Stochastik in der Schule, Heft 1, 7. Jg. (1987)

BEURTEILUNG VON ZWEI UNABHÄNGIGEN STICHPROBEN IM UNTERRICHT

von Alfred Müller, Coburg

Zusammenfassung: Die Nullhypothese, daß mit zwei verschiedenen Meßverfahren gleiche Haltbarkeitsdauern eines Materials gemessen werden, wurde mit parameterfreien Verfahren (Rangsummentest, U-Test, X-Test)geprüft. Diese Tests wurden mit den Schülern entwickelt und ausgeführt, dabei auftretende Fragen und Probleme sind angesprochen. Ausgangspunkt waren zwei unabhängige Stichproben, die von Haltbarkeitsmessungen nach den beiden Verfahren herrührten.

ZDM-Klassifikation: K74, C64

1. Im Stochastikunterricht der Sekundarstufe II wird im Vergleich mit der zur Verfügung stehenden Zeit immer noch zu großen Wert auf die Wahrscheinlichkeitsrechnung gelegt, obwohl sie nur im Hintergrund zur theoretischen Rechtfertigung für die Anwendung statistischer Methoden stehen sollte. Schüler lernen deshalb in der Regel nur das Testen von Hypothesen bei binomial- bzw. normalverteilten Grundgesamtheiten kennen, allenfalls noch das Testen auf Vorliegen einer Normalverteilung mit dem t - Test. Beobachtungen zeigen, daß sie deshalb auch versuchen, jeden vorgelegten Datensatz in eine der genannten Verteilungen hineinzupressen. Sicher ist dies näherungsweise möglich, wenn nur die Stichprobenlänge groß genug ist. Wie verfährt man aber dann, wenn die Anzahl der Messungen zu klein ist, um die Verteilung durch eine Normalverteilung genügend genau approximieren zu können? Wie man Schülern leicht klar machen kann, muß es auch Lösungsmöglichkeiten geben, wenn man von vornherein weiß, daß keine Normalverteilung vorliegt. Es gibt zwar Transformationen, die jede beliebige Verteilung in eine Normalverteilung überführen, aber sie verzerren die Skala der Beobachtungswerte häufig sehr stark. Deshalb wurden in den letzten fünfzig Jahren Prüfverfahren entwickelt, die Verteilungen miteinander vergleichen lassen, ohne daß ihre Form bekannt ist. Im allgemeinen wird nur vorausgesetzt, daß ihre Verteilungsfunktionen stetig sind. Da es sich um Vergleiche von Verteilungen und nicht von Parametern (wie z. B. μ, σ usw.) handelt, bezeichnet man diese Verfahren als parameterfreie oder nichtparametrische Testverfahren.

Einem LK-Mathematik wurde nach dem Besprechen des Testens von Hypothesen folgendes Beispiel, dessen Stichprobenwerte aus der Prüfabteilung eines eher kleinen Industrieunternehmens stammen, zur Beurteilung vorgelegt.

Beispiel: Ein zu einem mechanischen Getriebe gehörendes Zahnrad wurde bisher aus einer Metallegierung gefertigt. Wegen des Gewichtes und wegen des Herstellungspreises soll dieses Teil durch eines aus Kunststoff ersetzt werden. Vor dem Einbau des Teiles wurden auf den Meßplätzen des Unternehmens mit zwei unterschiedlichen Verfahren Haltbarkeitsmessungen vorgenommen. Aus dem gleichen Probeguß wurden je sechs Zahnräder ausgewählt und unabhängig voneinander geprüft. Wegen eines Meßplatzdefektes ergaben sich nach dem ersten Verfahren nur fünf Daten. Die in Tabelle 1 angegebenen Werte sind im Vergleich zu einer vom Betrieb vorgegebenen Lebensdauer angegeben. Da alle Messungen unabhängig voneinander ausgeführt wurden, sind die Werte in Tabelle 1 bereits in steigender Reihenfolge angeordnet.

Rangfolge der Messungen	Lebensdauer 1. Verfahren x - Werte	Lebensdauer 2. Verfahren y - Werte
1	1.80	1.72
2	1.85	1.74
3	1.91	1.86
4	1.92	1.93
5	2.02	1.95
6	-	1.98

Tab. 1

Da die x - Werte alle unter den gleichen Versuchsbedingungen beobachtet wurden, kann angenommen werden, daß sie alle die gleiche Verteilungsfunktion besitzen. Das gleiche kann für die y - Werte festgestellt werden.

Getestet werden soll nun die Nullhypothese H_0, daß mit beiden Verfahren im Mittel die gleiche Lebensdauer gemessen wird, d. h. daß die festgestellten Unterschiede in den Lebensdauern rein zufällig sind oder die x - und y - Werte alle die gleiche Verteilungsfunktion besitzen.

Wie kann man nun die Daten beurteilen, wie eine Entscheidung

über Ablehnung bzw. Nichtablehnung der Nullhypothese fällen? Schülerbeobachtungen und -vorschläge werden zusammen mit den erarbeiteten Testverfahren im folgenden dargestellt.

2. Es gab am Anfang zwei Hauptschwierigkeiten zu überwinden. Die erste kam von Schülern, die trotz einjährigen Unterrichts in Stochastik noch immer nicht die "besondere" Denkweise dieses Teilgebietes der Mathematik erspürt hatten. Sie schlugen vor, daß wegen $\bar{x} = 1.90$ und $\bar{y} = 1.86$ man mit dem 1. Verfahren höhere Lebensdauern messe als mit dem zweiten. Mitschüler selbst berichtigten diese Anschauung, indem sie darlegten, daß dies zwar für die vorliegenden Stichproben zutreffe, unsere Aufgabe es aber sei herauszufinden, ob aus diesen beiden Stichproben bis auf einen gewissen Fehler α auf die Allgemeingültigkeit dieser Aussage geschlossen werden könne.

Die zweite Schwierigkeit lag in der schon vorher erwähnten Auffassung, von einer vorliegenden Normalverteilung auszugehen, wenngleich einige aus früheren Überlegungen anmerkten, daß dies wegen der geringen Anzahl von Meßwerten problematisch sei. Da alle von vorherigen Aufgaben zum Vergleich einer empirischen Verteilung mit einer Normalverteilung noch im Besitz von Wahrscheinlichkeitspapier waren, bot sich als Hausaufgabe an, die Verteilung der x - und y - Werte in ein solches Papier einzutragen. Obwohl man sicher auch bei wenigen, stark streuenden Punkten eine Gerade einzeichnen kann, waren die meisten der Schüler überzeugt, daß mit diesen wenigen Werten die Annäherung durch eine Normalverteilung nicht genügend genau sei. Damit waren die Kursteilnehmer soweit, daß sie anfingen, die Stichproben genauer zu betrachten, um andere Möglichkeiten zur Entscheidung über die aufgestellte Behauptung zu finden.

3. Die Schüler erkannten dann als Besonderheit, daß in der Reihe der y - Werte die kleinste und in der Reihe der x - Werte die größte Lebensdauer vorkommen. Um dies deutlicher zu machen, ordneten wir die Messungen in einer Rangordnung an.

Rangnummer	1	2	3	4	5	6	7	8	9	10	11
Meßwert	1.72	1.74	1.80	1.85	1.86	1.91	1.92	1.93	1.95	1.98	2.02
x - oder y - Wert	y	y	x	x	y	x	x	y	y	y	x

Tabelle 2

Man stellt fest, daß zwei y - Werte kleiner als alle x - Werte und ein x - Wert größer als alle y - Werte sind. Reicht dies aus, um bereits eine Aussage machen zu können? Immerhin tritt das Ereignis E, daß zwei oder mehr y - Werte "vorne" und ein oder mehr x - Werte "hinten" stehen mit einer Wahrscheinlichkeit

$$P(E) = 1 - \frac{\binom{7}{3} + \binom{7}{4} + \binom{8}{4} + \binom{8}{5} + \binom{9}{5}}{\binom{11}{6}} = 30.3\ \%$$

auf, ist also kein seltenes Ereignis, d.h. ohne Berücksichtigung der wirklichen Werte (liegen sie knapp drüber oder drunter usw.) läßt sich schwerlich eine Aussage machen.

Hier wurde wenigstens darauf hingewiesen, daß es Schnelltests gibt, die auf der Basis der Überschreitungszahlen arbeiten. Voraussetzung für die Anwendung dieser Tests ist aber wie im Beispiel, daß die eine Stichprobe den höchsten und die andere Stichprobe den niedrigsten Wert der vereinigten Stichprobe enthält.

4. Die Schnelltests berücksichtigen nur die Ränder der gemeinsamen Verteilung. Die Kursteilnehmer erachteten es aber sofort für wichtig, daß man auch die Lage der Werte im "Innern" kennen müsse. Aus einer Diskussion heraus, wie man die Lage dieser Werte charakterisieren könne, kam der Vorschlag, eine Testmöglichkeit zu suchen, die nicht die Größe der Meßwerte, sondern nur ihre Rangfolge innerhalb der gemeinsamen Stichprobe verwendet. Da nur die Rangzahlen, nicht aber die Stichprobenwerte in die weitere Betrachtung eingehen, spricht man bei solchen Tests von Rangtests. Dem Schüler muß aber klar sein, daß wegen dieser Transformation von Meßwerten in Rangzahlen auf einen Teil der Informationen verzichtet wird; insbesondere können keine Ausreißer mehr erkannt werden. Wie sieht es daher mit der Wirksamkeit solcher Tests aus? Die Rangtests haben eine vergleichsweise hohe Effizienz; denn die Unterschiede in der Wirksamkeit im Vergleich zu den speziell auf eine vorliegende Normalverteilung abgestimmte Tests sind nur gering. Die Wirksamkeit ist sogar größer, falls keine Normalverteilung vorliegt. 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 in diesem Zusammenhang besprochen.

Die Schüler fanden heraus, daß dann, wenn die beiden Verfahren gleiche Lebensdauer messen, man erwartet, daß die x - und die y - Werte gut gemischt sind, d. h. man wird die Nullhypothese nur dann verwerfen, wenn die x - Werte zum "größten" Teil größer als die y - Werte sind oder umgekehrt. Daß sich die Summe der Platzzahlen, also die Rangsummen X_S bzw. Y_S als Testgröße eignen, kam als Schülervorschlag. Auch eine elementare Herleitung für die Entscheidungsregel mit Hilfe der Kombinatorik wurde von den Schülern selbständig erarbeitet.

Die x - Werte treten n_1-mal und die y - Werte n_2-mal auf. Jede Permutation der $n = n_1 + n_2$ Zahlen hat dann die gleiche Wahrscheinlichkeit, d. h. es gibt

$$\frac{(n_1+n_2)}{n_1!\ n_2!} = \binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2}$$

Möglichkeiten, sodaß jede Möglichkeit die Wahrscheinlichkeit

$$p = \frac{1}{\binom{n_1+n_2}{n_1}}$$

besitzt. Im Ablehnungsbereich der Nullhypothese zum Signifikanzniveau α liegen k Permutationen, wobei k die nächst

kleinere Zahlen von $\alpha \cdot \binom{n_1+n_2}{n_1}$ ist. Es gilt dann

$$\frac{k}{\binom{n_1+n_2}{n_1}} \leq \alpha \quad .$$

Im Beispiel gibt es $\binom{11}{5} = 462$ verschiedene Anordnungen, d. h. auf dem 5 % - Signifikanzniveau sind es bei zweiseitigem Test mit symmetrischer Anordnung höchstens 22 Möglichkeiten, die zur Ablehnung der Nullhypothese führen und zwar diejenigen, für die X_S kleinste bzw. größte Werte der Rangsumme besitzt. Zwei Dinge wurden an dieser Stelle in der Diskussion mit den Schülern geklärt; erstens, daß der auf Signifikanz zu prüfende Unterschied in der Lebensdauer keine bestimmte Richtung besitzt, d. h. ein zweiseitiger Test zu verwenden ist, und zweitens, daß man in der Praxis bei Beurteilung von Produkten im allgemeinen von einer statistischen Aussagesicherheit von 95 % ausgeht, es sei denn, die Entscheidung ist äußerst kapitalintensiv etwa in Form hoher Investitionen, was hier nicht der Fall ist.

Die Anordnungen mit größter bzw. kleinster Rangsumme lassen sich bei geringen Stichprobenumfängen leicht angeben, was Schüler besonders gerne ausführen; ansonsten sei auf das Computerprogramm in [1] verwiesen.

Für die Rangsummen X_S und Y_S gilt nach der Formel für die Summe der ersten n natürlichen Zahlen die einfache Beziehung

$$X_S + Y_S = \sum_{i=1}^{n_1+n_2} i = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} \quad .$$

Es genügt also eine der beiden Summen X_S oder Y_S zu berechnen. In der Praxis verwendet man die kleinere der beiden als Prüfgröße.

In der folgenden Tabelle 3 sind nur die x - Werte geschrieben, die y - Werte sind für die Leerstellen geschrieben zu denken.

1	2	3	4	5	6	7	8	9	10	11	X_S	U_x	X
x	x	x	x	x							15	0	-3.67
x	x	x	x		x						16	1	-3.46
x	x	x	x			x					17	2	-3.25
x	x	x		x	x						17	2	-3.24
x	x	x	x				x				18	3	-3.03
x	x	x		x		x					18	3	-3.03
x	x		x	x	x						18	3	-3.00
x	x	x			x	x					19	4	-2.82
x	x	x		x			x				19	4	-2.81
x	x	x	x					x			19	4	-2.79
x	x		x	x		x					19	4	-2.79
x		x	x	x	x						19	4	-2.70
					x	x	x	x		x	41	26	2.70
				x		x	x		x	x	41	26	2.79
		x					x	x	x	x	41	26	2.79
			x			x		x	x	x	41	26	2.81
				x	x			x	x	x	41	26	2.81
					x	x	x		x	x	42	27	3.00
				x		x		x	x	x	42	27	3.03
			x				x	x	x	x	42	27	3.03
					x	x		x	x	x	43	28	3.24
				x			x	x	x	x	43	28	3.25
					x		x	x	x	x	44	29	3.46
						x	x	x	x	x	45	30	3.67

Tabelle 3

Aus dem Stochastikunterricht weiß man, daß bei vorgegebenem Signifikanzniveau α der wirkliche Fehler α' kleiner oder höchstens gleich α sein soll. Im Beispiel werden wir die Nullhypothese auf dem 5 % - Signifikanzniveau ablehnen, wenn die Rangsumme X_S einen der Werte 15, 16, 17, 18 bzw. 42, 43, 44, 45 annimmt (siehe Tabelle 3). Der wirkliche Fehler α' beträgt dann aber nur $\alpha' = \frac{14}{462} = 3.03\%$. Würde man die Werte 19 bzw. 41 noch zum Ablehnungsbereich hinzunehmen, würde $\alpha' = \frac{24}{462} = 5.19\%$ die 5 % - Marke überschreiten. Die Folge ist, daß der Rangsummentest im Beispiel auf dem 5 % - Niveau keine größere Wirksamkeit besitzt als auf dem 3.03 % - Niveau.

Im Beispiel gilt X_S=31 (Y_S=35), d.h. H_o wird nicht verworfen, das Stichprobenergebnis zeigt keine "Überzufälligkeit". Die Problematik, wie vorsichtig man entscheiden will und soll und die Konsequenzen daraus sind beim Hypothesentest bei Binomial- und Normalverteilung ausführlich erörtert worden. Die Schüler

Zu den Größen $x_1, x_2, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ gehören die Rangzahlen 1, 2, ..., n in einer bestimmten Reihenfolge. Da es bei dieser Testart nur auf die Rangreihenfolge ankommt, kann eine Transformation mit Hilfe einer streng monotonen Funktion durchgeführt werden, die die Rangreihenfolge bekanntlich nicht ändert. Wegen der Verwendung der Standardnormalverteilungsfunktion Φ als Transformationsfunktion teilen wir deren Wertemenge $W =]0,1[$ mit n äquidistanten Punkten

$$p_i = \frac{i}{n+1}, \quad i = 1, 2, \ldots, n$$

in n + 1 gleich große Stücke, d. h. wir ordnen jeder Rangzahl i den Punkt p_i (als Ordinate) zu. Dann bestimmen wir zu jedem Punkt den zugehörigen u - Wert (als Abszisse) so, daß $\Phi(u_i) = p_i$ bzw. $u_i = \Phi^{-1}(p_i)$. Durch diese Abbildungen ist jedem Rang i ein Wert u_i zugeordnet, der je nach Rangzahl ein x - oder y - Wert sein kann. Ist X die Summe der den Rangzahlen der x - Werte zugeordneten u - Werte und Y entsprechend für die y - Werte, so gilt wegen der Symmetrie der Dichtefunktion φ der Standardnormalverteilungsfunktion Φ die Beziehung X + Y = 0, d.h. es genügt, z. B. die Summe X zu bestimmen. Als Entscheidungsregel bietet sich an, H_o zu verwerfen, falls X oder Y (d. h. $|X|$) einen kritischen Wert X_α überschreitet.

Wegen der Transformation der Rangzahlen mit Hilfe der Standardnormalverteilungsfunktion bzw. deren Umkehrung bezeichnet man den X - Test auch als Normalrangtest.

Am Beispiel sollen die einzelnen Schritte dieses Testverfahrens nochmals erläutert werden.

1. Schritt: Transformation der Ränge 1 bis 11 auf das Intervall $]0,1[$ gemäß $p_i = \frac{i}{n+1}$ (siehe Ordinatenwerte in Bild 1).

2. Schritt: Bestimmung der u - Werte aus $u_i = \Phi^{-1}(p_i)$ (siehe Tabelle 4 und Abszissenwerte in Bild 1).

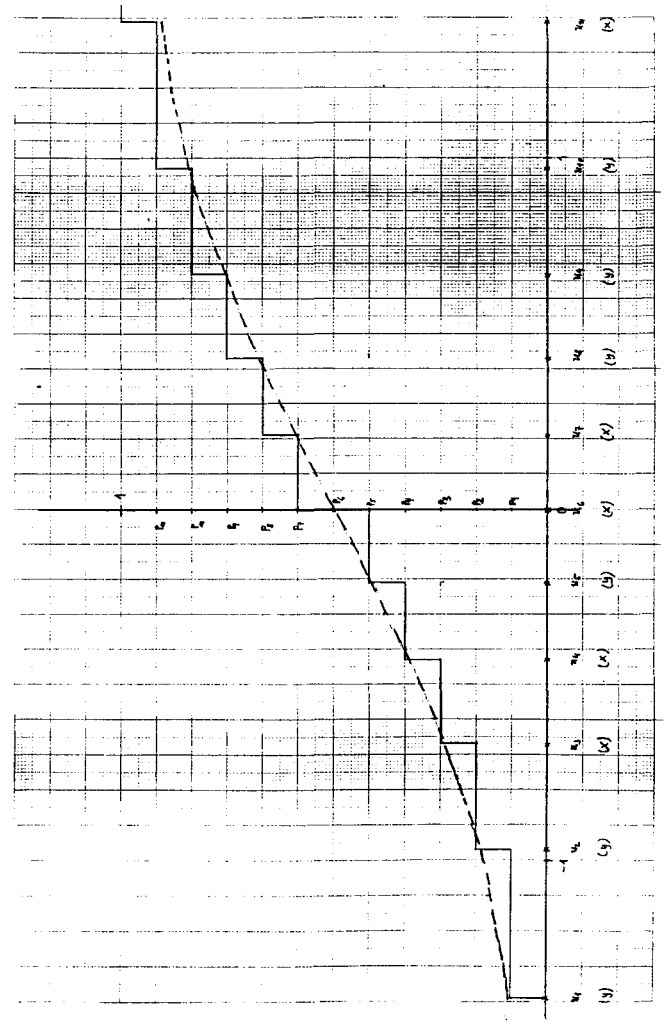

Bild 1

p_i	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{4}{12}$	$\frac{5}{12}$	$\frac{6}{12}$	$\frac{7}{12}$	$\frac{8}{12}$	$\frac{9}{12}$	$\frac{10}{12}$	$\frac{11}{12}$
$\Phi^{-1}(p_i)$	-1.39	-0.97	-0.67	-0.43	-0.21	0.00	0.21	0.43	0.67	0.97	1.39

Tabelle 4

3. Schritt: Berechnung der Summe X:

$$X = -0.67 - 0.43 + 0.00 + 0.21 + 1.39 = 0.5 \ .$$

4. Schritt: Festlegung der Schranke X_α und Entscheidung über Ablehnung oder Nichtablehnung der Nullhypothese H_o. Die Werte für X sind in der letzten Spalte der Tabelle 3 berechnet. Nach den Überlegungen in 4. liest man als Schranke $X_\alpha = 2.70$ ab. Wegen $X < X_\alpha$ kann H_o nicht verworfen werden. Für größere n_1, n_2 sind die Schranken für bestimmte Signifikanzniveaus tabelliert.

7. Anmerkungen

(1) Rangsummentest, U - Test und X Test basieren alle drei auf Rangtupel. Während der Rangsummentest und der U - Test (siehe 5.) äquivalent sind, unterscheidet sich der X - Test von beiden dadurch, daß durch die Transformation eine Gewichtung der Ränge vorgenommen wird, d. h. der X - Test ist wirksamer als die beiden anderen Testarten, da die Information, die man vom betrachteten System hat, besser ausgenutzt wird. Für große n_1, n_2 besitzt der X - Test eine Wirksamkeit wie Tests auf eine vorliegende Normalverteilung (vergleiche dazu auch die in Bild 1 gestrichelt eingezeichnete Standardnormalverteilungskurve und beachte, daß für $n = n_1 + n_2 = 11$ gilt, d. h. ein kleiner Wert für n vorliegt). In [4] wird für große n für den Rangsummentest und U - Test der Wert $\frac{3}{\pi} = 95.5\%$ und für den X - Test der Wert 1 als Wirksamkeit hergeleitet.

(2) Im Beispiel sieht man im Vergleich zu 4. und 5. ($\alpha' = 3.03\%$) die deutlich bessere Ausnutzung des Signifikanzniveaus mit $\alpha' = \frac{22}{462} = 4.76\%$.

(3) Die Transformation in Erwartungsnormalrangwerte mit der Funktion Φ^{-1}, d. h. in u - Werte der Standardnormalverteilung ist tabelliert und deshalb leicht vorzunehmen.

(4) Wenn man vermutet, daß wenigstens näherungsweise eine Normalverteilung vorliegt, sollte man wegen seiner hohen Effizienz den t - Test anwenden. Einen parameterfreien Test muß man benutzen, wenn eine Entscheidung oder Vermutung über eine vorliegende Verteilung nicht möglich ist, insbesondere dort, wo quantitative Messungen der Merkmale nicht möglich sind, d. h. nur Rangordnungen der Werte angegeben werden können, z. B. bei vielen Untersuchungen in der Medizin, Psychologie, Biologie etc.

(5) Aufgabe: Zeige mit den Überlegungen von 4. bis 6., daß die Nullhypothese H_o, daß die von zwei verschiedenen Firmen entwickelten Meßgeräte für elektrische Widerstände (unter sonst gleichen Voraussetzungen wie im Ausgangsbeispiel; siehe Tabelle 5) gleiche Widerstandswerte messen, nach dem Rangsummentest nicht, nach dem X - Test aber auf dem 5 % - Signifikanzniveau verworfen werden muß.

Rangfolge der Messung	Meßgerät 1 R in	Meßgerät 2 R in
1	7.999	7.993
2	8.003	7.995
3	8.007	7.998
4	8.008	8.000
5	8.010	8.002
6	-	8.006

Tabelle 5

8. Alle Schüler des Kurses arbeiteten in diesen sechs Unterrichtsstunden bereitwillig mit, weil sie wußten, daß es sich um reale Meßwerte handelte, und obwohl sie wußten, daß die besprochenen Testverfahren außerhalb des vorgeschriebenen Stoffes lagen. Es ist ein Dilemma des Stochastikunterrichts in der Schule, daß die Unterrichtsgegenstände in der Praxis nur selten benötigt werden, dagegen Methoden der Praxis umgekehrt nicht oder kaum (auch wegen des bevorstehenden Zentralabiturs in Bayern) unterrichtet werden. Es kam auch den

Schülern entgegen, daß sehr viel Wert auf eine ausführliche Diskussion gelegt wurde. Auch diejenigen, die nicht auf Lösungsmöglichkeiten kamen, gewannen in der Diskussion die notwendigen Einblicke, wie ein spezielles statistisches Problem angepackt werden kann. Eine solche Art der Anleitung ziehen die Schüler einer "Rezeptbuchstatistik" vor. Die Schüler waren so "begierig" auf weitere Verfahren, die etwas außerhalb des Stoffes liegen, daß sie bereitwillig auch Zeit außerhalb des Unterrichts opferten. Da sich erfahrungsgemäß bei Schülern immer Schwierigkeiten beim Begriff der statistischen Unabhängigkeit zeigen, wurden in dieser freiwilligen Arbeitsgruppe im Gegensatz zu den oben besprochenen Tests zwei Testverfahren (Vorzeichen - Test und Vorzeichen - Rang - Test von WILCOXON) für verbundene, d. h. abhängige Stichproben, die ebenfalls Meßwerte aus der Praxis enthielten, erarbeitet.

Literatur

1 DIFF: Schätzen und Testen SR 4. Tübingen, 1983.
2 DIFF: Stochastik MS 4. Tübingen, 1981.
3 STORM, R.: Wahrscheinlichkeitsrechnung/ Mathematische Statistik/ Qualitätskontrolle. Leipzig: Fachbuchverlag, 1974.
4 v. d. WAERDEN, B. L.: Mathematische Statistik. Berlin: Springer, 1971.
5 WEBER, H.: Einführung in die Wahrscheinlichkeitsrechnung und Statistik. Stuttgart: Teubner, 1983.