PARAMETERFREIE VERFAHREN ZUR BEURTEILUNG VON ZWEI ABHÄNGIGEN STICHPROBEN IM UNTERRICHT

von ALFRED MÜLLER, Coburg

Zusammenfassung: Bei einem Einstellungstest soll die Konzentrationsfähigkeit mit einer neuen Aufgabenstellung überprüft werden. Zehn Bewerber um einen Arbeitsplatz unterzogen sich an zwei aufeinanderfolgenden Tagen dem bisherigen und dem neuen Test. Mit Hilfe von parameterfreien Verfahren (Vorzeichentest, Vorzeichen - Rang - Test, Randomisierungstest von Fisher) wurde nun untersucht, ob aus den (abhängigen) Stichprobenwerten die Nullhypothese, daß beide Tests in gleicher Weise die Konzentrationsfähigkeit messen, aufrecht erhalten oder ob eine signifikante Abweichung festgestellt werden kann. Die stochastischen Verfahren, ihre Voraussetzungen und ihr Vergleich wurden mit Schülern diskutiert, entwickelt und ausgeführt.

ZDM- Klassifikation: K74

1. Bereits im Heft 1/1987 dieser Zeitschrift (siehe [4]) wurde bei der Beurteilung von unabhängigen Stichproben im Unterricht die Motivation dargelegt, warum verteilungsfreie Methoden zur Beurteilung von Stichproben gewählt wurden: Die Verteilung der Testgröße kann bei einer kleinen Zahl von Messungen in der Regel nicht genau genug durch eine Normalverteilung approximiert werden, so daß Verfahren angewendet werden müssen, die nicht auf eine spezielle Verteilung zugeschnitten sind. Diese parameterfreien Testverfahren sind umfassender anwendbar, da sie im allgemeinen nur an bescheidene Voraussetzungen gebunden sind, außer Meßwerten auch Rangdaten verarbeiten, für kleine Stichproben geeignet sind und nur einfache Formeln verwenden, d. h. den Rechenaufwand gering halten.

Die in einem Leistungskurs Mathematik gemachten Erfahrungen bei der Besprechung von Rangsummentest von WILCOXON, U-Test von MANN und WHITHEY und X-Test nach van der WAERDEN zur Beurteilung zweier unabhängiger Stichproben wurden im oben genannten Heft genau beschrieben. Dort wurde bereits angedeutet, daß auch in einer freiwilligen Arbeitsgruppe außerhalb des Unterrichts parameterfreie Verfahren zum Testen an verbundenen, d. h. abhängigen, Stichproben angewendet wurden.

2. Die spezielle Betrachtungsweise von Stichprobenergebnissen mit wenigen Werten war allen anwesenden Schülern nach der genannten Vorübung hinreichend geläufig, so daß gleich die Problematik der Entstehung verbundener, abhängiger Stichproben diskutiert werden konnte. Oft steht man in der Praxis vor der Aufgabe, zwei verschiedene Verfahren oder Methoden miteinander zu vergleichen. Sicher kann man so vorgehen, daß man beide an zwei Gruppen von Versuchsobjekten testet und dann diese beiden, allerdings unabhängigen Stichproben nach einer dafür geeigneten Methode (z. B. wie in [4]) miteinander vergleicht. Wie muß man vorgehen, um abhängige Stichproben zu erhalten, und wie kann man diese dann beurteilen?

Die Schülervorschläge trafen dann auch die in der Praxis am häufigsten angewandte Art, solche abhängigen Vergleichspaare zu gewinnen: Dieselbe Stichprobe nimmt an der Untersuchung zweimal teil, wobei man zwischen der ersten und der zweiten Meßwiederholung keinerlei Wechselwirkungen derart zuläßt, daß Lern-, Übungs- und Ermüdungseffekte oder Veränderungen des untersuchten Merkmals auftreten. Hier wurde besonders herausgestellt, daß die so gebildeten abhängigen Stichproben beim Testen von verschiedenen Verfahren oder Methoden einen großen Vorteil gegenüber unabhängien Stichproben zeigen, denn man kann verbundene Stichproben viel genauer vergleichen, weil die Streuung, die sonst zwischen den verschiedenen Objekten der Stichprobe besteht, wegfällt. Man modifiziert häufig die Versuchsausführung so, daß man die 1. Hälfte der Stichprobe dem Verfahren 1 und die 2. Hälfte der Stichprobe dem Verfahren 2 unterwirft und dann umgekehrt.

Besonders die möglichen Veränderungen führen dazu, daß es häufig nicht gelingt, ein Element einer Stichprobe zweimal zu untersuchen. Die Schüler erkannten auch hier die Möglichkeit, die zur Abhilfe angewendet wird: Man unterteilt die Stichprobe etwa durch Losen in Paare und führt dann die 1. Untersuchung an den ersten Elementen der Paare und die 2. Untersuchung an

den zweiten Elementen der Paare aus. In jedem der beiden Fälle erhält man zwei abhängige oder verbundene Stichproben, die wegen der Paarung immer den gleichen Umfang besitzen müssen.

3. Zur Untersuchung solcher abhängiger Stichproben wurde den Schülern folgendes Beispiel vorgelegt: Die Personalabteilung eines kleinen Industrieunternehmens plant ihre Einstellungstests selbst, unter anderem soll die Konzentrationsfähigkeit überprüft werden. Bisher mußten die Prüflinge über 60 Minuten die Ergebnisse vorgegebener einfachster Grundrechenarten durch die Einerziffer des Ergebnisses angeben. Ein neuer Test, der den Vergleich einfacher Grundrechenarten über 60 Minuten verlangt, soll diesen ersetzen, wobei als Ergebnis nur das richtige der Zeichen "<", "=" oder ">" anzugeben ist. Zehn Bewerber um einen Ausbildungsplatz mußten an zwei aufeinanderfolgenden Tagen beide Tests bearbeiten, wobei angenommen werden kann, daß die Versuchspersonen an beiden Tagen unter den gleichen Voraussetzungen an den Start gingen. Die Ergebnisse wurden als Prozentwerte im Vergleich zu einem erwarteten Wert gemessen. Es ergab sich folgendes Resultat:

Person	1	2	3	4	5	6	7	8	9	10
Testerg.1 in %	106	104	100	103	101	102	105	105	106	108
Testerg.2 in %	102	94	102	103	98	104	97	105	105	102

Tabelle 1

Kann auf Grund dieser Stichprobenwerte die Behauptung aufrecht erhalten werden, daß beide Tests in gleicher Weise die Konzentrationsfähigkeit messen, oder gibt es zwischen den beiden Meßreihen eine signifikante Abweichung?

4. Bevor man an eine nähere Untersuchung über eventuell auftretende Unterschiede herangeht, muß man sich überlegen, was

mit den Paaren geschehen soll, deren Differenz den Wert Null hat. In der Regel muß man davon ausgehen, daß diese Nulldifferenzen infolge ungenauer Messung des an sich stetig verteilten Merkmals aufgetreten sind, d. h. man könnte sich etwa durch einen Münzenwurf für ein Vorzeichen und damit für eine positive oder negative Abweichung entscheiden und etwa bei Rangvergaben diese Werte mit einbeziehen. Die Schülermeinung, daß diese Nulldifferenzen nichts zur Entscheidung "besser" oder "schlechter" beitragen und man solche Paare einfach wegläßt, d. h. den Stichprobenumfang von n auf n' reduziert, deckt sich mit der in der Praxis am häufigsten angewandten Methode der Reduzierung des Stichprobenumfanges. Sie ist allerdings nur angebracht, wenn nicht zu viele Nulldifferenzen vorliegen.

5. Die Erfahrung bei der Beurteilung von unabhängien Stichproben (siehe [4]) ließ nur sehr vorsichtig die Meinung aufkommen, daß bei sechs positiven und zwei negativen Abweichungen wohl der zweite Test eine höhere Anforderung an die Konzentrationsfähigkeit stellen wird als der erste. Sofort kam aber in der Entgegnung, daß man ja erst einmal wissen müsse, mit welcher Wahrscheinlichkeit zwei unter den acht Stichprobenwerten mit negativer Differenz auftreten und bei welcher Wahrscheinlichkeit man eine signifikante Abweichung bestätigen will; denn bei diesen beiden Stichproben treffe zwar das Überwiegen der positiven Differenzen zu, aber die Aufgabe des Testens sei es, aus diesem Ergebnis bis auf einen Fehler α auf die Allgemeingültigkeit dieser Aussage zu schließen. Die Problematik, wie vorsichtig man entscheiden will oder soll und die dazu gehörenden Konsequenzen sind bei den vorher besprochenen Test - verfahren ausführlich erörtert worden. Die Schüler wußten, daß der α-Fehler nur eine bedingte Wahrscheinlichkeit darstellt, d. h. daß die Nullhypothese H_0 nur in 100α % der Fälle verworfen wird, wenn sie zutrifft, und nicht, daß der Anteil der Fehlurteile insgesamt 100α % beträgt. Ferner war mit den Schülern schon ausführlich diskutiert worden, daß man in einem solchen Fall der Beurteilung von Stichproben von $\alpha = 5$ % ausgeht,

da H_0 weder einen theoretisch bedeutsamen Tatbestand impliziert, noch von überragender praktischer Bedeutung oder finanzieller Auswirkung ist, wo man als statistische Sicherheit 99 %, d. h. $\alpha = 1$ % verwenden würde. Außerdem war in einer Diskussion mit den Schülern geklärt worden, daß der zu prüfende Unterschied von vorneherein keine bestimmte Richtung besitzt, d. h. ein zweiseitiger Test zu verwenden ist.

6. Jetzt konnten die Stichproben näher untersucht werden. Wie bereits oben erwähnt, war der erste Ansatzpunkt die unterschiedlichen Vorzeichen der Differenzen. Die Schüler wußten sofort, daß jedes der beiden Vorzeichen mit der Wahrscheinlichkeit $p = 1/2$ auftreten kann. Man wird also dann unter der Annahme, daß beide Tests gleiche Anforderungen an die Konzentrationsfähigkeit stellen (Nullhypothese H_0), etwa gleichviele positive wie auch negative Vorzeichen erwarten, d. h. der Median der Differenzen hat den Wert Null. Ist der Unterschied in der Anzahl der Vorzeichen nicht mehr zufallsbedingt, so muß die Abweichung von der Gleichverteilung mit $p = 1 - p = 1/2$ signifikant sein, was mit Hilfe der zu $p = 1/2$ gehörenden Binomialverteilung berechnet werden kann. Als Testgröße Z verwendet man entweder die Anzahl Z_+ der positiven oder die Anzahl Z_- der negativen Vorzeichen, wobei in jeder Stichprobe $Z_+ + Z_- = n$ gilt. Wir haben vereinbart, diejenige mit der kleinsten Anzahl in der Stichprobe als Prüfgröße zu verwenden. Ohne Schwierigkeiten erarbeiteten die Kollegiaten, daß die Nullhypothese H_0 bei n Beobachtungspaaren ohne Nulldifferenzen in einem zweiseitigen Test auf dem Signifikanzniveau α genau dann abgelehnt wird, wenn für das Ereignis E, daß höchstens k positive oder negative Vorzeichen auftreten, gilt:

$$P(E) = 2 \cdot B^n_{0,5}(Z \leq k) = 2 \cdot \sum_{i=0}^{k} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i} = \frac{2}{2^n} \sum_{i=0}^{k} \binom{n}{i}$$

$$= \frac{1}{2^{n-1}} \sum_{i=0}^{k} \binom{n}{i} \leq \alpha \quad .$$

Für kleine n besteht die Möglichkeit, die Überschreitungswahrscheinlichkeit zu jeder Zahl Z = k auszurechnen und die Grenzzahl k_0 zum Signifikanzniveau α zu bestimmen. Es gilt dann, daß H_0 abgelehnt wird, falls $Z < k_0$ oder $N - Z > n - k_0$ gilt. Da nur ganzzahlige Grenzanzahlen k_0 auftreten können, kann das Signifikanzniveau α nicht voll ausgenutzt werden. In unserem Beispiel wurden die ohne die beiden Nulldifferenzen verbleibenden acht Paare in der alten Reihenfolge angeordnet, aber neu numeriert. Es ergab sich folgende Tabelle:

Person	1	2	3	4	5	6	7	8
Testerg.1 in %	106	104	100	101	102	105	106	108
Testerg.2 in %	102	94	102	98	104	97	105	102
Vorzeichen der Differenz (1)-(2)	+	+	-	+	-	+	+	+

Tabelle 2

Für die Prüfgröße $Z_- = 2$ gilt:

$$2 \cdot B^8_{0,5}\ (Z \leq 2) = \frac{1}{2^7} \sum_{i=0}^{2} \binom{8}{i} = \frac{1+8+28}{2^7} = \frac{37}{128} = 28{,}91\ \%.$$

H_0 kann auf dem 5 %-Signifikanzniveau nicht abgelehnt werden. Wenn man für unser Beispiel die Wahrscheinlichkeiten für alle k bestimmt, um die Grenzwahrscheinlichkeit zu erhalten, so findet man:

$2 \cdot B^8_{0,5}\ (Z=0) = 0{,}78\ \%$ $\qquad$ $2 \cdot B^8_{0,5}\ (Z \leq 1) = 7{,}03\ \%$

$2 \cdot B^8_{0,5}\ (Z \leq 2) = 28{,}91\ \%$ $\qquad$ usw.

Für unser Beispiel heißt dies, daß H_0 abgelehnt werden müßte, wenn $Z<1$ oder $Z>7$ gilt. Diese Grenzwerte sind für bestimmte Stichprobenlängen n und Signifikanzniveaus tabelliert, z. B. in [6].

Das eben beschriebene Prüfverfahren wird im allgemeinen als VORZEICHENTEST bezeichnet. Dieser ist wahrscheinlich der älteste Test überhaupt und wurde bereits 1710 von dem schottischen Arzt John ARBUTHNOT zur Widerlegung der Auffassung der Gleichhäufigkeit von Knaben- und Mädchengeburten verwendet, 1934 wiederentdeckt von R. A. FISHER und 1946 von DIXON und MOOD ausführlich beschrieben. Die wesentlichen Vorzüge des Vorzeichentestes wurden mit den Schülern diskutiert: Da der Vorzeichentest nur einen geringen rechnerischen Aufwand verlangt (die Werte der Binomialverteilung sind tabelliert bzw. können durch die Normalverteilung angenähert werden), wird er häufig als Schnelltest verwendet, zumal man auch bei grösseren Stichproben schnell einen Überblick gewinnen kann; denn wenn bereits ein wenig effizienter Test wie der Vorzeichentest Signifikanz ergibt, so kann man sich wirksamere Tests ersparen. Der Begriff der Wirksamkeit bzw. der Effizienz eines Tests, d. h. auf welchen Bruchteil man bei Verwendung eines für die Normalverteilung entwickelten Tests den Stichprobenumfang verringern kann, um die gleiche Trennschärfe (Gütefunktion) zu erhalten wie bei einem verteilungsfreien Test, wurde nochmals wiederholt. Der Vorzeichentest verlangt keine Voraussetzungen bezüglich der Verteilungsform der Differenzen oder über die Homogenität der Wertepaare, d. h. in unserem Beispiel dürfen sich die Paare in den nicht untersuchten Merkmalen wie z. B. Alter, Geschlecht, Intelligenz etc. unterscheiden. Die Wertepaare selbst müssen untereinander unabhängig sein, d. h. keines der Elemente eines Paares oder keines der Paare darf zwei- oder mehrmals in der Untersuchung auftreten. Der größte Vorteil des Vorzeichentests ist aber seine Anwendbarkeit auch dann, wenn eine quantitative Anordnung der Paare nicht möglich ist, d. h. gar keine Meßwerte vorliegen, sondern nur Daten, die die Anwendung der Größer - Kleiner - Relation innerhalb der Paare gestatten. Falls aber Maßzahlen wie in unserem Beispiel vorliegen, wird nur sehr wenig von den in den Maßzahlen steckenden Informationen ausgenutzt.

7. Wie kann man diese Informationen besser verwerten? Durch die Vorschulung bei der Beurteilung von unabhängigen Stichproben

mit zwei Rangtests (Rangsummentest und U-Test, siehe [4]) kam natürlich der Vorschlag, die Rangplätze der Differenzen zu berücksichtigen. Da negative und positive Differenzen auftreten, ist eine Anordnung nur sinnvoll, wenn man den Beträgen der Differenzen die Rangzahlen 1 bis n zuordnet. Ebenso wurde vorher geklärt, daß Nulldifferenzen wieder weggelassen werden und die dem Betrag nach gleichen Differenzen einen Durchschnittsrang zugeordnet bekommen. Als Prüfgröße S wurde die Summe S_- der Ränge, zu denen ein negatives Vorzeichen gehört, bzw. die Summe S_+ der Ränge der positiven Vorzeichen vorgeschlagen. Aus der Summenformel für die ersten n natürlichen Zahlen ergibt sich für die Zahlenwerte bei jeder Stichprobe $s_+ + s_- = \frac{n(n+1)}{2}$. Vereinbarungsgemäß beschränken wir uns beim Test auf die kleinere der beiden Rangsummen.

Wie sieht nun bei der Gültigkeit der Nullhypothese die Verteilung der Prüfgröße S aus? Wie muß zur Ablehnung oder Nichtablehnung von H_0 entschieden werden? Bis auf geringe Hinweise entwickelten die Kollegiaten das gewünschte Prüfverfahren selbst.

Jeder der n Ränge kann ein positives oder negatives Vorzeichen enthalten. Es gibt also insgesamt 2^n mögliche Vorzeichenanordnungen. S kann dabei alle Werte zwischen 0 und $\frac{n(n+1)}{2} = r$ annehmen, wobei wegen der Symmetrie der Binomialkoeffizienten $\binom{r}{k} = \binom{r}{r-k}$ für die Anzahl der möglichen Anordnungen mit k bzw. r-k Vorzeichen einer Art die Verteilung von S symmetrisch um den Erwartungswert $\frac{r}{2} = \frac{n(n+1)}{4}$ ist. Deswegen wird man den Ablehnungsbereich bei einem zweiseitigen Test aus gleich vielen der kleinsten und der größten Werte der Prüfgröße S bilden, d.h. bei einem Signifikanzniveau α liegen im Ablehnungsbereich $\frac{k}{2}$ kleinste bzw. $\frac{k}{2}$ größte Werte, wenn k die nächstkleinere ganze Zahl von $\alpha \cdot 2^n$ ist. Dann gilt nämlich $\frac{k}{2^n} \leq \alpha$.

In unserem Beispiel wurden zuerst die Rangplätze bestimmt. Es ergab sich folgende Tabelle:

Person	1	2	3	4	5	6	7	8
Testerg.1 in %	106	104	100	101	102	105	106	108
Testerg.2 in %	102	94	102	98	104	97	105	102
Differenz (1) - (2)	+4	+10	-2	+3	-2	+8	+1	+6
Rangplatz	5	8	2,5	4	2,5	7	1	6

Tabelle 3

Es gibt also $2^8 = 256$ verschiedene Anordnungen der Vorzeichen, d. h. auf dem 5 %-Signifikanzniveau sind es bei einem zweiseitigen Test mit symmetrischer Anordnung wegen $0{,}05 \cdot 256 = 12{,}8$ höchstens 12 Möglichkeiten, die zur Ablehnung der Nullhypothese führen, und zwar diejenigen, für die S kleinste oder größte Werte in der Rangsumme besitzt. Im Beispiel gilt $S_+ = 31$ und $S_- = 5$. Wir verwenden absprachegemäß S_- als Testgröße und bestimmen nun die in Frage kommenden Rangsummen. Für kleine n lassen sich diese leicht angeben, ansonsten findet man in [2] ein Computerprogramm zu deren Bestimmung.

1	2	3	4	5	6	7	8	S_-
+	+	+	+	+	+	+	+	0
-	+	+	+	+	+	+	+	1
+	-	+	+	+	+	+	+	2
+	+	-	+	+	+	+	+	3
-	-	+	+	+	+	+	+	3
+	+	+	-	+	+	+	+	4
-	+	-	+	+	+	+	+	4
+	+	+	+	-	+	+	+	5
-	+	+	-	+	+	+	+	5
+	-	-	+	+	+	+	+	5
.....								
-	-	-	-	+	-	-	-	31
+	-	-	+	-	-	-	-	31
-	+	+	-	-	-	-	-	31
+	-	+	-	-	-	-	-	32
-	-	-	+	-	-	-	-	32
+	+	-	-	-	-	-	-	33
-	-	+	-	-	-	-	-	33
-	+	-	-	-	-	-	-	34
+	-	-	-	-	-	-	-	35
-	-	-	-	-	-	-	-	36

Tabelle 4

Bei vorgegebenem Signifikanzniveau α soll der wirkliche Fehler α' höchstens gleich α sein. Wir sehen aus Tabelle 4, daß wir die Nullhypothese ablehnen werden, wenn die Rangsumme S_- einen der Werte 0, 1, 2, 3 bzw. 33, 34, 35, 36 annimmt. Der wirkliche Fehler α' beträgt dann nur $\alpha' = \frac{10}{256} = 3{,}91\ \%$, aber wenn man die Werte 4 bzw. 32 der Ransumme noch zum Ablehnungsbereich hinzunehmen würde, erhielte man dann $\alpha' = \frac{14}{256} = 5{,}47\ \%$, $5{,}47\ \% > 5\ \% = \alpha$. Der Test hat also auf dem 5 %-Niveau keine größere Wirksamkeit als auf dem 3,91 %-Signifikanzniveau.

Da in unserem Beispiel $S_- = 5$ gilt, kann H_0 nicht verworfen werden, d. h. auch nach diesem Test zeigt das Stichprobenergebnis keine signifikante Abweichung.

Genauso könnte man den Test mit der Überschreitungswahrscheinlichkeit P ausführen, d. h. man bestimmt die Wahrscheinlichkeit dafür, daß $S_- \leq 5$ bzw. $S_- \geq 31$ gilt. Dann erhält man mit $P = \frac{20}{256} = 7{,}81\ \%$ auch die Nichtablehnung von H_0. Für größere n kann man die Grenzrangsummen aus Tabellen wie z. B. in [5] ablesen.

Der mit den Schülern hergeleitete Test wurde 1945 erstmals von WILCOXON angegeben und trägt in der englischsprachigen Literatur den Namen "Wilcoxon-matched-pairs-signed-rank-test". Die in der deutschsprachigen Literatur meist gebrauchte Übersetzung "Wilcoxons-Vorzeichen-Rang-Test" scheint die richtigere zu sein im Vergleich zu der auch möglichen Übersetzung als "Wilcoxons-Rang-Vorzeichen-Test", da es sich ja mehr um einen Rang- bzw. Rangsummentest handelt als um einen mit Rangzahlen gewichteten Vorzeichentest. Die Anwendung des Testverfahrens setzt voraus, daß die Differenzen unabhängige, symmetrisch und stetig verteilte Zufallsgrößen sind, was im Beispiel erfüllt war.

8. Die Schüler bemerkten sehr richtig, daß die Betragsunterschiede zwischen den einzelnen Rangplätzen nicht gleichmäßig sind. Könnte nicht auch die absolute Größe der Differenzen bei der Entscheidungsfindung herangezogen werden? Wie müßte ein einfacher

Test aussehen, der diese Beträge nach ihrem Wert berücksichtigt? Ein Kollegiat schlug vor, die Beträge zu verwenden und dann genauso zu verfahren wie bei Vorzeichen-Rang-Test. Ein anderer bezweifelte, ob dann das gleiche Entscheidungsverfahren wie vorher beibehalten werden könnte. Damit war der Weg für das nächste Testverfahren bereits vorgezeichnet: Es sollten die Vorzeichen nicht mit Rangplätzen, sondern direkt mit den numerischen Werten der Differenzen gewichtet werden. Wir wußten wieder, daß bei n Paaren 2^n Differenzen mit unterschiedlichen Vorzeichen möglich und gleichwahrscheinlich sind. Als Testgröße S bot sich dabei die Summe der Differenzen an, d. h. $S = \sum_{i=1}^{n} d_i$. Bildet man für alle 2^n Vorzeichenkombinationen der n Differenzen diese algebraische Summe, so erhält man die Verteilung der Testgröße S.

Ein Testverfahren, das zu diesen Überlegungen paßte, wurde 1936 von R. A. FISHER angegeben und 1937 von E. J. G. PITMAN vervollständigt. Dieser Randomisierungstest von FISHER ist ein sogenannter "bedingter" Test, weil er auf der Voraussetzung beruht, daß die vorliegende Stichprobe ein genaues Abbild der zugehörigen Grundgesamtheit ist. Damit ist klar, daß die Prüfgröße nicht so wie bei Rangtests für bestimmte Stichprobenumfänge tabelliert werden kann, weil sich die Verteilung von S von Stichprobe zu Stichprobe ändert.

Nach FISHER werden wir H_0 ablehnen, wenn sich die Prüfgröße S unter den 5 % "Randwerten" der Prüfverteilung befindet oder wenn die Überschreitungswahrscheinlichkeit P der beobachteten Testgröße S unter 5 % liegt. Ist z_1 die Anzahl der Werte, die größer (kleiner) als der beobachtete S-Wert sind, und z_2 die Zahl der Werte, die gleich dem beobachteten S-Wert sind, so folgt nach dem Vorschlag von FISHER für den zweiseitigen Test eine Ablehnung der Nullhypothese H_0, falls $P = \frac{2 \cdot z_1 + z_2}{2^n} \leq \alpha$ gilt.

Für unser Beispiel bestimmen wir nach der folgenden Tabelle 5 die Randwerte der Prüfverteilung von S. Dazu sind der Übersichtlichkeit halber die Beträge der Differenzen der Größe nach ge-

ordnet. Der symmetrische Anteil kleinster S-Werte wurde weggelassen.

								S
10	8	6	4	3	2	2	1	36
10	8	6	4	3	2	2	-1	34
10	8	6	4	3	2	-2	1	32
10	8	6	4	3	-2	2	1	32
10	8	6	4	3	2	-2	-1	30
10	8	6	4	3	-2	2	-1	30
10	8	6	4	-3	2	2	1	30
10	8	6	4	-3	2	2	-1	28
10	8	6	4	3	-2	-2	1	28
10	8	6	-4	3	2	2	1	28
10	8	6	4	3	-2	-2	-1	26
.....								

Tabelle 5

In unserem Beispiel beträgt S = 28, d. h. wegen $P = \frac{2 \cdot 7 + 3}{256} =$ = 6,64 % kann H_0 nicht abgelehnt werden.

9. Von besonderer Wichtigkeit ist es, nach der Ausführung dieser Tests über die Verteilung der Differenzen nochmals auf die Unterschiede und die Anwendungsmöglichkeiten einzugehen. Da dies schon während der Herleitung der Testverfahren immer wieder geschehen ist, sollen deshalb nur noch einige wesentliche Punkte erwähnt werden.

(a) Der Vorzeichentest nimmt nur auf das Vorzeichen der Meßwertdifferenzen Bezug und prüft, ob die Zahl der positiven Vorzeichen mit der der negativen übereinstimmt oder nicht, d. h. der Vorzeichentest spricht allein auf den Median der Differenzverteilung und somit auf den Unterschied der Medianwerte der beiden abhängigen Stichproben an. Die Vorzeichen werden also alle mit dem Wert 1 gewichtet. Sind Ausreißerdifferenzen zu erwarten und will man diesen Tatbestand nicht stärker berücksichtigen, so empfiehlt sich der Vorzeichentest, da Ausreißer zwar den Durchschnitt, aber nicht den Median der Differenzenverteilung beeinflussen.

(b) Der Randomisierungstest von FISHER stellt gewissermaßen den

anderen Extremfall dar. Er geht von der algebraischen Summe der Meßwertdifferenzen, d. h. dem n-fachen des Durchschnittes aus und prüft nach, ob dieser in signifikanter Weise vom Erwartungswert Null abweicht. Die Vorzeichen werden mit den numerischen Größen der zugehörigen Differenzen gewichtet und erst dann zur Prüfgröße S addiert, d. h. der Test von FISHER spricht besonders gut den Unterschied der Durchschnitte der beiden abhängigen Stichproben an, Ausreißer werden entsprechend stark gewichtet.

(c) Der Vorzeichen-Rang-Test von WILCOXON nimmt zwischen den beiden schon angesprochenen Testverfahren eine vermittelnde Stellung ein. Zwar werden auch hier die Vorzeichen gewogen, ehe sie zur Prüfgröße aufaddiert werden, aber nicht mit den numerischen Werten der Differenzen, sondern mit ihren Rangzahlen, d. h. der Vorzeichen-Rang-Test wird auf einen Aspekt der Lageverteilung der Differenzen ansprechen, der, falls es solch einen Parameter geben würde, zwischen Median und Durchschnitt der Differenzverteilung liegen müßte. Auch bei Ausreißerdifferenzen nimmt der Vorzeichen-Rang-Test eine Mittelstellung ein, da er weder alle Differenzen gleichgewichtig macht, noch den Ausreißern ein verhältnismäßig großes Gewicht gibt.

(d) Der Begriff der Wirksamkeit eines Testverfahrens und der asymptotischen Wirksamkeit wurde schon mehrmals angesprochen. Für die drei besprochenen Testverfahren sind in [7] die asymptotischen Wirksamkeiten (für große n) hergeleitet, nämlich für den Vorzeichentest 64 %, für den Vorzeichen-Rang-Test 95 % und für den Randomisierungstest von FISHER 100 %. Diese angegebenen Wirksamkeiten beziehen sich natürlich nur auf solche Beispiele, in denen eine Normalverteilung vorliegt und Students t-Test angewendet werden kann. In unserem Beispiel war für die Differenzen sicher keine Normalverteilung erkennbar, so daß der t-Test unter Umständen sogar von geringerer Wirksamkeit wäre und ein falsches Ergebnis liefern würde.

10. Mit einem Beispiel, das im Vorzeichentest nicht, im Vorzeichen-Rang-Test auf dem 5 %-Niveau Signifikanz zeigt, wurde die Besprechung der Testverfahren abgeschlossen.
Beispiel: In einem Labor werden n = 9 Proben aus einer Mülldeponie auf einen ganz bestimmten Schadstoff hin mit Hilfe von zwei unterschiedlichen Verfahren überprüft und der Anteil in Promille angegeben. Auf dem 5 %-Signifikanzniveau soll überprüft werden, ob beide Verfahren im Mittel die gleichen Meßwerte liefern.

Probe	1	2	3	4	5	6	7	8	9
Meßerg.1 in ‰	3,1	2,8	3,5	2,9	3,1	3,5	2,9	3,5	3,4
Meßerg.2 in ‰	3,0	2,9	3,4	3,1	3,2	3,1	2,7	3,3	3,2

Tabelle 6

LITERATUR

[1] CLAUSS, G. und EBNER, H.: Grundlagen der Statistik. Frankfurt a. M.: Harri Deutsch 1982

[2] DIFF: Schätzen und Testen SR 4. Tübingen, 1983

[3] DIFF: Stochastik MS 4. Tübingen, 1981

[4] MÜLLER, A.: Beurteilung von zwei unabhängigen Stichproben im Unterricht. Stochastik in der Schule: Heft 1, 7. Jahrgang (1987)

[5] PFANZAGEL, J.: Allgemeine Methodenlehre der Statistik II. Berlin: Walter de Gruyter 1974

[6] RÖHR, M., LOHSE, H. und LUDWIG, R.: Statistische Verfahren. Frankfurt a. M.: Harri Deutsch 1983

[7] v. d. WAERDEN, B. L.: Mathematische Statistik. Berlin: Springer 1971

[8] WEBER, H.: Einführung in die Wahrscheinlichkeit und Statistik. Stuttgart: Teubner 1983