Zeitschrift für experimentelle und angewandte Psychologie
1984, Band XXXI, Heft 2, S. 214—231

Sind nonparametrische Tests parametrischen bei „beliebigen Verteilungen" vorzuziehen?

Empirische Untersuchungen zu einigen Flußdiagrammen, Entscheidungshilfen und Empfehlungen

Ronald Hübner und Willi Hager

Institut für Psychologie der Universität Göttingen

Es werden einige Empfehlungen aus der einschlägigen Literatur bzgl. der Wahl zwischen parametrischen und nonparametrischen Tests zur Prüfung von speziellen Lokationshypothesen mittels Computer-Simulationen an Zwei-Stichproben-Tests überprüft. Es erweist sich, daß einige Empfehlungen zu unpräzise sind, um für den Praktiker eine Hilfe darstellen zu können.

1. Einleitung und Fragestellung

Der in der Forschung tätige Psychologe sieht sich häufig einer Situation konfrontiert, die man wie folgt rekonstruieren kann (vgl. Hager & Westermann, 1983b): Er befaßt sich mit einer „theoretisch-psychologischen Hypothese", die sich auf theoretische Variablen bezieht und aus der für eine gewählte spezifische Operationalisierung (empirische Variable) Vorhersagen über (gerichtete oder ungerichtete Unterschiede) in zwei (experimentellen oder nicht-experimentellen) Treatmentgruppen ableitbar sind. Diese Vorhersagen („empirisch-psychologische Hypothesen") will der Forscher über den (i.a. notwendigen) „Umweg" statistischer Hypothesen, die der empirisch-psychologischen Hypothese „zugeordnet" bzw. aus ihr „abgeleitet" werden, prüfen (Bredenkamp, 1980; Westermann & Hager, 1982; Hager & Westermann, 1983a).

Die Festlegung der gegeneinander zu testenden statistischen Hypothesen trifft der Forscher allerdings nur in den seltensten Fällen durch eine „Zuordnung" oder „Ableitung", sondern fast stets durch die Wahl eines der gebräuchlichsten Standardtestverfahren. Im hier betrachteten Fall nur zweier Treatmentgruppen wird typischerweise in erster Linie der t-Test in Erwägung gezogen.

Mit diesem Test wird die statistische Nullhypothese (H_0) getestet, daß die Erwartungswerte μ_1 und μ_2 zweier normalverteilter und unabhängiger Zufallsvariablen mit homogenen Varianzen gleich sind (H_0: $\mu_1 = \mu_2$), und zwar gegen die Alternative (H_1), daß die Erwartungswerte ungleich sind (H_1: $\mu_1 \neq \mu_2$). Die beiden Hypothesen können bei entsprechend präzisen Vorhersagen auch gerichtet formuliert werden, bspw.: H_0: $\mu_1 \leq \mu_2$ und H_1: $\mu_1 > \mu_2$.

Hypothesen über die Erwartungswerte von normalverteilten Zufallsvariablen sind empirisch nur sinnvoll, wenn die Daten (mindestens) intervallskaliert sind. Dieser Aspekt gehört jedoch nicht zu den mathematischen Voraussetzungen des t-Tests; vielmehr muß die Frage des Skalenniveaus vor der Zuordnung von statistischer zu empirisch-psychologischer Hypothese unter Rückgriff auf inhaltliche Analysen und Überlegungen sowie entsprechende empirische Prüfungen abgeklärt werden (vgl. dazu Westermann, 1980, 1982; s.a. Hager & Westermann, 1983a). Das Ergebnis dieser Überlegungen und Prüfungen kann dann sein, daß das geforderte Skalenniveau nicht vorliegt oder nicht zu erwarten ist. In diesem Fall kann die empirisch-psychologische Hypothese über die mit dem t-Test getesteten statistischen Hypothesen nicht adäquat geprüft werden (a.a.O.). Unsere folgenden Erörterungen gehen davon aus, daß mindestens Intervallskalenniveau gegeben bzw. zu erwarten ist.

Unter dieser Voraussetzung hat sich der Forscher mit der Frage zu befassen, inwieweit die Gesamtheit oder Population der Daten pro Treatment durch das Normalverteilungsmodell adäquat repräsentiert werden kann. Typischerweise liegt eine Information über die Verteilungsfunktion der die Daten generierenden Zufallsvariablen in den Populationen nicht vor. Dem Forscher stehen angesichts dieses Informationsmangels im wesentlichen zwei alternative Verfahrensweisen zur Verfügung: Er testet die Normalverteilungsvoraussetzung statistisch (e.g. Wottawa, 1982), oder er stellt sich auf den Standpunkt, daß die entsprechenden Tests ihrerseits an bestimmte und zu überprüfende Voraussetzungen gebunden sind, so daß durch sie das zu lösende Problem nur auf eine andere Ebene verlagert wird (Gaensslen & Schubö, 1976, 59; Hager & Westermann, 1983a). Wir betrachten im folgenden nur die zweite Alternative und gehen davon aus, daß dem Forscher die Berechtigung der Normalverteilungsannahme zweifelhaft erscheint und daß er ausschließlich an der Testung gegen *spezielle* Lokationsalternativen interessiert ist, nämlich unter der H_1 annimmt, daß die Erwartungswerte der beiden Verteilungen gegeneinander verschoben sind. Bei dieser Art von Alternativen wird, ebenso wie im folgenden, davon ausgegangen, daß die betrachteten Verteilungen einen Erwartungswert besitzen, und die gleiche Annahme wird i.a. bzgl. der Varianz getroffen.

Durch diese wichtige Einschränkung werden allgemeinere Lokationsal-

ternativen, die sich bspw. auf Medianunterschiede beziehen, oder allgemeine Alternativen, die globale Verteilungsunterschiede betreffen, aus der unmittelbaren Betrachtung ausgeblendet. Dies erscheint uns aus Gründen der Übersichtlichkeit notwendig (vgl. dazu Hager, Lübbeke & Hübner, 1983) und aufgrund der Häufigkeit, mit der in der Literatur explizit oder implizit gegen derartige Hypothesen getestet wird (Edgington, 1974; Hager & Westermann, 1982), auch gerechtfertigt.

Welche Empfehlungen findet der Forscher für diesen Fall in den einschlägigen Lehrbüchern und den entsprechenden, leicht zugänglichen Veröffentlichungen?

Vorberg (1981, 272) empfiehlt für den Fall, daß die Verteilungen nicht normal, sondern „beliebig" sind, die Verwendung von nonparametrischen Tests (NPT) wie bspw. dem U-Test (s.u.) zur Testung gegen Alternativen bzgl. „Mittelwerte bzw. zentrale Tendenz" (a.a.O.; vgl. dazu auch Sachs, 1968, 131—132; Kreyszig, 1968, 337; Schüle, 1976, 311; Büning & Trenkler, 1978, 150; Bortz, 1979, 192). Vorberg (a.a.O.) läßt unerläutert, was er unter „beliebigen Verteilungen" versteht. Um Erwartungs- oder „Mittelwerts-"Hypothesen mit einem der traditionellen NPT testen zu können, muß vorausgesetzt werden, daß die Populationsverteilungen „homomer" sind, d.h. daß die Verteilungen unabhängig von der Form bis auf eine mögliche Lageverschiebung gleich sind (Lienert, 1973, 107), und daß die Form der Verteilungen unter der Gültigkeit der H_1 gleich bleibt. Nur unter diesen Bedingungen wird eine Verschiebung des Lokationsparameters „Median" von einer gleichsinnigen Verschiebung des Lokationsparameters „Erwartungswert" (E(X)) begleitet. Doch liegt dem Forscher i.a. eine derartig detaillierte Information ebenfalls nicht vor.

Zur Begründung der häufig anzutreffenden Empfehlung, bei Nonnormalität anstelle eines parametrischen Tests (PT) einen NPT zu verwenden, kann auf zahlreiche Simulationsuntersuchungen verwiesen werden, mit denen gezeigt werden konnte, wie „empfindlich" („unrobust") PT wie bspw. der t-Test auf (ausgeprägte) Verletzungen der Verteilungs- und anderer Annahmen reagieren können (vgl. vor allem Bradley, 1968, 1980a, b; Wike & Church, 1982a, b) und wie groß die Teststärkevorteile eines NPT gegenüber seinem parametrischen Homolog sein können (u.a. Blair, Higgins & Smitley, 1980; Blair & Higgins, 1980, 1981). Derartige Studien, bei denen häufig nicht verschiedene Verfahren unter denselben Verletzungen miteinander verglichen werden (bspw. Bradley, a.a.O.), sind jedoch dann von zweifelhaftem Wert als Entscheidungshilfe zwischen parametrischen und nonparametrischen Tests, wenn kein Vergleich zwischen NPT und PT unter denselben Bedingungen vorgenommen wird (bspw. Bradley, a.a.O.) und/oder wenn ausschließlich *sehr* extreme Annahmeverletzungen simuliert werden, die kaum auf Situationen übertragen

werden können, in denen weniger ausgeprägte Verletzungen vorliegen (vgl. dazu Hager et al., 1983).

Aus diesen Gründen kann es nicht verwundern, daß zahlreiche andere Autoren (meist unter weniger extremen Bedingungen) gefunden haben, daß die PT vglw. „robust" gegenüber (moderaten) Abweichungen von den Voraussetzungen sind (e.g. Glass, Peckham & Sanders, 1972; Havlicek & Peterson, 1974; Hager et al. 1983). Unter Berufung auf diese Robustheit empfehlen die genannten und zahlreiche weitere Autoren die Anwendung der PT auch bei berechtigtem Zweifel, ob normalverteilte Daten mit homogenen Varianzen usw. vorliegen, da die tatsächlichen Wahrscheinlichkeiten für Fehler 1. und 2. Art, p_α und p_β, unter moderaten Verletzungen der verschiedenen Voraussetzungen nur geringfügig von den nominellen Werten α und β abweichen. Implizit ist in dieser Empfehlung die Annahme enthalten, daß in der psychologischen Forschung eher mit moderaten als mit extremen Abweichungen von den für die valide Anwendung der PT notwendigen Voraussetzungen zu rechnen ist — siehe dazu Hager et al. (1983).

Die vorliegende Arbeit hat sich zum Ziel gesetzt, die verschiedenen Empfehlungen unter Verwendung der Technik der „Computer-Simulation" empirisch derart zu prüfen, daß verschiedene in Frage kommende parametrische und nonparametrische Tests unter verschiedenen Verletzungen von Voraussetzungen bei der Testung gegen die speziellen Lokationsalternativen „Unterschiede zwischen den Erwartungswerten" miteinander verglichen werden. Für diesen Vergleich erübrigt sich dann die Zugrundelegung eines Robustheitskriteriums, wie wir es in unserer Übersichtsarbeit verwendet haben (Hager, Lübbeke & Hübner, 1983). — Gleichzeitig sollen mit den darzustellenden Daten einige der in der genannten Übersicht offengebliebenen Fragen beantwortet werden.

2. Betrachtete Testverfahren

2.1 t-Test nach Gosset („Student", 1908)

Es seien X_{ik}, $i = 1, 2, \ldots, n_k$, $k = 1, 2$, unabhängige und normalverteilte Zufallsvariablen. Unter der Annahme homogener Varianzen ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) kann mit der Teststatistik t

$$t = \frac{\bar{X}_1 - \bar{X}_2 - E(\bar{X}_1 - \bar{X}_2)}{\sqrt{\hat{\sigma}^2}} \left(\frac{n_1 n_2}{n_1 + n_2} \right)^{1/2} \qquad (1)$$

die Hypothese der Gleichheit der Erwartungswerte μ_1 und μ_2 der beiden Normalverteilungen getestet werden:

$$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2 \quad (2)$$

(bzw. die entsprechenden gerichteten Hypothesen) (vgl. Hays, 1981). $\hat{\sigma}^2$ bezeichnet die Schätzung der beiden Populationen gemeinsamen Varianz und n_1 und n_2 die Stichprobenumfänge. $\bar{X}_1$ und $\bar{X}_2$ sind die empirischen Mittelwerte; der Erwartungswert der Mittelwertsdifferenz, $E(\bar{X}_1—\bar{X}_2)$, ist gleich der Differenz der Erwartungswerte unter Gültigkeit der H_0. Die Freiheitsgrade f_t des t-Tests betragen $f_t = n_1 + n_2 — 2$.

2.2 v-Test nach Welch (1947) und Aspin (1948)

Sollen die gleichen Hypothesen wie unter (2) angegeben ohne die Annahme gleicher Populationsvarianzen getestet werden, kann hierfür der t-Test nicht valide angewendet werden. Man nennt die dargestellte Testsituation das „Behrens-Fisher-Problem", für dessen Lösung in der Literatur zahlreiche Tests vorgeschlagen werden (vgl. die Übersichten von Mehta & Srinivasan, 1970, und Scheffé, 1970). Wir betrachten hier nur die am weitesten verbreiteten Lösungen, die auf die grundlegende Arbeit von Welch (1937) zurückgeführt werden können.

Welch (1937) schlägt als Alternative zum z-Test den bei ungleichen Varianzen und Stichprobenumfängen valideren V-Test vor, für den ebenso wie für den z-Test Bekanntheit der Populationsvarianzen vorausgesetzt werden muß und dessen Teststatistik

$$V = \frac{(\bar{X}_1—\bar{X}_2)—E(\bar{X}_1—\bar{X}_2)}{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]^{1/2}}, \quad (3)$$

mit f_V Freiheitsgraden t-verteilt ist:

$$f_V = \frac{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]^2}{\frac{\sigma_1^4}{n_1^2(n_1—1)} + \frac{\sigma_2^4}{n_2^2(n_1—1)}} \quad (4)$$

In seiner Arbeit von 1947 ersetzt Welch die Populationsvarianzen durch ihre Schätzungen $\hat{\sigma}^2$, und mit der Verteilung der auf diese Weise entstehenden Teststatistik v befaßt sich die Arbeit von Aspin (1948), die 1949 auch Tabellen für die Verteilung von v vorlegt (vgl. auch Pearson & Hartley, 1962, 136—137); zum analytischen Vergleich des Welch-Tests mit alternativen Verfahren siehe insbesondere Pfanzagl (1974).

2.3 Approximativer Welch-Test (v_t) nach Welch (1947, 1949)

Um die Benutzung des in Abschnitt 2.2 dargestellten Welch-Aspin-Tests zu erleichtern, hat Welch (1947, 32) eine Version dieses Tests entwickelt, bei der die Beurteilung der Teststatistik v approximativ über die t-Verteilung erfolgen kann, wozu die Freiheitsgrade f_{vtH} wie folgt zu bestimmen sind (vgl. Hays, 1981, 287):

$$f_{vtH} = \frac{\left[\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_1^2}{n_2}\right]^2}{\frac{\hat{\sigma}_1^4}{n_1^2(n_1+1)} + \frac{\hat{\sigma}_2^4}{n_2^2(n_2+1)}} - 2 \tag{5}$$

Unter Benutzung einer anderen Approximation gelangt Welch (1949) zu einer geringfügig anderen Version des Tests, für die die Freiheitsgrade f_{vt} wie folgt zu bestimmen sind:

$$f_{vt} = [c^2/(n_1-1) + (1-c)^2/(n_2-1)]^{-1} \tag{6}$$

mit

$$c = \frac{\frac{\hat{\sigma}_1^2}{n_1}}{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \tag{7}$$

Das vorstehend genannte Verfahren zur Bestimmung der Freiheitsgrade ist nicht schlechter als das unter (5) dargestellte (Welch, 1949, 295) und ist am weitaus häufigsten in den einschlägigen Lehrbüchern zu finden (vgl. auch Satterthwaite, 1946); wir legen es unseren Simulationsuntersuchungen zusätzlich zu der exakten Version des Welch-Aspin-Tests (vgl. oben) zugrunde.

Das Problem, daß die Freiheitsgrade meist nicht ganzzahlig und damit nicht tabellarisch verfügbar sind, kann durch die Anwendung einer von Wang (1971, 606) abgeleiteten Formel umgangen werden.

Die vorstehend besprochenen Tests zählen zu den parametrischen Verfahren, während die folgenden zu den nonparametrischen gehören.

2.4 U-Test nach Wilcoxon (1945) und Mann & Whitney (1947)

Es werden nicht mehr die Rohwerte selbst verwendet, sondern ihnen zugeordnete Ränge. Bezeichnet man mit T_1 und T_2 die Summe dieser Ränge in den beiden Stichproben, ergibt sich der Wert der Teststatistik U wie folgt:

$$U = \text{Minimum} \left[\begin{array}{l} n_1 \cdot n_2 + \dfrac{n_1(n_1 + 1)}{2} - T_1 \\ n_1 \cdot n_2 + \dfrac{n_2(n_2 + 1)}{2} - T_2 \end{array} \right] \quad (8)$$

Ohne Verteilungsannahmen werden mit dem U-Test die folgenden Hypothesen gegeneinander getestet:

$$H_0: P(X_1 > X_2) = P(X_2 > X_1); H_1: P(X_1 > X_2) < P(X_2 > X_1). \quad (9)$$

Um Lokationshypothesen testen zu können, sind einschränkende Annahmen erforderlich. Diese können bspw. sein (vgl. Hollander & Wolfe, 1973, 67):

a) $X_{1i} = e_i$, $i = 1, \ldots, n_1$
$X_{2j} = e_{n_1+1} + \Delta$; $j = 1, \ldots, n_2$

Die X_{1i} und die X_{2j} sind beobachtbare und die e_{n_1+1} bis $e_{n_1+n_2}$ sind nichtbeobachtbare Zufallsvariablen, und Δ ist ein unbekannter Lokationsparameter.

b) Die n_1 und n_2 Werte e sind unabhängig voneinander.

c) Jedes e stammt aus der gleichen kontinuierlichen Population.

Die Lokationshypothese lautet dann:

$$H_0: \Delta = 0; H_1: \Delta \neq 0. \quad (10)$$

Aus den Annahmen folgt dann, daß der einzige Unterschied zwischen der Population der X_1- und der der X_2-Werte ein Unterschied der Lokation ist, wobei Δ als die Differenz der beiden Mediane definiert ist: $\Delta = Mdn_2 - Mdn_1$ (Hollander & Wolfe, 1973, 70). „If we further assume that the mean of the *e* population exists then $\Delta = m_2 - m_1$ where m_2 is the mean of the ... (X_2) population and m_1 is the mean of the ... (X_1) population." (Hollander & Wolfe, 1973, 70—71). — Andere Modelle werden bspw. bei Lehmann (1975) diskutiert.

Im Falle symmetrischer Verteilungen mit heterogenen Varianzen gilt zwar auch die unter (8) angegebene allgemeine Nullhypothese, aber die Schätzung der Varianz weicht von der Schätzung unter gleichen Varianzen ab. Dies hat zur Folge, daß die tabellierten kritischen Werte nicht mehr zur validen Signifikanzbeurteilung herangezogen werden können (Pratt, 1964; Hilgers, 1982). Eine von Potthoff (1963) vorgeschlagene Modifikation des U-Tests ersetzt den üblichen Varianzschätzer durch einen Maximalwert; dies führt zu einem konservativen Test. Einen anderen Weg haben kürzlich Fligner & Policello (1981) gewählt.

2.5 Modifikation des U-Tests nach Fligner & Pollicello (1981)

Bei der modifizierten Version des U-Tests können Lokationshypothesen unter abgeschwächten Verteilungsvoraussetzungen getestet werden; es muß lediglich die Symmetrie der Verteilungen angenommen werden können, und die Annahme homogener Varianzen ist nicht erforderlich. Fligner & Policello (1981) betrachten die geordneten unabhängigen Zufallsstichproben $X_{1(1)} \leq X_{1(2)} \leq \ldots \leq X_{1(n_1)}$ und $X_{2(1)} \leq X_{2(2)} \leq \ldots \leq X_{2(n_2)}$ aus stetigen Populationen. Ihre Teststatistik ergibt sich nach

$$FP = \frac{\sum P_i - \sum S_m}{2(\sum (S_m - \bar{S})^2 + \sum (P_i - \bar{P})^2 + \bar{P} \cdot \bar{S})^{1/2}}, \qquad (11)$$

mit:

$$\bar{P} = \sum_{i=1}^{n_1} P_i/n_1 \text{ und } \bar{S} = \sum_{m=1}^{n_2} S_m/n_2;$$

$P_i = Q_i - i$ $(i = 1, \ldots, n_1)$: Anzahl der X_2 kleiner als $X_{1(i)}$
$S_m = R_m - m$ $(m = 1, \ldots, n_2)$: Anzahl der X_1 kleiner als $X_{2(m)}$
Q_i: Rang der $X_{1(i)}$ in der kombinierten Stichprobe
R_m: Rang der $X_{2(m)}$ in der kombinierten Stichprobe.

Für Stichproben der Größe $n_1 = 3$ und $n_2 = 3$ bis $n_1 = 12$ und $n_2 = 12$ haben Fligner & Policello (1981) die kritischen Werte der Teststatistik FP tabelliert. Für größere Stichproben können zur approximativ validen Signifikanzbeurteilung die Werte der Standard-Normalverteilung herangezogen werden (a.a.O.).

2.6 Modifikation des U-Tests nach Berchtold (1979)

Wenn die Voraussetzungen des t-Tests erfüllt sind, hat der U-Test eine asymptotische relative Effizienz (ARE) relativ zum t-Test von 0,9545. Die von Berchtold entwickelte Modifikation des U-Tests verbessert diesen Wert auf 0,9922. Berchtold (1979) definiert b als den ganzzahligen Teil von $(n_1 + n_2)/10$. Dann weist er jedem Wert der kombinierten Stichprobe seinen entsprechenden Rang zu (vgl. U-Test). Die Modifikation besteht darin, daß zwischen den b + 1 kleinsten und zwischen den b + 1 größten Werten die Rangabstände von 1 auf 5 erhöht werden. Für die Signifikanzbestimmung wird die U-Statistik verwendet, mit der die gleichen Hypothesen wie oben getestet werden.

2.7 Normal-Scores-Test nach van der Waerden (1953)

Unter den im Zusammenhang mit dem U-Test spezifizierten Modellvoraussetzungen können mit den im folgenden angesprochenen Normal-Scores-Tests Lokationshypothesen getestet werden. Bei der von van der Waerden (1953) vorgeschlagenen Version der Normal-Scores-Tests lautet die Teststatistik:

$$NS_{vdW} = \sum_{i=1}^{N} \Phi^{-1} \frac{R_i}{N+1} \qquad (12)$$

mit:

Φ^{-1}: Inverse der Standard-Normalverteilung
R_i: Rang von X_i, $i = 1, \ldots, N$, $N = n_1 + n_2$.

Unter den Voraussetzungen des t-Tests hat der NS_{vdW}-Test eine asymptotische relative Effizienz zum t-Test von 1.

2.8 Normal-Scores-Test nach Terry und Hoeffding (Terry, 1952)

Der Normal-Scores-Test von Terry und Hoeffding (Terry, 1952) (NS_{TH}) basiert auf dem gleichen Prinzip wie der NS_{vdW}-Test, mit dem Unterschied allerdings, daß anstelle der Inversen der Standard-Normalverteilung die Erwartungswerte der X_i unter der Normalverteilungsannahme eingesetzt werden (vgl. zu Einzelheiten etwa Marascuilo & McSweeney, 1977).

3. Methode

Die empirischen Fehlerwahrscheinlichkeiten p_α und p_β unter verschiedenen Annahmeverletzungen wurden mit Hilfe einer Computer-Simulation geschätzt, die unter Verwendung eines FORTRAN-Programms auf der UNIVAC-1100-Rechenanlage der Gesellschaft für wissenschaftliche Datenverarbeitung (GwD), Göttingen, durchgeführt wurde.

Die standard-normalverteilten Pseudozufallszahlen wurden über das Subroutine ggnml von IMSL (1982) erzeugt und anschließend mit dem Verfahren von Fleishman (1978) derart transformiert, daß sie sich wie folgt verteilten:

— N (0, 1): Standard-Normalverteilung mit $\gamma_1 = E(X—E(X))^3 = 0$ und $\gamma_2 = E(X—E(X))^4 — 3 = 0$;

— S (0, 1): schiefe Verteilung mit $\gamma_1 = 0{,}75$ und $\gamma_2 = 0$
— S^- (0, 1): schiefe Verteilung mit $\gamma_1 = -0{,}75$ und $\gamma_2 = 0$
— P (0, 1): platykurtische Verteilung mit $\gamma_1 = 0$ und $\gamma_2 = -1$
— L (0, 1): leptokurtische Verteilung mit $\gamma_1 = 0$ und $\gamma_2 = 3$

Durch die Transformation $y = \sigma \cdot x + m$ wurden die standardisierten Werte x in neue Werte y mit der Varianz σ^2 und dem Erwartungswert m transformiert. Die folgenden Varianzverhältnisse wurden untersucht:

— $\sigma_1^2/\sigma_2^2 = 1/4$
— $\sigma_1^2/\sigma_2^2 = 1/1$
— $\sigma_1^2/\sigma_2^2 = 4/1$

Zur Ermittlung der empirischen Teststärken $1 - \hat{p}_\beta$ wurden die Werte der jeweils zweiten Verteilung in jeder Verteilungskombination einmal mit $y = \sigma_2 x + 0{,}4$ und einmal mit $y = \sigma_2 x + 0{,}8$ transformiert.

Insgesamt wurden neun Verteilungskombinationen erzeugt, die mit drei verschiedenen Varianzverhältnissen insgesamt 27 Simulationsbedingungen ergaben (vgl. z.B. Tab. 1). Von diesen 27 Bedingungen repräsentieren 26 für den t-Test mindestens eine Annahmeverletzung, 24 für den W-Test, 23 für die U-, Berchtold- (B-) und die NS-Tests und 14 für den FP-Test.

Für jeden der betrachteten Tests wurden dieselben Rohwerte verwendet, um einen echten Vergleich zu ermöglichen. Die Tests erfolgten dabei links-, rechts- und beidseitig bei $\alpha = 0{,}05$ und jeweils links- und rechtsseitig bei $\alpha = 0{,}01$. Zur Bestimmung der empirischen Fehlerwahrscheinlichkeiten wurden jeweils $r = 5000$ Computerdurchläufe pro Bedingung realisiert, d.h. pro Bedingung konnten 5000 empirische Werte der einzelnen Teststatistiken berechnet werden. Bei dem nominellen Signifikanzniveau $\alpha = 0{,}05$ muß daher mit einem Standardfehler von $\sigma_{\hat{p}} = 0{,}00308$ gerechnet werden.

Da gleiche Stichprobengrößen nicht zu den Voraussetzungen einer validen Testanwendung gehören, wurden in unserer Untersuchung $n_1 = 8$ und $n_2 = 12$ sowie $n_1 = 15$ und $n_2 = 25$ gewählt. Da aus früheren Untersuchungen bekannt ist, daß einige Tests bei gleichen Stichprobenumfängen robuster sind als bei ungleichen (vgl. Hager et al., 1983), wurden hier also vglw. ungünstige Bedingungen geschaffen.

Die Durchführung des obligaten Probelaufs unter Standard-Normalbedingungen ergab keine nicht durch Zufall erklärbaren Abweichungen der empirischen Fehlerwahrscheinlichkeiten $\hat{p}_\alpha$ und $\hat{p}_\beta$ von den nominellen Werte α und β (vgl. Bedingung „NN 1." in Tabelle 1).

Tabelle 1

Robustheit des t-, Welch-Aspin-v-, Mann-Whitney-Wilcoxon-U-, Berchtold-B-, Fligner-Policello-FP und des Normal-Scores-Test von Terry-Hoeffding unter verschiedenen Verteilungsformen und Varianzverhältnissen bei $\alpha = 0{,}05$ (Angaben der relativen Häufigkeiten der Ablehnung der H_0 in Promille). (Weitere Erläuterungen im Text.)

σ_1/σ_2		t		v		U		B		FP		NS	
	.5	34	33	49	58	40	44	26	24	43	54	32	29
NN	1.	49	50	52	53	50	49	50	50	52	53	49	51
	2.	73	75	50	48	64	68	56	59	54	56	61	60
	.5	21	20	36	46	20	13	13	7	22	17	16	8
NS	1.	37	35	43	41	31	26	46	44	31	29	38	38
	2.	76	80	52	50	66	67	68	74	55	52	69	73
	.5	34	33	52	61	49	50	22	19	52	62	32	27
SN	1.	61	50	69	59	71	75	54	49	70	77	62	55
	2.	96	97	78	71	112	155	93	115	91	129	100	132
	.5	19	17	36	44	21	12	6	2	22	16	12	4
SS	1.	46	50	60	61	47	53	50	53	49	57	50	54
	2.	101	89	88	67	128	146	115	136	107	117	123	146
	.5	55	44	73	70	90	113	40	45	89	127	64	68
SS^-	1.	74	66	77	70	94	110	60	51	89	110	73	64
	2.	102	93	82	65	120	151	81	90	98	128	97	107
	.5	35	33	51	59	42	48	30	30	45	58	34	36
LL	1.	48	51	49	50	48	49	49	49	49	52	49	49
	2.	66	74	45	48	60	62	53	57	48	52	57	60
	.5	34	28	49	51	43	42	20	17	46	51	30	26
PP	1.	49	48	53	49	49	45	51	47	49	49	50	46
	2.	78	81	53	50	70	74	55	61	57	58	62	66
	.5	34	28	51	53	46	43	22	19	49	52	31	26
LP	1.	42	45	45	47	42	43	39	43	44	50	42	43
	2.	72	80	49	54	58	64	59	65	52	59	58	64
	.5	22	21	36	51	18	12	11	5	20	17	14	8
PS	1.	43	41	48	47	41	36	60	69	41	38	50	56
	2.	79	79	56	49	75	72	67	73	62	57	74	78

4. Ergebnisse

Die Ergebnisse der Simulationsuntersuchungen werden im folgenden auf zweierlei Arten, getrennt nach Robustheit und Teststärke, wiedergegeben. Zum einen sind in der Tabelle 1 die empirischen Schätzungen $\hat{p}_\alpha$ für die

Wahrscheinlichkeit für Fehler 1. Art, p_α, enthalten, und zwar für das Signifikanzniveau $\alpha = 0{,}05$ und für linksseitige Tests, wobei die Zahlenangaben die Anzahlen der Ablehnungen der Nullhypothese auf je Tausend Entscheidungen bedeuten. Die jeweils erste Spalte unter jedem Test bezieht sich dabei auf die kleineren Stichprobengrößen (8:12) und die jeweils zweite Spalte auf die größeren (15:25).

Die Werte aus dieser Tabelle 1 sind ebenso wie die nicht tabellierten Werte für $\alpha = 0{,}01$ und für rechts- sowie beidseitige Tests als absolute Abweichungen vom Signifikanzniveau über die 27 Simulationsbedingungen spaltenweise aufsummiert in der Tab. 2 enthalten. Aus dieser Tabelle kann daher gemäß unserer Frage entnommen werden, welche Rangreihenfolge die untersuchten Tests bzgl. der Robustheit über die einzelnen Bedingungen einnehmen.

Die gleichen Tabellentypen werden auch für die Teststärkeuntersuchungen realisiert.

Dabei werden die Ergebnisse des approximativen Welch-Tests (v_t) nicht dargestellt, da sie nur geringfügig von denen des exakten Welch-Aspin-Tests (v) abweichen. Von den Normal-Scores-Tests wird aus dem gleichen Grund auf die Auflistung der Werte für den van-der-Waerden-Test verzichtet, und die Resultate für den Terry-Hoeffding-Test rubrizieren unter „NS".

Die einzelnen Ergebnisse sprechen im wesentlichen für sich, so daß wir lediglich auf die folgenden Punkte besonders hinweisen wollen.

Bei allen simulierten varianzhomogenen Verteilungen sind die PT vglw. robust, und unter Varianzhomogenität führt die Verwendung der Teststatistiken v oder v_t zu besseren Resultaten als die Verwendung von t, und zwar auch bei nonnormalen Verteilungen.

Die NPT reagieren besonders empfindlich auf Kombinationen mit einer schiefen oder zwei entgegengesetzt schiefen Verteilungen. Der Grund hierfür ist darin zu sehen, daß unter diesen Bedingungen die allgemeine Alternativhypothese für diese Tests gilt, daß die Mediane unterschiedlich sind. Aufgrund dieser Überlegung wären die Ablehnungen der Nullhypothesen auf die Teststärke der NPT zurückzuführen und stellten daher keine mangelnde Robustheit dar. Da jedoch der Praktiker bei der hier betrachteten Anwendung der Tests ein signifikantes Ergebnis im Sinne eines Unterschiedes zwischen den Erwartungswerten interpretieren würde — der nicht vorliegt —, ordnen wir den gen. Befund unter dem Aspekt „Robustheit bei Erwartungswerthypothesen" ein.

Tab. 2 zeigt, daß der v- (und der v_t-) Test insgesamt mit Abstand am robustesten unter den simulierten Bedingungen ist und zudem als einziger Test mit zunehmendem Stichprobenumfang zunehmend robuster wird. Der Vorteil der v- und v_t-Tests bleibt im übrigen auch dann erhalten, wenn

Tabelle 2
Summe der absoluten Abweichungen vom Signifikanzniveau über 27 Simulationsbedingungen (Angabe in Promille) bei $\alpha = 0{,}05$ (vgl. Tab. 1) und $\alpha = 0{,}01$.

α	Richtung	$n_1:n_2$	t	v	U	B	FP	NS
0,05	linksseitig	8:12	536	262	581	498	420	516
		15:25	589	184	748	603	595	635
	zweiseitig	8:12	510	153	331	388	198	357
		15:25	633	127	540	468	545	492
	rechtsseitig	8:12	427	216	384	286	334	317
		15:25	466	178	571	351	536	411
0,01	linksseitig	8:12	283	155	167	160	139	163
		15:25	218	103	246	180	289	195
	rechtsseitig	8:12	131	88	108	88	92	92
		15:25	168	80	173	110	241	120

man die zugegebenermaßen nicht unproblematischen heteromeren Verteilungskombinationen unberücksichtigt läßt.

Obwohl für den FP-Test insgesamt die wenigsten Bedingungen mit verletzten Annahmen vorliegen, ist er im Vergleich mit den übrigen NPT nicht durchgängig am robustesten und erweist sich bei $\alpha = 0{,}01$ und den größeren Stichproben sogar als der unrobusteste Test.

Befassen wir uns anschließend mit den empirischen Schätzungen der Teststärke!

Tab. 3 gibt analog zur Tab. 1 die empirischen Schätzungen $1 - \hat{p}_\beta$ der Teststärken bei einem Unterschied zwischen den Erwartungswerten von $E(\bar{X}_1 - \bar{X}_2) = 0{,}4$ wieder. Bei diesem wie bei dem nicht-tabellierten größeren Unterschied (0,8) gilt, daß die Teststärke bei heterogenen Varianzen geringer ist als bei homogenen Varianzen, und zwar unabhängig davon, ob $\hat{p}_\alpha < \alpha$ oder ob $\hat{p}_\alpha > \alpha$. M.a.W. sind die untersuchten Tests nicht robust hinsichtlich der Wahrscheinlichkeiten für Fehler 2. Art, sofern die Voraussetzung der Varianzhomogenität nicht erfüllt ist. Allerdings erweist sich auch hierbei, daß der Welch-Aspin-Test insofern zu den besten Resultaten führt, als bei ihm die vglw. geringste Variabilität festzustellen ist. Demgegenüber ist die Variabilität in den Resultaten für die NPT um einiges größer.

In Tabelle 4 sind analog zur Tabelle 2 die über alle 27 Simulationsbedingungen *gemittelten* Schätzungen der Teststärke in Promille-Angaben enthalten. Die Daten zeigen erneut die Überlegenheit des v-Tests, der lediglich bei $\alpha = 0{,}01$ und $n_1:n_2 = 15:25$ vom FP-Test geringfügig übertroffen wird.

Tabelle 3

Teststärke der untersuchten Tests unter verschiedenen Verteilungsformen und Varianzverhältnissen bei $E(\bar{X}_1 - \bar{X}_2) = 0{,}4$ (Angaben in Promille).

	σ_1/σ_2	t		v		U		B		FP		NS	
	.5	97	135	135	218	120	170	80	109	125	196	95	129
NN	1.	208	319	208	325	198	306	205	308	202	321	203	314
	2.	177	242	132	171	153	210	139	191	134	184	148	199
	.5	71	108	108	195	60	63	44	46	65	79	53	52
NS	1.	178	319	188	329	154	244	205	355	155	254	185	318
	2.	175	229	128	161	147	191	147	201	126	160	152	204
	.5	101	133	142	210	127	166	71	82	130	189	95	111
SN	1.	224	335	236	344	251	391	215	321	242	393	232	343
	2.	203	276	161	209	228	347	185	290	193	294	204	312
	.5	73	119	112	206	60	73	32	31	63	87	44	41
SS	1.	210	330	227	348	227	349	263	446	223	351	252	420
	2.	204	270	166	204	225	334	202	318	187	284	218	338
	.5	141	166	177	237	201	303	106	151	204	333	152	197
SS⁻	1.	238	347	247	353	266	430	195	288	259	432	227	323
	2.	210	287	169	224	230	348	175	249	197	301	197	286
	.5	106	147	136	223	132	203	96	146	138	235	112	260
LL	1.	231	332	230	336	243	381	225	342	241	387	233	354
	2.	170	247	129	182	161	239	146	214	141	215	152	225
	.5	98	132	141	209	112	139	63	67	119	164	86	92
PP	1.	199	329	202	333	182	306	219	378	186	319	208	358
	2.	164	230	113	158	132	189	109	156	111	155	123	172
	.5	91	124	126	205	101	139	60	76	105	165	74	95
LP	1.	205	336	219	349	191	308	190	305	200	332	194	304
	2.	186	243	142	178	159	217	158	213	148	203	163	214
	.5	72	118	116	201	57	72	39	42	64	87	46	52
PS	1.	195	317	198	318	174	260	229	398	175	261	210	355
	2.	155	233	107	164	134	186	119	187	106	150	128	196

Diese empirischen Resultate stehen daneben im Einklang mit den analytischen von Pfanzagl (1974, 40), „... namely that Welch's test is, in the class of all translation invariant tests, asymptotically uniformly most powerful against one-sided alternatives".

Beim v-Test führt darüber hinaus eine Vergrößerung der Stichproben zum relativ größten Teststärkezuwachs.

Eine insgesamt geringe Teststärke hat die von Berchtold (1979) vorgeschlagene Modifikation des U-Tests, bei dem auch eine Vergrößerung der Stichproben die geringste Auswirkung zeitigt.

Tabelle 4

Über 27 Simulationsbedingungen gemittelte Schätzungen der Teststärke der untersuchten Tests bei $E(\bar{X}_1 - \bar{X}_2) = 0{,}4$ (vgl. Tab. 3) und bei $E(\bar{X}_1 - \bar{X}_2) = 0{,}8$.

α	$E(\bar{X}_1 - \bar{X}_2)$	$n_1:n_2$	t	v	U	B	FP	NS
0,05	0,4	8:12	159	163	164	145	157	155
		15:25	237	244	243	219	242	232
	0,8	8:12	362	362	350	322	337	339
		15:25	556	565	537	501	535	518
0,01	0,4	8:12	51	53	44	42	43	38
		15:25	86	88	88	73	107	77
	0,8	8:12	154	156	129	126	135	125
		15:25	314	320	300	267	335	279

Wider Erwarten ist auch mit den Normal-Scores-Tests eine vglw. geringe Teststärke verbunden, die zudem insgesamt geringer ist als die des U-Tests, obwohl die asymptotische relative Effizienz der NS-Tests mit 1 größer ist als die des U-Tests (0,955). Dieser Befund zeigt, daß die für unendlich große Stichproben berechneten Teststärkevergleiche über die A.R.E. nicht ungeprüft auf finite Fälle übertragen werden können.

5. Diskussion und Schlußbemerkungen

Die dargestellten Simulationsbefunde für einige der bei der Prüfung empirisch-psychologischer Hypothesen besonders wichtigen Zwei-Stichproben-Tests zeigen, daß etliche der in der gebräuchlichen psychologischen Literatur auffindbaren Empfehlungen bzgl. der Anwendung parametrischer und nonparametrischer Tests entweder unvollständig oder schlicht falsch sind; dies mag der Leser/die Leserin selbst durch einen Vergleich der Befunde mit den in der Einleitung wiedergegebenen Zitaten nachprüfen.

U.E. verdeutlichen die grundsätzlichen Divergenzen in den Empfehlungen zwei Schwierigkeiten, nämlich zum einen das Problem der Rezeption mathematisch-statistischer Befunde durch Nicht-Statistiker wie bspw. Psychologen und zum anderen das Problem, daß selbst bei Statistikern offenbar keine Einigkeit bzgl. der getesteten Hypothesen und Voraussetzungen bei bestimmten Tests bestehen (vgl. z.B. Wetherill, 1960, und Hilgers, 1982).

In einer derartigen Situation stellen Computersimulationen wie die von uns durchgeführte wohl die einzige Möglichkeit dar, die theoretisch noch

nicht beantworteten Fragen zumindest näherungsweise und möglicherweise nur vorläufig zu beantworten.

Die in der vorliegenden Arbeit gegebenen Antworten besitzen dabei nur Gültigkeit für die untersuchten Bedingungen und insbes. die betrachteten Erwartungswerthypothesen. Geht man dagegen von allgemeineren Hypothesen oder von einem niedrigeren Skalenniveau aus, wird man zu anderen Antworten und Empfehlungen gelangen als die oben gegebenen.

Summary

Some recommendations concerning the choice between parametric and nonparametric statistical tests aimed at testing specific location hypotheses are examined for some two-sample tests by means of computer simulations. It is shown that some of the wide-spread recommendations need reformulation.

Résumé

Les auteurs se demandent si, face à des distributions quelconques, le chercheur doit préférer les tests statistiques nonparamétriques. Les hypothèses qu'ils formulent sont fondées sur les recommandations de manuels faisant autorité en la matière. Un procédure de simulation permet de montrer que, dans le cas de la comparaison de deux échantillons, ces recommandations sont souvent trop peu précices pour être utiles; elles sont parfois même fausses.

Literatur

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

Aspin, Alice A.: Tables for use in comparisons whose accuracy involves two variances, separately estimated. Biometrika, 1949, **36**, 290—296.

Berchtold, H.: A modified Mann-Whitney test with improved asymptotic relative efficiency. Biometrical Journal, 1979, **21**, 649—655.

Blair, R. C. & 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, 1980 (a), **5**, 309—335.

Blair, R. C. & 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, 1980 (b), **4**, 645—656.

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

Bortz, J.: Lehrbuch der Statistik für Sozialwissenschaftler. Berlin: Springer, 1979 (Nachdruck der 1. Auflage).

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

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

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

Bredenkamp, J.: Theorie und Planung psychologischer Experimente. Darmstadt: Steinkopf, 1980.

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

Edgington, E. S.: A new tabulation of statistical procedures used in APA journals. American Psychologist, 1978, **29**, 25—26.

Fleishman, A. J.: A method for simulating non-normal distributions. Psychometrika, 1978, **43**, 521—532.

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

Gaensslen, H. & Schubö, W.: Einfache und komplexe statistische Analyse. München: Reinhardt, 1976[2].

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

Hager, W., Lübbeke, B. & Hübner, R.: Verletzung der Annahmen bei Zwei-Stichproben-Lokationstests: Eine Übersicht über empirische Resultate. Zeitschrift für experimentelle und angewandte Psychologie, 1983, **30**, 347—386.

Hager, W. & Westermann, R.: Die Elle — 10 Jahre danach. Zeitschrift für Sozialpsychologie, 1982, **13**, 250—252.

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

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

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

Hays, W. L.: Statistics. New York: Holt-Saunders, 1981[3].

Hilgers, R.: On the Wilcoxon-Mann-Whitney test as nonparametric analogue and extension of the t test. Biometrical Journal, 1982, **24**, 3—15.

Hollander, M. & Wolfe, D. A.: Nonparametric statistical methods. New York: Wiley, 1973.

IMSL: International Mathematical and Statistical Libraries. IMSL library 1. Houston: 1982.

Kreyszig, E.: Statistische Methoden und ihre Anwendungen. Göttingen: Vandenhoeck & Ruprecht, 1968.

Lehmann, E. L.: Nonparametrics: Statistical methods based on ranks. San Francisco: Holden-Day, 1975.

Lienert, G. A.: Verteilungsfreie Methoden in der Biostatistik. Meisenheim am Glan: Hain, 1973[2].

Mann, H. B. & Whitney, D. R.: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 1947, **18**, 50—60.

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

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

Pearson, E. S. & Hartley, H. O.: Biometrika tables for statisticians. Bd. 1. Cambridge: Cambridge University Press, 1962.
Pfanzagl, J.: On the Behrens-Fisher problem. Biometrika, 1974, **61**, 39—47.
Potthoff, R. F.: Use of the Wilcoxon statistic for a generalized Behrens-Fisher problem. Annals of Mathematical Statistics, 1963, **34**, 1596—1599.
Pratt, J. W.: Robustness of some procedures for the two-sample location problem. Journal of the American Statistical Association, 1964, **59**, 665—680.
Sachs, L.: Statistische Auswertungsmethoden. Berlin: Springer, 1968.
Satterthwaite, F. E.: An approximate distribution of estimates of variance components. Biometrics Bulletin, 1946, **2**, 110—114.
Scheffé, H.: Practical solutions of the Behrens-Fisher problem. Journal of the American Statistical Association, 1970, **65**, 1501—1508.
Schüle, W.: Flußdiagramm zur Bestimmung eines adäquaten statistischen Prüfverfahrens im Bereich der parametrischen und nicht-parametrischen Grundlagenstatistik. In: Siegel, S.: Nichtparametrische statistische Methoden. Frankfurt am Main: Fachbuchhandlung für Psychologie, 1976, 309—320 (Anhang).
Student (W. S. Gosset): The probable error of the mean. Biometrika, 1908, **6**, 1—25.
Terry, M. E.: Some rank order tests which are most powerful against specific parametric alternatives. Annals of Mathematical Statistics, 1952, **23**, 346—366.
van der Waerden, B. L.: Ein neuer Test für das Problem der zwei Stichproben. Mathematische Annalen, 1953, **26**, 93—107.
Vorberg, D.: Eine Entscheidungshilfe für die Auswahl statistischer Tests und Maße. Psychologische Rundschau, 1981, **31**, 267—277.
Wang, Ying Y.: Probabilities of the type I errors of the Welch test for the Behrens-Fisher problem. Journal of the American Statistical Association, 1971, **66**, 605—608.
Welch, B. L.: The significance of the difference between two means when the population variance are unequal. Biometrika, 1937, **29**, 350—362.
Welch, B. L.: The generalization of „Student's" problem when several populations are involved. Biometrika, 1947, **34**, 28—35.
Welch, B. L.: Further note on Mrs. Aspin's tables and on certain approximations to the tabled function. Biometrica, 1949, **36**, 293—296.
Westermann, R.: Die empirische Überprüfung des Niveaus psychologischer Skalen. Zeitschrift für Psychologie, 1980, **188**, 450—468.
Westermann, R.: Zur Messung von Einstellungen auf Intervallskalenniveau. Zeitschrift für Sozialpsychologie, 1982, **13**, 97—108.
Westermann, R. & Hager, W.: Entscheidung über statistische und wissenschaftliche Hypothesen: Zur Differenzierung und Systematisierung der Beziehungen. Zeitschrift für Sozialpsychologie, 1982, **13**, 13—21.
Wetherill, G. B.: The Wilcoxon test and non-null hypotheses. Journal of the Royal Statistical Society, B, 1960, **22**, 402—418.
Wike, E. L. & Church, J. D.: Nonrobustness in F tests: 1. A replication and extension of Bradley's study. Bulletin of the Psychonomic Society, 1982 (a), **20**, 165—167.
Wike, E. L. & Church, J. D.: Nonrobustness in F tests: 2. Further extensions of Bradley's study. Bulletin of the Psychonomic Society, 1982, **20**, 168—170.
Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin, 1945, **1**, 80—83.
Wottawa, H.: Zum Problem der Abtestung der Verteilungsvoraussetzung in Varianz- und Regressionsanalyse. Archiv für Psychologie, 1981/82, **134**, 257—263.

Anschrift der Verfasser: Dr. Willi Hager und Ronald Hübner, Institut für Psychologie, Goßlerstraße 14, 3400 Göttingen.