 26.2 ist das Ergebnis des Chi-Quadrat-Tests niedergelegt. Für die Parteien werden die empirischen („Beobachtetes N“) und erwarteten („Erwartete Anzahl“) Häufigkeiten sowie die Abweichungen dieser („Residuum“) aufgeführt. Die erwarteten Häufigkeiten unter H_0 ergeben sich durch Multiplikation der Fallanzahl mit dem Stimmenanteil für eine Partei. Werden mit n_i die empirischen und mit e_i die erwarteten Häufigkeiten einer Kategorie bezeichnet, so ergibt sich für die Prüfgröße Chi-Quadrat (die Summierung erfolgt über die Kategorien i = 1 bis k (hier: k = 5)

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} = 19{,}32 \qquad (26.1)$$

Aus der Formel wird ersichtlich, dass die Testgröße χ^2 umso größer wird, je stärker die Abweichungen zwischen beobachteten und erwarteten Häufigkeiten sind. Ein hoher Wert für χ^2 ist folglich ein Ausdruck für starke Abweichungen in den Verteilungen. Je größer der χ^2-Wert ist, umso unwahrscheinlicher ist es, dass die Stichprobe aus der Vergleichsverteilung stammt. Die Prüfgröße χ^2 ist asymptotisch chi-quadratverteilt mit k-1 Freiheitsgraden (df = degrees of freedom). Für eine gegebene Anzahl von Freiheitsgraden und einem Signifikanzniveau α (Irrtumswahrscheinlichkeit die H_0-Hypothese abzulehnen, obwohl sie richtig ist) lassen sich aus einer Chi-Quadrat-Verteilungstabelle[1] kritische Werte für χ^2 entnehmen. Für fünf Kategorien in unserem Beispiel ist df = 4. Bei einem

[1] Die Tabelle ist auf den Internetseiten zum Buch verfügbar.

Signifikanzniveau von $\alpha = 0{,}05$ (5 % Irrtumswahrscheinlichkeit) und df = 4, ergibt sich aus einer tabellierten Chi-Quadrat-Verteilung für $\chi^2_{krit} = 9{,}488$. Der empirische Wert von χ^2 fällt in den Ablehnungsbereich der H_0-Hypothese, da er mit 19,32 größer ist als der kritische. „Asymptotische Signifikanz" (= 0,001) ist die Wahrscheinlichkeit, bei df = 4 ein $\chi^2 \geq 19{,}32$ zu erhalten. Auch daraus ergibt sich, dass bei einem Signifikanzniveau von 5 % ($\alpha = 0{,}05$) die H_0-Hypothese abzulehnen ist (0,05 > 0,001). Die Stimmenverteilung auf die Parteien in der Stichprobe entspricht demnach nicht der tatsächlichen für 1991.

Tabelle 26.2. Ergebnisausgabe eines Chi-Quadrat-Tests

part_91

	Beobachtetes N	Erwartete Anzahl	Residuum
SPD	243	217,3	25,7
CDU	122	160,7	-38,7
Grüne/Gal	47	32,6	14,4
FDP	27	24,4	2,6
Sonstige	10	14,0	-4,0
Gesamt	449		

Statistik für Test

	part_91
Chi-Quadrat	20,160[a]
df	4
Asymptotische Signifikanz	,000

a. Bei 0 Zellen (,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 14,0.

Optionen. Durch Klicken auf „Optionen" öffnet sich die in Abb. 26.2 dargestellte Dialogbox mit der optionale Vorgaben festgelegt werden können:

- *Statistik.* Mit „Deskriptive Statistik" können das arithmetische Mittel, die Standardabweichung sowie das Minimum und das Maximum angefordert werden. Mit „Quartile" werden der Wert des 25., 50. (= Median) und 75. Perzentils berechnet.
- *Fehlende Werte.* Mit „Fallausschluss Test für Test" werden beim Testen mehrerer Variablen die fehlenden Werte jeweils für die einzelne Testvariable und mit „Listenweiser Fallausschluss" für alle Tests ausgeschlossen.

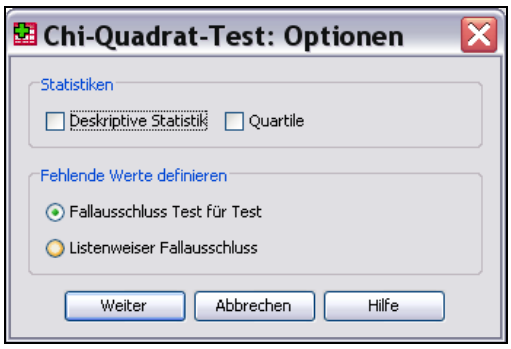

Abb. 26.2. Dialogbox „Chi-Quadrat-Test: Optionen"

Exakte Tests. Sollte man verwenden, wenn die Anwendungsbedingungen einen asymptotischen Chi-Quadrat-Test verbieten (⇨ Kap. 31).

Anwendungsbedingungen. Für den asymptotischen Chi-Quadrat-Test sollten folgende Anwendungsbedingungen beachtet werden: im Falle von df = 1 sollte e_i

≥ 5 für alle Kategorien i sein. Für df > 1 sollte $e_i \leq 5$ für nicht mehr als 20 % der Kategorien i und $e_i \geq 1$ für alle i sein (⇨ Fußnote in Tabelle 26.2 rechts).

Warnung. Der Chi-Quadrat-Test führt zu unsinnigen Ergebnissen, wenn die Fälle mit einer Variablen gewichtet werden, deren Werte Dezimalzahlen sind (z.B. 0,85, 1,20). In Version 17 wird eine Gewichtung ignoriert.

26.2.2 Binomial-Test

Eine Binomialverteilung ist eine Wahrscheinlichkeitsverteilung für eine diskrete Zufallsvariable, die nur zwei Werte annimmt (dichotome Variable). Mit Hilfe der Binomialverteilung lässt sich testen, ob ein prozentualer Häufigkeitsanteil für eine Variable in der Stichprobe mit dem der Grundgesamtheit vereinbar ist. Das oben verwendete Beispiel zur Wahlvorhersage (WAHLEN2.SAV, ⇨ Kap 26.2.1) soll dieses näher erläutern. Geprüft werden soll, ob der prozentuale Männeranteil in der Stichprobe mit dem in der Grundgesamtheit - alle Wahlberechtigten für die Hamburger Bürgerschaft - vereinbar ist oder nicht. Dazu gehen Sie wie folgt vor:

- ▹ Wählen Sie die Befehlsfolge „Analysieren“, „Nichtparametrische Tests ▹“, „Binomial···“· Es öffnet sich die in Abb. 26.3 dargestellte Dialogbox.
- ▹ Aus der Quellvariablenliste wird die Variable GESCHL in das Eingabefeld von „Testvariablen:“ übertragen. Sollen mehrere Variablen getestet werden, so sind diese alle in das Variableneingabefeld zu übertragen.
- ▹ In „Dichotomie definieren“ bestehen alternative Auswahlmöglichkeiten:
  - „Aus den Daten“ ist zu wählen, wenn - wie es in diesem Beispiel der Fall ist - die Variable dichotom ist.
  - „Trennwert“ ist zu wählen, wenn eine nicht-dichotome Variable mit Hilfe des einzugebenden Trennwertes dichotomisiert wird. Beispielsweise lässt sich die Variable ALTER durch „Trennwert“ = 40 in eine dichotome Variable verwandeln: bis einschließlich 40 haben alle Befragten den gleichen Variablenwert und ab 41 einen anderen Wert.
- ▹ In das Eingabefeld „Testanteil:“ wird der Anteilswert gemäß H_0-Hypothese für die Grundgesamtheit in dezimaler Form eingegeben. Die Männerquote für die Wahlberechtigten für die Bürgerschaft beträgt 48,3 % (einzugeben ist 0,483).

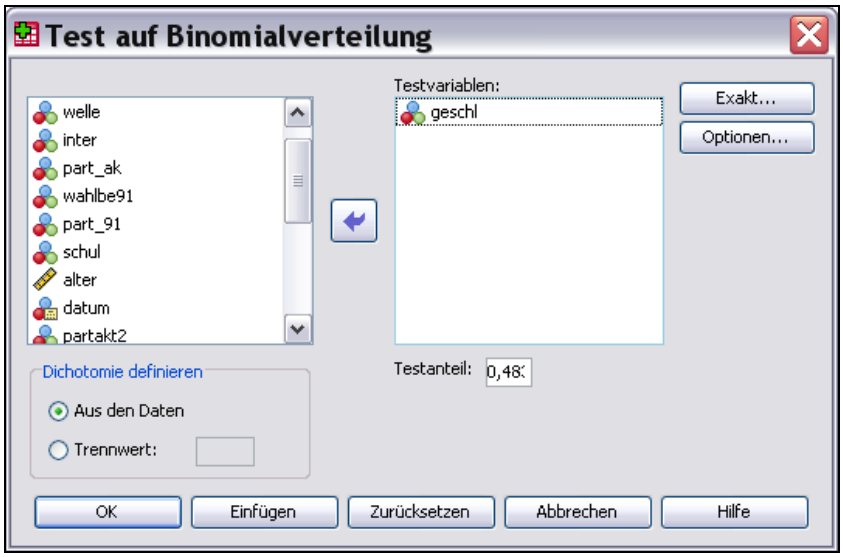

Abb. 26.3. Dialogbox „Binomial-Test“

In Tabelle 26.3 ist das Ergebnis des Binomial-Tests zu sehen. Die empirische Männerquote („Beobachteter Anteil“) beträgt 0,469 im Vergleich zur vorgegebenen Quote (0,483). Da der Stichprobenumfang hinreichend groß ist, wird die Binomialverteilung durch eine Normalverteilung approximiert. Der Test kann dann vereinfachend mittels der standardnormalverteilten Variable Z vorgenommen werden. Ergebnis ist, dass unter der H_0-Hypothese (eine Männerquote von 0,483 für die Wahlberechtigten) eine Wahrscheinlichkeit („Asymptotische Signifikanz, 1-seitig“) von 0,268 besteht, dass die Männerquote gleich bzw. kleiner als die beobachtete ist. Bei einem Signifikanzniveau von 5 % ($\alpha = 0{,}05$) wird wegen 0,268 > 0,05 die Hypothese H_0 nicht verworfen.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakte Tests. ⇨ Kap. 31.

Tabelle 26.3. Ergebnisausgabe des Binomial-Tests

Test auf Binomialverteilung

		Kategorie	N	Beobachteter Anteil	Testanteil	Asymptotische Signifikanz (1-seitig)
geschl	Gruppe 1	männlich	246	,469	,483	,268[a,b]
	Gruppe 2	weiblich	279	,531		
	Gesamt		525	1,000		

a. Nach der alternativen Hypothese ist der Anteil der Fälle in der ersten Gruppe < ,483.

b. Basiert auf der Z-Approximation.

26.2.3 Sequenz-Test (Runs-Test) für eine Stichprobe

Dieser Test ermöglicht es zu prüfen, ob die Reihenfolge der Werte einer Variablen in einer Stichprobe (und damit die Stichprobe) zufällig ist (H_0-Hypothese). Angewendet wird der Test z.B. in der Qualitätskontrolle und bei Zeitreihenanalysen.

Im folgenden Beispiel für eine Stichprobe mit einem Umfang von 20 sei eine (dichotome) Variable mit nur zwei Ausprägungen (hier dargestellt als + und −) in einer Reihenfolge gemäß Tabelle 26.4 erhoben. Diese Stichprobe hat eine Sequenz (runs) von 8, da achtmal gleiche (positive bzw. negative) Werte aufeinander folgen.

Tabelle 26.4. Beispiel für acht Sequenzen bei einem Stichprobenumfang von 20

++	- - -	+	- -	+ + + +	-	+ + +	- - - -
1	2	3	4	5	6	7	8

Wären die Merkmalswerte „+“ bzw. „-“ z.B. Zahl bzw. Wappen bei 20 aufeinander folgenden Würfen mit einer Münze, so kann die Sequenz der Stichprobe Hinweise hinsichtlich der „Fairness“ der Münze geben, die durch Feststellung einer „Wappen-Quote“ in der Stichprobe von ca. 50 % verdeckt bleiben würde. Die Erfassung von Sequenzen beschränkt sich nicht auf schon im Stadium der Messung dichotome Variablen, da Messwerte von Variablen in dichotome verwandelt werden können, indem festgehalten wird, ob die Messwerte kleiner oder größer als ein bestimmter Messwert (z.B. das arithmetische Mittel) sind.

Die Stichprobenverteilung der Anzahl von Sequenzen (= Prüfgröße) ist bekannt. Für große Stichproben ist die Prüfgröße approximativ standardnormalverteilt.

Beispiel. Im Folgenden soll getestet werden, ob die Stichprobe für die Wahlprognose (Datei WAHLEN2.SAV, ⇨ Kap. 26.2.1) zufällig ist. Als Testvariable wird das Alter der Wähler gewählt. Zur Durchführung des Tests gehen Sie wie folgt vor:

▹ Wählen Sie die Befehlsfolge „Analysieren“, „Nichtparametrische Tests ▹", „Sequenzen...“. Es öffnet sich die in Abb. 26.4 dargestellte Dialogbox.

▹ Aus der Quellvariablenliste wird die Variable ALTER in das Eingabefeld „Testvariablen:“ übertragen. Zur Dichotomisierung der Variablen stehen im Feld „Trennwert“ vier Optionen zur Verfügung:
- *Median:* Zentralwert.
- *Modalwert:* häufigster Wert.
- *Mittelwert :* arithmetisches Mittel.
- *Benutzerdefiniert:* vom Anwender vorgegebener Wert.

▹ In unserem Beispiel wird „Median“ gewählt. Dadurch erhält die Variable ALTER zur Ermittlung der Sequenz nur zwei Merkmalsausprägungen: kleiner als der Median und größer bzw. gleich dem Median.

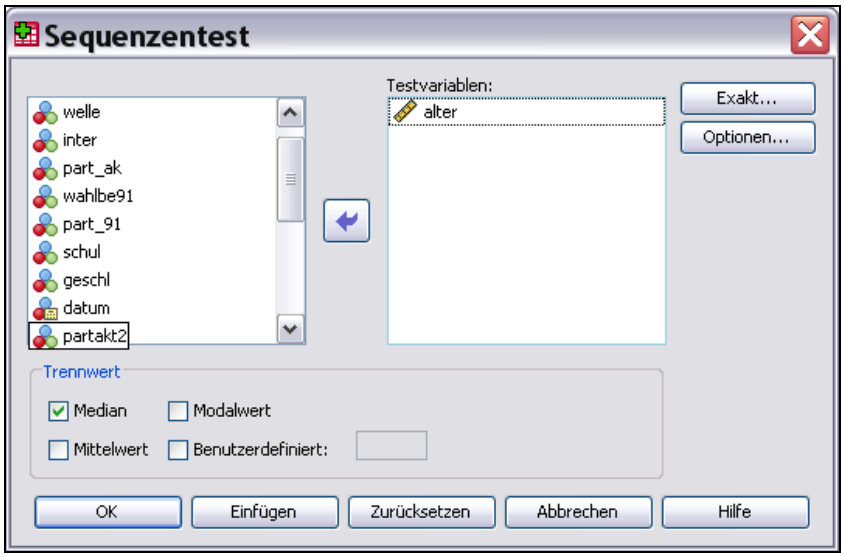

Abb. 26.4. Dialogbox „Sequenzanalyse“

In Tabelle 26.5 ist das Ergebnis des Tests zu sehen. Bei einem Stichprobenumfang in Höhe von 529 werden 158 Sequenzen ermittelt. 261 Befragte haben ein Alter kleiner und 268 größer bzw. gleich als der Median in Höhe von 51 Jahren. Für den Z-Wert der standardisierten Normalverteilung in Höhe von 9,354 ergibt sich die zweiseitige asymptotische Wahrscheinlichkeit in Höhe von 0,000. Die Anzahl der Sequenzen ist derart niedrig, dass die H_0-Hypothese (die Reihenfolge der Befragten ist zufällig) abgelehnt wird (wegen Irrtumswahrscheinlichkeit $\alpha = 0{,}05 > 0{,}000$).

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.5. Ergebnisausgabe eines Sequenzen-Tests

Sequenzentest

	alter
Testwert [a]	51
Fälle < Testwert	261
Fälle >= Testwert	268
Gesamte Fälle	529
Anzahl der Sequenzen	158
Z	-9,354
Asymptotische Signifikanz (2-seitig)	,000

a. Median

26.2.4 Kolmogorov-Smirnov-Test für eine Stichprobe

Wie der oben angeführte Chi-Quadrat-Test und der Binomial-Test hat auch der Kolmogorov-Smirnov-Test die Aufgabe zu prüfen, ob die Verteilung einer Stichprobenvariable die einer theoretischen Verteilung entspricht oder nicht (Anpassungstest). Der Kolmogorov-Smirnov-Test kann, im Unterschied zum Chi-Quadrat-Test, auch für kleine Stichproben angewendet werden (für kleine Stichproben ist meistens nicht gewährleistet, dass 20 % der Zellen eine erwartete Häufigkeit von mindestens 5 haben). Zudem ist der Kolmogorov-Smirnov-Test ein Anpassungstest für eine metrische Variable.

Zu beachten ist, dass für den Test die Parameter der theoretischen Verteilung (also Mittelwert und Standardabweichung der Grundgesamtheit für den Fall der Prüfung auf Normalverteilung) bekannt sein sollten. Per Syntax können diese für die Berechnung bereit gestellt werden.[2] Wird der Test per Menü und damit ohne Übergabe von Grundgesamtheitsparametern angewendet, so werden diese aus den Daten geschätzt. Aber dadurch verliert der Test an Trennschärfe (Testsstärke). Da in der Regel die Parameter unbekannt sind, sollte man zur Prüfung auf Normalverteilung den Kolmogorov-Smirnov-Test mit der Lilliefors-Korrektur verwenden (⇨ Kap. 9.3.2).

Dieser Test basiert auf der kumulierten empirischen sowie kumulierten erwarteten (theoretischen) Häufigkeitsverteilung. Die größte Differenz (D_{max}) zwischen beiden kumulierten Verteilungen und der Stichprobenumfang gehen in die Prüfgröße Z nach Kolmogorov-Smirnov ein ($KS - Z = \sqrt{n} * D_{max}$). Aus Tabellen kann man für einen gegebenen Stichprobenumfang n kritische Werte für D_{max} bei einem vorgegebenem Signifikanzniveau entnehmen (Siegel, 1956, S. 251).

Für die Befragung zur Wahlprognose für die Bürgerschaftswahl im Herbst 1993 (Datei WAHLEN2.SAV, ⇨ Kap. 26.2.1) soll geprüft werden, ob das Alter der Befragten vereinbar ist mit der Hypothese H_0: die Stichprobe stammt aus einer Grundgesamtheit mit normalverteiltem Alter (es wird hier ignoriert, dass die Grundgesamtheit der Wahlberechtigten tatsächlich nicht normalverteilt ist). Das Alter hat ein metrisches Messniveau. Der Kolmogorov-Smirnov-Test ist aber auch für ordinalskalierte Variablen anwendbar. Hier wenden wir zur Demonstration den Test mit Hilfe des Menüs an. Für die Verteilung des Alters der Wahlberechtigten ist uns der Mittelwert und die Standardabweichung nicht bekannt. Wir verweisen aber nochmals auf die geminderte Trennschärfe des Tests für diese Form der Anwendung.

Sie gehen wie folgt vor:

- ▹ Klicken Sie die Befehlsfolge „Analysieren“, „Nichtparamametrische Tests ▹“„K-S bei einer Stichprobe...“. Es öffnet sich die in Abb. 26.5 dargestellte Dialogbox.
- ▹ Die Testvariable ALTER wird in das Eingabefeld „Testvariablen“ übertragen.
- ▹ Die Testverteilung ist in diesem Beispiel die Normalverteilung. Daher wird in „Testverteilung“ diese ausgewählt. Als alternative theoretische Testverteilungen sind die Gleich-, die Poisson- und Exponentialverteilung wählbar.

[2] Siehe NPAR TESTS in der Command Syntax Reference in der Hilfe.

Abb. 26.5. Dialogbox „Ein-Stichproben-Kolmogorov-Smirnov-Test“

In Tabelle 26.6 ist das Ergebnis des Tests zu sehen. Das durchschnittliche Alter der Befragten beträgt 51,07 Jahre mit einer Standardabweichung von 18,48. Mit „Extremste Differenzen“ wird bei „Absolut“(und „Positiv“) D_{max} = 0,0762 angeführt. Die größte negative Abweichung beträgt –0,0418. Es ist $KS - Z = \sqrt{n} * D_{max} = = \sqrt{529} * 0{,}0762 = 1{,}7526$. Die zweiseitige (asymptotische) Wahrscheinlichkeit beträgt 0,004. Bei einem Signifikanzniveau von 5 % (α = 0,05) wird wegen 0,004 < 0,05 die Hypothese H_0 (das Alter ist normalverteilt) abgelehnt.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.6. Ergebnisausgabe des Kolmogorov-Smirnov-Tests zur Prüfung auf Normalverteilung

Kolmogorov-Smirnov-Anpassungstest

		alter
N		529
Parameter der Normalverteilung[a,,b]	Mittelwert	51,07
	Standardabweichung	18,481
Extremste Differenzen	Absolut	0,0762
	Positiv	0,0762
	Negativ	-0,0418
Kolmogorov-Smirnov-Z		1,7526
Asymptotische Signifikanz (2-seitig)		,004

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

26.3 Tests für 2 unabhängige Stichproben

Die folgenden Tests prüfen, ob eine Variable in zwei unabhängig voneinander erhobenen Stichproben aus einer gleichen Grundgesamtheit stammt.

26.3.1 Mann-Whitney U-Test

Dieser Test ist die Alternative zum parametrischen t-Test für den Vergleich von zwei Mittelwerten von Verteilungen (zentrale Tendenz bzw. Lage), wenn die Voraussetzungen für den t-Test nicht erfüllt sind: es liegt keine metrische Skala vor und/oder die getestete Variable ist nicht normalverteilt. Der Test prüft auf Unterschiede hinsichtlich der zentralen Tendenz von Verteilungen. Voraussetzung für den Mann-Whitney-Test ist, dass die getestete Variable mindestens ordinalskaliert ist. Bei dem Test werden nicht die Messwerte der Variablen, sondern Rangplätze zugrunde gelegt. An einem folgenden Beispiel sei das Test-Verfahren zunächst erläutert. Es werden zwei Schülergruppen A und B eines Jahrgangs mit unterschiedlichen Methoden in Mathematik unterrichtet. Schülergruppe B mit $n_1 = 5$ Schülern wird mit einer neuen Methode und die Kontroll-Schülergruppe A mit $n_2 = 4$ Schülern mit der herkömmlichen Methode unterrichtet. Zum Abschluss des Experiments werden Klausuren geschrieben. In der Tabelle 26.7 sind die Ergebnisse für beide Gruppen in erreichten Punkten aufgeführt.

Tabelle 26.7. Erreichte Leistungsergebnisse für zwei Testgruppen

A	21	14	10	24	
B	17	22	18	23	26

Geprüft werden soll, ob die Schülergruppe B eine bessere Leistung erbracht hat. Wegen der kleinen Stichproben und der ordinalskalierten Variable eignet sich hierfür der Mann-Whitney-Test. Da die beiden Gruppen als zwei unabhängige Stichproben aus Grundgesamtheiten interpretiert werden, lassen sich folgende Hypothesen gegenüberstellen:

- H_0-Hypothese: die Variable hat in beiden Grundgesamtheiten die gleiche Verteilung.
- H_1-Hypothese für die hier relevante einseitige Fragestellung: die Variable ist in der Grundgesamtheit B größer als in A.

Zur Prüfung der Nullhypothese werden die Werte beider Stichproben in aufsteigender Reihenfolge unter Aufzeichnung der Gruppenherkunft zusammengefasst (⇨ Tabelle 26.8). Aus der Reihenfolge von Werten aus den beiden Gruppen wird eine Testvariable U nach folgendem Messverfahren ermittelt: Es wird zunächst gezählt, wie viele Messwerte aus der Gruppe B vor jedem Messwert aus der Gruppe A liegen. U ist die Anzahl der Messwerte aus der Gruppe B, die insgesamt vor den Messwerten der Gruppe A liegen. Vor dem Messwert 10 der Gruppe A liegt kein Messwert der Gruppe B. Für den Messwert 14 der Gruppe A gilt gleiches. Vor dem

Messwert 21 der Gruppe A liegen zwei Messwerte der Gruppe B usw. Durch Addition erhält man

$$U = 0 + 0 + 2 + 4 = 6. \quad (26.2)$$

Tabelle 26.8. Rangordnung der Leistungsergebnisse

Messwerte	10	14	17	18	21	22	23	24	26
Gruppe	A	A	B	B	A	B	B	A	B
Rangziffer	1	2	3	4	5	6	7	8	9

Des Weiteren kann U´ ermittelt werden. Zur Ermittlung von U´ wird nach gleichem Schema gezählt, wie viele Messwerte der Gruppe A vor den Messwerten der Gruppe B liegen. Es ergibt sich

$$U' = 2 + 2 + 3 + 3 + 4 = 14. \quad (26.3)$$

Der kleinere Wert der beiden Auszählungen ist die Prüfvariable U. Wegen $U' = n_1 * n_2 - U$ und $U = n_1 * n_2 - U'$ lässt sich der kleinere Wert nach einer Auszählung leicht ermitteln. Der mögliche untere Grenzwert für U ist 0: alle Werte von A liegen vor den Werten von B. Insofern sprechen sehr kleine Werte von U für die Ablehnung der Hypothese H_0. Die Stichprobenverteilung von U ist unter der Hypothese H_0 bekannt. Für sehr kleine Stichproben $(n_1, n_2 < 8)$ gibt es Tabellen. Aus diesen kann man die Wahrscheinlichkeit - für H_0 ein U gleich/kleiner als das empirisch bestimmte U zu erhalten - entnehmen (Siegel, 1956). Für unser Beispiel mit $n_1 = 4$, $n_2 = 5$ und U = 6 ergibt sich eine Wahrscheinlichkeit von P = 0,206. Wenn das Signifikanzniveau auf $\alpha = 0{,}05$ festgelegt wird, kann die Hypothese H_0 nicht abgelehnt werden, da 0,206 > 0,05 ist. Für große Stichproben ist die standardisierte Testgröße U approximativ standardnormalverteilt.

Von *Wilcoxon* ist für gleiche Anwendungsbedingungen ein äquivalenter Test vorgeschlagen worden. Der Test von Wilcoxon ordnet ebenfalls die Werte der zusammengefassten Stichproben nach der Größe. Dann werden Rangziffern vergeben: der kleinste Wert erhält die Rangziffer 1 der nächstgrößte die Rangziffer 2 usw. (⇨ Tabelle 26.8). Schließlich werden für die Fälle einer jeden Gruppe die Rangziffern addiert. Wenn beide Gruppen die gleiche Verteilung haben, so sollten sie auch ähnliche Rangziffernsummen haben. Im obigen Beispiel ergibt sich für Gruppe A eine Rangsumme in Höhe von 16 und für B eine in Höhe von 29. Da die Rangziffernsummen in die Größen U bzw. U´ überführt werden können, führen beide Tests zum gleichen Ergebnis.

Nicht unproblematisch ist es, wenn Mitglieder verschiedener Gruppen gleiche Messwerte haben (im angelsächsischen Sprachraum spricht man von *ties*). Wäre z.B. der größte Messwert der Gruppe B auch 24, so wären für diese Fälle zwei Rangfolgen (zuerst A oder zuerst B) möglich mit unterschiedlichen Ergebnissen für die Höhe von U. Diesen Sachverhalt muss das Testverfahren natürlich berücksichtigen. Im Fall gleicher Messwerte wird zur Ermittlung von Rangziffernsummen das arithmetische Mittel der Rangordnungsplätze als Rangziffer vergeben: z.B. würden beim Messwert 24 für beide Gruppen die Rangordnungsplätze 8 und 9 belegt werden und der Mittelwert 8,5 als Rangziffer zugeordnet.

Die Befragungen von Männern und Frauen können als zwei unabhängige Stichproben angesehen werden. Die Messwerte „1“ bis „4“ der ordinalskalierten Variablen TREUE erfassen die Antworten „sehr schlimm“ bis „gar nicht schlimm“ auf die Frage nach der Bedeutung eines „Seitensprungs“. Die Variable TREUE ist ordinalskaliert. Zum Testen der Hypothese mit dem Mann-Whitney U-Test gehen Sie wie folgt vor:

- ▹ Klicken Sie die Befehlsfolge „Analysieren“, „Nichtparametrische Tests ▹", „Zwei unabhängige Stichproben···“· Es öffnet sich die in Abb. 26.6 dargestellte Dialogbox.
- ▹ Von den in „Welche Tests durchführen?“ auswählbaren Tests wird der Mann-Whitney U-Test durch Anklicken ausgewählt.
- ▹ Aus der Quellvariablenliste wird die Testvariable TREUE in das Eingabefeld „Testvariablen“ übertragen.
- ▹ Danach wird die Variable GESCHL, die die zwei unabhängigen Stichproben (Gruppen) definiert, in das Eingabefeld von „Gruppenvariable“ übertragen. Sie erscheint dort zunächst als „geschl(? ?)“.
- ▹ Durch Anklicken von „Gruppen definieren...“ öffnet sich die in Abb. 26.7 dargestellte Dialogbox. In die Eingabefelder werden die Variablenwerte „1“ und „2“ der Variablen GESCHL zur Bestimmung der beiden Gruppen Männer und Frauen eingetragen. Mit „Weiter“ und „OK“ wird die Testprozedur gestartet.

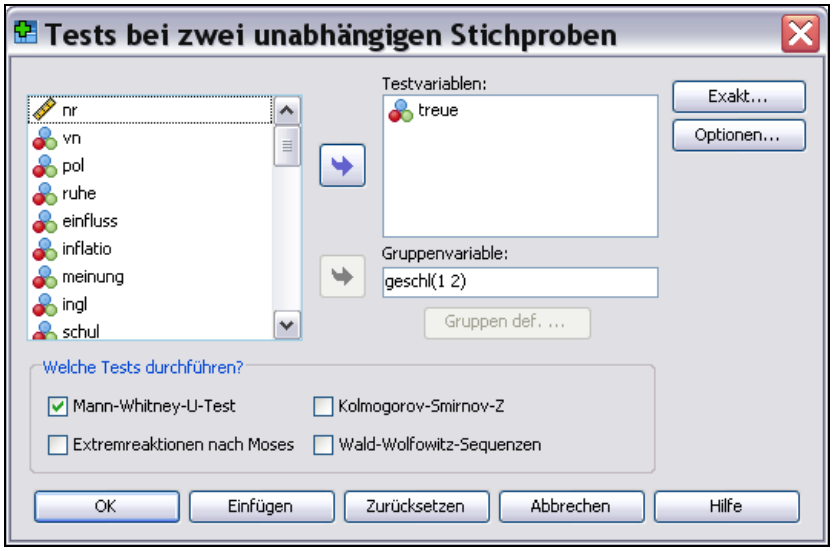

Abb. 26.6. Dialogbox „Tests bei zwei unabhängigen Stichproben“

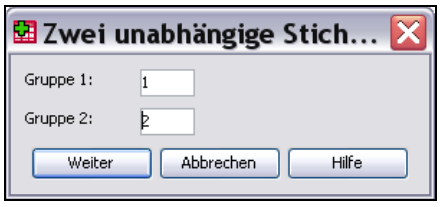

Abb. 26.7. Dialogbox „Zwei unabhängige Stichproben: Gruppen definieren“

Aus der Ergebnisausgabe (⇨ Tab. 26.9) kann man entnehmen, dass es insgesamt 153 gültige Fälle gibt mit 74 männlichen und 79 weiblichen Befragten. „Rangsumme“ gibt die Rangziffernsumme und „Mittlerer Rang“ die durchschnittliche Rangziffernsumme für jede Gruppe an. „Wilcoxon-W“ = 5394,5 ist die kleinste der Rangziffernsummen. „Mann-Whitney-U“ (= 2234,5) ist die Prüfgröße des Tests. Da für große Stichproben ($n_1 + n_2 \geq 30$) die Verteilung der Prüfgröße U durch eine Standardnormalverteilung approximiert werden kann, wird mit Z = -2,609 der empirische Wert der Standardnormalverteilung angegeben. Dem Z-Wert entspricht die zweiseitige Wahrscheinlichkeit von 0,009. Da diese Wahrscheinlichkeit kleiner ist als ein für den Test angenommenes 5-%- Signifikanzniveau (α = 0,05), wird die H_0-Hypothese einer gleichen Verteilung abgelehnt. Die Einstellung von Männer und Frauen ist demnach verschieden.

Der Test kann auch für die einseitige Fragestellung (H_1-Hypothese: Frauen bewerten einen Seitensprung als schlimmer als Männer) angewendet werden. Die durchschnittliche Rangziffernsumme für Frauen ist kleiner. Kleinere Rangziffern implizieren eine höhere Ablehnung eines Seitensprungs (sehr schlimm ist mit „1“, gar nicht schlimm mit „4“ codiert). Die einseitige exakte Signifikanz kann mit „Exakt Test“ berechnet werden.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.9. Ergebnisausgabe des Mann-Whitney U-Tests

Ränge

	geschl	N	Mittlerer Rang	Rangsumme
treue	MAENNLICH	74	86,30	6386,50
	WEIBLICH	79	68,28	5394,50
	Gesamt	153		

Statistik für Test[a]

	treue
Mann-Whitney-U	2234,500
Wilcoxon-W	5394,500
Z	-2,609
Asymptotische Signifikanz (2-seitig)	,009

26.3.2 Moses-Test bei extremer Reaktion

Dieser Test eignet sich dann, wenn man erwartet, dass bei experimentellen Tests unter bestimmten Testbedingungen manche Personen stark in einer Weise und andere Personen stark in einer entgegengesetzten Weise reagieren. Insofern stellt der Test auf Unterschiede in den Streuungen der Verteilungen ab.

Die Messwerte von zwei Vergleichsgruppen A und B (einer Kontroll- und einer Experimentiergruppe) werden in eine gemeinsame aufsteigende Rangfolge gebracht und erhalten Rangziffern. Unter der H_0-Hypothese (die Stichproben A und B kommen aus einer gleichen Grundgesamtheit) kann man erwarten, dass sich die Messwerte in der Kontroll- und Experimentiergruppe gut mischen. Unter der Hypothese H_1 (die Stichproben stammen aus unterschiedlichen Grundgesamtheiten bzw. unter den Testbedingungen haben die Testpersonen reagiert) kann man für die Experimentiergruppe sowohl relativ mehr höhere als auch niedrigere Messwerte erwarten. Der Test von Moses prüft, ob sich die Spannweite der Rangziffern (höchster minus kleinster plus eins) der Kontrollgruppe von der aller Probanten unterscheidet.

Beispiel. Es soll geprüft werden, ob sich die Einstellung zur Treue (hinsichtlich ihrer Streuung) bei jungen (18-29-jährige) und älteren (60-74-jährige) Menschen unterscheidet (Datei ALLBUS90.SAV). Vermutet wird, dass bei älteren eine höhere Variation in der Einstellung zur Treue besteht. Testvariable ist TREUE und Gruppenvariable ist ALT2 in der die Altersgruppen codiert sind. Zur Durchführung des Tests geht man wie in Kap. 26.3.1 erläutert vor. Im Unterschied dazu wird der Test von Moses sowie „1“ und „4“ als Gruppen der Gruppenvariable ALT2 gewählt.

In Tabelle 26.10 ist die Ergebnisausgabe niedergelegt. Es werden in der ersten Tabelle die gültigen Fallzahlen für beide Altersgruppen und in der zweiten Tabelle die Spannweite für die Kontrollgruppe (= Gruppe 1) sowie das exakte Signifikanzniveau für die einseitige Fragestellung („Signifikanz“) angegeben. Die Spannweite und das Signifikanzniveau wird auch unter Ausschluss von Extremwerten bzw. Ausreißern („getrimmte Kontrollgruppe“) aufgeführt. Als Testergebnis kann festgehalten werden, dass die H_0-Hypothese - die Altersgruppen unterscheiden sich nicht hinsichtlich ihrer Einstellung zur Treue - abgelehnt wird, da der Wert von „Signifikanz“ (0,00 bzw. 0,021) kleiner ist als ein vorgegebenes Signifikanzniveau von z.B. 5 % ($\alpha = 0{,}05$).

Optionen. ⇨ **Erläuterungen** zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.10. Ergebnisausgabe des Tests von Moses

Häufigkeiten

	alt2	N
treue	18 - 29 JAHRE (Kontrolle)	36
	60 - 74 JAHRE (Experimentell)	33
	Gesamt	69

Statistik für Test[a,b]

		treue
Beobachtete Spannweite der Kontrollgruppe		57
	Signifikanz (1-seitig)	,000
Spannweite der getrimmten Kontrollgruppe		57
	Signifikanz (1-seitig)	,021
Ausreißer an beiden Enden entfernt		1

a. Moses-Test
b. Gruppenvariable: alt2

26.3.3 Kolmogorov-Smirnov Z-Test

Dieser Test hat die gleichen Anwendungsvoraussetzungen wie der Mann-Whitney U-Test: zwei unabhängige Zufallsstichproben, das Messniveau der Variable ist mindestens ordinalskaliert. Auch die H_0-Hypothesen entsprechen einander: beide Stichproben stammen aus Grundgesamtheiten mit gleicher Verteilung.

Im Vergleich zum Mann-Whitney U-Test prüft der Test jegliche Abweichungen der Verteilungen (zentrale Tendenz, Streuung etc.; deshalb auch Omnibus-Test genannt). Soll lediglich geprüft werden, ob sich die zentrale Tendenz der Verteilungen unterscheidet, so sollte der Mann-Whitney U-Test bevorzugt werden.

Analog zum Kolmogorov-Smirnov-Test für den 1-Stichprobenfall (⇨ Kap. 26.2.4) basiert die Prüfgröße auf der maximalen Differenz (D_{max}) zwischen den kumulierten Häufigkeiten der beiden Stichprobenverteilungen. Wenn die Hypothese H_0 gilt (die Verteilungen unterscheiden sich nicht) so kann man erwarten, dass die kumulierten Häufigkeiten beider Verteilungen nicht stark voneinander

abweichen. Ist D_{max} größer als unter der Hypothese H_0 zu erwarten ist, so wird H_0 abgelehnt.

Zur Anwendung des Kolmogorov-Smirnov Z-Tests im 2-Stichprobenfall wird wie zur Durchführung des Mann-Whitney U-Tests (⇨ Kap. 26.3.1) vorgegangen. Im Unterschied dazu wird aber der Kolmogorov-Smirnov Z-Test gewählt. Ein Test auf Unterschiede zwischen Männern und Frauen in der Einstellung zur Treue führt zu zwei Ausgabetabellen. In der ersten (hier nicht aufgeführten) Tabelle wird die Häufigkeit der Variable TREUE nach dem Geschlecht untergliedert (⇨ Tabelle 26.9 links). In der zweiten Tabelle steht das Testergebnis (⇨ Tabelle 26.11).

Als größte (positive) Differenz D_{max} der Abweichungen in den kumulierten Häufigkeiten wird 0,180 ausgewiesen. Aus der Differenz ergibt sich nach Kolmogorov und Smirnov für die Prüfgröße Z = 1,11 gemäß Gleichung 26.4.

$$KS - Z = D_{max}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 0{,}18\sqrt{\frac{74 * 79}{74 + 79}} = 1{,}11 \qquad (26.4)$$

Von Smirnov sind Tabellen entwickelt worden, in denen den Z-Werten zweiseitige Wahrscheinlichkeiten zugeordnet sind. Dem Wert Z = 1,11 entspricht die zweiseitige Wahrscheinlichkeit 0,17. Eine maximale absolute Differenz gemäß der bestehenden kann demnach mit einer Wahrscheinlichkeit von 17 % auftreten. Legt man das Signifikanzniveau auf α = 0,05 fest, so kann wegen 0,17 > 0,05 die Hypothese H_0 (es gibt keinen Unterschied in der Einstellung zur Treue) nicht abgelehnt werden.

Führt man aber einen exakten Test mit dem Monte Carlo-Verfahren durch, so ergibt sich eine (2-seitige) Signifikanz = 0,038 (⇨ Tabelle 26.11). Demgemäß würde die Hypothese H_0 abgelehnt werden. Hier zeigt sich, dass man nicht immer auf die Ergebnisse asymptotischer Tests vertrauen kann.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.11. Ergebnisausgabe des Kolmogorov-Smirnov Z-Tests für zwei Stichproben

Statistik für Test[b]

			treue
Extremste Differenzen	Absolut		,180
	Positiv		,180
	Negativ		,000
Kolmogorov-Smirnov-Z			1,110
Asymptotische Signifikanz (2-seitig)			,170
Monte-Carlo-Signifikanz(2-seitig)	Signifikanz		,038[a]
	99%-Konfidenzintervall	Untergrenze	,033
		Obergrenze	,043

a. Basiert auf 10000 Stichprobentabellen mit einem Startwert von 622500317.

b. Gruppenvariable: geschl

26.3.4 Wald-Wolfowitz-Test

Der Wald-Wolfowitz-Test testet die H_0-Hypothese - beide Stichproben stammen aus gleichen Grundgesamtheitsverteilungen - gegen die Hypothese verschiedener Verteilungen in jeglicher Form (zentrale Lage, die Streuung etc., deshalb auch Omnibus-Test genannt). Er ist eine Alternative zum Kolmogorov-Smirnov Z-Test. Vorausgesetzt werden mindestens ein ordinales Skalenniveau sowie zwei unabhängige Stichproben.

Ganz analog zum Mann-Whitney U-Test werden die Messwerte beider Stichproben in eine Rangordnung gebracht, wobei mit dem kleinsten Wert begonnen wird. Dann wird – anolog zum Sequenzen-Test für eine Stichprobe – die Anzahl der Sequenzen gezählt. Es handelt sich also um einen Sequenzen-Test in Anwendung auf den 2-Stichprobenfall.

Am Beispiel zur Erläuterung des Mann-Whitney U-Tests (⇨ Tabelle 26.7) kann dieses gezeigt werden. Die Anzahl der Sequenzen beträgt 6 (⇨ Tabelle 26.12). Im Fall von Bindungen (gleiche Messwerte in beiden Gruppen) wird der Mittelwert der Ränge gebildet.

Tabelle 26.12. Beispiel zur Ermittlung von Sequenzen

Messwerte	10	14	17	18	21	22	23	24	26
Gruppe	A	A	B	B	A	B	B	A	B
Sequenz	1.	1.	2.	2.	3.	4.	4.	5.	6.

Das Beispiel Einstellung zur Treue aus der Datei ALLBUS90.SAV eignet sich nicht für den Test, weil die Variable TREUE nur vier Werte hat und es deshalb zu viele Bindungen (ties) gibt. Es wird das Beispiel zur Tabelle 26.7 zur Berechnung genommen (Datei MATHE.SAV). Die Vorgehensweise entspricht - bis auf die Auswahl des Tests - der in Kapitel 26.3.1 erläuterten. In Tabelle 26.13 ist die Ergebnisausgabe zu sehen.

Für Stichprobengrößen $n_1 + n_2 \leq 30$ wird ein einseitiges exaktes Signifikanzniveau berechnet. Für Stichproben > 30 wird eine Approximation durch die Standardnormalverteilung verwendet. In der ersten Ausgabetabelle (hier nicht aufgeführt) werden die Häufigkeiten für die Gruppen genannt. In der zweiten Ausgabetabelle (Tabelle 26.13) werden die Z-Werte mit der damit verbundenen einseitigen Wahrscheinlichkeit für die Anzahl der exakten Sequenzen [bzw. minimale und maximale Anzahl im Fall von Bindungen (ties)] angegeben. Sind die ausgewiesenen Wahrscheinlichkeiten kleiner als das gewählte Signifikanzniveau (z.B. α = 0,05), so wird die Hypothese H_0 abgelehnt. Da „Exakte Signifikanz (1-seitig)“ mit 0,786 größer ist als α =0,05, wird H_0 (kein Unterschied in den Mathematik-Lehrmethoden) angenommen.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.13. Ergebnisausgabe des Wald-Wolfowitz-Tests

Statistik für Test[b,c]

		Anzahl der Sequenzen	Z	Exakte Signifikanz (1-seitig)
punkte	Exakte Anzahl der Sequenzen	6[a]	,763	,786

a. Es wurden keine Bindungen zwischen Gruppen gefunden.

b. Test nach Wald-Wolfowitz

c. Gruppenvariable: methode

26.4 Tests für k unabhängige Stichproben

Bei diesen Tests wird in Erweiterung der Fragestellung für den Fall von zwei unabhängigen Stichproben geprüft, ob sich k (drei oder mehr) Gruppen (Stichproben) unterscheiden oder nicht. Es wird die H_0-Hypothese (alle Gruppen stammen aus der gleichen Grundgesamtheit) gegen die H_1-Hypothese (die Gruppen entstammen aus unterschiedlichen Grundgesamtheiten) geprüft. Die übliche parametrische Methode für eine derartige Fragestellung ist der F-Test der einfaktoriellen Varianzanalyse. Voraussetzung dafür aber ist, dass die Messwerte unabhängig voneinander aus normalverteilten Grundgesamtheiten mit gleichen Varianzen stammen. Des Weiteren ist Voraussetzung, dass das Messniveau der abhängigen Variablen mindestens intervallskaliert ist. Wenn die untersuchte Variable ordinalskaliert ist oder die Annahme einer Normalverteilung fragwürdig ist, sind die folgenden nichtparametrischen Tests einsetzbar.

26.4.1 Kruskal-Wallis H-Test

Der Kruskal-Wallis-Test eignet sich gut zur Prüfung auf eine unterschiedliche zentrale Tendenz von Verteilungen. Er ist eine einfaktorielle Varianzanalyse für Rangziffern. Die Messwerte für die k Stichproben bzw. Gruppen werden in eine gemeinsame Rangordnung gebracht. Aus diesen Daten wird die Prüfgröße H wie folgt berechnet:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} R_i^2 / n_i - 3(n+1) \qquad (26.5)$$

R_i = Summe der Rangziffern der Stichprobe i
n_i = Fallzahl der Stichprobe i
n = Summe des Stichprobenumfangs aller k Gruppen.

Für den Fall von Bindungen (englisch: ties), wird die Gleichung mit einem Korrekturfaktor korrigiert (⇨ Bortz/Lienert/Boehnke, S. 223). Die Prüfgröße H ist approximativ chi-quadratverteilt mit k-1 Freiheitsgraden.

Beispiel. Mit Daten der Datei ALLBUS90.SAV soll untersucht werden, ob die Einstellung zur Treue in einer Partnerschaft unabhängig vom Alter ist. Die Personen verschiedener Altersgruppen (codiert in der Variable ALT2) können als vier unabhängige Stichproben angesehen werden. Zum Testen der Hypothese wird der Kruskal-Wallis H-Test wie folgt angewendet:

- ▹ Klicken Sie die Befehlsfolge „Analysieren“, „Nichtparametrische Tests ▹", „K unabhängige Stichproben···“. Es öffnet sich die in Abb. 26.8 dargestellte Dialogbox.
- ▹ In „Welche Tests durchführen?“ wird „Kruskal-Wallis H“ angeklickt.
- ▹ Aus der Quellvariablenliste wird die Testvariable TREUE in das Eingabefeld „Testvariablen“ übertragen.
- ▹ Danach wird die Variable ALT2, deren Altersgruppen als unabhängige Stichproben anzusehen sind, in das Eingabefeld von „Gruppenvariable“ übertragen. Sie erscheint dort zunächst als „alt2(? ?)“.
- ▹ Durch Anklicken von „Bereich definieren“ öffnet sich die in Abb. 26.9 dargestellte Dialogbox. In die Eingabefelder „Minimum“ und „Maximum“ werden die Variablenwerte „1“ und „4“ der Variablen ALT2 zur Bestimmung des Wertebereichs der Variable ALT2 eingetragen. Der Test prüft dann auf Unterschiede für die ersten vier Altersgruppen. Mit „Weiter“ und „OK“ wird die Testprozedur gestartet.

In Tabelle 26.14 ist die Ergebnisausgabe des Tests zu sehen. „Mittlerer Rang“ gibt die durchschnittlichen Rangziffern und „N“ die Fallzahlen der vier Altersgruppen an. Der Wert der approximativ chi-quadratverteilten Prüfgröße ist mit 4,044 kleiner als ein aus einer Chi-Quadrat-Tabelle für $k-1=3$ Freiheitsgrade (df) bei einer Irrtumswahrscheinlichkeit von α = 0,05 entnehmbarer kritischer Wert von 7,82 Demnach wird die Hypothese H_0 (es gibt für die Altersgruppen keinen Unterschied in der Einstellung zur Treue) angenommen. Diese Schlussfolgerung ergibt sich auch aus dem angegebenem Signifikanzniveau 0,257 („Asymptotische Signifikanz“), das die mit α = 0,05 vorgegebene Irrtumswahrscheinlichkeit übersteigt.

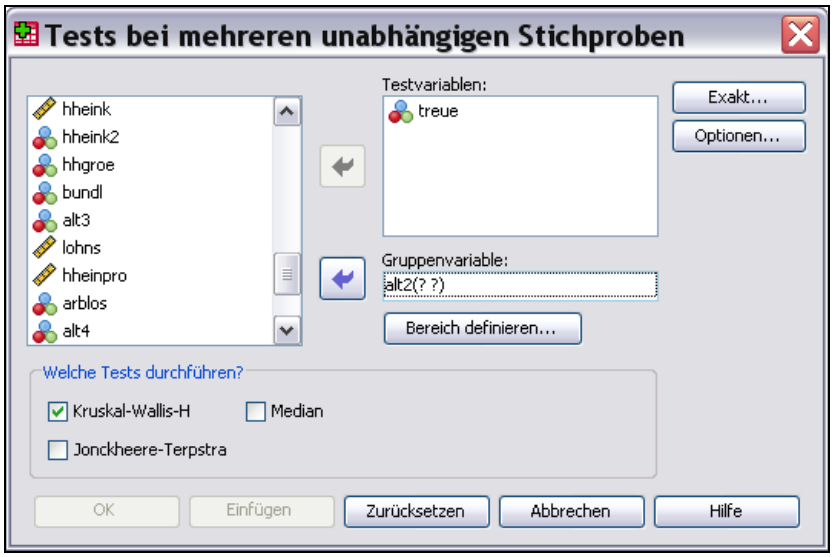

Abb. 26.8. Dialogbox „Tests bei mehreren unabhängigen Stichproben“

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

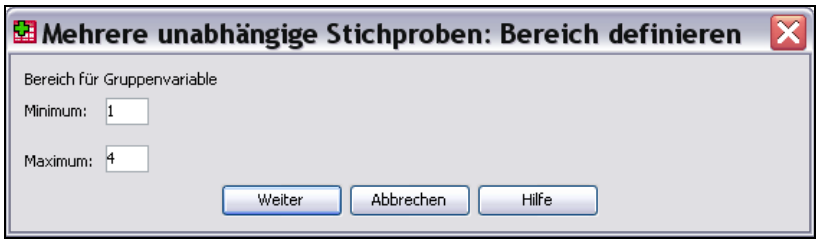

Abb. 26.9. Dialogbox „Mehrere unabh. Stichproben: Bereich definieren“

Tabelle 26.14. Ergebnisausgabe des Kruskal-Wallis H-Tests

Ränge

	alt2	N	Mittlerer Rang
treue	18 - 29 JAHRE	36	77,39
	30 - 44 JAHRE	43	74,03
	45 - 59 JAHRE	27	64,72
	60 - 74 JAHRE	33	61,00
	Gesamt	139	

Statistik für Test[a,b]

	treue
Chi-Quadrat	4,044
df	3
Asymptotische Signifikanz	,257

a. Kruskal-Wallis-Test

b. Gruppenvariable: alt2

26.4.2 Median-Test

Auch der Median-Test verlangt, dass die untersuchte untersuchte Variable mindestens ordinalskaliert ist. Geprüft wird, ob die Stichproben aus Grundgesamtheiten mit gleichen Medianen stammen.

Der Test nutzt nur Informationen über die Höhe eines jeden Beobachtungswertes im Vergleich zum Median. Daher ist er ein sehr allgemeiner Test.
Bei diesem Testverfahren wird zunächst für die Messwerte aller k Gruppen der gemeinsame Median bestimmt. Im nächsten Schritt wird jeder Messwert als kleiner bzw. größer als der gemeinsame Median eingestuft und für alle Gruppen werden die Häufigkeiten des Vorkommens von kleiner bzw. größer als der Median ausgezählt. Es entsteht für k Gruppen eine $2*k$-Häufigkeitstabelle. Falls $n > 30$ ist, wird aus der Häufigkeitstabelle eine Chi-Quadrat-Prüfgröße ermittelt und für $k-1$ Freiheitsgrade ein approximativer Chi-Quadrat-Test durchgeführt. Für kleinere Fallzahlen wird mit Fischer´s exact Test die genaue Wahrscheinlichkeit berechnet.

Das folgende Anwendungsbeispiel (Datei ALLBUS90.SAV) ist das gleiche wie in Kap. 26.4.1: es soll geprüft werden, ob die Einstellung zur Treue in einer Partnerschaft unabhängig vom Alter ist. Zum Testen der Hypothese geht man wie dort beschrieben vor. Im Unterschied dazu wird aber der Test „Median“ durch Klicken gewählt.

In Tabelle 26.15 ist die Ergebnisausgabe dargestellt. Da $k = 4$ ist, wird eine $2*4$-Häufigkeitstabelle dargestellt. In der ersten Ausgabetabelle werden für die vier Altersgruppen die Häufigkeiten für die Variable TREUE mit den Ausprägungen größer als der Median und gleich-kleiner als der Median aufgeführt. Mit „Chi-Quadrat“ = 6,226 wird der ermittelte empirische Chi-Quadrat-Wert ausgewiesen. Für $k-1 = 3$ Freiheitsgrade („df“) und einem Signifikanzniveau von 5 % (α =

0,05) ergibt sich aus einer Chi-Quadrat-Tabelle[3] ein kritischer Wert von 7,82. Da der empirische Wert kleiner ist als der kritische, wird die Hypothese H_0 (die Einstellung zur Treue ist unabhängig vom Alter) angenommen. Dieses Testergebnis ergibt sich einfacher auch daraus, dass die von SPSS ausgewiesene „Asymptotische Signifikanz“ = 0,101 größer ist als die gewählte in Höhe von $\alpha = 0{,}05$.

Optionen. ⇨ Erläuterungen Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.15. Ergebnisausgabe des Median-Tests

Häufigkeiten

		alt2			
		18 - 29 JAHRE	30 - 44 JAHRE	45 - 59 JAHRE	60 - 74 JAHRE
treue	> Median	20	21	11	9
	< = Median	16	22	16	24

Statistik für Test[b]

	treue
N	139
Median	2,00
Chi-Quadrat	6,226[a]
df	3
Asymptotische Signifikanz	,101

a. Bei 0 Zellen (,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 11,8.

b. Gruppenvariable: alt2

26.4.3 Jonckheere-Terpstra-Test

Dieser Test ist nur nach Installation des SPSS-Moduls „Exakt Test“ verfügbar.

Weder der Kruskal-Wallis- noch der Median-Test sind geeignet, Annahmen über die Richtung des Unterschiedes zwischen den Gruppen zu prüfen. In manchen Untersuchungen (speziell bei experimentellen Untersuchungsdesigns) hat man die Situation, dass die Wirkungen mehrerer Aktivitäten oder Maßnahmen simultan geprüft werden sollen und eine Rangfolge in der Wirkungsrichtung angenommen werden kann. In unserem Anwendungsbeispiel haben wir oben geprüft, ob mit wachsendem Alter die Einstellung zur Treue unterschiedlich ist. Es kam zur Annahme der H_0-Hypothese: kein Unterschied. Geht man aber davon aus, dass mit wachsendem Alter die Einstellung zur Treue sich in eine Richtung verändert (je höher das Alter, umso größer wird die Wertschätzung von Treue), kann man mit dem Jonckheere-Terpstra-Test eine bessere Trennschärfe zum Testen auf Unterschiede der Altersgruppen in der Einstellung zur Treue erzielen. Der Test ermöglicht ein Testen von geordneten Alternativen. Ein anderes Beispiel dafür wäre, wenn für mehrere Versuchsgruppen die Wirkung eines Medikaments mit jeweils einer höheren Dosis geprüft wird.

[3] Die Tabelle ist auf den Internetseiten zum Buch verfügbar.

Zum Testen der Hypothese geht man wie in Kap. 26.4.1 beschrieben vor. Im Unterschied dazu wird der Test „Jonckheere-Terpstra“ gewählt. In Tabelle 26.16 ist die Ergebnisausgabe dargestellt. Für die 139 gültigen Fälle („N“) wird die empirische („Beobachtete“) Testgröße „J-T-Statistik“, ihr Mittelwert, ihre Standardabweichung, ihr standardisierter Wert (in z-Werte transformiert) sowie ein asymptotisches 2-seitiges Signifikanzniveau ausgewiesen. Da der Wert von „Asymptotische Signifikanz (2-seitig)“ mit 0,051 größer ist als ein vorzugebendes Signifikanzniveau von z.B. 0,05 (α = 0,05 %) wird die H_0-Hypothese (ein Zusammenhang zwischen der Einstellung zur Treue und dem Alter besteht nicht) angenommen. Damit werden die Ergebnisse in Kap. 26.4.1 und 26.4.2 (Kruskal-Wallis- und Median-Test) bestätigt.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.16. Ergebnisausgabe des Jonckheere-Terpstra-Tests

Jonckheere-Terpstra-Test[a]

	treue
Anzahl der Stufen in alt2	4
N	139
Beobachtete J-T-Statistik	3092,500
Mittelwert der J-T-Statistik	3589,500
Standardabweichung der J-T-Statistik	255,136
Standardisierte J-T-Statistik	-1,948
Asymptotische Signifikanz (2-seitig)	,051

a. Gruppenvariable: alt2

26.5 Tests für 2 verbundene Stichproben

Bei diesem Testtyp möchte man prüfen, ob eine Maßnahme oder Aktivität wirksam ist oder nicht und bildet zwei Stichprobengruppen: eine Experiment- und eine Kontrollgruppe (matched pairs, ⇨ Kap. 26.1).

Die Grundhypothese (auch H_0-Hypothese genannt) postuliert, dass keine Unterschiede zwischen beiden Gruppen bestehen. Mit dieser Hypothese wird die Wirkung einer Maßnahme (z.B. die Wirksamkeit eines Medikaments oder der Erfolg einer neuen Lehr- oder Lernmethode) nicht anerkannt. Die Gegenthese H_1 geht von der Wirksamkeit aus.

26.5.1 Wilcoxon-Test

Der Test eignet sich, wenn Unterschiede in der zentralen Tendenz von Verteilungen geprüft werden sollen. Der Test beruht auf Rängen von Differenzen in den Variablenwerten. Der Wilcoxon-Test ist dem Vorzeichen(Sign)-Test (⇨ Kap. 26.5.2) vorzuziehen, wenn die Differenzen aussagekräftig sind.
Im Folgenden wird zur Anwendungsdemonstration ein Beispiel aus dem Bereich der Pädagogik gewählt. Zur Überprüfung einer neuen Lehrmethode werden Schü-

lerpaare gebildet, die sich hinsichtlich ihres Lernverhaltens und ihrer Lernfähigkeiten gleichen. Eine Schülergruppe mit jeweils einem Schüler der Paare wird nach der herkömmlichen Lehrmethode (Methode A genannt) und die andere Gruppe mit dem zweiten Schüler der Paare nach der neuen (Methode B genannt) unterrichtet. Die Lernergebnisse wurden bei Leistungstests in Form von erreichten Punkten erfasst und als Variable METH_A und METH_B in der Datei LEHRMETH.SAV gespeichert (⇨ Ausschnitt in Abb. 26.10). Geprüft werden soll, ob die beiden Methoden sich unterscheiden oder nicht.

Es handelt sich hier um ordinalskalierte Variablen, wobei aber Differenzen von Variablenwerten eine gewisse Aussagekraft haben.

	nr	meth_a	meth_b	meth_c
1	1	11	14	9
2	2	15	13	17
3	3	12	14	13
4	4	14	15	16
5	5	12	14	15
6	6	13	13	17

Abb. 26.10. Ausschnitt aus der Datei LEHRMETH.SAV

Bei dem Testverfahren werden im ersten Schritt die Differenzen der Messwerte für die Paare berechnet. Im nächsten Schritt werden die absoluten Differenzen (also ohne Vorzeichenbeachtung) in eine gemeinsame Rangziffernreihen-Ordnung gebracht. Haben Paare gleiche Messwerte, so werden diese Fälle aus der Analyse ausgeschlossen. Schließlich werden diesen Rangziffern die Vorzeichen der Differenzen zugeordnet. Unter der Hypothese H_0 (kein Unterschied der beiden Methoden) kann man erwarten, dass aufgetretene große Differenzen sowohl durch die Methode A als auch durch die Methode B bedingt sind. Summiert man jeweils die positiven und negativen Rangziffern, so ist unter H_0 zu erwarten, dass die Summen sich zu Null addieren. Unter H_1 wäre dementsprechend zu erwarten, dass sich die Summen unterscheiden. Von Wilcoxon liegen Tabellen vor, aus denen man für die Prüfgröße (die kleinere der Rangziffernsummen) für ein vorgegebenes Signifikanzniveau von z.B 5 % ($\alpha = 0{,}05$) kritische Werte entnehmen kann (Siegel, 1956, S. 79 f.).

Zum Testen, ob die Lehrmethoden A und B unterschiedlichen Erfolg haben oder nicht, kann der Wilcoxon-Test wie folgt angewendet werden:

- ▹ Klicken Sie die Befehlsfolge „Analysieren“, „Nichtparametrische Tests ▹“, „Zwei verbundene Stichproben···“· Es öffnet sich die in Abb. 26.11 dargestellte Dialogbox.
- ▹ In „Welche Tests durchführen?“ wird „Wilcoxon“ angeklickt.
- ▹ Aus der Quellvariablenliste werden die Variablen METH_A und METH_B mit dem Pfeil in das Eingabefeld „Testpaare“ übertragen. Mit „OK“ wird die Testprozedur gestartet.

In Tabelle 26.17 ist die Ergebnisausgabe des Tests zu sehen. In der ersten Tabelle werden für die negativen (METH_B < METH_A) und positiven (METH_B > METH_A) Rangziffern die Summe, die Durchschnitte („Mittlerer Rang“) und

Fallzahlen („N“) aufgeführt. In einem Fall sind die Messwerte gleich. Dieser Fall wird als „Bindungen“ ausgewiesen (METH_B = METH_A). Die negative Rangsumme ist mit 59 am kleinsten. Aus der Tabelle von Wilcoxon (Siegel, 1956, S. 254) ergibt sich z.B. für ein Signifikanzniveau von 5 % (bei einem zweiseitigen Test) und für n = 19 ein kritischer Wert von 46 für die kleinere Rangziffernsumme. Da der empirische Wert mit 59 diesen übersteigt, wird die Hypothese H_0 angenommen. Die Differenz der Rangziffernsummen ist nicht hinreichend groß, um einen Unterschied der Methoden zu begründen.

Für Stichprobenumfänge n > 25 kann die Tabelle von Siegel nicht genutzt werden. Da die Prüfgröße der kleineren Rangziffernsumme in derartigen Fällen approximativ normalverteilt ist, kann der Test mit Hilfe der Standardnormalverteilung durchgeführt werden. Von SPSS werden der empirische Z-Wert der Standardnormalverteilung sowie das zugehörige zweiseitige Signifikanzniveau ausgegeben. Da dieses (zweiseitige) Signifikanzniveau („Asymptotische Signifikanz = 0,138“ für „Z = - 1,483“) das vorgegebene $\alpha = 0{,}05$ übersteigt, kann auch hieraus der Schluss gezogen werden, dass die Hypothese H_0 (keine signifikanten Unterschiede der Lehrmethoden) angenommen wird.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Abb. 26.11. Dialogbox „Tests bei zwei verbundenen Stichproben“

Tabelle 26.17. Ergebnisausgabe des Wilcoxon-Tests

Ränge

		N	Mittlerer Rang	Rangsumme
meth_b - meth_a	Negative Ränge	6[a]	9,83	59,00
	Positive Ränge	13[b]	10,08	131,00
	Bindungen	1[c]		
	Gesamt	20		

a. meth_b < meth_a

b. meth_b > meth_a

c. meth_b = meth_a

Statistik für Test[b]

	meth_b - meth_a
Z	-1,483[a]
Asymptotische Signifikanz (2-seitig)	,138

a. Basiert auf negativen Rängen.

b. Wilcoxon-Test

26.5.2 Vorzeichen-Test

Der Vorzeichen-Test (englisch: sign) stützt sich - wie der Wilcoxon-Test (⇨ Kap. 26.5.1) - auf Differenzen von Messwerten zwischen Paaren von Gruppen bzw. im „vorher-nachher"-Stichprobendesign. Im Unterschied zum Wilcoxon-Test gehen nur die Vorzeichen der Differenzen, nicht aber die Größen der Differenzen in Form von Rangziffern in das Testverfahren ein. Dieser Test bietet sich immer dann an, wenn (bedingt durch die Datenlage) die Höhe der Differenzen nicht aussagekräftig ist. Fälle, bei denen die Differenzen der Paare gleich Null sind, werden nicht in das Testverfahren einbezogen. Gezählt werden die Anzahl der positiven und die Anzahl der negativen Differenzen.

Unter der Hypothese H_0 (keine unterschiedliche Wirkung einer Maßnahme bzw. Aktivität) ist zu erwarten, dass die Fallzahlen mit positiven und negativen Vorzeichen etwa gleich sein werden. Für $n < 25$ kann die Wahrscheinlichkeit für die Häufigkeit der Vorzeichen mit Hilfe der Binomialverteilung berechnet werden.

Zur Durchführung des Vorzeichen-Tests geht man wie beim Wilcoxon-Test vor (⇨ Kap. 26.5.1). Im Unterschied dazu wird aber der Vorzeichen-Test in der Dialogbox der Abb. 26.11 angeklickt. Für das obige Beispiel der in Abb. 26.10 ausschnittsweise dargestellten Datei LEHRMETH.SAV erhält man die in Tabelle 26.18 dargestellte Ergebnisse. Es werden in der ersten Tabelle die Fallzahlen mit negativen und positiven Differenzen angeführt. Die Wahrscheinlichkeit für das Auftreten von sechs negativen und 13 positiven Vorzeichen wird mit 0,167 („Exakte Signifikanz") angegeben. Bei einem Signifikanzniveau von 5 % (α = 0,05) für den Test wird die Hypothese H_0 angenommen, da 0,167 > 0,05 ist. Das Testergebnis entspricht dem von Wilcoxon.

Für große Fallzahlen ($n > 25$) wird die Binomialverteilung durch die Normalverteilung approximiert. SPSS gibt in diesen Fällen (wie bei dem Wilcoxon-Test) den Z-Wert der Standardnormalverteilung sowie die zugehörige Wahrscheinlichkeit an.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.18. Ergebnisausgabe des Vorzeichen-Tests

Häufigkeiten

		N
meth_b - meth_a	Negative Differenzen[a]	6
	Positive Differenzen[b]	13
	Bindungen[c]	1
	Gesamt	20

a. meth_b < meth_a

b. meth_b > meth_a

c. meth_b = meth_a

Statistik für Test[b]

	meth_b - meth_a
Exakte Signifikanz (2-seitig)	,167[a]

a. Verwendetete Binomialverteilung.

b. Vorzeichentest

26.5.3 McNemar-Test

Der McNemar-Test eignet sich für ein „vorher-nachher"-Testdesign mit dichotomen Variablen und testet Häufigkeitsunterschiede. Anhand eines Beispiels sei der Test erklärt. Um zu prüfen, ob zwei Aufgaben den gleichen Schwierigkeitsgrad haben, können diese nacheinander Probanten zur Lösung vorgelegt werden. Das Ergebnis in Form von Häufigkeiten kann in einer 2∗2-Tabelle festgehalten werden. Die Häufigkeiten n_A und n_D in Tabelle 26.19 erfassen Veränderungen im Lösungserfolg durch den Wechsel der Aufgaben. Die Häufigkeiten n_C und n_B geben die Fälle mit gleichem Lösungserfolg an. Je weniger sich diese Häufigkeiten unterscheiden, um so wahrscheinlicher ist es, dass die H_0-Hypothese (durch den Wechsel der Aufgaben tritt keine Veränderung im Lösungserfolg ein) zutrifft. Die Wahrscheinlichkeit kann mit Hilfe der Binomialverteilung berechnet werden.
Zur Anwendungsdemonstration werden Daten der ausschnittsweise in Abb. 26.12 zu sehenden Datei TESTAUFG.SAV verwendet.

In der Datei sind Lösungsresultate für von Studierenden bearbeitete Testaufgaben erfasst. Die Variablen AUFG1 und AUFG2 sind nominalskalierte Variable mit dichotomen Ausprägungen: Der Variablenwert „1" steht für „Aufgabe nicht gelöst" und „2" für „Aufgabe gelöst". Zur Durchführung des Tests geht man wie bei den anderen Tests bei zwei verbundenen Stichproben vor (⇨ Kap. 26.5.1). Im Unterschied dazu wird der McNemar-Test gewählt.

Tabelle 26.19. 4-Felder Tabelle zur Erfassung von Änderungen

	Aufgabe 2	
Aufgabe 1	nicht gelöst	gelöst
gelöst	n_A	n_B
nicht gelöst	n_C	n_D

	nr	aufg1	aufg2	aufg3
1	1	2	1	2
2	2	1	1	1
3	3	2	1	2
4	4	1	1	2
5	5	2	1	1
6	6	2	1	2

Abb. 26.12. Ausschnitt aus der Datei TESTAUFG.SAV

In der folgenden Tabelle 26.20 wird das Ausgabeergebnis für den Test dokumentiert. Die Ausgabeform entspricht der Tabelle 26.19 bei einer Fallzahl von 15. Die Wahrscheinlichkeit wird wegen der kleinen Fallzahl (n < 25) auf der Basis einer Binomialverteilung ausgegeben.

Für große Fallzahlen wird approximativ ein Chi-Quadrat-Test durchgeführt. Testergebnis ist, dass die H_0-Hypothese (kein Unterschied im Schwierigkeitsgrad der Aufgaben) angenommen wird, da die angeführte zweiseitige Wahrscheinlichkeit („Exakte Signifikanz“) das Signifikanzniveau von 5 % ($\alpha = 0{,}05$) übersteigt.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.20. Ergebnisausgabe des McNemar-Tests

aufg1 & aufg2

aufg1	aufg2 gelöst	aufg2 nicht gelöst
gelöst	3	3
nicht gelöst	8	1

Statistik für Test[b]

	aufg1 & aufg2
N	15
Exakte Signifikanz (2-seitig)	,227[a]

a. Verwendetete Binomialverteilung.

b. McNemar-Test

26.5.4 Rand-Homogenitäts-Test

Dieser Test ist nur nach Installation des SPSS-Moduls „Exakt Test“ verfügbar.

Er ist eine Verallgemeinerung des McNemar-Tests. Anstelle von zwei (binären) Kategorien (vorher - nachher) werden mehr als zwei Kategorien berücksichtigt. Dabei muss es sich um geordnete Kategorien handeln. Auf die theoretischen Grundlagen des Tests kann hier nicht eingegangen werden (⇨ Kuritz, Landis und Koch, 1988).

Beispiel. Ein Arzt verabreicht 25 Personen ein Präparat zur Erhöhung der allgemeinen Leistungsfähigkeit und im Abstand von drei Monaten ein Placebo. Anstelle der binären Kategorien „Wirkung“ - „keine Wirkung“ (codiert mit „1“ und „2“), der einen McNemar-Test auf Prüfung der Wirksamkeit ermöglichen würde, werden die Merkmale „keine Wirkung“, „geringe Wirkung“ und „starke Wirkung“ (kodiert mit „1“, „2“ und „3“) erfasst. In Abb. 26.13 ist ein Ausschnitt aus der Datei PATIENT.SAV zu sehen.

Anstelle einer 2*2-Kreuztabelle (⇨ Tabelle 26.19) für den McNemar-Test würde nun eine 3*3-Kreuztabelle entstehen.

	patient	praepara	placebo
1	1	1	2
2	2	2	1
3	3	3	2
4	4	2	1
5	5	3	1

Abb. 26.13. Ausschnitt aus der Datei PATIENT.SAV

In Tabelle 26.21 ist die Ergebnisausgabe (bei Auswahl des Variablen PRAEPA und PLACEBO als Testpaare) zu sehen. Außerhalb der Diagonalen der in der SPSS-Ausgabe nicht aufgeführten 3*3-Kreuztabelle gibt es 15 Fälle. Der empirische Wert für die Prüfgröße beträgt 35. Die Prüfgröße (MH = marginale Homogenität) hat einen Durchschnittswert von 29,5 mit einer Standardabweichung von 2,291. Daraus ergibt sich die standardisierte Prüfgröße in Höhe von 2,40. Da die ausgegebene 2-seitige Wahrscheinlichkeit („Asymptotische Signifikanz“ = 0,016) kleiner ist als ein vorzugebendes Signifikanzniveau von z.B. $\alpha = 0{,}05$, kann von der Wirksamkeit des Präparats ausgegangen werden.

Optionen. ⇨ Erläuterungen zu Abb. 26.2.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.21. Ergebnisausgabe des Rand-Homogenitäts-Test

Rand-Homogenitätstest

	praepara & placebo
Unterschiedliche Werte	3
Fälle außerhalb der Diagonalen	15
Beobachtete MH-Statistik	35,000
Mittelwert der MH Statistik	29,500
Standardabweichung der MH-Statistik	2,291
Standardisierte MH-Statistik. MH Statistic	2,400
Asymptotische Signifikanz (2-seitig)	,016

26.6 Tests für k verbundene Stichproben

Bei diesen Testverfahren geht es um die simultane Prüfung von Unterschieden zwischen drei und mehr Stichproben bzw. Gruppen, wobei es sich um abhängige bzw. verbundene Stichproben handelt (⇨ Kap. 26.1). Die H_0-Hypothese lautet, dass die Stichproben aus identischen Grundgesamtheiten stammen.

26.6.1 Friedman-Test

Der Friedman-Test ist eine Zwei-Weg-Varianz-Analyse für Rangziffern zur Prüfung der Frage, ob die Stichproben aus einer gleichen Grundgesamtheit kommen. Es handelt sich um einen allgemeinen Test, der auf Unterschiede prüft ohne aufzudecken, um welche Unterschiede es sich handelt.

Der Test wird am Beispiel der Prüfung von drei Lehrmethoden auf den Lernerfolg von drei Studentengruppen demonstriert (⇨ Datei LEHRMETH.SAV, Abb. 26.10). Die drei Stichprobengruppen wurden dabei aus Sets von jeweils drei Studenten mit gleicher Fähigkeiten, Lernmotivation u.ä. zusammengestellt, um die Wirkung der Lehrmethoden eindeutiger zu messen. In Tabelle 26.22 werden die Messwerte der ersten vier Zeilen (Sets) aus der Datei der Abb. 26.10 angeführt. Im Testverfahren werden für jede Reihe (Zeile) der Tabelle Rangziffern vergeben. Unter der Hypothese H_0 - kein Unterschied im Erfolg der Methoden - verteilen sich die Rangziffern auf die drei Spalten zufällig, so dass auch die spaltenweise

aufsummierten Rangziffernsummen sich kaum unterscheiden. Der Friedman-Test prüft, ob sich die Rangziffernsummen signifikant voneinander unterscheiden. Zum Testen dient folgende Prüfgröße:

$$\chi^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \qquad (26.6)$$

n = Anzahl der Fälle (= Anzahl der sets)
k = Anzahl der Variablen (= Spaltenanzahl = Anzahl der Gruppen)
R_j = Summe der Rangziffern in der Spalte j, d.h. der Gruppe j
Die Prüfgröße ist asymptotisch chi-quadratverteilt mit k-1 Freiheitsgraden.

Tabelle 26.22. Messwerte und Rangziffern der ersten vier Sets der Datei

Methode	**Meth. A**		**Meth. B**		**Meth. C**	
Set	*Messwert*	*Rangziffer*	*Messwert*	*Rangziffer*	*Messwert*	*Rangziffer*
Set 1	11	2	14	1	9	3
Set 2	15	2	13	3	17	1
Set 3	12	3	14	1	13	2
Set 4	14	3	15	2	16	1
Rangsumme R		10		7		7

Zum Testen der Hypothese - haben die Lehrmethoden A, B und C unterschiedlichen Erfolg oder nicht - gehen Sie wie folgt vor:

- ▹ Klicken Sie die Befehlsfolge „Analysieren“, “Nichtparamametrische Tests ▹“, „K verbundene Stichproben...“. Es öffnet sich die in Abb. 26.14 dargestellte Dialogbox.
- ▹ In „Welche Tests durchführen?“ wird „Friedman“ angeklickt.
- ▹ Aus der Quellvariablenliste werden die Variablen METH_A, METH_B und METH_C markiert und durch Klicken auf den Pfeilschalter in das Eingabefeld „Testvariablen“ übertragen. Mit „OK“ wird die Testprozedur gestartet.

In Tabelle 26.23 ist die Ergebnisausgabe zu sehen. Es werden die durchschnittliche Rangziffernsumme („Mittlerer Rang“), die Fallzahl („N“), der empirische Wert der Prüfgröße Chi-Quadrat mit der Anzahl der Freiheitsgrade („df“) sowie der zugehörigen Wahrscheinlichkeit („Signifikanz“) angegeben. Wird für den Test z.B. ein Signifikanzniveau von 5 % ($\alpha = 0{,}05$) gewählt, so wird die H_0-Hypothese abgelehnt, da 0,006 < 0,05 ist. Dieses Testergebnis wird plausibel, wenn man feststellt, dass der Wilcoxon-Test (⇨ Kap. 26.5.1) ergibt, dass sich die Methode C signifikant sowohl von A als auch von B unterscheidet.

Statistiken. Durch Klicken auf die Schaltfläche „Statistiken...“ öffnet sich eine (Unter-)Dialogbox in der deskriptive statistische Maßzahlen (Mittelwert, Standardabweichung) sowie Quartile (25., 50. 75. Perzentil) angefordert werden können.

Exakter Test. ⇨ Kap. 31.

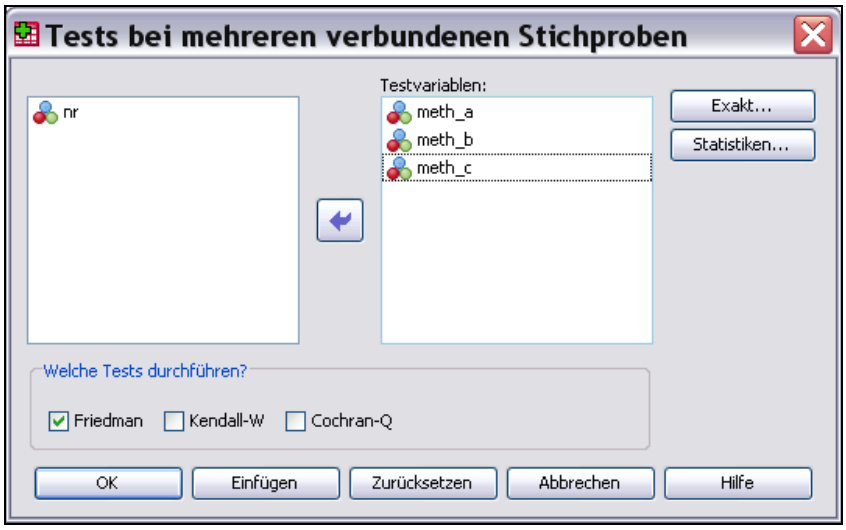

Abb. 26.14. Dialogbox „Tests bei mehreren verbundenen Stichproben“

Tabelle 26.23. Ergebnisausgabe des Friedman-Tests

Ränge

	Mittlerer Rang
meth_a	1,60
meth_b	1,85
meth_c	2,55

Statistik für Test[a]

N	20
Chi-Quadrat	10,347
df	2
Asymptotische Signifikanz	,006

a. Friedman-Test

26.6.2 Kendall's W-Test

Der Test ist dem von Friedman äquivalent. Er beruht im Unterschied zum Friedman-Test auf dem Maß W. Kendall's Koeffizient der Konkordanz W ist ein Maß für die Stärke des Zusammenhangs von mehr als zwei ordinalskalierten Variablen. Er misst, in welchem Maße Rangziffern für k Gruppen übereinstimmen.

Es sei angenommen, drei Lehrer bewerten die Klassenarbeit von 20 Schülern. Für die Klassenarbeiten der Schüler entsteht pro Lehrer eine Rangfolge in Form von Rangziffern, wobei für die beste Arbeit die Rangziffer 1 vergeben worden sei. Zur Bestimmung des Maßes W werden die Rangziffern für jeden Schüler aufsummiert. Aus diesen Summen wird das Ausmaß der unterschiedlichen Bewertung deutlich. Bewerten alle drei Lehrer die Arbeiten gleich, so hat der beste Schüler von allen Lehrern die Rangziffer 1, der zweitbeste die Rangziffer 2 usw. erhalten. Daraus ergeben sich die Rangziffersummen 3, 6, 9 usw. Die Unterschiedlichkeit - die Variation - der Rangziffernsummen ist demgemäß ein Maß für die Übereinstimmung der Bewertung. In Gleichung 26.7 ist das Maß W definiert. Im Zähler des Bruches steht die Variation der Rangziffernsumme in Form ihrer quadratischen Abweichung vom Mittelwert. Im Nenner steht diese Variation für den Fall völlig gleicher Bewertung: er reduziert sich dann auf den im Nenner angegebenen Ausdruck.

$$W = \frac{\sum_{j=1}^{n}(R_j - \frac{\sum_{j=1}^{n} R_j}{n})^2}{(1/12)k^2(n^3 - n)} \qquad (26.7)$$

R_j = Rangziffernsumme der Objekte oder Individuen j

k = Anzahl der Sets von Bewertungen bzw. Bewerter

n = Anzahl der bewerteten Objekte bzw. Individuen.

Aus der Formel ergibt sich, dass mit der Höhe von W das Ausmaß der Übereinstimmung bei der Rangziffernvergabe wächst. W kann zwischen 0 und 1 liegen.

Für Stichprobenumfänge größer sieben ist $k(n-1)W$ annähernd chi-quadratverteilt mit $n-1$ Freiheitsgraden (Siegel, 1956, S. 236). Zur praktischen Demonstration werden die in Kap. 26.5.1 genutzten Daten (⇨ Abb. 26.10) in anderer Interpretation verwendet. Es soll sich bei den Variablen jetzt um Bewertungen von Schülerarbeiten durch drei Lehrer A, B und C handeln. Dafür wurden die Variablen in LEHR_A, LEHR_B und LEHR_C umbenannt (Datei LEHRER.SAV). Die Durchführung des Tests entspricht der Vorgehensweise in Kap. 26.6.1 mit dem Unterschied, dass nun „Kendall W“ gewählt wird.
In Tabelle 26.24 ist die Ergebnisausgabe des Tests zu sehen. Sie unterscheidet sich nicht von der in Tabelle 26.23, so dass auf eine Erläuterung verzichtet werden kann. Da „Signifikanz“ mit 0,006 kleiner ist als das gewählte Signifikanzniveau von z.B. 5 % ($\alpha = 0{,}05$), wird die Hypothese H_0 - die Bewertungen stimmen überein - abgelehnt.

Statistiken. ⇨ Kap. 26.6.1.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.24. Ergebnisausgabe des Kendall W-Test

Ränge

	Mittlerer Rang
lehr_a	1,60
lehr_b	1,85
lehr_c	2,55

Statistik für Test

N	20
Kendall-W[a]	,259
Chi-Quadrat	10,347
df	2
Asymptotische Signifikanz	,006

a. Kendalls Übereinstimmungskoeffizient

26.6.3 Cochran Q-Test

Dieser Test entspricht dem McNemar-Test mit dem Unterschied, dass er für mehr als zwei dichotome Variablen (z.B. „2“ = Erfolg einer Aktivität bzw. eines Einflusses, „1“ = nicht Erfolg) angewendet werden kann.

Die Prüfgröße Q wird - ausgehend von der Datenmatrix - aus den Häufigkeiten des Eintretens von „Erfolg“ ermittelt. Q ist wie folgt definiert:

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} ss_j^2 - (\sum_{J=1}^{k} ss_j)^2\right]}{k\sum_{i=1}^{n} zs_i - \sum_{i=1}^{n} zs_i^2} \qquad (26.8)$$

ss_j = Spaltensumme für Variable j (Häufigkeit des Erfolges, also z.B. von „2“)
zs_i = Zeilensumme für den Fall i (Häufigkeit des Erfolges, also z.B. von „2“)
k = Anzahl der Stichproben (Variablen)

Q ist asymptotisch chi-quadratverteilt mit $k-1$ Freiheitsgraden.

Das folgende Beispiel verwendet die Variablen aus der in Abb. 26.12 ausschnittsweise dargestellten Datei TESTAUFG.SAV. In den Variablen AUFG1, AUFG2 und AUFG3 ist erfasst, ob drei verschiedene Aufgaben von Studenten gelöst worden sind oder nicht. Zur Anwendung des Tests geht man wie in Abschnitt 26.6.1 beschrieben vor mit dem Unterschied, dass der Cochran Q-Test angeklickt wird.

In Tabelle 26.25 wird die Ergebnisausgabe dargestellt. Für die drei Variablen werden die Häufigkeiten des Auftretens der Werte „2“ (= Aufgabe gelöst) und „1“ (= Aufgabe nicht gelöst) aufgelistet. Es wird die Zahl der Fälle („N“), Cochrans Q, die Zahl der Freiheitsgrade (df) sowie das Signifikanzniveau für den Test angegeben. Da das ausgegebene Signifikanzniveau mit 0,076 größer ist als ein z.B. mit 5 % (α = 0,05) gewähltes, wird die Hypothese H_0 - der Lösungserfolg und somit der Schwierigkeitsgrad der Aufgaben unterscheiden sich nicht - beibehalten.

Statistiken. ⇨ Kap. 26.6.1.
Exakter Test. ⇨ Kap. 31.

Tabelle 26.25. Ergebnisausgabe des Cochran Q-Tests

Häufigkeiten

	Wert	
	1	2
aufg1	6	9
aufg2	11	4
aufg3	5	10

Statistik für Test

N	15
Cochrans Q-Test	5,167[a]
df	2
Asymptotische Signifikanz	,076

a. 2 wird als Erfolg behandelt.