Psychotherapie Forum Bd. 4, Heft 3, 1996 (Originalpublikation Springer Wien)

B. Rüger

Fragen und Anmerkungen zu einigen statistischen Methoden in der Psychotherapieforschung

Zusammenfassung *Die Beachtung der methodischen Grundsätze statistischer Verfahren entscheidet über Wert und Glaubwürdigkeit eines statistisch gewonnenen Ergebnisses und nicht der Inhalt des Ergebnisses. Auf der Grundlage dieses methoden-orientierten Standpunktes eines Statistikers werden die klassischen statistischen Verfahren der empirischen Psychotherapieforschung untersucht. Es wird aufgezeigt, an welch strenge Voraussetzungen die gängigen „harten" Inferenzverfahren (vor allem statistische Tests) gebunden sind und daß in den meisten Fällen empirischer Studien der Einsatz „weicher" Methoden deskriptiver oder explorativer Art angezeigt ist. Untersucht werden auch Prinzipien zur Evaluation, Prüfungen von Validität und Reliabilität und insbesondere statistische Probleme bei der empirischen Beurteilung der Therapiedauer.*

Schlüsselwörter: *Evaluation, Validität, Reliabilität, geeignete statistische Methoden, multiple Signifikanztests, Therapiedauer.*

A few questions and remarks concerning the application of specific statistical methods in psychotherapy research

Abstract *The value and credibility of a statistically obtained result is not judged by its content, but by the methodological principles of the statistical procedures by which the result was obtained. Based on this methodological point of view (as is usual in statistics) the classical statistical procedures in empirical psychotherapy studies are considered. It is pointed out, how restrictive the assumptions are on which the usual "hard" inference procedures (viz.: statistical tests) are based. It turns out that in most cases the use of "weaker" methods as developed in descriptive and explorative data analysis is more adequate. Methods of evaluation, measurements of validity and reliability, in particular statistical problems concerning the effect of the duration of a therapy are also inquired.*

Keywords: *Evaluation, validity, reliability, proper statistical methods, multiple tests, duration of therapy.*

Questions et remarques concernant certaines méthodes statistiques utilisées dans le domaine de la recherche en psychothérapie

Résumé *Ce n'est pas le contenu des résultats qui définit la valeur et la fiabilité des chiffres obtenus par le biais de méthodes statistiques, mais le respect de principes méthodiques fondamentaux. Partant du point de vue d'un statisticien axé sur la méthode, nous examinons les procédures statistiques classiques employées au niveau de la recherche par la psychothérapie: les principes permettant d'évaluer un type de thérapie, la vérification de la validité et de la fiabilité des données, les procédures "dures" d'inférence (les tests statistiques surtout) et (en particulier) les problèmes statistiques que présente l'évaluation empirique de la durée de la thérapie. Nous nous concentrons sur des indications ayant trait aux questions et rapports suivants:*

1) l'échelonnement de la "taille de l'effet (effect size)" (la répartition d'un item) est en général très libre et peut en partie être fixée arbitrairement. Dans quelle mesure les modifications (par la thérapie) subies par "l'importance de l'effet" sont-elles dépendantes de l'échelle choisie?

2) la validité de cette même variable est mesurée en règle générale par le calcul d'une corrélation, à savoir celle entre la taille de l'effet et un critère (quantité visée) permettant d'évaluer de manière absolue les effets d'une thérapie (validité du critère). Si un critère "idéal" de ce type existe, sa substitution par une variable "importance de l'effet" dans les études d'évaluation ne se justifie que si les données peuvent être collectées beaucoup plus aisément que si l'on

employait un véritable critère. Souvent, ces études se contentent d'employer en tant que critère une autre variable, plus ou moins bien testée, sans en indiquer la validité; en effet la validité d'une nouvelle variable "taille de l'effet" n'est que valeur relative.

3) on mesure la fiabilité des résultats en enregistrant plusieurs fois (en règle générale deux fois) la variable "taille de l'effet" (effect size) de l'instrument d'évaluation (re-test, test parallèle, division du test en deux) et on l'indique en tant que proportion de co-variance par rapport à la variance ou plus simplement, par l'alpha de Cronbach. Cette manière de procéder requiert certaines conditions (hypothèse d'une homogénéité de l'échantillon et d'un manque de corrélation entre les erreurs de mesure), dont la plupart des études d'évaluation négligent de vérifier la présence. Lorsque celles-ci sont absentes, l'alpha de Cronbach ne constitue plus forcément une mesure comparative de fiabilité; il peut arriver que deux valeurs alpha différentes ne soient plus comparables.

4) du fait que la plupart des recherches en psychothérapie utilisent des variables "taille de l'effet" (et des critères) qualitatifs, il n'est pas possible d'utiliser l'habituel coefficient de corrélation selon Bravais-Pearson lorsque validité et fiabilité sont mesurées comme décrit plus haut. Il faut plutôt utiliser le coefficient de corrélation selon le rang de Spearman lorsque les variables sont échelonnées de manière ordinale et un coefficient de contingence lorsque les variables ne sont échelonnées que de manière nominale. Ces coefficients peuvent être influencés par une modification de la graduation choisie pour les cotes ordinal et nominal. Les mesures de validité et de fiabilité dépendent donc de l'exactitude de l'échelonnement de la variable " taille de l'effet" (effect size) étudiée (du nombre de réponses prévues pour un item donné).

5) les procédures "dures" d'inférence normalement utilisées, et surtout les tests statistiques, sont basées sur un échantillonnage aléatoire et des hypothèses liées à l'indépendance des variables; dans de nombreux cas elles ne peuvent se pratiquer qu'avec des variables échelonnées de manière cardinale ou même présentant une répartition normale – en règle générale, concernant la recherche empirique en psychothérapie ces conditions ne sont pas remplies. Il n'est donc pas possible d'utiliser ces procédures. Il est en particulier absurde (bien qu'en accord avec les conventions) de procéder à un test statistique et d'indiquer des valeurs p si l'on ne dispose pas d'un échantillonnage aléatoire.

6) pour la plupart des études effectuées dans la domaine de la psychothérapie il est recommandé d'utiliser des procédures statistiques "molles": par ceci nous entendons les méthodes de l'analyse descriptive ou exploratrice des données. Elles permettent de confirmer des hypothèses, mais non de les démontrer statistiquement (preuve par le seuil de signification).

7) l'utilisation de tests multiples (plusieurs tests étant effectués sur un même échantillon) doit s'accompagner d'un ajustement de l'alpha permettant de réduire nettement les alpha des différents tests (ajustement de Bonferroni).

8) des problèmes (parfois statistiques) particuliers se manifestent par rapport à l'évaluation empirique des effets de la durée de la thérapie si l'on n'effectue pas une très nette distinction entre les deux questions suivantes: comment les changements provoqués par la thérapie chez un seul patient dépendent-ils en moyenne de la durée de la thérapie? comment le nombre moyen de patients pour lesquels la thérapie a apporté une amélioration dépend-il de la durée de la thérapie? Chacune de ces deux questions requiert ses propres conditions d'étude. Des exemples sont présentés, qui permettent de débattre des problèmes qui se présentent.

Die folgenden kritischen Betrachtungen über statistische Methoden in der empirischen Psychotherapieforschung stehen unter zwei Leitlinien.

Die erste betrifft den wissenschaftlichen Standpunkt eines Statistikers, demzufolge die Beachtung der methodischen Grundsätze statistischer Verfahren über den Wert und die Glaubwürdigkeit eines statistischen Ergebnisses entscheidet und nicht der Inhalt des Ergebnisses. Die Ausrichtung (der „Rückzug") auf die Methoden ist für einen verantwortungsvoll arbeitenden Statistiker absolut notwendig: Er arbeitet nicht ergebnis- sondern verfahrensorientiert. Um es verschärft auszudrücken: Wenn man den statistischen Nachweis eines Ergebnisses nicht methodengerecht durchführt, ist das Ergebnis selbst (statistisch!) wertlos, unabhängig davon, wie plausibel es ansonsten sein mag. Nur unter der Wahrung dieses Standpunktes kann empirisch-statistischen Untersuchungen Vertrauen entgegengebracht werden.

Die zweite Leitlinie betrifft eine Eingrenzung des Stoffes: Es sollen nur „klassische" statistische Verfahren angesprochen werden. Einmal werden diese Verfahren immer noch am häufigsten eingesetzt, zum anderen sind ihre Grundlagen und Voraussetzungen besonders gut theoretisch abgesichert und damit allgemeinverbindlich. Die „modernen" statistischen Verfahren, die auch in der Psychotherapieforschung angewandt werden, z.B. Methoden aus der Zeitreihenanalyse, der Theorie stochastischer Prozesse oder der Chaostheorie, besitzen zwar ebenso gut abgesicherte theoretische Grundlagen, sind aber oft für ganz andere Anwendungsgebiete entwickelt worden und verwenden daher als Voraussetzungen Modelle, die sich nur sehr bedingt auf psychotherapeutische Prozesse übertragen lassen. Zweifellos befindet sich die Statistik hier wie auch woanders in einem Dilemma, das sie als Methodenlehre weder beneidenswert noch attraktiv macht: Ihre Verfahren sollen Erkenntnisse, die in der klinischen Forschung und therapeutischen Praxis oft schon gewonnen wurden, empirisch verbindlich absichern, die Entwicklung der dazu geeigneten statistischen Methoden eilt dem Erkenntnisprozeß der Forschung hinterher, die Ergebnisse der oft genug sehr komplizierten statistischen Verfahren erwecken dann keine Neugier mehr.

Auf eine ganze Reihe von Beispielen und auch auf die statistischen Probleme, die typischerweise in Meta-

Analysen auftreten, kann hier aus Platzgründen nicht eingegangen werden; ich verweise dazu auf meinen umfangreicheren Beitrag in Fäh-Barwinski et al.

Evaluation, Validität, Reliabilität

Eine allgemeinverbindliche Methodik zur Evaluation (Bewertung) von sozialpolitischen oder psychotherapeutischen Maßnahmen existiert nicht – mit jeder Evaluationsstudie wird immer ein Stück Neuland betreten, auch in methodischer Hinsicht. Gleichwohl enthalten Evaluationsstudien auch Gemeinsamkeiten, die hier im Mittelpunkt stehen sollen. Dazu gehören im Bereich der Psychotherapieforschung vor allem die Verwendung von Effektgrößen, mit denen die Wirkung einer Therapie oder Intervention erfaßt werden soll, und die Beobachtungen dieser Effektgrößen in einem Patientenkollektiv.

Die Effektgrößen resultieren aus Rating-Instrumenten (Manuals, Inventaren, Fragebögen), die zur Diagnose und Beurteilung psychischer Störungen eingesetzt werden. In der mir zugänglichen Literatur habe ich Hinweise auf insgesamt mehr als 60 verschiedene solcher Instrumentarien gefunden; sie stellen ein breit gefächertes und sehr vielseitiges Spektrum mit ganz verschiedenen Schwerpunkten dar. Darunter gibt es Instrumente, die sich auf ganz spezielle Symptombereiche konzentrieren und jeweils nur ein oder zwei Effektgrößen abgeben, und andere, die wesentlich umfassender und differenzierter aufgebaut sind, körperliche, seelische und soziale Beeinträchtigungen enthalten und jeweils eine ganze Reihe von Effektgrößen liefern.

Die Wahl oder Konstruktion geeigneter Effektgrößen und deren Skalierung sind wesentliche Bestandteile in einer Evaluationsstudie. Insbesondere müssen die Effektgrößen mit Hilfe eigener oder früherer Patientenstichproben auf ihre Validität und Reliabilität hin überprüft werden. Die Validität ist auf die Frage gerichtet, ob die Effektgröße wirklich das mißt, was gemessen werden soll, und die Reliabilität auf die Frage, wie genau oder zuverlässig mit der Effektgröße gemessen wird. Die allgemeinen Grundsätze und Methoden zu Evaluation, Validität und Reliabilität findet man bei Bergin und Garfield (1994), Bortz und Döring (1995), Faller und Frommer (1994), Koch und Wittmann (1990), Lienert und Raatz (1994), Steyer und Eid (1993), ihre Anwendungen und Probleme sowie Beispiele speziell auf dem Gebiet der Psychotherapieforschungen bei Kächele und Kordy (1995), Lang (1990), Leuzinger-Bohleber (1995), Manz und Schepank (1993), Rudolf (1991), Rudolf et al. (1994), Rüger und Senf (1994), Senf und von Rad (1995), Tschuschke und Czogalik (1990), Zielke (1993). Die entscheidenden Schritte einer Evaluationsstudie sind: *Auswahl der Effektgrößen, Skalierung der Effektgrößen, Validitätsprüfung und Reliabilitätsprüfung.* Dabei sind folgende Probleme und kritische Punkte zu beachten:

1) Oft wird die Wirkung einer Therapie mit den durch sie verursachten Veränderungen der Effektgrößen einfach gleichgesetzt.
2) Die Skalierung einer Effektgröße (die Abstufung eines Items) ist in der Regel sehr frei und teilweise willkürlich festlegbar. Wie stark hängen die (durch eine Therapie verursachten) Veränderungen der Effektgröße von der gewählten Skala ab?
3) Die Validität einer Effektgröße wird in der Regel durch eine Korrelation gemessen, nämlich derjenigen zwischen der Effektgröße und einer Kriteriumsgröße (Zielgröße), mit der die Wirkung einer Therapie verbindlich beurteilt werden kann (Kriteriumsvalidität). Mit dem betreffenden Korrelationskoeffizienten r wird nach Konvention die folgende Validitätsanforderung aufgestellt: Bei Werten von r zwischen 0.4 und 0.6 gilt die Effektgröße als mittelmäßig und bei Werten oberhalb von 0.6 als ausreichend valide. Wenn es ein solches „ideales" Kriterium gibt, so ist die Verwendung einer Effektgröße an seiner Stelle in Evaluationsstudien nur zu rechtfertigen, wenn die Effektgröße wesentlich einfacher erhoben werden kann als die Kriteriumsgröße. Oft wird als Kriterium nur eine andere, mehr oder weniger gut erprobte Effektgröße verwendet, deren Validität nicht angegeben wird; dann handelt es sich bei der Validität der neuen Effektgröße nur um einen relativen Wert.
4) Die Reliabilität wird durch mehrfache, in der Regel nur zweifache Erhebungen der Effektgröße (des Rating-Instrumentes) gemessen (Retestverfahren, Paralleltestverfahren, Testhalbierungsverfahren) und durch den dabei festgestellten Anteil der Kovarianz an der Varianz oder einfach durch Cronbachs Alpha angegeben. Reliabilitätswerte zwischen 0.8 und 0.9 gelten als mittelmäßig, solche oberhalb von 0.9 als hoch. Diese Vorgangsweise ist an bestimmte Voraussetzungen (Homogenitätsannahme, Annahme über die Unkorreliertheit der Meßfehler) gebunden, die in den meisten Evaluationsstudien nicht überprüft werden. Bei Verletzung der Voraussetzungen ist nicht mehr gewährleistet, daß Cronbachs Alpha ein komparatives Maß für die Reliabilität darstellt; zwei verschiedene Alpha-Werte sind dann u.U. nicht mehr vergleichbar.
5) Da es sich wohl in nahezu allen Fällen einer Psychotherapieforschung um qualitative Effektgrößen (und Kriteriumsgrößen) handelt, darf bei der oben beschriebenen Art der Validitäts- und Reliabilitätsmessung nicht der übliche Korrelationskoeffizient nach Bravais-Pearson verwendet werden. Vielmehr ist an seiner Stelle der Spearmansche Rangkorrelationskoeffizient zu benutzen, wenn ordinal skalierte Größen vorliegen, oder ein Kontingenzkoeffizient bei lediglich nominal skalierten Größen. Diese Koeffizienten sind nicht unempfindlich gegenüber einer Veränderung der Anzahl der zugrundegelegten Stufen der Ordinalskala bzw. Ausprägungen der Nominalskala. Damit wird die Validitäts- und Reliabilitätsmessung von der Feinheit der Skalierung der untersuchten Effektgröße (von der Anzahl der für ein Item vorgesehenen Antworten) abhängig.

Mit meinen Bemerkungen habe ich mich ausschließlich auf die weit verbreiteten „klassischen" Evaluationsstudien bezogen, die sich allein auf Effektgrößen

stützen. In einer „modernen" Evaluationsmethodik steht das Prozeßgeschehen einer Therapie stärker im Mittelpunkt; als Analyseinstrumente stehen u.a. zur Verfügung: Direktbeobachtungen des Behandlungsprozesses durch Tonband- oder Videoaufzeichnungen, Analysen der therapeutischen Beziehungsentwicklung, dynamische Modelle zur Beschreibung des Prozesses, laufende Forschungsinterviews mit den Patienten, Untersuchungen der psychischen Struktur (strukturellen Störung), Kombinationen von On-Line- und Off-Line-Forschung, Einbeziehung von Supervisionen. Man vergleiche dazu etwa: Bergin und Garfield (1994) (Part III), Hartkamp und Heigl-Evers (1995), Leuzinger-Bohleber (1994, 1995), Rudolf et al. (1995), Schiepek (1994), Schiepek und Kowalik (1994), Shapiro und Emde (1995) und Tress et al. (1994). Auch hier werden aber zur Untersuchung von Einzelfragen immer wieder Effektgrößen herangezogen.

Anwendbarkeit statistischer Tests

Innerhalb einer Evaluationsstudie werden die in einer oder mehreren Patientenstichproben festgestellten Werte der Effektgrößen in der Regel mit statistischen Verfahren ausgewertet. Welche Verfahren sind dafür (in einer gegebenen Situation) geeignet? Mit „geeignet" ist dabei zweierlei gemeint: „erlaubt" und „dem Studienziel gerecht werdend". Hier wenden wir uns der ersten Bedeutung zu, die zweite ist Gegenstand des nächsten Abschnittes.

Jedes statistische Verfahren beruht auf mehr oder weniger stark einschneidenden Voraussetzungen, und nur, wenn diese (wenigstens hinreichend genau) erfüllt sind, darf das Verfahren angewandt werden, ist es erlaubt. Man vergleiche dazu die einschlägige Literatur zur statistischen Methodenlehre, z.B. Bortz (1993), Bortz et al. (1990), Büning und Trenkler (1994), Lehmann (1986, 1991), Menges (1982), Rüger (1996) sowie auch Bredenkamp (1972), Morrison und Henkel (1970).

Die Voraussetzungen richten sich einerseits an die Qualität der Stichprobe und andererseits an die Skalenart und die (Wahrscheinlichkeits-)Verteilung der Effektgrößen. Tendenziell gilt: Je schärfer die Aussagen sind, die das statistische Verfahren ermöglicht, desto einschneidender sind die Voraussetzungen an die Stichproben und Effektgrößen. Am häufigsten werden in Studien zur Psychotherapieforschung die schärfsten statistischen Verfahren eingesetzt, nämlich Signifikanztests, obwohl sie auf Voraussetzungen beruhen, die in diesen Studien praktisch nie erfüllt sind. Generell setzen statistische Tests

(1) Zufallsstichproben

voraus und bauen darüber hinaus auf ganz bestimmte

(2) Unabhängigkeitsannahmen

auf, die im wesentlichen besagen, daß die in die Prüfgröße des Tests eingehenden Werte der Effektgrößen unabhängig sein müssen (d.h. zumindest: von verschiedenen Patienten stammen müssen). Die beliebten t-Tests und F-Tests sind zusätzlich an

(3) kardinal skalierte Effektgrößen und

(4) normalverteilte Effektgrößen

gebunden. Schließlich benötigen die häufig eingesetzten Verfahren der einfachen und multivariaten Varianzanalyse (ANOVA und MANOVA) über die bereits genannten Voraussetzungen hinausgehend noch die Erfüllung der

(5) Varianzhomogenität

in den verschiedenen Gruppen. Ich nehme an, daß fast jeder empirisch arbeitende Psychotherapieforscher diese Voraussetzungen kennt. Kaum einer von ihnen scheint sie aber ernst zu nehmen – und wird in dieser Haltung nicht nur von der Mehrzahl der Veröffentlichungen seines Forschungsbereiches immer wieder bestärkt, sondern auch von einschlägigen methodisch orientierten Lehrbüchern seines Gebietes. Bei Cohen (1988) wird z.B. die entscheidende Voraussetzung (1) an zwei Stellen (Seiten 2 und 19) eher beiläufig erwähnt und ansonsten (stillschweigend) als erfüllt vorausgesetzt, und bei Bortz und Döring (1995) findet man dazu auf Seite 375 den bemerkenswerten Satz: „Prinzipiell ist für jede irgendwie geartete Stichprobe bzw. für eine ‚Ad-hoc'-Stichprobe im nachhinein eine (fiktive) Population konstruierbar, für die die Stichprobe repräsentativ bzw. zufällig ist."

Andererseits gibt es aber auch nicht nur in theoretischen Statistik-Lehrbüchern warnende Gegensteuerungen. So zeigten z.B. Scarino und Davenport (1987), wie stark im Fall der ANOVA die α-Wahrscheinlichkeit und erst recht die β-Wahrscheinlichkeit anwachsen, wenn in Verletzung von (2) die Intraklasskorrelation nicht Null (sondern z.B. nur 0.1) ist. Ähnliche Verfälschungen entstehen, wenn eine ANOVA durchgeführt wird, ohne daß (5) erfüllt ist. Zu diesen speziell die ANOVA betreffenden Fragen vergleiche man auch Miller (1986).

Die Probleme, die entstehen, wenn eine geforderte Verteilungsannahme, beispielsweise die Normalverteilungsannahme (4), nicht erfüllt ist, führen zu Untersuchungen der Robustheit des betreffenden Verfahrens; man nennt ein Verfahren robust, wenn es relativ unempfindlich ist gegenüber Abweichungen von der vorausgesetzten Verteilung. Beispielsweise sind der t-Test und der F-Test robust gegenüber Abweichungen von der Normalverteilungsannahme – diese positive Eigenschaft betrifft allerdings nur den Fehler 1. Art (α-Fehler) und nicht die Power und damit den Fehler 2. Art (β-Fehler) dieser Tests. Robustheitsuntersuchungen haben sich innerhalb der Statistik zu einer eigenen Disziplin entfaltet, man vergleiche dazu Büning (1991), Hampel et al. (1986) und vor allem Huber (1981).

Auch eine Verletzung von (3) könnte man unter die Rubrik der Robustheit einordnen, nämlich der relativen Unempfindlichkeit des Verfahrens gegenüber Veränderungen der vorausgesetzten Skalierung. Hat man dabei eine Abänderung der Kardinalskala in eine Ordinal- oder sogar nur Nominalskala vor Augen, so wird das Problem anders gestellt und gelöst: Statt der auf der Kardinalskala aufbauenden Testverfahren (z.B. t-Tests oder F-Tests) werden sogenannte verteilungsfreie oder nichtparametrische statistische Tests (z.B. Wilcoxon-

Tests oder Kruskal-Wallis-Tests) verwendet; man vergleiche dazu Bortz et al. (1990), Büning und Trenkler (1994), Conover (1980) und Wolf (1980). Andererseits haben wir bereits darauf hingewiesen, daß bei Ordinal- oder Nominalskalen oft die Anzahl der Abstufungen bzw. Ausprägungen relativ willkürlich ist und z.B. Korrelations- oder Kontingenzmaße davon empfindlich berührt werden. Dieser Sachverhalt trifft auch auf viele nichtparametrische Verfahren zu. Ein spezielles aber doch sehr erhellendes Beispiel für solche Empfindlichkeiten auch „in umgekehrter Richtung" findet man bei Maxwell und Delany (1993), wo nachgewiesen wird, daß eine Dichotomisierung stetiger Prädiktorgrößen zu einer Überschätzung der Stärke des Zusammenhanges und damit zu Scheinsignifikanzen führen kann.

Wenn man auch, wie wir angedeutet haben, bei Verletzung einiger der genannten Voraussetzungen auf andere Verfahren ausweichen oder sich mit Robustheitsargumenten begnügen kann: Voraussetzung (1) gestattet generell keine Abschwächung, wenn man statistische Tests verwenden will. (Unter geringfügigen Einschränkungen gilt für Voraussetzung (2) ähnliches.)

Liegen keine Zufallsstichproben vor, so ist die Durchführung von statistischen Tests mit Angaben von P-Values eine sinnleere Konvention. Ein P-Value ist dann weder absolut (im Sinne einer Irrtumswahrscheinlichkeit) noch relativ (im Sinne eines komparativen Signifikanzmaßes) interpretierbar.

Mit „komparativem Signifikanzmaß" ist die übliche Verwendung von P-Values gemeint im Sinne von: „Je kleiner der P-Value, desto höher die Signifikanz". Diese Eigenschaft eines statistischen Tests und seiner P-Values geht im Fall von nichtzufälligen Stichproben verloren, weil dann die Wahrscheinlichkeit dafür, daß der Test zur Ablehnung der Nullhypothese führt, obwohl diese richtig ist (also die Wahrscheinlichkeit für den Fehler 1. Art), nicht nur prinzipiell unberechenbar ist, sondern schlicht eine sinnlose Größe darstellt. Dieser Sachverhalt liegt in den meisten Studien zur Psychotherapieforschung vor.

Geeignete statistische Verfahren

Die Antwort auf die für jede empirische Studie ganz zentrale Frage, welche statistischen Verfahren zur Auswertung der beobachteten Daten geeignet sind, hängt von drei Kriterien ab: der Skalierung der Effektgrößen, der Qualität der Stichprobe und der Art der Untersuchung (dem gesteckten Untersuchungsziel).

Skalierung der Effektgrößen

Für Effektgrößen wie ganz allgemein für Untersuchungsmerkmale werden die folgenden drei Skalenarten (in aufsteigender Reihenfolge) unterschieden: *Nominalskala* (die Ausprägungen des Merkmals stellen Bezeichnungen dar und dienen der Klassifizierung; Beispiele: Geschlecht, Familienstand), *Ordinalskala* (zwischen den Ausprägungen des Merkmals herrscht eine Rangabstufung oder Ordnungsrelation; Beispiele: Beurteilung einer Leistung, Schwere einer Erkrankung, nahezu alle Items eines Rating-Instrumentes) und *Kardinalskala* (der Unterschied zwischen zwei Ausprägungen läßt sich durch ihre Differenz (*Intervallskala*) oder ihren Quotienten (*Verhältnisskala*) wiedergeben; Beispiele: Temperatur, Länge, Gewicht, Dauer einer Behandlung). Zu jeder Skala gibt es die passenden statistischen Verfahren und Parameter; diese beruhen (unmittelbar oder mittelbar, d.h. unter Anwendung geeigneter Transformationen) bei nominalskalierten Daten auf Häufigkeitsverteilungen der Stichprobe (Beispiele: Kontingenztafeln, Chi-Quadrat-Tests), bei ordinal skalierten Daten auf den Rängen der Beobachtungswerte (Beispiele: Rangkorrelationen, Wilcoxon-Tests) und nur bei kardinal skalierten Daten auf den beobachteten Werten selbst (Beispiele: Übliche Korrelationen, t-Tests, F-Tests).

Qualität der Stichprobe

Auch für Stichproben möchte ich drei Qualitätsstufen (in aufsteigender Reihenfolge) unterscheiden: *Beliebige* oder *„ad-hoc" Stichproben* (an die keinerlei Voraussetzungen gestellt werden), *repräsentative Stichproben* (mit gewissen, mehr oder weniger starken Repräsentativitätsanforderungen) und *Zufallsstichproben* (bei denen jedes Element der Population eine berechenbare und von Null verschiedene Wahrscheinlichkeit besitzt, in die Stichprobe zu gelangen). – In diesem allgemeinen Sinn fallen unter die Zufallsstichproben auch *geschichtete Stichproben* und *Klumpenstichproben*. Die *uneingeschränkte* oder *reine Zufallsauswahl*, bei der jede gleich große Teilmenge (und damit insbesondere jedes Element) der Population die gleiche Wahrscheinlichkeit besitzt, in die Stichprobe zu gelangen, ist eine spezielle Zufallsstichprobe. Auch die Beobachtungen einer Effektgröße in einem oder mehreren Patientenkollektiven stellt eine Zufallsstichprobe dar, wenn man davon ausgehen kann, daß die Beobachtungen voneinander unabhängig sind und jeweils bei jedem Patienten des Kollektivs derselben Verteilung gehorchen, die auch in der betreffenden Population herrscht. – Die Repräsentativität einer Stichprobe wird in der Regel dadurch nachgewiesen, daß gewisse Merkmale wie Alter, Geschlecht, Krankheitsbilder usw. in der Stichprobe mit etwa denselben relativen Häufigkeiten auftreten wie in der Population, aus der die Stichprobe gewählt wurde. Bei der *Quotenstichprobe* werden solche Anteilswerte („Quoten") der Population für die Stichprobe vorgeschrieben, wobei die Erfüllung der Quoten nach Gutdünken vorgenommen wird. Es bleibt anzumerken, daß es sich bei der „Repräsentativität" um einen recht vagen Begriff (jedenfalls keinen statistischen Fachbegriff) handelt.

Art und Ziel der Untersuchung

In der Statistik werden drei Typen von Untersuchungen mit den ihnen zugehörigen statistischen Verfahren betrachtet: deskriptive, explorative und konfirmative Untersuchungen. *Deskriptive Untersuchungen* dienen einer möglichst informativen Beschreibung der beobachteten Datenmenge; sie sind an keine Voraussetzungen über die ihnen zugrundeliegenden Stichproben ge-

bunden; streng genommen beziehen sich alle Aussagen deskriptiver Art nur auf die vorliegenden Beobachtungswerte. *Explorative Untersuchungen* haben den Zweck, Gesetze oder Hypothesen aus den beobachteten Daten zu erkennen, Sachverhalte also, die über die Beobachtungsdaten hinausgehen und typisch oder charakteristisch für den Untersuchungsgegenstand sind; daher setzen solche Untersuchungen nicht zu kleine und (in irgendeiner Form) repräsentative Stichproben voraus. *Konfirmative (explanative) Untersuchungen* dienen der Prüfung, dem statistischen Nachweis (der „Signifikanz") von Gesetzen oder Hypothesen, die in einer Population herrschen, aus der die Stichprobe mit ihren Beobachtungswerten stammt; diese Untersuchungen und ihre statistischen Verfahren (Tests, Schätzungen, Konfidenzintervalle) sind an Zufallsstichproben gebunden. Jede der drei Untersuchungsarten hat ihre eigenen statistischen Methoden; es gibt deskriptive, explorative und konfirmative statistische Verfahren.

Zusammenfassend ist festzustellen, daß bereits bei der Planung einer empirischen Studie dafür zu sorgen ist, daß das gesteckte Untersuchungsziel und die Qualität der Stichprobe zusammenpassen, damit überhaupt adäquate statistische Verfahren existieren und in Abhängigkeit von der vorliegenden Skalenart ausgewählt werden können. Innerhalb der Psychotherapieforschung überwiegen bei weitem solche Studiensituationen, in denen deskriptive und explorative Methoden die adäquaten statistischen Verfahren sind. Diese Methoden findet man z.B. bei Enke et al. (1992), Ferschl (1985), Hoaglin et al. (1983, 1985, 1991), Lebart (1984), Menges (1982), Polasek (1994), Tukey (1977).

Multiple Tests (Simultaneous Inference)

In empirischen Studien zur Psychotherapieforschung, in denen statistische Tests zur Anwendung gelangen, werden in nahezu allen Fällen nicht ein, sondern mehrere, oft sehr viele Tests auf dieselbe Erhebung (Stichprobe) angewandt und die dazugehörigen P-Values bestimmt, wobei in jedem Einzeltest ein Ergebnis als signifikant bzw. hochsignifikant ausgewiesen wird, wenn der betreffende P-Value kleiner als 0.05 bzw. 0.01 ausfällt.

Diese dargestellte Vorgangsweise der Durchführung mehrerer Tests in ein und derselben Stichprobe (multiple Tests, simultane Inferenz) bedarf einer Adjustierung der Signifikanzniveaus (Schranken für die Wahrscheinlichkeiten der Fehler 1. Art) der Einzeltests, die dazu führt, daß das Ergebnis eines Einzeltests erst bei einem wesentlich kleineren P-Value als signifikant bzw. hochsignifikant anzusehen ist als den dafür üblichen Schranken 0.05 bzw. 0.01. Eine solche Korrektur wird in keiner der mir bekannten Studien vorgenommen. Sie ist notwendig, weil die einzelnen Tests mit derselben Stichprobe durchgeführt werden und damit voneinander abhängig sind. Diesem Umstand trägt man in der Theorie multipler Tests dadurch Rechnung, daß man die Einzeltests zu einem Gesamttest (einer ganzen Testprozedur) zusammenfaßt und an den Gesamttest die Forderung stellt, daß seine globale oder multiple Wahrscheinlichkeit für den Fehler 1. Art eine vorgegebene Schranke α nicht überschreitet. Dies wird dadurch erreicht, daß die Einzeltests mit einer deutlich kleineren Schranke für die betreffenden Irrtumswahrscheinlichkeiten als das globale α durchgeführt werden. Die einfachste derartige Vorgangsweise ist die Testprozedur nach Bonferroni: Werden k Einzeltests durchgeführt, so wird nach dieser Prozedur die globale Schranke α zu gleichen Teilen α / k auf die Einzeltests aufgeteilt; dadurch ist (wenn auch auf sehr konservative Weise) gewährleistet, daß der Gesamttest das Signifikanzniveau α einhält. Ein Einzeltest weist also erst dann ein signifikantes bzw. hochsignifikantes Ergebnis auf, wenn sein P-Value kleiner als $0.05/k$ bzw. $0.01/k$ (Bonferroni-Adjustierung) ausfällt. Es gibt andere, ausgefeiltere multiple Testprozeduren, die nicht so konservativ wie die Bonferroni-Testprozedur sind und mit einer weniger starken Adjustierung auskommen; man vergleiche dazu die einschlägige Literatur zu multiplen Tests, z.B. Bauer et al. (1988), Horn und Vollandt (1995), Hsu (1996), Miller (1981).

Über den Zeitfaktor in der Psychotherapieforschung

Es ist nicht meine Aufgabe, auf die tieferen Zusammenhänge einzugehen, in denen die verschiedenen Wirkfaktoren einer Psychotherapie stehen, wie sie auf den Besserungsprozeß eines Patienten einwirken, wie sich dieser Prozeß im Verlauf der Therapie entwickelt und welchen Einfluß dabei insbesondere die Therapiedauer selbst ausübt. Man vergleiche dazu Henseler und Wegner (1993), Leuzinger-Bohleber (1994, 1995), Mertens (1995), Rudolf (1991), Rudolf et al. (1994), Shapiro und Emde (1995), Senf und von Rad (1995), Tschuschke und Czogalik (1990), Tschuschke et al. (1994) und Zielke (1993). Hier geht es vielmehr nur um einige Bemerkungen zur statistischen Erfassung und Beurteilung des Zeitfaktors in der Psychotherapieforschung.

Zunächst steht einmal außer Zweifel, daß „the amount of therapeutic benefit is positively associated with amount of treatment", wie bei Howard et al. (1986), Seite 159, festgestellt und belegt wird. In dieser Arbeit, in der es auf Seite 159 weiterhin heißt: „ To date there has been no systematic attempt to specify the mathematical form of this dose-effect relationship or to determine its accuracy", wird zum erstenmal eine Bestimmung (Schätzung) einer Dosis-Effekt-Beziehung durchgeführt. Seitdem wird als Ergebnis dieser und anderer Studien immer wieder festgestellt, daß zwischen der Dauer (Dosis) einer Therapie und deren Wirkung (Benefit: Nutzen, den der Patient aus der Therapie zieht) ein logarithmischer Zusammenhang besteht: Die Wirkung einer Therapie wächst proportional zum Logarithmus ihrer Dauer. Dieser Zusammenhang wird wohl auch deswegen so gern herangezogen, weil er mit dem in der Ökonomie bekannten Gesetz vom fallenden Grenznutzen übereinstimmt. Es muß festgestellt werden, daß ein Gesetz dieser Form von Howard et al. (1986) so nicht aufgestellt wurde und auch nicht in darauffolgenden Arbeiten empirisch nachgewiesen wird. Mißverständnisse über den Inhalt von Dosis-Effekt-Beziehungen in der Psychotherapieforschung ent-

stehen, wenn man die beiden folgenden Fragen an die Wirkung der Therapiedauer nicht auseinanderhält:

(1) Wie hängen die durch eine Therapie bewirkten Veränderungen bei einem Patienten im Mittel von der Dauer der Therapie ab?
(2) Wie hängt die mittlere Anzahl der Patienten, die in einer Therapie Besserung erfahren, von der Therapiedauer ab?

Die erste Fragestellung ist die schwieriger zu untersuchende. Hier steht die Psychotherapie eines Patienten als ein dynamischer Prozeß im Mittelpunkt sowie die vergleichende Zusammenfassung verschiedener derartiger Prozesse. Zweckmäßige Untersuchungsbedingungen sind Längsschnittanalysen, Einzelfallstudien, prospektive oder auch retrospektive zeitabhängige (auch katamnestische) Erhebungen, die den Verlauf einer Therapie und ihre Wirkung bei einem Patienten erfassen, und die zusammenfassende Beurteilung eines Samples von Therapieverläufen. Einen ersten groben, aber aus statistischen Gründen nicht unproblematischen Einblick kann auch eine einfache ex post Befragung von Patienten (z.B. nach Dauer und Erfolg der Therapie) geben.

Die zweite Fragestellung hat vornehmlich eine Kosten-Nutzen-Analyse (auch im gesundheitspolitischen Sinne) zum Gegenstand. Es wird danach gefragt, nach jeweils wievielen Sitzungen einer Therapie wieviel Prozent der behandelten Patienten eine Besserung erfahren haben. In einer entsprechenden Untersuchung ist zunächst festzulegen, was unter einer Besserung zu verstehen ist (Erfolgskriterium); danach wird dann für verschiedene Zeitpunkte der Behandlung festgestellt, wie groß der Anteil der Patienten ist, die das Kriterium erfüllen; diese Erhebung kann durch Therapeuteneinschätzung oder Patientenselbsteinschätzung während der Behandlung durchgeführt werden. Hier liegen mehrere, verschiedenen Zeitdauern einer Therapie zugeordnete Querschnittserhebungen vor.

In der Studie von Howard et al. (1986) wird der Zeitfaktor (Anzahl der Sitzungen) als „Therapiedosis" in Beziehung gesetzt zum Anteil der Patienten, die eine Besserung erfahren haben. Das Erfolgskriterium „Besserung" wird nach zwei Methoden erhoben, einmal mit Hilfe einer retrospektiven Einschätzung am Ende der Therapie durch Psychotherapieforscher (Auswertung von Krankenblättern), zum anderen durch eine nach jeder Sitzung durchgeführten Patientenselbsteinschätzung. Offensichtlich wird hier eine Untersuchung zur Fragestellung (2) durchgeführt; als Ergebnis erhalten Howard et al. die Dosis-Effekt-Beziehung: Der Anteil der gebesserten Patienten wächst ungefähr proportional zum Logarithmus der Therapiedauer (genauer: „... a log-normal transformation would produce a linear function"; Howard et al. [1986], Seite 160). Dabei wird nur nach „gebessert" und „nicht oder noch nicht gebessert" mit einem festen Besserungskriterium unterschieden. Auch in einer durchgeführten Meta-Analyse innerhalb ihrer Studie kommen Howard et al. zu vergleichbaren Ergebnissen. Danach ergeben sich z.B. bei etwa 75% aller Patienten eine Besserung nach 26 Sitzungen und bei etwa 80% nach 52 Sitzungen. Howard et al. geben diese Ergebnisse an mit dem Zusatz auf Seite 163: „This of course, does not mean that such patients have achieved maximum treatment benefits."

Bei Grawe et al. (1994) werden auf Seite 697 die Ergebnisse von Howard et al. folgendermaßen wiedergegeben: „... bei 75% aller Patienten treten bis zur 26. Therapiesitzung deutliche Besserungen ein, und bei 52 wöchentlichen Therapiesitzungen, also nach einem Jahr, haben die Patienten im Durchschnitt die *maximale Wirkung* [Hervorhebung: B.R.] erreicht." Auf der gleichen Seite heißt es weiterhin bei Grawe et al.: „Allerdings handelt es sich ... um einen logarithmischen Zusammenhang, d.h. der *Zuwachs an Besserung* [Hervorhebung: B.R.] wird mit zunehmender Therapiedauer immer geringer." Im Gegensatz zu Howard et al. wird der entscheidende Unterschied zwischen den beiden Fragestellungen (1) und (2) bei Grawe et al. verwischt: Der Zuwachs an gebesserten Patienten wird mit dem Zuwachs an Besserung verwechselt. Auf dieser Verwechslung beruht unter anderem auch das Urteil Grawes (Grawe et al., 1994, Seite 698): „Therapeuten, die für sich selbst feststellen müssen, daß die Mehrzahl ihrer Therapien länger als 40 Therapiesitzungen dauert, müssen über die Bücher. Sie sind Opfer einer falschen Ausbildung und/oder einer selbst produzierten Realitätsverzerrung." Auch bei Orlinsky et al. (1994) klingt eine solche Verwechslung an, wenn es, wiederum unter Bezug auf Howard et al. (1986) auf Seite 352 heißt: „Overall, the findings indicate that patients tend to improve more the longer they stay in treatment, although the relationship between duration and outcome is clearly far from linear."

In der Arbeit von Seligman (1995) findet man eine Untersuchung, die der Fragestellung (1) nachgeht. Darin wird die Wirkung einer Therapie nach drei globalen Gesichtspunkten unterschieden (Seligman, 1995, Seiten 967f): „Specific improvement (‚How much did treatment help with the specific problem that led you to therapy?'), Satisfaction (‚Overall how satisfied were you with this therapist's treatment of your problems?') and Global improvement (how respondents described their ‚overall emotional state' at the time of the survey compared with the start of treatment)". Die Antworten von 2.846 Patienten auf diese drei Fragen wurden jeweils auf eine 100-Punkte-Skala transformiert und additiv zu einer 300-Punkte-Skala zusammengefaßt und anschließend der Dauer der Therapie gegenübergestellt. In dem Ergebnis (vgl. z.B. Figur 1, Seite 968, in Seligman, 1995) zeigt sich eine deutlich größere Therapiewirkung solcher Therapien, die länger (auch mehr als zwei Jahre) dauern, gegenüber solchen, die eine kürzere Dauer aufweisen. Dieses Ergebnis steht nicht im Widerspruch zu dem Resultat von Howard et al. (1986), da hier im Vergleich zu dort eine andere Fragestellung behandelt wird, hier das Ausmaß der Therapiewirkung in Abhängigkeit von ihrer Dauer, dort der Zuwachs an gebesserten Patienten in Abhängigkeit von der Therapiedauer. Einige mögliche Einwände gegen die Consumer-Report-Studie bringt Seligman in dem genannten Artikel selbst vor und entkräftet sie gleich dort (Seligman, 1995, Seiten 971–974). Weitere kritische Punkte sollen hier nur

angedeutet werden: Der erste betrifft die Frage, ob sich die Skalen für die drei verschiedenen globalen Wirkungsbereiche einfach addieren lassen, der zweite die Einbeziehung völlig unterschiedlicher Therapieformen und Therapeuten in die Erhebung, so daß eine sehr heterogene Stichprobe entsteht, und der dritte den Umstand, daß unter den befragten Patienten (eventuell?) auch solche sind, die sich noch in Therapie befinden, und das damit verbundene Problem einer möglichen (statistisch bedingten) Verzerrung: Unter den in Therapie befindlichen Patienten werden solche mit einer sehr langen Therapiedauer überproportional vertreten sein. Wie stark sich diese Verzerrung bemerkbar macht, hängt nicht nur von dem Anteil der bei der Befragung in Therapie befindlichen Patienten ab, sondern auch von der Frage, wann innerhalb einer Therapie die verschiedenen Veränderungen (Besserungen) stattfinden. Diese Frage läßt sich aber durch die (im wesentlichen) ex post Befragung der Consumer-Report-Studie nicht untersuchen.

Die beiden verschiedenen Fragen an die Wirkung einer Therapiedauer stehen nicht unverbunden nebeneinander. Eine erste Verbindung entsteht durch die Möglichkeit, eine komplizierte Längsschnittanalyse durch eine zeitliche Folge von Querschnittserhebungen zu ersetzen; eine solche Ersetzung ist jedoch an sehr restriktive Voraussetzungen über den stochastischen Prozeß gebunden, der als Modell des therapeutischen Prozesses herangezogen werden kann. Eine zweite, tiefergehende Verbindung kommt durch den Sachverhalt zustande, daß sich die Wirkfaktoren einer Therapie (darunter auch ihre Dauer), die sich im Behandlungsverlauf auf *einen* Patienten auswirken, auf irgendeine Weise auch auf den Anteil der gebesserten Patienten eines ganzen Kollektivs durchschlagen müssen. Mir ist aber unklar, wie dies geschieht, und mir erscheint problematisch, ob dafür eine quantitative Beziehung gefunden werden kann, die erforderlich wäre, um die Ergebnisse der einen auf die der anderen Fragestellung umrechnen zu können. Müßte doch eine solche Beziehung eine Transformationsformel darstellen, die zum Inhalt hat, wie sich der Behandlungsverlauf und Veränderungsprozeß vom Einzelfall (oder der Beobachtung mehrerer Einzelfälle) auf ein Gesamtkollektiv auch quantitativ übertragen läßt.

Literatur

Bauer P, Hommel G, Sonnemann E (1988) Multiple hypothesis testing. Springer, Berlin Heidelberg New York Tokyo

Bergin AE, Garfield SL (1994) Handbook of psychotherapy and behavior change. Wiley, New York

Bortz J (1993) Statistik. Springer, Berlin Heidelberg New York Tokyo

Bortz J, Döring N (1995) Forschungsmethoden und Evaluation. Springer, Berlin Heidelberg New York Tokyo

Bortz J, Lienert GA, Boehnke K (1990) Verteilungsfreie Methoden in der Biostatistik. Springer, Berlin Heidelberg New York Tokyo

Bredenkamp J (1972) Der Signifikanztest in der psychologischen Forschung. Akademische Verlagsanstalt, Frankfurt/Main

Büning H, Trenkler G (1994) Nichtparametrische statistische Methoden. de Gruyter, Berlin

Büning H (1991) Robuste und adaptive Tests. de Gruyter, Berlin

Cohen J (1988) Statistical power analysis for the behavioral sciences. Erlbaum, New York

Conover WJ (1980) Practical nonparametric statistics. Wiley, New York

Enke H, Gölles J, Haux R, Warnecke K-D (Hrsg) (1992) Methoden und Werkzeuge für die exploratorische Datenanalyse in den Biowissenschaften. Fischer, Stuttgart Jena New York

Faller H, Frommer J (Hrsg) (1994) Qualitative Psychotherapieforschung. Grundlagen und Methoden. Asanger, Heidelberg

Ferschl F (1985) Deskriptive Statistik. Physica, Würzburg

Grawe K, Donati R, Bernauer F (1994) Psychotherapie im Wandel. Von der Konfession zur Profession. Hogrefe, Göttingen

Hampel FR, Rousseeuw PJ, Ronchetti EM, Stahel WA (1986) Robust statistics, the approach based on influence functions. Wiley, New York

Hartkamp N, Heigl-Evers A (1995) Feinstrukturen einer analytischen Supervision. Z Psychosom Med Psychoanal 41: 253–267

Henseler H, Wegner P (Hrsg) (1993) Psychoanalysen, die ihre Zeit brauchen. Zwölf klinische Darstellungen. Westdeutscher Verlag, Opladen

Hoaglin D, Mosteller F, Tukey JW (1983) Understanding robust and exploratory data analysis. Wiley, New York

Hoaglin DC, Mosteller F, Tukey JW (Hrsg) (1985) Exploring data, tables, trends, and shapes. Wiley, New York

Hoaglin DC, Mosteller F, Tukey JW (1991) Fundamentals of exploratory analysis of variance. Wiley, New York

Horn M, Vollandt R (1995) Multiple Tests und Auswahlverfahren. Fischer, Stuttgart

Howard KI, Kopta SM, Krause MS, Orlinsky DE (1986) The dose-effect relationship in psychotherapy. Am Psychol 41: 159–164

Hsu J (1996) Multiple comparisons. Theory and methods. Chapman & Hall, London

Huber P (1981) Robust statistics. Wiley, New York

Kächele H, Kordy H (1995) Ergebnisforschung in der psychosomatischen Medizin. In: Uexküll T v (Hrsg) Psychosomatische Medizin. Urban & Schwarzenberg, München Wien

Koch U, Wittmann WW (1990) Evaluationsforschung. Bewertungsgrundlage für Sozial- und Gesundheitsprogramme. Springer, Berlin Heidelberg New York Tokyo

Lang H (Hrsg) (1990) Wirkfaktoren der Psychotherapie. Springer, Berlin Heidelberg New York Tokyo

Lebart L (1984) Multivariate Descriptive statistical analysis. Wiley, New York

Lehmann EL (1986) Testing statistical hypotheses. Wiley, New York

Lehmann EL (1991) Theory of point estimation. Wiley, New York

Leuzinger-Bohleber M (1994) Veränderungen kognitiv-affektiver Prozesse in Psychoanalysen. Versuch einer Kombination von (qualitativer) On-Line- und (quantitativer) Off-Line-Forschung bei der Untersuchung psychoanalytischer Prozesse. In: Faller H, Fromme J (Hrsg) Qualitative Psychotherapieforschung. Asanger, Heidelberg, S 195–228

Leuzinger-Bohleber M (1995) Die Einzelfallstudie als psychoanalytisches Forschungsinstrument. Psyche 49: 434–480

Lienert GA, Raatz U (1994) Testaufbau und Testanalyse. Beltz, Weinheim

Manz R, Schepank H (1993) Köps: Ein Selbstrating-Instrument zu Erfassung körperlicher, psychischer und sozial-kommunikativer Beeinträchtigungen. Z Psychosom Med Psychoanal 39: 1–13

Maxwell SE, Delaney HD (1993) Bivariate median splits and spurious statistical significance. Psychol Bull 113: 181–190
Menges G (1982) Die Statistik. Zwölf Stationen des statistischen Arbeitens. Gabler, Wiesbaden
Mertens W (1995) Warum Psychoanalysen lange dauern. Psyche 49: 405–433
Miller R (1981) Simultaneous statistical inference. Springer, Berlin Heidelberg New York
Miller R (1986) Beyond ANOVA, basics of applied statistics. Wiley, New York
Morrison DE, Henkel RE (eds) (1970) The significance test controversy. Aldine, Chicago
Orlinsky DE, Grawe K, Parks BK (1994) Process and outcome in psychotherapy – Noch einmal. In: Bergin AE, Garfield SL (eds) Handbook of psychotherapy and behavior change. Springer, Berlin Heidelberg New York Tokyo, pp 270–376
Polasek W (1994) EDA – explorative Datenanalyse. Springer, Berlin Heidelberg New York Tokyo
Rudolf G (1991) Die therapeutische Arbeitsbeziehung. Untersuchungen zum Zustandekommen, Verlauf und Ergebnis analytischer Psychotherapie. Springer, Berlin Heidelberg New York Tokyo
Rudolf G, Manz R, Öri Ch (1994) Ergebnisse psychoanalytischer Therapien. Z Psychosom Med Psychoanal 40: 25–40
Rudolf G, Buchheim P, Ehlers W, Küchenhoff J, Muhs A, Pouget-Schors D, Rüger U, Seidler GH, Schwarz F (1995) Struktur und strukturelle Störung (Franz Heigl, einem der Pioniere strukturellen Denkens in der Psychoanalyse zum 75. Geburtstag gewidmet.) Z Psychosom Med Psychoanal 41: 197–212
Rüger B (1996) Induktive Statistik. Oldenbourg, München
Rüger B. Über statistische Methoden in der Psychotherapieforschung. In: Fäh-Barwinski M, Fischer G (Hrsg) Sinn und Unsinn in der Psychotherapieforschung (erscheint 1997)
Rüger U, Senf W (1994) Klinische Bedeutung von Psychotherapie-Katamnesen. Z Psychosom Med Psychoanal 40: 103–116
Scariano SM, Davenport JM (1987) The effects of violations of independence assumptions in the one-way ANOVA. The American Statistician 41: 123–129
Schiepek G (1994) Ist eine systemische Psychotherapieforschung möglich? Z Klin Psychol Psychopathol Psychother 42: 297–318
Schiepek G, Kowalik ZJ (1994) Dynamik und Chaos in der psychotherapeutischen Interaktion. Verhaltenstherapie und psychosoziale Praxis 4: 503–527
Seligman MEP (1995) The effectivness of psychotherapy. The Consumer Reports Study. Am Psychol 50: 965–974
Senf W, Rad M v (1995) Ergebnisforschung in der psychosomatischen Medizin. In: Uexküll T v (Hrsg) Psychosomatische Medizin. Urban & Schwarzenberg, München Wien
Shapiro T, Emde RN (Hrsg) (1995) Research in psychoanalysis: process, development, outcome. Int. Univ. Press, Madison
Steyer R, Eid M (1993) Messen und Testen. Springer, Berlin Heidelberg New York Tokyo
Tress W, Hildenbrand G, Junkert-Tress B, Hartkamp N (1994) Zum Verhältnis von Forschung und Praxis in der analytischen Psychotherapie. Z Psychosom Med Psychoanal 40: 341–352
Tschuschke V, Czogalik D (Hrsg) (1990) Psychotherapie – welche Effekte verändern? Zur Frage der Wirkmechanismen therapeutischer Prozesse. Springer, Berlin Heidelberg New York Tokyo
Tschuschke V, Kächele H, Hölzer M (1994) Gibt es unterschiedlich effektive Formen von Psychotherapie? Psychotherapeut 39: 281–297
Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, MA
Wolf GK (1980) Klinische Forschung mittels verteilungsunabhängiger Methoden. Springer, Berlin Heidelberg New York
Zielke M (1993) Wirksamkeit stationärer Verhaltenstherapie. Beltz, Weinheim

Korrespondenz: Prof. Dr. Bernhard Rüger, Institut für Statistik der Universität München, Ludwigstraße 33, D-80539 München, Bundesrepublik Deutschland.

Rüger Bernhard, Prof. Dr. rer. nat., Diplom-Mathematiker, München: Professor für Statistik am Institut für Statistik der Universität München. Forschungsschwerpunkte und Veröffentlichungen auf den Gebieten Grundlagen statistischer Inferenz, Test- und Schätztheorie, Zeitreihenanalyse und Biometrie.