Zeitschrift für experimentelle und angewandte Psychologie
1983, Band XXX, Heft 3, S. 347—386

Verletzung der Annahmen bei Zwei-Stichproben-Lokationstests: Eine Übersicht über empirische Resultate*)

Willi Hager, Bernhard Lübbeke und Ronald Hübner

In der psychologischen Forschungspraxis werden sehr häufig Zwei-Stichproben-Lokationstests angewendet. Sowohl die parametrischen wie die nonparametrischen Varianten dieser Tests sind an bestimmte Voraussetzungen gebunden, die gegenübergestellt und in ihrer Bedeutung diskutiert werden. Die Frage nach einem Testverfahren, das selbst dann noch hinreichend valide Resultate erbringt, also „robust" ist, wenn eine oder mehrere der Voraussetzungen nicht erfüllt sind, wird zu beantworten versucht, indem theoretische und empirische Studien zum t-, U-, Welch- und zu den Normal-Scores-Tests nach Terry-Hoeffding und van der Waerden zusammengefaßt, verglichen und bewertet werden. Die Ergebnisse favorisieren den stärkeren Gebrauch des Welch-Tests zur Prüfung von parametrischen Mittelwertshypothesen anstelle des t-Tests. Ein theoretischer Vergleich der Teststärken beider Tests zeigt, daß der Welch-Test selbst unter homogen varianten Normalverteilungen kaum testschwächer als der t-Test ist, während er bei heterogenen Varianzen dem t-Test an Robustheit überlegen ist. Die Unterschiede zwischen den nonparametrischen Tests zur Prüfung von Mittelwertsunterschieden bzw. im allgemeineren Fall von Lokationsunterschieden sind vglw. gering und nicht einheitlich.

Der empirisch arbeitende Psychologe steht nach der Datenerhebung häufig vor der Frage, ob er seine Daten mittels eines parametrischen oder eines nonparametrischen oder verteilungsfreien[1]) Tests auswerten soll. In der Regel wird dabei zuerst ein parametrisches Verfahren routinemäßig in Er-

*) Wir danken Frau Viola Gremmel und Frau Gabriele Reimann für ihre Unterstützung beim Anfertigen des Manuskripts, und Frau Elke Heise gilt unser Dank für ihr sorgfältiges Korrekturlesen des Manuskripts.

1) Die Begriffe „nonparametrisch" und „verteilungsfrei" werden hier austauschbar benutzt; vgl. zur Unterscheidung Marascuilo & McSweeney (1977) und Büning & Trenkler (1978).

wägung gezogen, und nonparametrische Tests werden in den meisten Statistik-Lehrbüchern lediglich als Ersatz für erstere in den Fällen empfohlen, in denen es zumindest fraglich ist, ob die parametrischen Annahmen erfüllt sind (etwa Siegel, 1976; Bortz, 1979; Vorberg, 1981). Bradley (1968) dagegen spricht sich für die grundsätzliche Bevorzugung nonparametrischer Techniken aus, weil die parametrischen Tests nicht „robust" gegenüber Annahmenverletzungen sind. Autoren wie Glass, Peckham & Sanders (1972) und Havlicek & Peterson (1974) berufen sich auf der anderen Seite auf die „bemerkenswerte Robustheit" der parametrischen Tests und plädieren für deren durchgängige Anwendung.

Beide Empfehlungen erfolgen unter Berufung auf quasi-empirische, durch Computersimulationen erhobene Daten, die offenbar nicht einheitlich sind. In dieser Arbeit soll daher versucht werden, die Simulationsstudien zu den meistempfohlenen Zwei-Stichproben-Lokationstests zusammenzustellen und die Ergebnisse miteinander zu vergleichen, um ggf. Empfehlungen für den Praktiker ableiten zu können.

1. Robustheit und Voraussetzungen von Tests

Im Anschluß an Box (1953; vgl. auch Box & Andersen, 1955) heißt ein Signifikanztest zur Prüfung einer bestimmten Hypothese „robust", wenn die Verletzung einer bestimmten Annahme (oder einer bestimmten Kombination von Annahmen) zu keinen bzw. zu keinen nennenswerten Abweichungen der faktischen von den nominellen Fehlerwahrscheinlichkeiten führt (vgl. auch Bradley, 1968; Marascuilo & McSweeney, 1977; Büning & Trenkler, 1978). Der offenkundige Mangel dieser nur qualitativen Definition hat bislang jedoch offenbar nur zu zwei Versuchen einer Präzisierung geführt. Ramsey (1980, 338) schlägt vor, einen Test dann noch als robust anzusehen, wenn bei vorgegebenem Signifikanzniveau α die faktische Wahrscheinlichkeit für einen Fehler 1. Art, p_α, der Ungleichung $0{,}04 \leq p_\alpha \leq 0{,}06$ bei $\alpha = 0{,}05$ und $0{,}007 \leq p_\alpha \leq 0{,}015$ bei $\alpha = 0{,}01$ genügt. — Bradley (1978, 146) schlägt wahlweise ein strenges und ein liberales quantitatives Robustheitskriterium vor. Er hält Differenzen zwischen p_α und α dann noch für „vernachlässigbar" (a.a.O.), wenn die Ungleichung $0{,}9\alpha \leq p_\alpha \leq 1{,}1\alpha$ erfüllt ist, und sein liberales Kriterium sieht vor, daß $0{,}5\alpha \leq p_\alpha \leq 1{,}5\alpha$ ist. — Wir legen bei unserer Zusammenfassung das liberale Kriterium von Bradley (1978) zugrunde[2]).

2) In einigen Studien (etwa Pearson & Please, 1975) wird die Beurteilung der Resultate unter Einbezug des Standardfehlers $\sigma_{\hat{p}}$ vorgenommen, der von der Anzahl r der gezogenen Stichproben abhängig ist: $\sigma_{\hat{p}} = [p(1 - p)/r]^{1/2}$.

In Ergänzung der vorstehenden Autoren sei betont, daß auch unter Gültigkeit von Alternativhypothesen Fehlentscheidungen getroffen werden können (Fehler 2. Art), weswegen wir es für wichtig ansehen, bei Robustheitsuntersuchungen auch die Teststärkefunktionen der betreffenden Tests zu berücksichtigen.

Darüber hinaus sollte unseres Erachtens stets von der „Robustheit eines Tests bei der Prüfung einer bestimmten Hypothese" gesprochen werden. Beispielsweise prüft der U-Test die recht allgemeine Nullhypothese identischer Verteilungen unter der Annahme der Unabhängigkeit der Daten und der Kontinuität der Verteilungsfunktionen. Will man mit dem gleichen Test dagegen die wesentlich spezifischere Hypothese prüfen, daß die beiden Verteilungsfunktionen $F_1(y)$ und $F_2(y)$ der betrachteten Zufallsvariablen Y um einen bestimmten Betrag δ gegeneinander verschoben sind, muß vorausgesetzt werden, daß die Verteilungsfunktionen von Y die gleiche Form aufweisen (Annahme der Homomerität; vgl. Sawrey, 1958; Wetherill, 1960; Lubin, 1962; Lienert, 1973). Der U-Test ist bei der Prüfung *dieser* Hypothese nicht robust gegenüber Varianz- und allgemeinen Verteilungsunterschieden (vgl. Birnbaum & Klose, 1957; Wetherill, 1960; van der Vaart, 1961; Edgington, 1964; Pratt, 1964; Trommer, 1967). Nehmen wir zur weiteren Verdeutlichung an, zwei Verteilungen seien heteromer, d.h. nicht von gleicher Form, hätten jedoch den gleichen Erwartungswert E(Y), und die Wahrscheinlichkeit signifikanter Testergebnisse ist bei Verwendung des U-Tests größer als das nominelle Signifikanzniveau α. Ein signifikantes Resultat läßt in dieser Situation zwei Interpretationen zu: (1) Es gilt die Nullhypothese gleicher Erwartungswerte, und der U-Test ist bei der Prüfung *dieser* Hypothese nicht robust gegenüber allgemeinen Verteilungsunterschieden. (2) Es gilt die Alternativhypothese unterschiedlicher Verteilungsfunktionen F_1 und F_2, und das Ergebnis ist auf die Teststärke des U-Tests gegenüber derartigen Alternativen zurückzuführen. Die Entscheidungen zwischen den beiden möglichen Interpretationen kann nur getroffen werden, wenn man spezifiziert, an welcher der prüfbaren statistischen Hypothesen man interessiert ist. In der folgenden Übersicht gehen wir davon aus, daß *Lokationshypothesen* (Hypothesen über Erwartungswerte und Mediane) getestet werden sollen, daß also derjenige Hypothesentyp betrachtet wird, an dem Psychologen aufgrund der Häufigkeit der Anwendung der betreffenden Tests offenbar in erster Linie interessiert sind.

Mit dem t-Test nach W. S. Gosset („Student", 1908; vgl. Hays, 1977; Bortz, 1979) wird die statistische Nullhypothese der Gleichheit zweier Mittelwerte aus normalverteilten Populationen getestet. Für seine valide Anwendung sind die folgenden Annahmen oder „statistischen Oberhypothesen" (Stegmüller, 1973) als gültig vorauszusetzen (vgl. Hays, 1977; Marascuilo & McSweeney, 1977):

(1) Unabhängigkeit der Beobachtungen innerhalb der und zwischen den beiden Stichproben;
(2) Gleichheit der beiden Populationsvarianzen σ_1^2 und σ_2^2;
(3) Normalverteilung der Rohwerte in beiden Populationen;
(4) Aus Voraussetzung (3) folgt die Kontinuität der Populationsverteilungen.

Die Verallgemeinerung dieser Tests für den Fall ungleicher Populationsvarianzen („Behrens-Fisher-Problem") stellt der W-Test nach Welch (1937, 1947, 1949) und Aspin (1948) dar (vgl. Pfanzagl, 1978). Zwar dürfte die Anwendung der exakten Prüfverteilung der Teststatistik für viele praktische Belange zu kompliziert sein, jedoch kann eine akzeptable Annäherung an die exakte Lösung über die t-Verteilung mit modifizierten Freiheitsgraden erfolgen (Pfanzagl, 1978). Über andere Lösungen zum Behrens-Fisher-Problem berichten zusammenfassend Scheffé (1970) und Mehta & Srinivasan (1970).

Das bekannteste nonparametrische Homologon des t-Tests ist der U-Test nach Wilcoxon (1945) und Mann & Whitney (1947) (vgl. Lienert, 1973; Marascuilo & McSweeney, 1977). Weniger bekannt sind die Normal-Scores- oder Normal-Rang-Tests von Terry (1952) und Hoeffding (1951) (vgl. jedoch bereits Fisher & Yates, 1938) und von van der Waerden (1953) (zur Darstellung siehe Lehmann, 1975; Marascuilo & McSweeney, 1977)[3]). Der anstelle des t-Tests seit einiger Zeit gehäuft empfohlene Randomisierungs-t-Test nach Fisher (vgl. Bradley, 1968; Lehmann, 1975) benutzt zwar die gleiche Teststatistik wie der parametrische t-Test, aber die Ableitung der Prüfverteilung erfolgt bei ihm — ebenso wie beim U- und den Normal-Scores-Tests — ausschließlich unter Zugrundelegung des Randomisierungsprinzips (Bradley, 1968; Lehmann, 1975; Marascuilo & McSweeney, 1977). Um mit den Randomisierungstests Lokationshypothesen testen zu können, müssen unter Zugrundelegung des Populationsmodells die folgenden Oberhypothesen als gültig angenommen werden (vgl. Marascuilo & McSweeney, 1977):

(1) Unabhängigkeit der Beobachtungen innerhalb der und zwischen den Populationen;
(2) Homomerität der Verteilungen[4]);
(3) Kontinuität der Verteilungen.

3) Auf den Normalrangtest nach Bell & Doksum (1965), dessen Ergebisse über verschiedene Auswerter nicht übereinstimmen müssen, gehen wir nicht ein. — Die beiden anderen Normal-Scores-Tests werden im folgenden stets zusammen betrachtet und als „NS-Tests" bezeichnet.

4) Die Asymmetrie einer Verteilung kann über die Schiefe (engl. „skewness") erfaßt werden, für die gilt: $\gamma_1 = E(Y - E(Y))^3/\sigma^3$. Bei positiver Schiefe („Linkssteilheit"; $\gamma_1 > 0$) ist der

Die Randomisierungstests prüfen die exakt gleiche Hypothese wie der t-Test nur dann, wenn die Voraussetzungen des t-Tests erfüllt sind; allerdings ist die relative Effizienz (A.R.E.; vgl. Büning & Trenkler, 1978) des U-Tests unter diesen Bedingungen geringer als die des t-Tests ($A.R.E._{U,t} = 0{,}955$), während die übrigen Tests die gleiche Effizienz wie der t-Test aufweisen (Hodges & Lehmann, 1956, 1961; Bradley, 1968; Büning & Trenkler, 1978). Sind die Voraussetzungen für den t-Test nicht erfüllt, prüft er nicht die Hypothese gleicher Mittelwerte aus zwei homogen varianten Normalverteilungen, sondern die allgemeinere Hypothese identischer Normalverteilungen (Bradley, 1968, 41), wobei er entsprechend seinen nonparametrischen Homologen am teststärksten auf Mittelwerts- und damit auch Medianunterschiede reagiert (a.a.O.). Entsprechendes gilt für den W-Test.

In dem Ausmaß, in dem die zur Prüfung bestimmter spezieller Hypothesen notwendigen Voraussetzungen im konkreten Anwendungsfall nicht erfüllt sind, weichen die faktischen Fehlerwahrscheinlichkeiten p_α für Fehler 1. Art und p_β für Fehler 2. Art von den nominellen oder theoretischen Werten α und β ab. Mehr oder minder starke Abweichungen von den Idealbedingungen dürften dabei in der Empirie die Regel sein (Hey, 1938; Geary, 1947; Bradley, 1968; Hays, 1977; Kraemer, 1981).

Mittels Computersimulationen (sog. Monte-Carlo-Studien; vgl. dazu Kleijnen, 1974, 1975; Bauknecht, Kohlas & Zehnder, 1976) können für die finiten Fälle (endliche Stichproben), für die theoretische Analysen i.a. noch nicht gelungen sind, alle möglichen Abweichungen von den notwendigen Voraussetzungen sowie Kombinationen von Abweichungen hergestellt werden. Zieht man, vereinfacht ausgedrückt, aus den bei vorgegebenen Parametern oder Verteilungsformen konstruierten Populationen dann r Stichproben und berechnet die interessierende Teststatistik, können dies[illegible]werte in einer Häufigkeitsverteilung zusammengefaßt werden, die mit den entsprechenden theoretischen Verteilungen verglichen werden kann. Auf diese Weise wird ermittelt, wie groß der Anteil falscher Entscheidungen bezüglich der jeweils interessierenden Hypothese bei Verwendung eines bestimmten Tests unter Annahmeverletzungen ist. Während jedoch die beiden vorliegenden Übersichten (Bradley, 1968; Glass, Peckham & Sanders, 1972) lediglich das Verhalten eines oder einer Gruppe sehr eng verwandter Tests (z-, t- und F-Test bzw. nur F-Test)

Median kleiner als der Erwartungswert, bei negativer Schiefe („Rechtssteilheit"; $\gamma_1 < 0$) dagegen größer; im Falle der Symmetrie ist $\gamma_1 = 0$ und der Median gleich dem Erwartungswert (vgl. im einzelnen Mood, Graybill & Boes, 1974; Bortz, 1979, 41, 49; Sachs, 1968, 98). — Sind zwei Verteilungen symmetrisch, sind *nur* die dritten zentralen Momente gleich (Null); bei Homomerität dagegen sind *alle* zentralen Momente außer möglicherweise dem ersten gleich.

betrachten, soll hier — wie erwähnt — eine vergleichende Betrachtung mehrerer „alternativer" Testverfahren erfolgen, da wir vermuten, daß die gegensätzlichen Empfehlungen bzgl. des Gebrauchs parametrischer und nonparametrischer Verfahren im wesentlichen eine Folge der einseitigen Betrachtungsweisen der genannten Autoren darstellen dürfte. Ein anderer Grund für die uneinheitlichen Empfehlungen liegt u.E. darin begründet, daß bislang offenbar keine Einigkeit darüber erzielt worden ist, wann ein Test als (noch) „robust" (gegenüber der Verletzung einer bestimmten Annahme) angesehen werden sollte.

2. Ergebnisse der empirischen Robustheitsuntersuchungen

Um dem Leser eine Kontrolle unserer Sucharbeit in mathematisch-statistischen, sozialwissenschaftlichen und psychologischen Fachzeitschriften zu erleichtern, haben wir alle von uns ermittelten empirischen Robustheitsuntersuchungen der letzten ca. 20 Jahre in der Tabelle 1 zusammengefaßt. Sie enthält neben der Angabe des Untersuchungszieles, der verwendeten Signifikanzniveaus, Stichprobengrößen und der Anzahl r der Durchläufe auch eine Kurzbeschreibung der simulierten Annahmeverletzungen. In den letzten beiden Spalten finden sich Informationen über die Validierung der Methode (Probelauf unter bestimmten Standardbedingungen) und über die Möglichkeit, die Methode nachzuvollziehen.[5])

Betrachten wir nun die Ergebnisse der Studien getrennt nach den einzelnen möglichen Annahmeverletzungen und deren Kombinationen.

2.1 Skalenniveau

Eine Reihe von Autoren (etwa Siegel, 1976; Schüle, 1976) zählt ein bestimmtes Skalenniveau zu den *mathematischen* Voraussetzungen statistischer Testverfahren. Das Skalenniveau besagt etwas über die Beziehungen zwischen Meßwerten und den gemessenen empirischen Größen, jedoch nichts über die Meßwerte selbst. Ein statistischer Test dagegen benutzt nur diese Meßwerte, um Hypothesen über Erwartungswerte mittlerer Ränge, über Mittelwerte oder über Verteilungsfunktionen zu prüfen, mithin ist die Verteilung einer Teststatistik unabhängig vom Skalenniveau der Daten

5) Nicht aufgenommen wurden die Arbeiten von Hemelrijk (1961) (Teststärkevergleich zwischen t- und U-Test bei $n_1 = n_2 = 10$ und $\alpha = 0{,}025$ sowie Anzahl der Durchläufe $r = 50$) sowie von Kempthorne & Doerfler (1969) (Teststärkevergleich für einige Randomisierungstests für *abhängige* Stichproben). Ebenfalls nicht aufgenommen wurden die Arbeiten von Baker, Hardyck & Petrinovich (1966) und von Trachtman, Giambalvo & Dippner (1978) zur Bedeutung des Skalenniveaus für die betrachteten Tests.

(Hager & Westermann, 1983b). Letzteres ist dann von zentraler Bedeutung, wenn von den Eigenschaften der Meßwerte auf empirische Sachverhalte zurückgeschlossen werden soll, und dieser Rückschluß läßt sich nicht mit Mitteln der Inferenzstatistik bewerkstelligen. Nach der hier vertretenen Auffassung legt das Skalenniveau fest, welche statistischen Kennwerte und welche Aussagen über statistische Kennwerte und Verteilungen als „empirisch sinnvoll" („meaningful") im Sinne von Suppes & Zinnes (1963) angesehen werden können (vgl. hierzu Hager & Westermann, 1983a,b; Westermann, 1980, 1982).

Baker, Hardyck & Petrinovich (1966), Havlicek & Peterson (1978) sowie Trachtman, Giambalvo & Dippner (1978) haben versucht, aus per definitionem intervallskalierten Daten durch nicht-lineare Transformationen ordinalskalierte Daten zu erzeugen. Durch die Transformationen wurden jedoch lediglich die Formen der Populationsverteilungen geändert, so daß die angeblichen Verletzungen einer das Skalenniveau betreffenden Voraussetzung tatsächlich lediglich die Auswirkungen der Verletzung einer mathematischen Voraussetzung darstellen, nämlich einer bestimmten Verteilungsannahme (vgl. zur Kritik der Untersuchungen auch Krauth, 1980). Zwar kann man das Skalenniveau von Meßdaten empirisch bestimmen (Westermann, 1980, 1982), aber man kann u.E. nicht mit Hilfe von Computersimulationen die Bedeutung des Skalenniveaus für die valide Anwendung von statistischen Tests untersuchen.

Sollen mit den hier betrachteten t-, W-, U- und NS-Tests Mittelwertshypothesen getestet werden, so ist dabei Intervallniveau der Daten vorauszusetzen, und zwar aus mindestens zwei Gründen: Erstens sind Aussagen über Mittelwerte i.a. nur ab Intervallniveau empirisch sinnvoll. Zweitens sind Aussagen über die „Symmetrie" oder die „Homomerität" von Verteilungen ebenfalls nur dann sinnvoll, wenn die betrachtete Variable mindestens eindeutig bis auf positiv-lineare Transformationen ist. Auch Aussagen über Varianzen sind i.a. nur für intervallskalierte Variablen sinnvoll (a.a.O.).

2.2 Stetigkeit der Populationsverteilungen

Die Forderung nach stetigen Verteilungen der Populationen kann der Mathematiker leicht durch Setzung erfüllen, wodurch sich seine Ableitun-

---

6) Der Anstieg einer Dichtekurve kann über den Exzeß γ_2 (engl. „kurtosis") erfaßt werden, der oft wie folgt definiert wird: $\gamma_2 = E(Y - E(Y))^4/\sigma^4 - 3$. Bei positivem Exzeß ($\gamma_2 > 0$; „Leptokurtosis") steigt die Dichtekurve der betr. Zufallsvariablen steiler an als die einer normal verteilten Variablen („Mesokurtosis"), bei negativem Exzeß ($\gamma_2 < 0$; „Platykurtosis") dagegen flacher (siehe zu Einzelheiten Mood, Graybill & Boes, 1974; ferner Sachs, 1968; Bortz, 1979).

Tabelle 1

Übersicht über die verwendeten Robustheitsstudien

	Art der Tests	Robustheit	Teststärke	Art d. Verletzungen	Stichprobengröße	Signifikanzniveau (in Prozent)	Anzahl der Durchläufe	Methode Angabe z. Probelauf	Methode replizierbar
Bennett und Hsu (1961)	W, Behrens-Fisher-Test	+	+	N×V	2 bis 13	1/2,5/5	1000	–	+
Bevan, Denton und Myers (1974)	F (K≥2)	+	–	kategoriale Daten	4 / 8 (g)	1/5/10	1000	+	+
Blair und Higgins (1980a,b)	t, U	–	+	NN	3 bis 81	0,5/1/2,5/5	5000	–	+
Blair, Higgins und Smitley (1980)	t, U	–	+	NN	3 bis 81	1/2,5/5	5000	–	+
Boneau (1960)	t	+	–	N×V, NN, NN×V, U	5/15	1/5	1000	+	+
Boneau (1962)	t, U	+	+	N×V, NN, NN×V, U	5/15	1/5	1000	+	+
Bowman, Beauchamp und Shenton (1977)	t	+	–	NN	5/95	>25	25000 100000	–	+
Bradley (1980a,b)	t	+	–	N×V, NN, NN×V, U, U×V	8 bis 1024	0,1/1/5	30000	+	+
Conover, Wehmanen und Ramsey (1978)	U, NS u.a.	–	+	NN	4/5	1/2,5/10	–	–	–
Donaldson (1968)	F (K≥2)	+	+	N×V, NN, NN×V	4/8/16/32/(g)	1/5/10	10000	+	+
Fligner und Pollicello (1981)	t, W, U u.a.	+	+	N×V, NN×V, U	11/10; 25/20	5	10000	+	+
Havlicek und Peterson (1974)	t	+	–	N×V, NN, NN×V, U×V	5/15/30	1/5	5000	+	+
Hsu und Feldt (1969)	F (K≥2)	+	–	kategoriale Daten	11/51 (g)	1/5/10	5000 10000	+	+
Hübner, Lübbeke und Hager (1982)	t, W, NS	+	+	N×V, NN×V, NN, U, U×V	8/12/10/15/20/25	0,1/1/5	20000	+	+
de Jonge (1961)	t	+	–	NN, U	5 (g)	0,5/1/2/2,5/5/10	500 1000	–	–

				Daten			10000		
Hübner, Lübbeke und Hager (1982)	t, W, NS	+	+	N×V, NN×V, NN, U, U×V	8/12/10/15/20/25	0,1/1/5	20000	+	+
de Jonge (1961)	t	+	–	NN, U	5 (g)	0,5/1/2/2,5/5/10	500 1000	–	–
Kemp und Conover (1973)	U, NS u.a.	+	+	NN	5/6/10/19/20	5/10	200	–	+
van der Laan und Oosterhoff (1965, 1967)	t, U, NS	–	+	N	5/6/8/10/15	1/5/10	2000	+	+
van der Laan und Weima (1978)	t, U	+	+	NN	3/6/10/(g)	1/5/10	5000	+	+
van der Laan und Weima (1980)	U	–	+	NN	6/10/15/(g)	1/5	5000	+	+
Lee, Desu und Gehan (1975)	F (K=2), U u.a.	–	+	NN, U	20/50/(g)	5	500	–	+
Lissitz und Chardos (1975)	t	+	–	N, abhängige Daten	31/(g)	1/5/10 u.a.	1000	+	–
Lunney (1970)	F (K≥2)	+	+	dichotome Daten	3 bis 31 (g)	1/2,5/5/10	1000	+	+
Neave und Granger (1968)	t, U, NS u.a.	+	+	N×V, NN, NN×V	20/40	5	500	+	+
Pearson und Please (1975)	t	+	–	NN	10/25/(g)	1/5	2000	–	+
Posten (1978)	t	+	+	NN	5/10/15/20/25/30/(g)	1/5	100000	+	+
Ramsey (1971)	U, NS u.a.	+	+	NN	5 (g)	5 u.a.	–	–	–
Ramsey (1980)	t (U, W)	+	–	N×V	2 bis 40	1/5/ u.a.	–	–	–
Toothaker (1972)	t, U, Random.	+	+	N, NN	2/3/4/5	2,86 bis 10	1000	+	+
Woodward und Overall (1977)	t, U u.a.	+	+	kategoriale Daten	15/21/31 (g)	1/2,5/5/10	10000	+	+
Young und Veldman (1963)	t	+	+	N, NN, N×V, NN×V	10 (g)	5	5000	–	–

Die Erläuterungen zu den einzelnen Angaben befinden sich überwiegend im Text. Die dort nicht erläuterten Abkürzungen bedeuten:
2. Spalte: u.a.: Außer den hier behandelten Tests wurden weitere untersucht; F (K=2): F-Test über 2 Bedingungen, entspricht einem zweiseitigen t-Test.
5. Spalte: N: beide Populationen normal und homogen variant; N×V: beide Populationen normal und heterogen variant; NN: beide Populationen nicht-normal, aber homomer; NN×V: nicht-normale Verteilungen mit heterogenen Varianzen; U: beide Verteilungen unterschiedlich, aber homogen variant; U×V: unterschiedliche Verteilungen mit heterogenen Varianzen.
6. Spalte: (g): nur gleiche Stichprobengrößen untersucht.

gen vereinfachen, etwa weil (Rang-)Bindungen zwischen Werten vermieden werden. Der Psychologe hingegen dürfte es wohl stets mit diskreten (Populations-)Verteilungen zu tun haben, die zudem i.a. endlich sind (Sachs, 1968). Unter diesem Aspekt kann die Stetigkeit der Verteilungen mit Gutjahr (1971, 98) als „ein theoretisches Modell" dieser diskreten Häufigkeitsverteilungen des Psychologen interpretiert werden. Wegen der engen Beziehung der diskreten Bi- und Multinomialverteilungen zu den kontinuierlichen Normal- und χ^2-Verteilungen insbesondere für große Stichproben ($n \to \infty$) ist grundsätzlich zu erwarten, daß die Anwendung parametrischer t- und F-Tests auch bei diskreten Verteilungen gerechtfertigt ist, sofern die Stichproben nicht zu klein und die Abweichungen von der Stetigkeit nicht allzu ausgeprägt sind (vgl. Cochran, 1950; ferner u.a. Seeger & Gebrielsson, 1968; d'Agostino, 1971, 1972; Light & Margolin, 1971). Lediglich unter inhaltlichen Aspekten (Skalenniveau) könnte daher diese Art der Auswertung inadäquat sein. Diesen Gesichtspunkt lassen wir in diesem Abschnitt jedoch ebenso außer acht wie die Tatsache, daß es adäquatere Auswertungstechniken für nominale Daten gibt (Lienert, 1973, 1978; Smith, 1976; Küchler, 1979; Langeheine, 1980).

Gefragt werden kann nun nach dem Ausmaß des Fehlers, der begangen wird, wenn bei relativ ausgeprägten Abweichungen von der Stetigkeit der Populationen mit vglw. kleinen Stichproben, wie sie in der Psychologie üblich sind, gearbeitet wird. Von vglw. ausgeprägten Abweichungen von der Stetigkeit kann man sicher im Fall des Vorliegens von nur zwei, drei oder fünf Antwortkategorien sprechen, wie sie in psychologischen Fragebögen häufig anzutreffen sind (dicho- bzw. polytome Daten).

Hsu & Feldt (1969) haben zur Beantwortung der gestellten Fragen Populationen konstruiert, die nur aus zwei, drei, vier oder aber fünf möglichen Werten bestanden und die sich hinsichtlich der Parameter Erwartungswert μ, Varianz σ^2, Schiefe γ_1 und Exzeß γ_2 unterschieden[6]); vgl. Tabelle 2, zusammengestellt nach den Angaben von Hsu & Feldt (1969, 518—522). Aus den paarweise zusammengestellten Populationen (vgl. letzte Spalte der Tabelle) wurden Stichproben der Größe $n_1 = n_2 = 11; 51$ gezogen und die Werte für t berechnet, wobei sowohl homomere wie heteromere Populationen zugrundelagen (bei der Zwei-Kategorien-Skala nur homomere Populationen).

Die Gesamtheit der Ergebnisse zeigt für den zweiseitigen t-Test für zwei experimentelle Behandlungen (und für den F-Test bei mehr als zwei Behandlungen), daß dieser unter allen untersuchten Bedingungen robust bzgl. des hier zugrundegelegten Kriteriums von Bradley (1978; s.o.) ist. Bei Verwendung der aus dichotomen Werten bestehenden Populationen (Zwei-Punkte-Skala) ergaben sich die relativ größten Abweichungen der tatsächlichen von den nominellen Fehlerwahrscheinlichkeiten, während für die aus

Tabelle 2
Untersuchte Populationen bei Hsu und Feldt (1969)

Anzahl der Kategorien (S) (und der Durchläufe (r))	Nr. der Population	μ	σ^2	γ_1	γ_2	Untersuchte Populationen für K=2
S = 5	Pop 1	2,0	1,04	0,0	−0,49	Pop 2, Pop 2
(r = 5.000)	Pop 2	2,0	0,54	0,0	+0,09	Pop 3, Pop 3
	Pop 3	2,35	1,15	−0,39	−0,27	Pop 1, Pop 2
S = 4	Pop 1	1,5	0,89	0,0	−0,9	Pop 2, Pop 2
(r = 5.000)	Pop 2	1,5	0,45	0,0	−0,22	Pop 3, Pop 3
	Pop 3	1,75	0,66	−0,08	−0,68	Pop 1, Pop 2
S = 3	Pop 1	1,0	0,66	0,0	−1,85	Pop 2, Pop 2
(r = 10.000)	Pop 2	1,0	0,33	0,0	+0,03	Pop 3, Pop 3
	Pop 3	0,85	0,43	0,16	−0,71	Pop 1, Pop 2
S = 2	Pop 1	0,5	0,25	0,0	−2,0	Pop 1, Pop 1
(r = 10.000)	Pop 2	0,4	0,24	0,41	−1,83	Pop 2, Pop 2
	Pop 3	0,25	0,19	0,15	−0,67	Pop 3, Pop 3

Werte zusammengestellt nach Hsu und Feldt (1969, S. 518—522).
Die Abweichungen von der Symmetrie ($\gamma_1 = 0$) und von der Mesokurtosis ($\gamma_2 = 0$) sind als geringfügig anzusehen.

fünf unterschiedlichen Werten bestehenden Populationen (Fünf-Punkte-Skala) die vglw. geringsten Abweichungen resultierten. Ein zusätzlich durchgeführter Vergleich des zweiseitigen t-Tests mit dem üblicherweise bei Kontingenztafeln angewandten χ^2-Test erbrachte erwartungsgemäß eine hohe Übereinstimmung der empirischen Fehlerwahrscheinlichkeiten.

Diese Ergebnisse konnte Lunney (1970) für dichotome Daten, einen weiteren Bereich von gleichen Stichprobengrößen (vgl. Tab. 1) und für verschiedene varianzanalytische Versuchspläne bestätigen. Glass, Peckham & Sanders (1972, 252) weisen ergänzend jedoch darauf hin, daß Lunneys Ergebnisse nicht problemlos auf ungleiche Stichprobengrößen übertragen werden können.

Lunney (1970) fand in seiner Simulationsstudie ferner, daß die empirisch bestimmte Teststärke für kleine Stichproben sehr viel geringer war als

die theoretisch geschätzte, insbesondere bei sehr unterschiedlichen relativen Häufigkeiten der einzelnen Antwortkategorien. In einer Reanalyse dieses Befundes weisen jedoch Glass et al. (1972, 269ff.) einige mathematische Unrichtigkeiten bei Lunney (1970) nach, deren Korrektur einen wesentlich geringeren Unterschied zwischen theoretisch geschätzter und empirisch ermittelter Teststärke ergibt.

Bevan, Denton & Myers (1974) untersuchten die gleiche Fragestellung für drei, fünf und sieben Kategorien unter Zugrundelegung von Normal-, Exponential-, Gamma- und Rechteckverteilungen bei gleichen Stichproben der Größe $n_1 = n_2 = 4; 8$ und moderaten Abweichungen von der „normalen" Schiefe und dem „normalen" Exzeß : γ_1 variiert von 0 bis 1,45 und γ_2 von $-1{,}50$ bis $+0{,}95$. Für die untersuchten Bedingungen erwies sich der zweiseitige t-Test erneut als robust, und die Abweichungen der tatsächlichen von den nominellen Fehlerwahrscheinlichkeiten wurden um so geringer, je mehr Kategorien verwendet wurden und je größer das angesetzte Signifikanzniveau und die Stichproben waren. Bei den 24 extremsten Abweichungen ist das Robustheitskriterium nur ein einziges Mal verletzt — ein Ergebnis, das mit einiger Rechtfertigung als „zufällig" bezeichnet werden kann.

Woodward & Overall (1977) bezogen in ihre Untersuchung zur gleichen Frage außer dem t-Test verschiedene Versionen des χ^2-Tests und den U-Test ein. Sie verwendeten Normal-, Rechteck- und schiefe Verteilungen, aus denen die Daten gezogen und anschließend in vier Kategorien eingeteilt wurden. Für alle untersuchten Tests, Stichprobengrößen ($n_1 = n_2 = 15; 21; 31$) und alle simulierten Verteilungen ergaben sich robuste Resultate, und ein Vergleich der empirisch ermittelten Teststärken der einzelnen Tests ergab keine erwähnenswerten Unterschiede zwischen dem t- und dem U-Test, die sich zudem beide als teststärker erwiesen als die modifizierten χ^2-Tests.

Insgesamt belegen die durchgeführten Untersuchungen in Übereinstimmung mit den wenigen theoretischen Resultaten (s.o.), daß offenbar der Voraussetzung der Stetigkeit der Populationen keine besondere Bedeutung bei der validen Anwendung der parametrischen wie der nonparametrischen Tests beizumessen ist, selbst wenn nur kleine Stichproben verwendet werden.

2.3 Unabhängigkeit der Daten

Die Daten müssen für die valide Anwendung aller Tests innerhalb der und zwischen den Populationen stochastisch unabhängig voneinander sein. Diese Voraussetzung wird üblicherweise dazu vereinfacht, daß Nullkorrelationen gefordert werden (lineare Unabhängigkeit der Daten).

Eine von Null verschiedene Korrelation innerhalb der Behandlungsgruppen kann bspw. entstehen, wenn die ersten Versuchspersonen eines Experiments ihren Nachfolgern über Details der experimentellen Prozedur berichten, woraufhin die Nachfolger mit systematisch anderen Vorkenntnissen und Einstellungen bzgl. des Experiments an diesem teilnehmen, als wenn sie unvoreingenommen wären.

Korrelation der Daten zwischen den Behandlungsbedingungen kann in einer Situation entstehen, in der zunächst die Werte auf der abhängigen Variablen für eine (Kontroll-)Gruppe von Vpn gemessen werden und anschließend auch für die andere, bspw. die Experimentalgruppe, in der sich jedoch etwa aus Mangel an geeigneten Vpn auch *einige* Vpn aus der Kontrollgruppe befinden, die damit zweimal der Messung unterzogen werden (vgl. zu diesem und anderen Beispielen Hollander, Pledger & Lin, 1974). (Die vorstehende Situation darf im übrigen nicht mit dem Fall verwechselt werden, in dem im Rahmen eines Zwei-Gruppen-Experiments *alle* Vpn zweimal gemessen werden („Meßwiederholungen", „abhängige Messungen"). In diesem Fall kann die Auswertung auf Differenzwerten aufbauen, die von der Korrelation bereinigt sind; vgl. den t- und den Wilcoxon-Test für abhängige Stichproben.)

In seiner Übersichtsarbeit hat Cochran (1947) theoretisch ableiten können, daß eine konstante positive Korrelation zwischen den Werten aller Versuchspersonen zu viele ungerechtfertigte Ablehnungen der Nullhypothese nach sich zieht („liberaler Test"), während eine negative Korrelation in einen „konservativen Test" mündet (zu viele ungerechtfertigte Ablehnungen der Alternativhypothese). Diese theoretischen Befunde haben später Box (1954) und Scheffé (1959) bestätigen können, vgl. dazu auch Glass et al. (1972). Mit ähnlichen Auswirkungen von korrelierten Werten ist für den U-Test zu rechnen (vgl. Serfling, 1968; Hollander et al., 1975; Lehmann, 1975) und ebenso für die anderen Tests, für die u.W. jedoch keine entsprechenden Untersuchungen vorliegen.

Üblicherweise wird das Problem korrelierter Werte für den komplexeren Fall der Varianzanalysen behandelt, und daher haben wir nur eine Studie aufgefunden, in der die Auswirkung korrelierter Daten auf die Fehlerwahrscheinlichkeit 1. Art für den t-Test untersucht wird. Lissitz & Chardos (1974) haben sich in ihrer Monte-Carlo-Simulation mit den folgenden vier Arten von Abhängigkeiten innerhalb und zwischen den Gruppen befaßt:

1) Es besteht eine konstante Korrelation zwischen allen Paaren von Versuchspersonen. Die Resultate stehen im Einklang mit den theoretischen Ableitungen, d.h.: „The effect of positive dependence is to increase the size of the tails and the effect of negative dependency is to decrease the size of tails of the empirically derived distribution" (Lissitz & Chardos, 1975, 356).

2) Es besteht eine serielle Korrelation derart zwischen den Versuchspersonen, daß jeweils zwei benachbarte Vpn voneinander abhängig, aber unabhängig von allen anderen sind. Die Auswirkungen dieser seriellen Korrelation auf die Fehlerwahrscheinlichkeit 1. Art sind denen bei der konstanten Korrelation unter 1) vergleichbar, allerdings weniger stark ausgeprägt. Auch dieses Ergebnis steht im Einklang mit den theoretischen Vorhersagen von Box (1954b) und Scheffé (1959).

3) Die serielle Korrelation wird verallgemeinert, so daß nunmehr jede Versuchsperson mit jeder anderen zusammenarbeitet. Dies führt zu sehr ausgeprägten Unterschieden zwischen faktischen und nominellen Fehlerwahrscheinlichkeiten, bspw. $p_\alpha = 0{,}50$ bei $\alpha = 0{,}05$ und zweiseitigem Testen.

4) Die serielle Korrelation entsteht dadurch, daß jede Vp mit fünf weiteren Vpn kommuniziert, und zwar derart, daß diese Kommunikation mit der Entfernung der Vpn voneinander abnimmt. Die Resultate sind mit denen der seriellen Korrelation unter 3) erwartungsgemäß vergleichbar, und der t-Test erweist sich erneut als nicht robust.

Die Bedeutung von Abhängigkeiten für die Teststärke ist weder von Lissitz & Chardos (1975) noch — soweit wir wissen — von anderen Autoren untersucht worden.

Insgesamt bewegen sich die Abweichungen der tatsächlichen von den nominellen Fehlerwahrscheinlichkeiten bei korrelierten Daten teilweise in Größenordnungen, wie sie bei den im folgenden anzusprechenden Annahmeverletzungen kaum anzutreffen sind. Allerdings wird in der Literatur immer wieder darauf hingewiesen (etwa von Box, 1954; Glass et al., 1972), daß sorgfältiges Randomisieren als das beste Mittel zur Vermeidung abhängiger Daten innerhalb und zwischen den Gruppen anzusehen ist.

2.4 Normalverteilungen mit homogenen Varianzen

Sind die Grundgesamtheiten homogen variant und normal verteilt, ist für keinen der hier betrachteten Tests eine Voraussetzung verletzt. In diesem Fall stellt jedoch der t-Test die beste Wahl zur Prüfung von Mittelwertshypothesen dar, weil er sowohl „unverzerrt" ist als auch den „gleichmäßig teststärksten" von allen möglichen Tests darstellt (vgl. zu den Testgütekriterien bspw. Fisz, 1971). Robustheitsuntersuchungen erübrigen sich daher, und die einzige interessierende Frage richtet sich darauf, wieviel besser der t-Test unter „seinen" Optimalbedingungen ist als die Vergleichstests, wenn es um die Entdeckung „wahrer" Alternativhypothesen geht (Vergleich der Teststärken; siehe dazu die Angaben zur A.R.E. der Tests im Teil 1, die sich auf unendlich große Stichproben beziehen).

2.4.1 Teststärkevergleich: t- und W-Test

Rouanet & Lépine (1974) haben abgeleitet, daß für große Stichproben der exakte Welch-Test nach Welch (1937) und Aspin (1948, 1949) vernachlässigbar weniger teststark ist als der t-Test, während bei kleinen Stichproben dieser Unterschied als etwas ausgeprägter zu erwarten ist, und zwar zugunsten des t-Tests. Ramsey (1980) stellt ebenfalls fest, daß die Teststärke des Welch-Tests geringfügig niedriger ist als die des t-Tests, macht jedoch keine Angaben über die Größenordnung.

Golhar (1972) hat für einseitige Tests bei $\alpha = 0{,}01$ und 0,05 die theoretischen Werte der Teststärkefunktionen für den exakten Welch-Test bestimmt, der sich nach den Befunden von Scheffé (1970) und Wang (1971) für alle praktischen Belange nicht von dem approximativen, auf die t-Verteilung bezogenen W-Test (Welch, 1949; s.o.) unterscheidet, was die Wahrscheinlichkeiten für Fehler 1. Art anbelangt.

Zwar vergleicht Golhar selbst seine Resultate für gleiche Varianzen und Stichprobenumfänge nicht mit den entsprechenden Werten der Teststärkefunktion des t-Tests, doch ist dies unter Verwendung entsprechender Tabellen vglw. einfach möglich. Golhar (1971, 210 und 212) benutzte als Effektmaß den Nonzentralitätsparameter δ:

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

In den für den t-Test angelegten Teststärke-Tabellen von Cohen (1977) wird dagegen der folgende Parameter zugrundegelegt:

$$d = \frac{\mu_1 - \mu_2}{\sigma} \tag{2}$$

Für den hier interessierenden Fall, daß $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ist und $n_1 = n_2 = n$, gilt folgende Beziehung zwischen beiden Effektmaßen:

$$d = \delta \cdot \sqrt{\frac{2}{n}} \tag{3}$$

Mithilfe von Gleichung (3) läßt sich daher Golhars δ in Cohens d umrechnen, und somit kann ein hinreichend genauer Vergleich der Teststärken für den t- und den exakten Welch-Test unter der Annahme der homogen varianten Normalverteilungen unter Verwendung von Golhars und Cohens Tabellen erfolgen. Für die Werte von δ, für die es in Cohens Tabellen kein Äquivalent d gibt, kann die folgende recht genaue Annäherungsformel verwendet werden (Cohen, 1977, 53):

$$n_{\text{tabelliert}} = (n-c) \cdot 100 \cdot d^2, \tag{4}$$

wobei c = 1,5 für $\alpha = 0{,}01$ (einseitig) und c = 0,7 für $\alpha - 0{,}05$ (einseitig); $n_{tabelliert}$ bezieht sich auf die Spalte „$n_{.10}$" in Cohens Tabellen (zu den Einzelheiten siehe Cohen, 1977).

Cohens Tabellen sollen zudem bei ungleichen Stichprobengrößen approximativ und konservativ benutzbar sein, wenn man zur Berechnung der Teststärke das tabellierte n' wie folgt bestimmt (Cohen, 1977, 42):

$$n' = \frac{2n_1n_2}{n_1 + n_2} \quad \text{(harmonisches Mittel der Umfänge)} \qquad (5)$$

Tabelle 3
Vergleich einiger theoretischer Teststärken-Werte für den t- und den Welch-Test, berechnet nach den Tabellen von Golhar (1972) und Cohen (1977). Erläuterungen im Text.

		$\alpha = 0{,}01$			
		$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
$n = 4$	Cohen	0,07	0,22	—	—
	Golhar	0,0623	0,2114	—	—
$n' = 5{,}1$	Cohen	0,07	0,25	—	—
$(n_1 = 4; n_2 = 7)$	Golhar	0,0683	0,229	—	—
$n = 7$	Cohen	0,07	~0,285	—	—
	Golhar	0,07417	0,28629	—	—
$n' = 8{,}5$	Cohen	0,07	0,295	~0,61	—
$(n_1 = 7; n_2 = 11)$	Golhar	0,07637	0,29161	0,62132	—
$n' = 10{,}5$	Cohen	0,075	0,30	~0,655	—
$(n_1 = 7; n_2 = 21)$	Golhar	0,07322	0,27034	0,58149	—
$n' = 11{,}4$	Cohen	0,075	0,32	~0,67	—
$(n_1 = 7; n_2 = 31)$	Golhar	0,07043	0,25725	0,55845	—
$n = 21$	Cohen	0,08	0,345	~0,70	0,935
	Golhar	0,08691	0,34676	0,71608	0,93772
$n' = 25{,}04$	Cohen	0,084	0,346	~0,75	~0,92
$(n_1 = 21; n_2 = 31)$	Golhar	0,08726	0,34817	0,71783	0,93847

Tabelle 3 (Fortsetzung) $\alpha = 0{,}05$

		$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
$n = 4$	Cohen	0,23	0,57	—	—
	Golhar	0,2198	0,5377	—	—
$n' = 5{,}1$	Cohen	0,235	0,58	—	—
$(n_1 = 4; n_2 = 7)$	Golhar	0,2251	0,543	—	—
$n = 7$	Cohen	0,24	0,60	—	—
	Golhar	0,24073	0,59356	—	—
$n' = 8{,}5$	Cohen	0,25	0,605	0,88	—
$(n_1 = 7; n_2 = 11)$	Golhar	0,24223	0,59575	0,88076	—
$n' = 10{,}5$	Cohen	~0,25	0,615	0,885	—
$(n_1 = 7; n_2 = 21)$	Golhar	0,23788	0,58289	0,86932	—
$n' = 11{,}4$	Cohen	~0,25	0,615	0,89	—
$(n_1 = 7; n_2 = 31)$	Golhar	0,23471	0,57454	0,86218	—
$n = 21$	Cohen	~0,25	0,625	0,905	0,99
	Golhar	0,25405	0,624	0,90384	0,98888
$n' = 25{,}04$	Cohen	0,28	0,63	0,985	>0,995
$(n_1 = 21; n_2 = 31)$	Golhar	0,25441	0,62673	0,9435	0,98899

Für ungleiche Varianzen dagegen lassen sich Cohens Tabellen nicht sinnvoll einsetzen (Cohen, 1977, 44).

Wir haben die entsprechenden Berechnungen vorgenommen und in der Tabelle 3 zusammengestellt. Dieser Tabelle ist zu entnehmen, daß die Übereinstimmung der Teststärke-Werte für den exakten Welch- und den t-Test hervorragend ist — mit Ausnahme der Fälle, in denen die Stichprobenumfänge sehr unterschiedlich sind.

Hübner, Lübbeke & Hager (1982) sind der Frage empirisch nachgegangen, inwieweit sich die vorstehenden Ergebisse auf den approximativen W-Test, bei dem die Signifikanzbeurteilung über die t-Verteilung mit adjustierten Freiheitsgraden erfolgt, übertragen lassen. Ihre Ergebnisse zeigen, daß zwischen den Tests nur verschwindend geringe Unterschiede hinsicht-

lich der Teststärke bestehen (vgl. dazu beispielhaft Abbildung 1); es ist daher zu rechtfertigen, die für den exakten W-Test gewonnenen Resultate von Golhar (1972) auf den approximativen W-Test zu übertragen.

Des weiteren zeigt die theoretische Arbeit von Mehta & Srinivasan (1970), daß der (exakte) Welch-Test unter homogen varianten Normalverteilungen stets teststärker ist als seine zahlreichen Konkurrenten.

2.4.2 Teststärkenvergleich: U- und NS-Tests

Van der Laan & Oosterhoff (1965, 1967) fanden bei gleichen Stichprobengrößen und einseitigen Tests so gut wie keine Unterschiede zwischen dem U- und den Normal-Scores-Tests nach van der Waerden und nach Terry-Hoeffding. Dieses Ergebnis konnten Neave & Granger (1968) für etwas größere Stichproben bestätigen. In dem untersuchten Bereich von Stichproben (6 bis 40 pro Gruppe; vgl. Tab. 1) macht sich offenbar der leichte Vorteil der NS-Tests gegenüber dem U-Test hinsichtlich der relativen Effizienz nicht bemerkbar.

2.4.3 Teststärkenvergleich: parametrische und nonparametrische Tests

Boneau (1962) fand in seiner Studie nur eine geringfügige Überlegenheit des t- vor dem U-Test; das gleiche Resultat erhielten van der Laan &

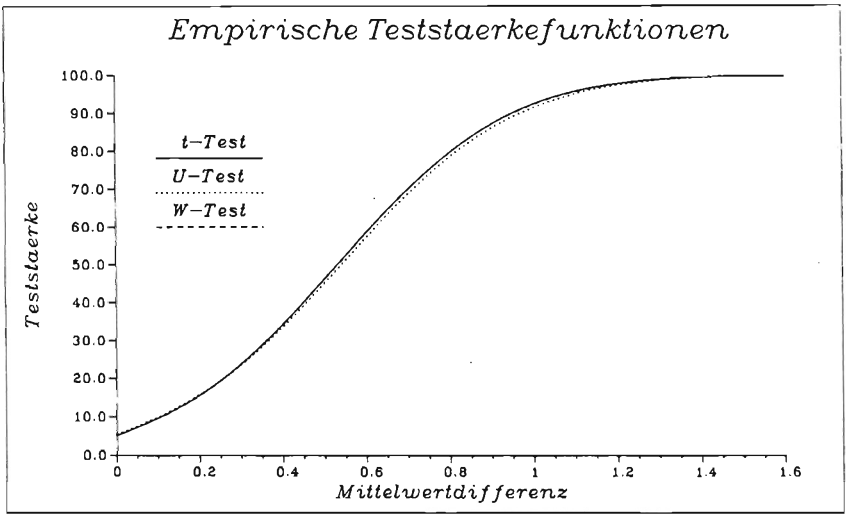

Abb. 1
Empirischer Teststärkenvergleich zwischen einseitigen t-, U- und W-Tests unter homogen varianten Normalverteilungen bei $n_1 = n_2 = 20$ und $\alpha = 0{,}05$; Kurven geglättet (aus Hübner, Lübbeke und Hager, 1982).

Oosterhoff (1965, 1967) sowie Hübner, Lübbeke & Hager (1982), die zusätzlich den W-Test untersuchten. Die der Arbeit der letztgenannten Autoren entnommene Abbildung 1 verdeutlicht diese Befunde beispielhaft.

Die einzige uns bekannte empirische Studie zum Vergleich von t-, U- und Randomisierungs-t-Test stammt von Toothaker (1972), der gleiche und ungleiche Stichproben untersucht (vgl. Tab. 1), wobei er zur Bestimmung der Teststärken ein Populationsmodell mit wiederholtem Stichprobenziehen zugrundelegt (vgl. zu den möglichen Modellen Toothaker, 1972, 85; ferner Kempthorne & Doerfler, 1969; Lehmann, 1975). Die Teststärke des t-Tests ist durchgängig geringfügig besser als die der beiden anderen Tests, deren Teststärken bei kleinen Stichproben ($n_1 \leq 3$; $n_2 \leq 5$) gleich sind. Lediglich bei den etwas größeren Stichproben zeigen sich leichte Vorteile des Randomisierungs-t-Tests vor dem U-Test. Es muß betont werden, daß insgesamt die Unterschiede nur selten größer als $2\sigma_{\hat{p}}$ sind, wobei der größte Wert für $\sigma_{\hat{p}} = 0{,}0145$ ist (vgl. Toothaker, 1972, 90).

2.5 Normalverteilungen mit heterogenen Varianzen

Die Bedeutung heterogener Varianzen für die Stichprobenverteilung der Teststatistik t wird bereits recht lange untersucht. Hsu (1938) ist es dabei als einem der ersten Mathematiker gelungen, theoretische Resultate zu der Frage abzuleiten, welche Konsequenzen die Verletzung der Varianzhomogenitätsvoraussetzung auf den t-Test hat. Im wesentlichen bestimmte Hsu (1938) für den zweiseitigen t-Test und für unterschiedliche Varianzverhältnisse (σ_1^2/σ_2^2) die Differenz zwischen tatsächlichen (p_α) und nominellen (α) Fehlerwahrscheinlichkeiten 1. Art. Seine Ergebnisse wurden in den Folgejahren von zahlreichen Autoren bestätigt und erweitert (u.a. Gronow, 1951; Box, 1954a,b; Scheffé, 1959; Pratt, 1964; Bhattacharjee, 1968; ferner zusammenfassend Glass, Peckham & Sanders, 1972).

Der wesentliche theoretische Befund ist wohl darin zu sehen, daß die Auswirkungen unterschiedlicher Populationsvarianzen bei *gleichen* Stichprobenumfängen i.a. vernachlässigt werden können. Bei *ungleichen* Stichproben ist dagegen mit ausgeprägten Unterschieden zwischen p_α und α zu rechnen. Das Ausmaß dieser Unterschiede kann ungefähr bestimmt werden, wenn man eine von van der Vaart (1961) abgeleitete Formel verwendet, die sich in leicht handhabbarer Form etwa bei Glass et al. (1972, 245) findet. Danach läßt sich zusammenfassend folgendes ausführen: Entstammt bei ungleichen Stichproben die kleinere Stichprobe der Population mit der kleineren Varianz, *unterschreitet* das tatsächliche Signifikanzniveau das nominelle ($p_\alpha < \alpha$). Wurde dagegen die kleinere Stichprobe der Population mit der größeren Varianz entnommen, *übersteigt* das tatsächliche Signifikanzniveau das nominelle ($p_\alpha > \alpha$).

2.5.1 Gleiche Stichprobenumfänge

2.5.1.1 Robustheit

In den empirischen Untersuchungen von Boneau (1962), Havlicek & Peterson (1974) sowie von Bradley (1980a) erweisen sich ein- und zweiseitige t-Tests bei gleichen Stichproben ($n_1=n_2=5$ bis 16) selbst bei einem Varianzverhältnis von $\sigma_1^2:\sigma_2^2 = 1:4$ bei $\alpha = 0{,}01$ und 0,05 als robust im Sinne des Bradley-Kriteriums (s.o.). Bei dem extremeren Varianzverhältnis von 1:16 und $n_1 = n_2 = 5$ ist der t-Test bei Havlicek & Peterson (1974) nur noch ab $\alpha = 0{,}05$ robust. Für noch extremere Varianzverhältnisse (bis $\sigma_1^2:\sigma_2^2 \rightarrow \infty$) muß nach der Untersuchung von Ramsey (1980) der Gesamtstichprobenumfang mindestens $N = 14$ ($n = 7$) sein, damit der t-Test robust bleibt. Um diese Robustheit auch auf dem niedrigeren Niveau $\alpha = 0{,}01$ zu gewährleisten, muß $n_1 = n_2 \geq 16$ sein (Ramsey, 1980). Ähnliche Resultate fanden sich auch bei Young & Veldman (1963), deren Daten teilweise in Glass et al. (1972) abgedruckt sind.

In Übereinstimmung mit den oben erfolgten theoretischen Erörterungen fanden Boneau (1962) sowie Hübner, Lübbeke & Hager (1982), daß der U-Test zur Prüfung von Mittelwertshypothesen bei einem realistischen Varianzverhältnis von 1:4 sich nicht anders verhält als der t-Test; das gleiche Ergebnis fanden Neave & Granger (1968) für den t-, den U- und die NS-Tests.

Mehta & Srinivasan (1970) haben zudem einen *theoretischen* Vergleich des Welch-Test mit fünf seiner Alternativen vorgenommen und dabei teilweise sehr extreme Varianzverhältnisse in der maximalen Größenordnung von 1:10.000 untersucht. Sie weisen nach, daß der (exakte) Welch-Test unter allen Bedingungen eine ausgezeichnete Kontrolle der Fehlerwahrscheinlichkeiten 1. Art gewährleistet, die von keinem der Alternativverfahren übertroffen wird, sofern $n_1 = n_2 \geq 7$. — In einer ebenfalls theoretischen Studie weist Wang (1971) ferner nach, daß sich der approximative W-Test in bezug auf das Signifikanzniveau nicht von dem exakten Welch-Test unterscheidet.

2.5.1.2 Teststärke

Die theoretische Bestimmung von Teststärkefunktionen ist bei heterogenen Varianzen nicht exakt möglich. Die Berechnung entsprechender approximativer Werte beruht im Regelfall auf einem Mittelwert aus den heterogenen Varianzen (Horsnell, 1953; Donaldson, 1968; Glass, Peckham & Sanders, 1972).

Den Teststärke-Untersuchungen von Boneau (1962) bei $n_1 = n_2 = 5$ und 15 ist zu entnehmen, daß sich die tatsächlichen Teststärken von t- und

U-Test praktisch nicht unterscheiden. Während Boneau das Varianzverhältnis 1:4 untersuchte, befassen sich Neave & Granger (1968) mit dem kleineren Verhältnis 1:2 und finden bei $n_1 = n_2 = 20$ kaum Teststärkeunterschiede zwischen dem t-, dem U- und den NS-Tests.

In der empirischen Untersuchung von Bennett & Hsu (1961) erwies sich des weiteren der exakte Welch-Test bei gleichen Stichproben als teststärker als der Behrens-Fisher-Test. — Diesen Befund konnten Mehta & Srinivasan (1970) in ihrer theoretischen Untersuchung bestätigen und verallgemeinern; danach ist der Welch-Test teststärker als seine Äquivalente, sofern $n_1 = n_2 \geq 7$.

2.5.2 Ungleiche Stichprobenumfänge

2.5.2.1 Robustheit

Wie bereits oben ausgeführt, resultiert ein liberaler Test, d.h. $p_\alpha > \alpha$, wenn die größere Stichprobe aus der Population mit der kleineren Varianz stammt, und im umgekehrten Fall wird der Test konservativ, d.h. $p_\alpha < \alpha$. Bei einem wohl als realistisch, wenn auch bereits als vglw. extrem anzusehenden Varianzverhältnis $\sigma_1^2 : \sigma_2^2 = 1 : 4$ ist der einseitige t-Test selbst bei nur geringfügigen Unterschieden zwischen den beiden Stichprobenumfängen ($n_1 = 5$, $n_2 = 7$) nur dann robust, wenn die kleinere Stichprobe aus der Population mit der kleineren Varianz stammt ($\alpha = 0{,}05$; 0,01) (Havlicek & Peterson, 1974); für die Kombinationen $n_1 : n_2 = 7 : 5$; $15 : 5$; $16 : 8$ ist der t-Test auf keinem der untersuchten Signifikanzniveaus robust (vgl. Boneau, 1960, 1962; Havlicek und Peterson, 1974; Bradley, 1980a). Das gleiche gilt grundsätzlich auch für das extremere Varianzverhältnis von 1 : 16, das Havlicek und Peterson (1974) untersucht haben.

Die im Abschnitt 2.5.1.1 zum Welch-Test und zum approximativen W-Test erfolgten Ausführungen, basierend auf den theoretischen Arbeiten von Mehta & Srinivasan (1970) und von Wang (1971) (vgl. auch Scheffé, 1970), lassen sich uneingeschränkt übertragen: Auch beim Vorliegen ungleicher Stichprobenumfänge gewährleisten die beiden Formen des Welch-Tests eine optimale Kontrolle der Typ-1-Fehlerwahrscheinlichkeit unter Varianzheterogenität.

Der U-Test reagiert bei Varianzheterogenität unter ungleichen Stichproben in der gleichen Weise unrobust wie der t-Test (Boneau, 1960; Edgington, 1965). Um diese mangelnde Robustheit auszugleichen, haben u.a. Potthoff (1963), Trommer (1967) und Fligner & Pollicello (1981) nonparametrische Äquivalente zum Welch-Test vorgeschlagen.

2.5.2.2 Teststärke

Nach den vorliegenden Befunden verringert sich unter Normalverteilungen mit heterogenen Varianzen die tatsächliche Teststärke im Vergleich zur nominellen unter homogen varianten Normalverteilungen. Die relative Teststärke zweier verschiedener Tests hängt jedoch davon ab, wie groß die Werte p_α für die beiden Tests relativ zueinander sind. In den Fällen, in denen bspw. p_α für den t-Test geringer ist als für den U-Test, ist aufgrund der theoretischen Befunde von Wetherill (1960) eine überlegene tatsächliche Teststärke des U-Tests relativ zum t-Test zu erwarten; die empirischen Ergebnisse von Boneau (1962) bestätigen diese Erwartung.

Zur Teststärke der NS-Tests unter Varianzheterogenität liegen unseres Wissens keine empirischen Untersuchungen vor. Dagegen haben Bennett & Hsu (1961) die Teststärke des exakten Welch-Tests mit dem Behrens-Fisher-Test verglichen und fanden auch bei ungleichen Stichproben durchgängige Vorteile zugunsten des Welch-Tests; Mehta & Srinivasan (1970) konnten in ihrem theoretischen Vergleich der Teststärken ebenfalls die Überlegenheit des Welch-Tests nachweisen.

Von den oben erwähnten robusten Modifikationen des U-Tests ist lediglich die von Fligner & Pollicello (1981) bei größeren Stichproben ($n_1 = 25$, $n_2 = 20$) nicht weniger teststark als der U-Test.

2.6 Nicht-normale homomere Verteilungen

Betrachtet man nicht-normale homomere Verteilungen, empfiehlt es sich aus Gründen der Übersichtlichkeit, zwischen homomeren symmetrischen und homomeren asymmetrischen Verteilungen zu differenzieren. — Bei nicht-normalen Verteilungen ist die Normalitätsvoraussetzung für die parametrischen Tests nicht erfüllt, während bei homomeren Verteilungen keine Voraussetzungen für die nicht-parametrischen Tests verletzt sind.

2.6.1 Symmetrische homomere Verteilungen

Zu den symmetrischen Verteilungen zählen die Rechteck- (oder Gleich-), die Doppelt-Exponential- und die logistische Verteilung, die auch recht häufig in Simulationsuntersuchungen benutzt werden, um von der Normalverteilung abweichende symmetrische Verteilungen zu repräsentieren.

Da lediglich eine Voraussetzung der parametrischen Tests verletzt ist, sind auch nur diese unter Robustheitsaspekten zu vergleichen.

2.6.1.1 Robustheit

Bei Vorliegen zweier Rechteckverteilungen erweist sich der t-Test selbst bei kleinen Stichproben der Größe $n_1 = n_2 = 5$ und bei $\alpha = 0{,}01$ und $0{,}05$

als robust (Boneau, 1960). Das gleiche Resultat erhielt Toothaker (1972) bei seinem Vergleich des t- mit dem U- und dem Randomisierungs-t-Test unter zwei Rechteckverteilungen bei α = 0,0286; 0,0317; 0,0357; 0,0476; 0,05; 0,0667; 0,10 und bei noch kleineren Stichprobenumfängen.

Hübner et al. (1982) variierten den Exzeß γ_2 zweier homomerer Verteilungen von -1 bis $+3$ und erhielten sowohl für den t- wie für den W-Test unter diesen lepto- und platykurtischen Verteilungen bei α = 0,05 und 0,01 robuste Resultate. Bei $\gamma_2 = +3$ und α = 0,001 führten beide Tests zu konservativen Entscheidungen.

2.6.1.2 Teststärke

Blair & Higgins (1980a) finden in ihrer Simulationsstudie, daß der t-Test unter Rechteckverteilungen ($A.R.E._{U,t}$ = 1,0) einen geringfügigen Vorteil gegenüber dem U-Test aufweist ($n_1 : n_2$ = 3 : 9; 6 : 6; 9 : 27; 18 : 18; 27 : 81; 54 : 54). Über vergleichbare Resultate berichtet Boneau (1962) für den U-Test und Toothaker (1972) für den t-, den U- und den Randomisierungs-t-Test.

Unter Doppelt-Exponential-Verteilungen ist die $A.R.E._{U,t}$ = 1,5; und Blair & Higgins (1980) weisen entsprechende Vorteile des U-Tests nach. — Unter dem gleichen Verteilungstyp haben Conover, Wehmanen & Ramsey (1978) die Teststärke und des Normal-Scores-Tests nach van der Waerden sowie einiger weiterer Tests miteinander verglichen. In dieser Untersuchung erwies sich der U-Test im Regelfall als geringfügig teststärker als die übrigen Tests, wobei ihm der Normal-Scores-Test nur wenig nachstand (vgl. auch Ramsey, 1971).

Bei logistischen Verteilungen beträgt die $A.R.E._{U,t}$ = 1,1. Van der Laan & Weima (1978) haben diesen Verteilungstyp untersucht. Bei $n_1 = n_2$ = 3; 6 erwies sich der U-Test als durchgängig testschwächer als der t-Test. Bei größeren Stichprobenumfängen (ab $n_1 = n_2$ = 10) hängen die insgesamt geringen Unterschiede vom gewählten Signifikanzniveau ab: Bei α = 0,01 ist der U-Test schwächer, bei α = 0,05; 0,10 dagegen stärker als der t-Test.

2.6.2 Asymmetrische homomere Verteilungen

Zu den asymmetrischen Verteilungen, die vglw. häufig in Monte-Carlo-Studien zugrundegelegt werden, zählen Exponential-, die Lognormal-, die Poisson-, die zusammengesetzten Normal- sowie die χ^2-Verteilungen.

2.6.2.1 Robustheit

Posten (1978) untersuchte die Robustheit des t-Tests bei gleichen Stichprobenumfängen unter einer Vielzahl von nicht-normalen Verteilungen, die

einerseits als „moderate Abweichungen" von der Normalverteilung angesehen werden können und die andererseits wohl einen realistischen Bereich von Abweichungen repräsentieren (hierzu Pearson & Please, 1975). Posten variierte die Schiefe γ_1 über einen Bereich von $-\sqrt{2}$ bis $+\sqrt{2}$ und den Exzeß γ_2 von $-1{,}6$ bis $+4{,}8$. Der t-Test erwies sich stets und ausnahmslos als robust unter Verwendung des hier zugrundegelegten Bradley-Kriteriums. Das gleiche Resultat berichten Pearson & Please (1975) für einen realistischen Bereich von Abweichungen, während Bowman, Beauchamp & Shenton (1977) teilweise Werte für p_α fanden, die dem Kriterium nicht mehr genügen. Dagegen bestätigen die Ergebnisse von Havlicek & Peterson (1974) ($\gamma_1 = 0{,}422$ und $\gamma_2 = -0{,}914$; $\gamma_1 = 0{,}978$ und $\gamma_2 = 0{,}543$) und von Toothaker (1972) (χ^2-Verteilungen mit df = 3 und $\gamma_1 = 1{,}633$ und $\gamma_2 = 4{,}0$) die Befunde von Posten (1978) und Pearson & Please (1975). Neave & Granger (1968) konnten die Robustheit des t-Tests selbst bei zwei bimodalen Verteilungen mit $\gamma_1 = 0{,}75$ und $\gamma_2 = -0{,}376$ nachweisen.

Diese Robustheit verschwindet jedoch allmählich mit anwachsendem Ausmaß der Verschiedenheit der untersuchten Verteilungen relativ zur Normalverteilung. Boneau (1962) verwendete Exponentialverteilungen mit $\gamma_1 = 2$ und $\gamma_2 = 6$ und stellte fest, daß der t-Test bei $\alpha = 0{,}01$ nicht mehr robust ist, während Donaldson (1968) auch für $\alpha = 0{,}01$ noch über robuste Werte für p_α berichtet. Noch extremere nicht-normale Verteilungen hat Bradley (1968, 1977, 1980a, b) ausführlich untersucht; er verwendet vorwiegend eine L-förmige Verteilung mit $\gamma_1 = 3{,}18$ und $\gamma_2 = 10{,}85$ (vgl. Bradley, 1977, 1980c), die zwar als untypisch, nicht jedoch als unplausibel oder gar unmöglich in der psychologischen Forschung anzusehen ist (Bradley, 1977). Unter diesen extremen Bedingungen wird das Ausmaß deutlich, in dem die Schiefen der Populationen die Schiefe der Prüfverteilung beeinflussen können, deren genaue Form zudem nicht unmaßgeblich vom Verhältnis der beiden Stichprobenumfänge bestimmt wird (s.u.).

Bradley (1980a) untersucht nur eine Kombination gleicher Umfänge ($n_1 = n_2 = 16$). Bei dieser erweisen sich links-, rechts- und zweiseitige t-Tests unter den oben beschriebenen extrem rechtsschiefen Populationen als zwar konservativ, aber robust, solange bei $\alpha = 0{,}05$ getestet wird. Unter Verwendung der niedrigeren Signifikanzniveaus $\alpha = 0{,}01$; $0{,}001$ dagegen stellen sich teilweise sehr ausgeprägte Unterschiede zwischen den tatsächlichen und den nominellen Fehlerwahrscheinlichkeiten ein, bspw. $p_\alpha = 0{,}1\alpha$ (Bradley, 1980a, 277). Donaldson (1968) untersuchte zwei lognormale Verteilungen mit $\gamma_1 = 4$ und $\gamma_2 = 38$ bei gleichen Stichprobengrößen ($n_1 = n_2 = 4$; 8; 16; 32) und den Niveaus $\alpha = 0{,}10$; $0{,}05$; $0{,}01$ für zweiseitige Tests; seine Resultate können im wesentlichen als Bestätigung der Befunde

Bradleys angesehen werden, allerdings mit der Ergänzung, daß bei n = 32 der t-Test auch bei $\alpha = 0{,}01$ robust ist.

Werden ungleiche Stichprobenumfänge verwendet (bei Bradley, 1980a: $n_1 \neq n_2$, $n_1 = 8; 16; 24$, $n_2 = 8; 16; 24$), ist der t-Test um so weniger robust, je stärker das Verhältnis $n_1 : n_2$ vom Wert 1 abweicht, wobei einseitige Tests stärker betroffen sind als zweiseitige, die bei $\alpha = 0{,}05$ noch dem Robustheitskriterium genügen können. Als Beispiel für die recht komplexen Zusammenhänge zwischen der Schiefe der Populationsverteilungen, den Stichprobengrößen und den Stichprobenverteilungen mögen die folgenden Zahlen aus Bradley (1980a, 277) dienen: Der linksseitige t-Test verhält sich bei den o.gen. extrem rechtsschiefen Verteilungen bei $n_1 : n_2 = 8 : 16$ konservativ (und nicht robust) ($p_\alpha = 0{,}0101 < \alpha = 0{,}05$) und bei $n_1 : n_2 = 16 : 8$ liberal und robust ($p_\alpha = 0{,}0638 > \alpha = 0{,}05$), während für den rechtsseitigen t-Test genau das Umgekehrte gilt und der zweiseitige Test konservativ und robust ist ($p_\alpha = 0{,}0350 < \alpha = 0{,}05$).

Über ganz ähnliche Resultate berichten Havlicek & Peterson (1974), die zusätzlich zu Bradley die Parameter Schiefe und Exzeß getrennt voneinander variierten. Hierbei stellte sich heraus, daß in erster Linie die Schiefe für die mangelnde Robustheit des t-Tests verantwortlich zeichnet, während der Einfluß des Exzeß' vernachlässigbar ist. Die Befunde von Pearson & Please (1975) weisen in die gleiche Richtung.

2.6.2.2 Teststärke

Bradley (1980a) behauptet einen Zusammenhang zwischen Änderungen der Wahrscheinlichkeiten für Fehler 1. und 2. Art: „It is sometimes claimed that it is less serious for p to fall below α than to fall above it because in the former case the actual probability of a Type 1 error is on the „safe" or „conservative" side of the nominal probability α. However, this is no cause for complacency, since, in that case, the power of the test is reduced accordingly and the probability of a Type II error must therefore be on the „unsafe" or „radical" side of '" (Bradley, 1980a, 278). Diese Behauptung ist jedoch unzutreffend: Bei Exponentialverteilungen ist $p_\alpha < \alpha$ für den t-Test, der damit konservativ ist. Dessen ungeachtet ist die Teststärke für fast alle Werte des Nonzentralitätsparameters größer als bei Normalverteilungen, wie Boneau (1962) und Donaldson (1968) aufzeigen konnten; dieser Teststärkevorteil ist unter log-normalen Verteilungen sogar noch ausgeprägter. In allen genannten Fällen beginnt die Teststärkenkurve zwar bei einem Wert, der kleiner als α ist, wächst jedoch anschließend erheblich steiler an als unter Normalitätsbedingungen, wie die Tabelle 4 ausweist.

In der Untersuchung von Boneau (1962) zeigen sich keine durchgängi-

Tabelle 4

Vergleich der Teststärken (in Prozent) des zweiseitigen t-Tests unter zwei Normal-, Exponential- oder Lognormalverteilungen für verschiedene Nichtzentralitätswerte d (vgl. Formel (3)), gleiche Varianzen und Stichprobenumfänge ($n_1 = n_2 = 8$) (zusammengestellt nach Donaldson, 1968, 665).

	Normalverteilung		Exponentialverteilung		Lognormalverteilung	
	α		α		α	
d	5%	1%	5%	1%	5%	1%
0	4,9	0,9	4,4	0,7	3,5	0,5
.3536	9,8	2,5	11,5	3,0	14,5	6,0
.7072	26,0	9,2	32,5	12,8	41,7	20,1
1.0608	50,5	24,4	57,3	32,8	66,3	44,9
1.4144	74,9	46,5	77,3	55,3	81,5	65,3
1.7680	90,9	70,9	88,7	73,8	89,9	78,9
2.1216	97,8	87,5	95,1	85,6	94,9	87,7

gen Teststärkeunterschiede zwischen dem t- und dem U-Test bei Exponentialverteilungen und kleinen Stichproben ($n_1 = n_2 = 5$). Blair, Higgins & Smitley (1980) bemängeln diese Umfänge als zu klein und weisen für größere einen teilweise erheblichen Teststärkevorteil des U- vor dem t-Test nach. Dieser beträgt bei $n_1 = n_2 = 18$ unter Exponentialverteilungen 50% und bei $n_1 = n_2 = 54$ sogar 100% bei ein- und zweiseitigen Tests. In diesem Fall macht sich also die hohe $A.R.E._{U,t} = 3{,}0$ des U- im Vergleich zum t-Test schon bei vglw. geringen Stichproben bemerkbar.

Blair & Higgins (1980a,b) haben zusammengesetzte Normalverteilungen (vgl. dazu Bradley, 1977) mit einer $A.R.E._{U,t} = 45$ untersucht. Auch hier erweist sich der U-Test bei $n_1 = n_2 = 18$ als wesentlich stärker als der t-Test, während bei kleineren Stichproben ($n_1 = n_2 = 6$; $n_1 = 3$, $n_2 = 9$) teilweise der t-Test bessere Resultate erbrachte.

In diesem Zusammenhang soll noch auf eine Eigenart der Studien von Blair & Higgins (1980a,b) und Blair, Higgins & Smitley (1980) hingewiesen werden: Diese Autoren lassen Fehler 1. Art außer acht und benutzen für die empirischen Teststärkenschätzungen kritische Werte, die nicht mit den tabellierten übereinstimmen, sondern die ebenfalls empirisch unter den betreffenden nicht-normalen Bedingungen ermittelt werden. Da

man jedoch bei der praktischen Anwendung der Testverfahren auf tabellierte Werte angewiesen ist, erscheint uns die Frage der Übertragbarkeit der Ergebnisse von Blair et al. auf reale Situationen nur schwer zu beantworten zu sein.

Ein Vergleich der Teststärke von t- und W-Test findet sich bei Hübner et al. (1982), die bei einer Schiefe von $\gamma_1 = 0{,}75$ als moderater Abweichung von der Normalverteilung selbst bei etwas größeren Stichproben ($n_1 = n_2 = 20$) keine nennenswerten Vorteile des U- vor dem t- und dem W-Test feststellen konnten, obwohl der U-Test stets geringfügig teststärker war als die anderen Tests.

Neave & Granger (1968) haben die Teststärke von t-, U- und den NS-Tests unter zusammengesetzten Normalverteilungen mit $\gamma_1 = 0{,}75$ und $\gamma_2 = -0{,}375$ untersucht und gefunden, daß der U-Test teststärker ist als der t-Test, aber deutlich schwächer als der Normal-Scores-Test nach Terry-Hoeffding. Kemp & Conover (1973) kommen unter Poisson-Verteilungen zu dem gleichen Resultat und gehen auf das Problem der Rangbindungen ein, für das sie die Bildung von Durchschnittswerten empfehlen.

Der Vergleich der Teststärken von t-, U- und Randomisierungs-t-Test, über den Toothaker (1972) berichtet, ergab bei zwei χ^2-Verteilungen für 3 Freiheitsgrade mit $\gamma_1 = 1{,}633$ und $\gamma_2 = 4{,}0$ bei sehr kleinen Stichproben, daß der Randomisierungs-t-Test nur bei ungleichen Umfängen geringfügig besser ist als der t-Test. Der U-Test ist nur in Ausnahmefällen der stärkste der drei untersuchten Tests; vermutlich handelt es sich jedoch um zufallsbedingte Resultate, da keine Regelhaftigkeit (Abhängigkeit von der Parameterdifferenz δ bzw. d oder der Stichprobengröße) auszumachen ist.

2.7 Nicht-normale heteromere Verteilungen

Folgt man den Empfehlungen u.a. von Schüle (1976) und Vorberg (1981), sollte der U-Test insbesondere bei mangelnder Kenntnis der Populationsverteilungen anstelle des t-Tests Verwendung finden. Diese allgemeinen Empfehlungen erscheinen jedoch kaum gerechtfertigt, wenn man sich der folgenden theoretischen Resultate von Wetherill (1960), erinnert, der den t- mit dem U-Test u.a. unter heteromeren Populationsverteilungen vergleicht: Der einseitige U-Test „... is not ... a test of medians or of means. Differences of skewness could obscure differences of means or medians which it was desired to test." (S. 415) „Thus, both skewness and kurtosis have a[n] ... effect on the size of the Wilcoxon test, the effect of skewness being particularly severe." (S. 416) Der einseitige U-Test „... is much more sensitive to skewness and kurtosis than the t-test." (Wetherill, 1960, 418). Dagegen ist der zweiseitige U-Test „... insensitive to departures of skewness and kurtosis..." (S. 416).

2.7.1 Heteromere Verteilungen mit homogenen Varianzen

2.7.1.1 Robustheit

Bradley (1980b) vergleicht in seiner Untersuchung eine normale mit einer L-förmigen Verteilung ($\gamma_1 = 3{,}18$; $\gamma_2 = 10{,}85$; s.o.), deren Varianzen gleich sind. Hier erweist sich der t-Test auf keinem der benutzten Signifikanzniveaus und weder bei gleichen noch bei ungleichen Stichprobenumfängen als robust. Ausgeprägte Unterschiede zwischen p_α und α treten dabei sowohl bei ein- wie bei zweiseitigen Tests auf.

Havlicek & Peterson (1974) untersuchen die Robustheit des t-Tests unter verschiedenen Arten der Heteromerität bei links- und rechtsseitigen Tests:

1) Eine normale ($\gamma_1 = 0$; $\gamma_2 = 0$) und eine rechtsschiefe ($\gamma_1 = 1{,}720$; $\gamma_2 = 4{,}256$) Population: Bei $\alpha = 0{,}05$ ist der linksseitige Test unabhängig von den beiden Stichprobengrößen stets robust und dabei konservativ, während der rechtsseitige t-Test liberal und nicht robust ist — mit Ausnahme der größeren Stichproben. Bei $\alpha = 0{,}01$ ist i.a. weder der links- noch der rechtsseitige Test robust, und die Abweichungen liegen in der gleichen Richtung wie bei $\alpha = 0{,}05$.

2) Eine rechts- ($\gamma_1 = 1{,}720$; $\gamma_2 = 4{,}256$) und eine linksschiefe ($\gamma_1 = -1{,}720$; $\gamma_2 = 3{,}976$) Population: Hier sind die einseitigen t-Tests nur in sehr wenigen — und meist mutmaßlich eher zufällig bedingten — Fällen robust. Allerdings werden die Abweichungen $p_\alpha - \alpha$ mit steigenden Stichprobenumfängen stets geringer, so daß — als einzige systematische Ausnahme — für $n_1 = n_2 = 30$ bei $\alpha = 0{,}05$ der links- und der rechtsseitige t-Test robust ist.

3) Eine rechtsschiefe (vgl. vorstehende Angaben) und eine leptokurtische ($\gamma_1 = -0{,}043$; $\gamma_2 = 10{,}646$) Population: Hier erweist sich der einseitige t-Test als überwiegend robust mit leichten Tendenzen in konservativer Richtung bei rechts- und liberaler Richtung bei linksseitigen Tests.

Boneau (1962) untersucht den Fall, daß die eine Population normal, die andere dagegen exponential verteilt ist. Im einzelnen sind seiner Arbeit (Boneau, 1962, 254ff.) die folgenden Resultate für den einseitigen t- und den einseitigen U-Test zu entnehmen, die zeigen, daß beide Tests unter der angegebenen Populationssituation nicht robust und konservativ reagieren ($\alpha = 0{,}05$): $n_1 = n_2 = 5$: t: $p_\alpha = 0{,}014$; U: $p_\alpha = 0{,}022$; $n_1 = n_2 = 15$: t: $p_\alpha = 0{,}037$; U: $p_\alpha = 0{,}016$; zur Interpretation dieser interessanten Befunde siehe Abschnitt 2.7.2.

In Ergänzung der vorstehenden Befunde fanden Hübner, Lübbeke& Hager (1982) in ihrer Studie, daß bei geringfügigen Unterschieden zwischen den Verteilungen (normal und rechtsschief mit $\gamma_1 = +0{,}75$) der t- und der W-Test bei $\alpha = 0{,}01$ und 0,05 robust sind, nicht jedoch bei $\alpha = 0{,}001$;

dagegen ließ der U-Test Robustheit durchgängig vermissen. Bei ihm nahm darüber hinaus die Diskrepanz zwischen p_α und α mit zunehmender Stichprobengröße ebenfalls zu (siehe dazu Abschnitt 2.7.2 unten). — Das gleiche Ergebnismuster erhielten die Autoren auch in den Fällen, in denen anstelle der normalen eine platykurtische ($\gamma_2 = -1{,}0$) oder eine leptokurtische Verteilung ($\gamma_2 = +3{,}0$) mit einer schiefen Verteilung ($\gamma_1 = +0{,}75$) gepaart wurde.

Die Resultate belegen insgesamt, daß der ausschließliche Nachweis der mangelnden Robustheit des t-Test nicht dazu benutzt werden sollte, unkritisch für den allgemeinen Einsatz des U- oder anderer nonparametrischer Tests zu plädieren, die nicht notwendigerweise robuster sind.

2.7.1.2 Teststärke

Boneau (1962) vergleicht die Teststärke von t- und U-Test unter einer Normal- und einer Exponentialverteilung (siehe den vorigen Abschnitt) und findet, daß bei $n_1 = n_2 = 5$ der t-Test weniger teststark als der U-Test ist, während sich diese Relation für $n_1 = n_2 = 15$ umkehrt. In diesem Fall trifft die im Abschnitt 2.6.2.2 zitierte Aussage von Bradley (1980a) zu. — Weitere Untersuchungen unter dieser Art von Verletzungen, insbesondere auch ein Vergleich mit anderen Tests, sind uns nicht bekannt.

2.7.2 Heteromere Verteilungen mit heterogenen Varianzen

Treten heteromere Verteilungen mit heterogenen Varianzen zusammen auf, sind zwei Voraussetzungen für den t-Test nicht erfüllt.

2.7.2.1 Robustheit

Havlicek & Peterson (1974) befassen sich mit der Robustheit des t-Tests unter der Kombination von Annahmeverletzungen „Heteromerität" und „Varianzheterogenität". Die simulierten Verteilungsformen sind die gleichen, die bereits im Abschnitt 2.7.1.1 ausführlicher dargestellt wurden, geändert werden lediglich die Varianzverhältnisse von $\sigma_1^2 : \sigma_2^2 = 16 : 16$ zu $16 : 64$ und $16 : 256$. Der einseitige t-Test erweist sich in der überwiegenden Mehrzahl der Fälle als nicht robust unter den angegebenen Bedingungen, und die Abweichungen $p_\alpha - \alpha$ sind in Abhängigkeit vom Ausmaß der Unterschiedlichkeit der beiden Populationen und vom Verhältnis der Stichprobengrößen zueinander und relativ zur Varianz der jeweiligen Population teilweise sehr beträchtlich (vgl. dazu auch Abschnitt 2.5).

Lediglich Neave & Granger (1968) haben u.W. die hier betrachtete Kombination von verletzten Annahmen in ihrer Wirkung auf den t-, den U- und die NS-Tests untersucht. Diese Autoren vergleichen die Tests bei $n_1 = n_2 = 20$ und bei $\alpha = 0{,}05$ unter zwei bimodalen, asymmetrischen Verteilungen mit $\sigma_1^2 : \sigma_2^2 = 1 : 2$ und finden in diesem Fall, in dem die Heteromerität aufgrund der vglw. geringen Varianzunterschiede entsteht, keine erwähnenswerten Diskrepanzen zwischen α und p_α.

Glass et al. (1972, 273) behaupten, daß sich die Wirkungen der beiden hier betrachteten Annahmeverletzungen auf den F-Test für Mittelwerte lediglich aufaddieren, daß also keine Interaktion (oder Wechselwirkung) zwischen den Faktoren „Nicht-Normalität" und „Varianzheterogenität" stattfindet. Diese Behauptung ist zunächst unter theoretischen Gesichtspunkten wenig plausibel, und darüber hinaus lassen sich auch empirische Gegenbeweise auffinden, etwa bei Havlicek & Peterson (1974) und bei Bradley (1980a) für den zweiseitigen t-Test, der — wie oben erwähnt

Tabelle 5

Empirische Fehlerwahrscheinlichkeiten für t-, W- und U-Test bei verschiedenen Varianzverhältnissen ($\sigma_1^2 : \sigma_2^2$) und Populationen ($Po_1 : Po_2$) bei gleichen Stichprobengrößen ($n_1 = n_2 = 20$) (aus Hübner et al., 1982). Angabe in Prozent; N = Normalverteilung; S = rechts-schiefe Verteilung; $\gamma_1 = 0{,}75$; unterstrichene Werte liegen außerhalb des Robustheits-Kriterium).

σ_1^2	σ_1^2	Po_1	Po_2		$\alpha = 5\%$			$\alpha = 1\%$			$\alpha = 0{,}1\%$		
					L	Z	R	L	Z	R	L	Z	R
				t:	5,00	5,21	5,05	1,11	1,05	1,01	0,14	0,17	0,15
1	4	N	N	U:	5,73	5,58	5,74	1,16	1,14	1,03	0,13	0,15	0,11
				W:	4,84	4,98	4,92	1,04	0,92	0,95	0,11	0,14	0,14
				t:	4,16	4,77	5,50	0,68	1,06	1,23	0,06	0,12	0,16
1	1	N	S	U:	2,94	5,23	7,82	0,39	1,03	1,62	0,04	0,11	0,17
				W:	4,14	4,70	5,48	0,67	1,04	1,22	0,06	0,11	0,16
				t:	3,93	5,51	6,43	0,68	1,25	1,71	0,04	0,20	0,33
1	4	N	S	U:	2,10	8,87	13,86	0,31	2,13	3,50	0,02	0,30	0,52
				W:	3,72	5,32	6,30	0,60	1,14	1,65	0,03	0,20	0,31
				t:	5,10	5,56	5,48	1,10	1,28	1,32	0,15	0,18	0,18
4	1	N	S	U:	5,30	6,23	6,77	1,09	1,32	1,53	0,13	0,15	0,14
				W:	5,00	5,29	5,32	1,03	1,14	1,23	0,14	0,14	0,15

— dem varianzanalytischen F-Test für zwei experimentelle Gruppen entspricht, und bei Hübner et al. (1982) für den t- und den U-Test (vgl. Tabelle 5). Greifen wir zur Erläuterung eines der Beispiele heraus!

Bei zwei Normalverteilungen ($\gamma_1 = 0$) bedingt ein Varianzverhältnis von $\sigma_1^2 : \sigma_2^2 = 1 : 4$ bei $n_1 = n_2 = 20$ beim rechtsseitigen U-Test eine Fehlerwahrscheinlichkeit $p_\alpha = 0{,}0574 = 0{,}05 + 0{,}0074$, während bei homogenen Varianzen unterschiedliche Schiefen ($\gamma_1 = 0$ und $\gamma_1 = +0{,}75$) zu $p_\alpha = 0{,}0782 = 0{,}05 + 0{,}0282$ führen. Treten beide Verletzungen dagegen simultan auf, steigt die Fehlerwahrscheinlichkeit auf $p_\alpha = 0{,}1386$ anstatt nur auf $p_\alpha = 0{,}05 + 0{,}0074 + 0{,}0282 = 0{,}0856$.

Auf der anderen Seite finden sich auch zahlreiche Fälle, in denen sich die Wirkungen von Annahmeverletzungen gegenseitig aufheben, wodurch die betr. Tests formal robust bleiben. Dies geschieht bspw. dann, wenn bei gleichen Stichprobenumfängen die größere Varianz mit einer symmetrischen und die kleinere mit einer rechtsschiefen Population verknüpft ist — vgl. dazu ebenfalls Tabelle 5. Diese Tabelle enthält ebenfalls den immer wieder auftretenden Befund, daß bei Vorliegen schiefer Verteilungen ausgeprägte Unterschiede bei der Beurteilung der Robustheit in Abhängigkeit vom benutzten Rejektionsbereich in den betr. Stichprobenverteilungen (links-, rechts- oder zweiseitige Tests) bestehen.

An dieser Stelle seien noch einige auf den ersten Blick überraschende Befunde kurz erläutert: In Abschnitt 2.7.1.1 war das Resultat als „interessant" bezeichnet worden, daß sich trotz ansteigender Stichprobenumfänge der U-Test als immer weniger robust herausstellte, während bei den parametrischen Testverfahren insbesondere eine Erhöhung der Anzahl von Beobachtungseinheiten stets zu erhöhter Robustheit der Tests geführt hatte. Dieses Phänomen zeigte sich auch in der Studie von Hübner et al. (1982) unter heteromeren Verteilungen mit ungleichen Varianzen.

Der festgestellte Mangel an Robustheit des U-Tests ist im Falle heteromerer Verteilungen darauf zurückzuführen, daß zwar die Nullhypothese gleicher Mittelwerte für den t-Test zutrifft, nicht jedoch die notwendige Oberhypothese bestimmter Verteilungsfunktionen. Für den U-Test gilt dagegen die H_1 (s.o.) über Unterschiede zwischen den Verteilungsfunktion. Der U-Test ist dabei hinsichtlich „seiner" Alternative teststärker als der t-Test in der gleichen Situation. Die Teststärke wächst jedoch unter sonst gleichen Bedingungen mit der Stichprobengröße. Allerdings ist der U-Test weniger *robust* als der t-Test im Hinblick auf die Klasse der speziellen Hypothesen über Mittelwerte (s.o.). Da hier das Verhalten der Tests unter *dieser* Klasse von Hypothesen betrachtet wird — dies bedeutet, daß ein signifikantes Resultat vom Forscher als „Mittelwertsunterschied" interpretiert wird —, ist es gerechtfertigt, die empirischen Ergebnisse im Sinne einer mangelnden Robustheit des U-Tests zu interpretieren.

2.7.2.2 Teststärke

Neave & Granger (1968) haben die Teststärke von t-, U- und den NS-Tests unter den im vorigen Abschnitt angegebenen Simulationsbedingungen untersucht und insgesamt recht uneinheitliche Ergebnisse erhalten: Bei $\alpha = 0{,}05$ und einem geringen Mittelwertsunterschied (von 0 auf 0,5 Einheiten bei $\sigma_1^2 = 1$ und $\sigma_2^2 = 2$) erwies sich der t-Test als teststärker als die NS- und der U-Test, während bei einem größeren Unterschied (0 auf 1,0) die NS-Tests stärker als der t-Test waren, wobei der U-Test erneut am relativ testschwächsten operierte. Bei $\alpha = 0{,}01$ und dem geringen Mittelwertsunterschied zeigt der U-Test eine höhere Teststärke als der t-Test und die NS-Tests, die sich nicht voneinander unterschieden, und bei dem größeren Mittelwertsunterschied war die entsprechende Reihenfolge NS, U und t. Insgesamt waren alle Unterschiede nur schwach ausgeprägt; sie stellen möglicherweise nur zufällige Diskrepanzen dar, da der Standardfehler $\sigma_{\hat{p}}$ in der Studie von Neave & Granger (1968) vglw. groß gewesen ist ($r = 500$).

Hübner et al. (1982) haben den t-, W- und den U-Test unter heteromeren Verteilungen (normal mit $\gamma_1 = 0$ und rechtsschief mit $\gamma_1 = 0{,}75$) mit verschiedenen Varianzverhältnissen ($\sigma_1^2 : \sigma_2^2 = 1:4; 4:1$) und unterschied-

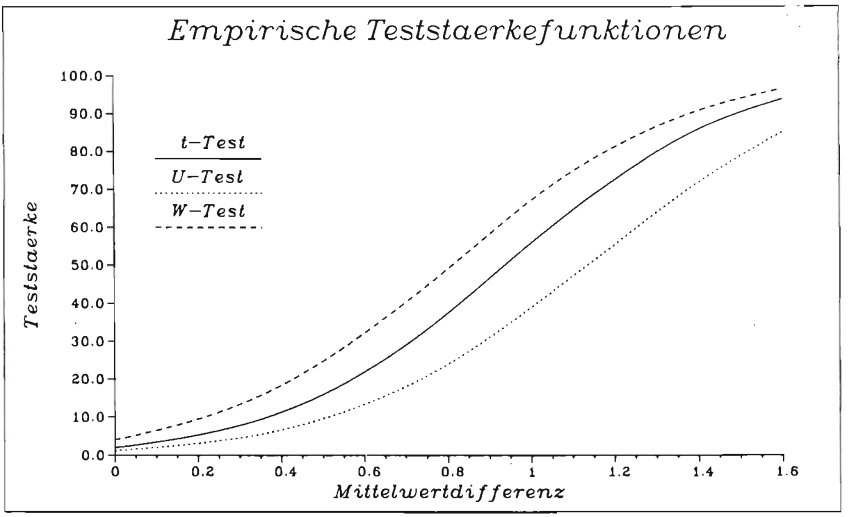

Abb. 2

Empirischer Teststärkenvergleich zwischen einseitigen t-, U- und W-Tests unter heteromeren Populationen bei $n_1 = 15$; $n_2 = 25$ und $\alpha = 0{,}05$; Kurven geglättet. Population 1: normal, $\mu_1 = 0$, $\sigma_1^2 = 1$, $\gamma_1 = 0$; Population 2: rechts-schief, $\mu_2 \geq 0$, $\sigma_2^2 = 4$, $\gamma_1 = 0{,}75$ (aus Hübner, Lübbeke und Hager, 1982).

lichen Stichprobenumfängen ($n_1 : n_2 = 15 : 25; 25 : 15$) bei $\alpha = 0{,}05$ verglichen und einige Teststärkekurven empirisch bestimmt; eine dieser Kurven ist als Abbildung 2 hier enthalten. Die Teststärke des W-Tests erwies sich als durchgängig höher als die der beiden übrigen Tests, von denen der U-Test unter einigen Bedingungen deutlich schwächer als die parametrischen Tests war (vgl. Abbildung 2).

Insgesamt liegen leider bislang nur sehr wenige Untersuchungen zum Verhalten der Tests bei heteromeren Verteilungen vor.

3.8 Zusammenfassung der Ergebnisse

Welche Empfehlungen lassen sich aus den zusammengefaßten Robustheitsuntersuchungen für den Praktiker ableiten? Zur Beantwortung dieser Frage unterscheiden wir zwei Ausgangssituationen:

1) Man kann von intervallskalierten Beobachtungsdaten ausgehen und ist an der Prüfung von Mittelwertshypothesen interessiert.

In diesem Fall sollten i.a. keine nonparametrischen Tests herangezogen werden, sondern eher die parametrischen, die sich über einen weiten Bereich von nicht-normalen Verteilungen und anderen nicht-erfüllten Voraussetzungen sowohl als hinreichend robust als auch als genügend teststark im Vergleich zu den nonparametrischen Konkurrenten erwiesen haben. Der U-Test zur Prüfung von Mittelwertshypothesen ist sehr anfällig bei Verletzung der Homomeritätsannahme; die Normal-Scores-Tests und der Randomisierungs-t-Test sind zu selten untersucht worden, um über ihre Leistungsfähigkeit hinreichend sicher urteilen zu können. Bei den parametrischen Tests kann zwischen dem t- und dem (approximativen) Welch-Test gewählt werden. Letzterer ist oft robuster als der t-Test und selten testschwächer. Im einzelnen kann bei gleichen Stichprobenumfängen der t-Test Verwendung finden, während bei ungleichen Stichprobengrößen besser der W-Test herangezogen werden sollte.

2) Es liegen lediglich ordinalskalierte Daten vor, oder man ist auf Intervallskalenniveau nicht an Lokations-, sondern anderen (etwa Varianz-) Hypothesen interessiert. In diesem Fall kommen die parametrischen Mittelwertstests kaum in Frage, weil sie i.a. auf Ordinalskalenniveau nicht zu empirisch sinnvoll interpretierbaren Resultaten führen und weil zur Prüfung anderer als Mittelwertshypothesen die für diese Klasse von Hypothesen geeigneteren Prüfverfahren heranzuziehen sind (etwa gegenüber Nicht-Normalität robuste Varianztests usw.). Zur Prüfung der auf Ordinalskalenniveau empirisch sinnvollen statistischen Hypothesen (Hager & Westermann, 1983b) über allgemeine Verteilungsunterschiede bieten sich die Normal-Scores-Tests an, wobei die alternative Verwendung des U-Tests (etwa wegen der leichteren Zugänglichkeit der tabellierten kritischen Wer-

te) allerdings nur in Ausnahmefällen zu nennenswert anderen faktischen Fehlerwahrscheinlichkeiten führen dürfte. Bei dieser Empfehlung ist jedoch zu berücksichtigen, daß die Normal-Scores- und der U-Test bislang nur recht selten zusammen untersucht worden sind. Auf die Frage der Ableitung der statistischen Hypothesen aus den inhaltlichen Forschungshypothesen sind wir an anderer Stelle eingegangen (Hager & Westermann, 1983a, b).

Eine nicht ganz unwesentliche Frage wurde bisher stets ausgeklammert: Soll man die Voraussetzungen statistisch überprüfen? Die Voraussetzungen sind häufig im Prinzip einer statistischen Überprüfung etwa mittels Tests für die Anpassungsgüte oder für Varianzhomogenität zugänglich. Derartige Überprüfungen dürften jedoch im Regelfall nicht indiziert sein, weil die zur Überprüfung zu benutzenden Tests ihrerseits wieder an Voraussetzungen gebunden sind, die meist ebenfalls einer Überprüfung bedürften (vgl. Gaensslen & Schubö, 1976; Hager & Westermann, 1983a). Eines der interessantesten Beispiele hierfür liefert der Bartlett-Test zur Überprüfung der Varianzhomogenität bei t- und F-Tests, der auch heute noch in kaum einem Lehrbuch *nicht* für diese Zwecke empfohlen wird. Dieser Test ist jedoch seinerseits — ebenso wie auch der F-Test zur Prüfung von *Varianzhypothesen* — extrem anfällig gegenüber der Verletzung der Normalverteilungsannahme (vgl. bereits Box, 1953; Scheffé, 1959; Bradley, 1980a; Hager & Westermann, 1983a), so daß man mit einiger Sicherheit insgesamt weniger Fehlentscheidungen bzgl. statistischer Hypothesen treffen dürfte, wenn man auf die Anwendung dieser und ähnlicher Tests im Zweifelsfalle eher verzichtet. Dessen ungeachtet erscheint es allerdings notwendig, bei jedem Datensatz rein deskriptiv sowohl die Daten pro Bedingung graphisch zu veranschaulichen als auch die Varianzen pro Bedingung anzugeben — diese Forderung vieler Lehrbücher wird in der psychologischen Fachliteratur immer noch zu wenig beachtet. Nur durch ihre Berücksichtigung wird es jedoch überhaupt erst ermöglicht werden können, im notwendigen Ausmaß zu ermitteln, welche Verteilungen und Verteilungskombinationen und welche Annahmeverletzungen in welchen Ausmaßen als realistisch und typisch für psychologische Fragestellungen angesehen werden können. Erst mit diesen Informationen können die vorliegenden theoretischen und empirischen Befunde zur Robustheit und Teststärke von statistischen Tests hinreichend differenziert beurteilt werden.

Summary

A very commonly encountered class of statistical tests concerns the two-sample location problem. The parametric as well as the nonparametric pro-

cedures are based on certain assumptions which are compared and discussed. The question relevant to the practioner is which test is most robust to violations of one or more assumptions under a wide range of psychologically relevant distributions or population situations. One way to answer this question is by means of Monte Carlo sampling studies. The results of various investigations dealing with t, Welch, U and Normal Scores tests are compared and evaluated. For testing parametric hypotheses about mean differences these results favor the more general use of the Welch instead of the t test. The theoretical powers of these tests are compared under normal theory conditions, showing only negligible advantages of the t test, which is less robust with respect to type I errors under heterogeneous variances than the Welch test. The differences found between the three nonparametric procedures for testing the more general location problem are small in general, without clearly favoring any particular test.

Résumé

Le psychologue chercheur a souvent recours aux tests statistiques qui permettent de comparer les résultats de deux échantillons. Ces tests, qu'ils soient paramétriques ou non, sont liés à certains présupposés qui sont rappelés et discutés ici. La question d'un test qui resterait valide même lorsqu'un ou plusieurs de ces présupposés sont violés, donc la question de sa »robustesse«, fait l'objet d'une analyse comparative empirique et théorique entre les tests t, U, celui de Welch, celui de Terry-Hoeffding et celui de van der Waerden. Les résultats conduisent les auteurs à recommander l'utilisation du test de Welch, de préférence au t, même lorsqu'il s'agit de comparer des données paramétriques, parce que le test de Welch est pratiquement aussi puissant que le t, lorsque les distributions sont normales et homogènes, mais il est plus robuste lorsque les variances sont hétérogènes.

Literatur

Aspin, A. A.: An examination and further development of a formula occurring in the problem of comparing two mean values. Biometrika **35** (1948), 88—96.

Aspin, A. A.: Tables for use in comparisons whose accuracy involves two variances separately estimated. Biometrika **36** (1948), 290—296.

Baker, B. O., Hardyck, C. D. und Petrinovich, L. F.: Weak measurements vs. strong statistics: An empirical critique to S. S. Stevens' proscription on statistics. Educational and Psychological Measurement **26** (1966), 291—309.

Bauknecht, K., Kohlas, J. und Zehnder, C. A.: Simulationstechnik. Berlin: Springer, 1976.

Bell, C. B. und Doksum, K. A.: Some new distribution-free statistics. Annals of Mathematical Statistics **36** (1965), 203—214.

Bennett, B. U. und Hsu, P. L.: Sampling studies on the Behrens-Fisher problem. Metrika **4** (1961), 89—104.

Bevan, M. F., Denton, J. Q. und Myers, J. L.: The robustness of the F-Test to violations of continuity and form of treatment population. British Journal of Mathematical and Statistical Psychology **27** (1974), 199—204.

Bhattacharjee, G. P.: Non-normality and heterogeneity in two-sample t-test. Annals of the Institute of Statistical Mathematics **20** (1968), 239—254.

Birnbaum, Z. W. und Klose, O. M.: Bounds for the variance of the Mann-Whitney statistic. Annals of Mathematical Statistics **28** (1957), 933—945.

Blair, R. C. und Higgins, J. J.: A comparison of the power of Wilcoxon's rank sum statistic to that of Student's t statistic under various nonnormal distributions. Journal of Educational Statistics **5** (1980a), 309—335.

Blair, R. C. und Higgins, J. J.: A comparison of the power of the t and Wilcoxon statistics when samples are drawn from a certain mixed normal distribution. Evaluation Review **4** (1980b), 645—656.

Blair, R. C., Higgins, J. J. und Smitley, W. D. S.: On the relative power of the U and t tests. British Journal of Mathematical and Statistical Psychology **33** (1980), 114—120.

Boneau, C. A.: The effects of violations of assumptions underlying the t-test. Psychological Bulletin **57** (1960), 49—64.

Boneau, C. A.: A comparison of the power of the U and t tests. Psychological Review **69** (1962), 246—256.

Bortz, J.: Lehrbuch der Statistik für Sozialwissenschaftler. Nachdruck der 1. Aufl. Berlin: Springer, 1979.

Bowman, K. O., Beauchamp, J. J. und Shenton, L. R.: The distribution of the t-statistic under non-normality. International Statistical Review **45** (1977), 233—242, 256.

Box, G. E. P.: Non-normality and tests on variances. Biometrika **40** (1953), 318—335.

Box, G. E. P.: Some theorems on quadratic forms applied in the study of analysis of variance problems. Pt. I: Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics **25** (1954a), 290—302.

Box, G. E. P.: Some theorems on quadratic forms applied in the study of analysis of variance problems. Pt. 2: Effect of inequality of variance and correlation of errors in the two-way classification. Annals of Mathematical Statistics **25** (1954b), 284—298.

Box, G. E. P. und Andersen, S. L.: Permutation theory in the derivation of robust criteria and the study of departures from assumptions. Journal of the Royal Statistical Society, Series B, **17** (1955), 1—34.

Bradley, J. V.: Distribution-free statistical tests. Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Bradley, J. V.: A common situation conducive to bizarre distribution shapes. American Statistician **31** (1977), 147—150.

Bradley, J. V.: Robustness? British Journal of Mathematical and Statistical Psychology **31** (1978), 144—152.

Bradley, J. V.: Nonrobustness in classical tests on means and variances: A large-scale sampling study. Bulletin of the Psychonomic Society **15** (1980a), 275—278.

Bradley, J. V.: Nonrobustness in z, t, and F tests at large sample sizes. Bulletin of the Psychonomic Society **16** (1980b), 333—336.

Bradley, J. V.: Nonrobustness in one-sample z and t tests: A large-scale sampling study. Bulletin of the Psychonomic Society **15** (1980c), 29—32.

Bradley, R. A.: Corrections for non-normality in the use of the two-sample t- and F-tests at high significance levels. Annals of Mathematical Statistics **23** (1952), 103—113.

Büning, H. und Trenkler, G.: Nichtparametrische statistische Methoden. Berlin: de Gruyter, 1978.

Cochran, W. G.: Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics **3** (1947), 22—38.

Cochran, W. G.: The comparison of percentages in matched samples. Biometrika **37** (1950), 256—266.

Cohen, J.: Statistical power analysis for the behavioral sciences. 2. Aufl. New York: Academic Press, 1977.

Conover, W. J., Wehmanen, O. und Ramsey, F. L.: A note on the small-sample power function for nonparametric tests of location in the double exponential family. Journal of the American Statistical Association **73** (1978), 188—190.

D'Agostino, R. B.: A second look at analysis of variance on dichotomous data. Journal of Educational Measurement, **8** (1971), 327—333.

D'Agostino, R. B.: Relation between the chi-squared and Anova tests for testing the equality of k independent dichotomous populations. American Statistician **26** (3) (1972), 30—32.

Donaldson, T. S.: Robustness of F-test to errors of both kinds and the correlation between numerator and denominator of the F-ratio. Journal of the American Statistical Association **63** (1968), 660—676.

Edgington, E. S.: The assumption of homogeneity of variance for the t test and nonparametric tests. Journal of Psychology **59** (1965), 177—179.

Fisher, R. A. und Yates, F.: Statistical tables for biological, agricultural, and medical research. Edinburgh: Oliver & Boyd, 1938.

Fisz, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik. 5. Aufl. Berlin: Deutscher Verlag der Wissenschaften, 1970.

Fligner, M. A. und Pollicello, G. E.: Robust rank procedures for the Behrens-Fisher problem. Journal of the American Statistical Association **76** (1981), 162—168.

Geary, R. C.: Testing for normality. Biometrika **34** (1947), 209—242.

Glass, G. V., Peckham, P. D. und Sanders, J. R.: Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. Review of Educational Research **42** (1972), 237—288.

Golhar, M. B.: The errors of first and second kinds in Welch-Aspin's solution of the Behrens-Fisher problem. Journal of Statistical Computation and Simulation **1** (1972), 209—224.

Gronow, D. G. C.: Test for significance of the difference between means in two normal populations having unequal variances. Biometrika **38** (1951), 252—256.

Gutjahr, W.: Die Messung psychischer Eigenschaften. Berlin: Deutscher Verlag der Wissenschaften, 1971.

Hager, W. und Westermann, R.: Planung und Auswertung von Experimenten. In Bredenkamp, J. und Feger, H. (Hrsg.): Hypothesenprüfung (= Enzyklopädie der Psychologie. Themenbereich B. Serie I. Band 5). Göttingen: Hogrefe, 1983a, 24—238.

Hager, W. und Westermann, R.: Zur Wahl und Prüfung statistischer Hypothesen in psychologischen Untersuchungen. Zeitschrift für experimentelle und angewandte Psychologie, **30** (1983b), 67—94.

Havlicek, L. L. und Peterson, Nancy: Robustness of the t test: A guide for researchers on effect of violation of assumptions. Psychological Reports **34** (1974), 1095—1114.

Hays, W. L.: Statistics for the social sciences. 2. Aufl. London: Holt, Rinehart & Winston, 1977.

Hemelrijk, J.: Experimental comparison of Student's and Wilcoxon's two sample tests. In: Jonge, H. de (Hrsg.): Quantitative methods in pharmacology. Amsterdam: North-Holland, 1961, 118—134.

Marascuilo, L. A. und McSweeney, Maryellen: Nonparametric and distribution-free methods for the social sciences. Monterey, Cal.: Brooks/Cole, 1977.

Mehta, J. S. und Srinivasan, R.: On the Behrens-Fisher problem. Biometrika **57** (1970), 649—655.

Menges, G.: Grundriß der Statistik. Teil I: Theorie. Opladen: Westdeutscher Verlag, 1968.

Mood, A. M., Graybill, F. und Boes, D. C.: Introduction to the theory of statistics. 3. Aufl. Tokyo: McGraw-Hill, 1974.

Neave, H. R. und Granger, C. W. J.: A Monte Carlo study comparing various two-sample tests for differences in mean. Technometrics **10** (1968), 509—522.

Pearson, E. S. und Please, N. W.: Relations between the shape of population distribution and the robustness of four simple test statistics. Biometrika **62** (1975), 223—241.

Pfanzagl, J.: Allgemeine Methodenlehre der Statistik. 5. Aufl. Berlin: de Gruyter, 1978.

Posten, H. O.: The robustness of the two-sample t-test over the Pearson system. Journal of Statistical Computation and Simulation **6** (1978), 295—311.

Potthoff, R. F.: Use of the Wilcoxon statistic for a generalized Behrens-Fisher problem. Anuals of Mathematical Statistics **34** (1963), 1596—1599.

Pratt, J. W.: Robustness of some procedures for the two-sample location problem. Journal of the American Statistical Association **59** (1964), 665—680.

Ramsey, F. L.: Small sample power functions for nonparametric tests of location in the double exponential family. Journal of the American Statistical Association **66** (1971), 149—151.

Ramsey, P. H.: Exact type I error rates for robustness of Student's t test with unequal variances. Journal of Educational Statistics **5** (1980), 337—349.

Rouanet, H. und Lépine, D.: Problèmes de méthodologie statistique. II. Etude d'un conflit robustesse-efficacité dans le problème de la comparaison de deux moyennes (groupes indépendants). Mathematiques et Sciences Humaines **12** (1974), 61—71.

Sachs, L.: Statistische Auswertungsmethoden. Berlin: Springer, 1968.

Sawrey, W. L.: A distinction between exact and approximate non-parametric methods. Psychometrika 23 (1958), 171—177.

Scheffé, H.: The analysis of variance. New York: Wiley, 1959.

Scheffé, H.: Practical solutions of the Behrens-Fisher problem. Journal of the American Statistical Association **65** (1970), 1501—1508.

Seeger, P. und Gabrielsson, A.: Applicability of the Cochran Q test and the F test for statistical analysis of dichotomous data for dependent samples. Psychological Bulletin **69** (1968), 269—277.

Serfling, R. J.: The Wilcoxon two-sample statistic on strongly mixing processes. Annals of Mathematical Statistics **39** (1968), 1202—1209.

Siegel, S.: Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Smith, J. E. K.: Analysis of qualitative data. Annual Review of Psychology **27** (1976), 487—499.

Stegmüller, W.: „Jenseits von Popper und Carnap": Die logischen Grundlagen des statistischen Schließens. (= Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band IV: Personelle und Statistische Wahrscheinlichkeit. Studienausgabe, Teil D). Heidelberg: Springer, 1973.

Student (= W. S. Gosset): The probable error of the mean. Biometrika **6** (1908), 1—25.

Suppes, P. und Zinnes, J. L.: Basic measurement theory. In: Luce, R. D., Bush, R. R. und Galanter, E. (Hrsg.): Handbook of mathematical psychology. Vol. I. New York, 1963, 1—76.

Terry, M. E.: Some rank order tests which are most powerful against specific parametric alternatives. Annals of Mathematical Statistics **23** (1952), 346—366.

Toothaker, L. E.: An empirical investigation of the permutation t test. British Journal of Mathematical and Statistical Psychology **25** (1972), 83—94.

Trachtman, J. N., Giambalvo, V. und Dippner, R. S.: On the assumptions concerning the assumptions of a t test. Journal of General Psychology **99** (1978), 107—116.

Trommer, R.: Untersuchungen zur Robustheit des Wilcoxon-Tests gegenüber Streuungsungleichheit. Biometrische Zeitschrift **9** (1967), 14—21.

Van der Laan, P. und Oosterhoff, J.: Monte Carlo estimation of the powers of the distribution-free two-sample tests of Wilcoxon, van der Waerden and Terry and comparison of these powers. Statistica Neerlandica **19** (1965), 265—275.

Van der Laan, P. und Oosterhoff, J.: Experimental determination of the power functions of the two-sample rank tests of Wilcoxon, van der Waerden and Terry by Monte Carlo techniques. I. Normal parent distributions. Statistica Neerlandica **21** (1967), 55—68.

Van der Laan, P. und Weima, J.: Experimental comparison of the powers of the two-sample tests of Wilcoxon and Student under logistic parent distributions. Journal of Statistical Computation and Simulation **8** (1978), 133—144.

Van der Laan, P. und Weima, J.: Asymptotic power of the two-sample test of Wilcoxon for logistic shift alternatives, and comparison with simulation results. Statistica Neerlandica **34** (1980), 117—121.

Van der Vaart, H. R.: On the robustness of Wilcoxon's two-sample test. In: Jonge, H. de (Hrsg.): Quantitative methods in pharmacology. Amsterdam: North-Holland, 1961, 140—158.

Van der Waerden, B. L.: Ein neuer Test für das Problem der zwei Stichproben. Mathematische Annalen **26** (1953), 93—107.

Vorberg, D.: Eine Entscheidungshilfe für die Auswahl statistischer Tests und Maße. Psychologische Rundschau **31** (1981), 267—277.

Wang, Y. Y.: Probabilities of the type I errors of the Welch test for the Behrens-Fisher problem. Journal of the American Statistical Association **66** (1971), 605—608.

Welch, B. L.: The significance of a difference between two means when the population variances are unequal. Biometrika **29** (1937), 350—362.

Welch, B. L.: The generalization of ‚Student's' problem when several populations are involved. Biometrika **34** (1947), 28—35.

Welch, B. L.: Further notes on Mrs. Aspin's tables and on certain approximations to the tabled function. Biometrika **36** (1949), 293—296.

Westermann, R.: Die empirische Überprüfung des Niveaus psychologischer Skalen. Zeitschrift für Psychologie **188** (1980), 450—468.

Westermann, R.: Zur Messung von Einstellungen auf Intervallskalenniveau. Zeitschrift für Sozialpsychologie **13** (1982), 97—108.

Wetherill, B. G.: The Wilcoxon test and non-null hypotheses. Journal of the Royal Statistical Society, Series B, **22** (1960), 402—418.

Wilcoxon, F.: Individual comparison by ranking methods. Biometrics Bulletin **1** (1945), 80—83.

Woodward, J. A. und Overall, J. E.: The significance of treatment effects in ordered category data. Journal of Psychiatric Research **13** (1977), 169—177.

Young, R. K. und Veldman, D. J.: Heterogeneity and skewness in analysis of variance. Perceptual and Motor Skills **16** (1963), 588.

Anschrift der Verfasser: Dr. Willi Hager, Bernhard Lübbeke und Ronald Hübner, Institut für Psychologie der Universität Göttingen, Goßlerstraße 12, 3400 Göttingen.