Methods of Psychological Research Online 1996, Vol.1, No.4
Internet: http://www.pabst-publishers.de/mpr/

Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen

Peter Sedlmeier*

Zusammenfassung

Die Dominanz des Signifikanztests in der psychologischen Datenanalyse könnte den Eindruck erwecken, daß er in der überwiegenden Mehrzahl der Fälle das adäquate Analyseverfahren ist, und/oder daß keine Alternativen existieren. Beide Schlußfolgerungen wären nicht richtig. Probleme bei der Anwendung des Signifikanztests in der psychologischen Forschung werden seit langem ausführlich diskutiert, doch dies hatte erstaunlicherweise kaum Auswirkungen auf die Forschungspraxis. Der hauptsächliche Grund hierfür scheint eine mangelnde Vertrautheit mit alternativen Verfahren zu sein. In diesem Beitrag werden nach einem kurzen Überblick über die kritisierten Unzulänglichkeiten des Signifikanztests solche Verfahren vorgestellt. Darunter fallen Error-Bar-Plots (mit Einschränkungen), Verfahren der Explorativen Datenanalyse, die Berechnung von Effektgrößen und die Metaanalyse. Die einzelnen Verfahren werden anhand von Beispielen illustriert und zum Gebrauch empfohlen.

Schlüsselwörter: Signifikanztest, EDA, Effektgröße, Metaanalyse, Datenanalyse

Abstract

Beyond the Ritual of Significance Testing: Alternative and Supplementary Methods

The prevalence of significance testing in psychological research implies that it suits almost every purpose and/or that alternative procedures for data analysis do not exist. Neither implication is correct. A considerable amount of criticism of significance testing has been published, but so far it has had little impact on actual practice in psychological research. The reason for this seems to be that researchers are in general unfamiliar with alternative procedures. This article briefly reviews the problems of significance testing and then proposes some alternatives: plot-plus-error-bars (with restrictions), exploratory data analysis (EDA), effect sizes, and meta analysis. Examples illustrate the usefulness of these alternatives.

Keywords: Significance test, Exploratory Data Analysis, meta analysis, data analysis.

*Ich danke den Studentinnen und Studenten an der Universität Salzburg und der University of Chicago für ihre kritischen Fragen, sowie Edgar Erdfelder, Gerd Gigerenzer, Wolfgang Hell, Anita Hewer, Detlef Köhlers, Jürgen Locher, Ralph Hertwig, Manfred Wettler und zwei anonymen Reviewern für hilfreiche Rückmeldungen. Diese Arbeit wurde unterstützt durch ein Feodor-Lynen Stipendium der Alexander-von-Humboldt Stiftung und durch ein Habilitationsstipendium der Deutschen Forschungsgemeinschaft.

1 Einleitung

Das Testen von Nullhypothesen dominiert nach wie vor die Auswertung psychologischer Forschungsresultate. Angesichts einer etwa 35 Jahre andauernden und größtenteils unwidersprochenen, substantiellen Kritik der Anwendung des Verfahrens (Carver, 1993) ist das höchst merkwürdig. Meist ist es allerdings bei der Kritik geblieben, d. h., es wurden kaum konstruktive Alternativen angeboten. Eine in den USA weithin rezipierte Ausnahme ist ein kürzlicher Herausgeberbeitrag von Geoffrey Loftus in *Memory & Cognition* (Loftus, 1993a). Loftus (1993a,b) argumentiert, daß p-Werte nahezu nutzlos für die Interpretation psychologischer Daten sind. Statt dessen schlägt er vor, graphische Methoden zu verwenden, die Aussagen über *Muster* und *Größe* von Effekten erlauben. Loftus ist allerdings nicht der erste Herausgeber, der sich gegen den übermäßigen Gebrauch von Signifikanztests ausspricht und Alternativen einfordert. Schon 1970 bemerkten Jürgen Bredenkamp und Hubert Feger in der *Zeitschrift für Sozialpsychologie*, daß das Ergebnis eines Signifikanztests "oftmals inadäquat" sei (Bredenkamp & Feger, 1970). Sie ermuntern explizit zu exakten Replikationen von Experimenten und zur Einsendung von Manuskripten mit nichtsignifikanten Ergebnissen und schlagen vor, Effektgrößen zu berechnen. Letztendlich kommen sie jedoch zu dem etwas deprimierenden Schluß, "...an der augenblicklichen Bevorzugung des Signifikanztests als *dem* statistischen Verfahren nichts ändern zu können..." (S. 45). Gut zehn Jahre später konnten Hager & Westermann (1982) diesen Schluß empirisch untermauern.

Alternativen und Ergänzungen zum Signifikanztesten sind seit langer Zeit bekannt und werden zumindest sporadisch in fast allen neueren Statistikpaketen für PC und Macintosh angeboten (vgl. Butler & Neudecker, 1989). Warum werden sie dann nicht angewandt? Meine hier schon vorweggenommene (Haupt-) Antwort auf diese Frage ist, daß die meisten Psychologen zwar schon von Alternativen gehört haben, aber wenig Konkretes darüber wissen. Dies hat natürlich zur Folge, daß solche Alternativen auch nicht gelehrt werden. In diesem Beitrag sollen die Vorschläge von Bredenkamp, Feger und Loftus aufgegriffen, erweitert und konkretisiert werden. Die Adressaten sind nicht in erster Linie methodisch versierte Leser, denen vieles bekannt sein wird, sondern vor allem "Nicht-Methodiker", der Großteil der an Universitäten oder Forschungsinstituten tätigen Psychologen. Dieser Personenkreis ist möglicherweise mit der Problematik des Signifikanztestens nicht eingehend vertraut. Deswegen wird zunächst die Kritik am Nullhypothesen-Testen in Grundzügen rekapituliert. Sodann, und das ist der Hauptteil des Beitrags, werden einige Vorschläge dazu gemacht, wie Signifikanztesten ergänzt oder auch ganz ersetzt werden kann. Die vorgestellten Methoden (Konfidenzintervalle, Verfahren der Explorativen Datenanalyse und Verfahren zur Berechnung und Integration von Effektgrößen) werden jeweils anhand von Beispielen illustriert. Aus Platzgründen ist dieser Überblick eher knapp gehalten. Die Beispiele sollten aber das Prinzip der Verfahren verständlich machen, und die angegebene Literatur wird in den meisten Fällen weiterhelfen können.

2 Was bedeutet das Ergebnis eines Signifikanztests?

Viele Studenten (und auch einige etablierte Forscher) sind überrascht, wenn sie bemerken, daß es nicht *einen* Signifikanztest, sondern mehrere unterschiedliche gibt (siehe Ostmann & Wutke, 1994, für einen Überblick). Wenn man Statistiklehrbücher für Psychologen liest, scheint es allerdings oft so, wie wenn nur ein einziges Verfahren existierte. Dieses Verfahren ist eine Mixtur verschiedener Ansätze, meist gemischt aus dem von R. A. Fisher entwickelten Signifikanztesten und dem Hypothesentesten von J. Neyman and E. S. Pearson, oft garniert mit Bayesiani-

schen Interpretationen (Acree, 1979; Gigerenzer & Murray, 1987). Der Ansatz von Fisher unterscheidet sich von dem Neyman-Pearson'schen in vielerlei Hinsicht (siehe hierzu Gigerenzer, 1993; Oakes, 1986). Einige dieser Unterschiede seien hier kurz wiederholt: Während bei Fisher nur eine statistische Hypothese, die Nullhypothese (H_0), spezifiziert wird, ist die Alternativhypothese, meist als H_1 bezeichnet, ein fester Bestandteil des Neyman-Pearson Ansatzes. Konzepte wie "β-Fehler" oder "Teststärke" sind somit nur im zweiten Ansatz sinnvoll zu interpretieren. Bei Neyman und Pearson wird klar unterschieden zwischen dem Signifikanzniveau α und dem p-Wert. Während α die Wahrscheinlichkeit dafür bezeichnet, die H_0 ungerechtfertigt zu verwerfen, ist p die Wahrscheinlichkeit dafür, daß das empirische Datum oder ein extremeres Datum gefunden werden kann, wenn die H_0 wahr ist. Den p-Wert erhält man *nach* dem Experiment, α wird *vor* dem Experiment festgelegt. Beim Fisher'schen Verfahren kann hingegen das "Signifikanzniveau" vor und nach dem Test bestimmt werden. Der interessanteste Unterschied zwischen den beiden Ansätzen liegt aber in der Interpretation des Ergebnisses des Signifikanztests. In beiden Ansätzen ist das Ergebnis eines Signifikanztests die Auftretenswahrscheinlichkeit eines Datums unter der Gültigkeit der Nullhypothese – $p(D|H_0)$, der oben erwähnte p-Wert. Ist p kleiner als α so ist das Ergebnis signifikant, andernfalls ist es nicht signifikant. Im Ansatz von Neyman und Pearson erhält das Testergebnis eine "Verhaltens-Interpretation". Bei einem signifikanten Ergebnis sollte man sich so verhalten, als ob die Alternativhypothese wahr sei, bei einem nicht signifikanten Ergebnis, als ob die Nullhypothese zuträfe (vgl. Blackwelder, 1982). Im Ansatz von Fisher kann die H_0 nur verworfen, nicht aber angenommen werden – bei Nichtsignifikanz kann keine Entscheidung getroffen werden (siehe Gigerenzer et al., 1989, über den Wandel in Fisher's eigener Interpretation des p-Werts). Die im folgenden Paragraph besprochenen Interpretationen von p-Werten würde Fisher jedoch nicht unterstützt haben.

2.1 Wie kann man p-Werte mißinterpretieren? – Einige beliebte Varianten

Die Wahrscheinlichkeit eines empirischen Datums (D) unter Gültigkeit der H_0 ist in den meisten Fällen nicht besonders interessant. Weit interessanter wäre es, aufgrund der Kenntnis des Resultats eines Experiments die Antwort auf die Frage nach der Wahrscheinlichkeit von H_0 oder H_1 zu erhalten. Weitere interessante Fragen wären etwa "Wie bedeutsam ist der Effekt?" oder "Wie wahrscheinlich ist es, daß ich in einem zweiten Experiment wieder ein signifikantes Ergebnis erhalte?". Der p-Wert liefert leider auf keine dieser Fragen eine Antwort, wird aber nicht selten so interpretiert, wie wenn er dies täte (z.B. Tversky & Kahneman, 1971; Oakes, 1986). So wird häufig die tatsächlich gefundene Wahrscheinlichkeit, $p(D|H_0)$ mit der inversen Wahrscheinlichkeit $p(H_0|D)$ verwechselt. Gigerenzer (1993) nennt dies "Bayesian wishful thinking", da man mithilfe des Bayes Theorems eine solche inverse Wahrscheinlichkeit berechnen könnte (siehe Kleiter, 1981, für eine Einführung in die Bayes Statistik). Die Bedeutsamkeit eines Effekts, die zweite interessante Frage, hängt neben inhaltlichen Kriterien in erster Linie von seiner Größe ab (siehe Absatz über Effektgrößen). Und schließlich kann die Wahrscheinlichkeit dafür, bei der Wiederholung eines Experiments ein signifikantes Resultat bei identischem α zu replizieren, nur geschätzt werden, wenn vorher eine Schätzung des Populationseffekts vorliegt. Außerdem muß hierzu die Größe der Stichprobe spezifiziert werden. Ein p-Wert alleine liefert diese Informationen nicht.

Meist korrespondiert die substantielle- oder Forschungshypothese mit der H_1, d. h., ein signifikantes Ergebnis wird (in unterschiedlichen Varianten) als Unterstützung dieser Forschungshypothese interpretiert. Wenn aber die Forschungshypothese lautet, daß "kein Unterschied" vorliegt (z.B. zwischen einer Kontrollgruppe und ei-

ner Experimentalgruppe *vor* der experimentellen Manipulation), oder daß "kein Zusammenhang" besteht, wenn sie also mit der H_0 korrespondiert, dann ist besondere Vorsicht geboten. Tatsächlich ist in solchen Fällen die Teststärke, die a priori Wahrscheinlichkeit, ein signifikantes Ergebnis zu erhalten, wenn ein Effekt in der Population vorliegt, oft sehr niedrig; ein nicht signifikantes Ergebnis wird aber trotzdem in solchen Fällen oft als eine Bestätigung der Forschungshypothese interpretiert (Sedlmeier & Gigerenzer, 1989).[1]

2.2 Wann und wie kann der Signifikanztest sinnvoll benutzt werden?

Wenn Forscher das Ergebnis des Signifikanztests falsch interpretieren, so kann das nicht dem Verfahren an sich angelastet werden. Gehen wir einmal von einer korrekten Interpretation aus – wann macht es im Prinzip Sinn, einen Signifikanztest zu rechnen? Oder, anders gefragt – wann gibt uns ein p-Wert die Information, die wir benötigen? Ein p-Wert ist entweder "signifikant" oder "nicht signifikant". Ein Signifikanztest liefert also eine "Ja/Nein" Information. Diese Ja/Nein Information wird in der heutzutage dominierenden Neyman Pearson'schen Fassung des Signifikanztestens als Basis für eine *Handlungsentscheidung*[2] benutzt (Gigerenzer et al., 1989, S. 98ff). Sicherlich haben auch Psychologen Handlungsentscheidungen zu treffen – Soll Therapie A oder Therapie B angewandt werden? Soll die Lernmethode C in den Lehrplan aufgenommen werden? Die Neyman Pearson'sche Fassung des Signifikanztestens beinhaltet allerdings auch, daß man sich über den zu erwartenden Effekt (die Größe des Effekts in der Population) Gedanken macht und abhängig davon die Risiken falscher Entscheidungen abwägt. Das Ergebnis einer solchen Kosten-Nutzen Analyse schlägt sich dann in der Wahl der Stichprobengrößen und der Werte für α und β nieder. Das ist die Theorie – würde sie befolgt, könnte diese Art des Signifikanztestens auch in der Psychologie in manchen Fällen sinnvoll angewandt werden. Wenn allerdings mehrere Studien zu einem Gegenstandsbereich vorliegen – der Regelfall in der Psychologie – dann sollten die Ergebnisse aller relevanten Studien für eine Entscheidung benutzt werden. Die dazu benötigte Methode ist *nicht* das Auszählen von Signifikanzen, sondern die Analyse von *Effektgrößen* (siehe unten).

In der Fisher'schen Version des Signifikanztestens kann der p-Wert als Maß dafür benutzt werden, wie stark der gefundene Wert von der Nullhypothese abweicht (z.B. Gigerenzer et al. 1989, S. 95). Ceteris paribus ist der p-Wert tatsächlich ein Indikator für die Größe eines Effekts, aber zum einen ist seine Interpretation sehr problematisch (siehe oben), zum anderen kann man die gesuchte Information, die Größe eines Effekts, viel einfacher bekommen. Eine Möglichkeit, unter Zuhilfenahme des Ergebnisses eines Signifikanztests Effektgrößen zu berechnen, wird aus dem folgenden allgemein gültigen Gleichungs-Gerüst (Rosenthal & Rosnow, 1991) ersichtlich:

$$\text{Signifikanztest} = \text{Effektgröße x Größe der Studie.}$$

[1] Eine Poweranalyse (Cohen, 1988) ist in einer solchen Situation unabdingbar. Eine solche Analyse sollte auch in allen anderen Fällen, in denen Signifikanztests benutzt werden, durchgeführt werden. Mittlerweile liegt ein kostenlos erhältliches, sehr komfortables Programm hierfür vor (Erdfelder, Faul & Buchner, 1996).

[2] Den zahlreichen Diskussionen und Kontroversen über den Einsatz des (erweiterten) Signifikanztests zum *Test von Theorien* will ich keinen neuen Beitrag hinzufügen (siehe hierzu etwa Bredenkamp, 1972; Westermann & Hager, 1982; Westermann & Hager, 1984; und die entsprechenden Beiträge im 1991er Jahrgang der *Psychologischen Rundschau).* Selbst wenn man der Meinung ist, daß ein Signifikanztest zum Zwecke der Theorienprüfung unbedingt notwendig ist, sind die im Folgenden besprochenen Verfahren als Ergänzungen von großem Wert. Ein weiteres Problem, das hier nicht diskutiert wird, ist die Beurteilung der Repräsentativität von Stichproben. Dieses Problem ist jedoch nicht mit speziellen Verfahren verbunden, sondern tritt immer auf, wenn man generelle Schlußfolgerungen aufgrund von Stichprobenergebnissen zieht.

Weiß man die "Größe" einer Studie, die jeweils als Funktion der Freiheitsgrade oder der Stichprobengröße ausgedrückt werden kann, so kann man aufgrund der Kenntnis des p-Werts die Effektgröße berechnen. Die spezifischen Formeln unterscheiden sich natürlich in Abhängigkeit der verwendeten Teststatistik und Effektgrößenmaße. Der in dem Gleichungs-Gerüst ausgedrückte Zusammenhang ist sehr nützlich, da die meisten Statistikpakete wenig Möglichkeiten für die Berechnung von Effektgrößen bieten, aber Ergebnisse von Signifikanztests sehr ausführlich darstellen. Die Beziehung zwischen dem Ergebnis eines Signifikanztests und Effektgrößen wird später noch ausführlich besprochen.

2.3 Warum ist Signifikanztesten so beliebt?

Ein großer Teil der Kritik am Signifikanztesten ist seiner fehlerhaften Verwendung und Interpretation anzulasten. Allerdings bleibt selbst bei einer korrekten Anwendung und Interpretation nicht allzu viel an Informationsgehalt (siehe für weitere Kritikpunkte zur Theorie und Praxis des Signifikanztestens: Cohen, 1990; 1994; Dar, Serlin & Omer, 1994; Falk & Greenbaum, 1995; Greenwald, 1975; Meehl, 1967; 1978; Morrison & Henkel, 1970; Rosnow & Rosenthal, 1989; Wottawa, 1990). Trotzdem ist Signifikanztesten unglaublich beliebt - warum? Eine pragmatische Nutzung des Signifikanztests für die Berechnung von Effektgrößen scheint nicht der Grund zu sein – man findet Effektgrößen bisher eher selten in psychologischen Fachzeitschriften. Eine plausible Erklärung ist wohl "Tradition" – "jedermann benutzt Signifikanztests, das *muß* einen guten Grund haben". Eine weitere Ursache für die Beliebtheit des Signifikanztestens ist sicher auch die gängige Publikationspraxis. Signifikanz oder Nichtsignifikanz kann über Leben und Tod eines Artikels entscheiden (z.B. Atkinson, Furlong & Wampold, 1982; Bredenkamp, 1972; L. H. Cohen, 1979; Coursol & Wagner, 1986). Der wichtigste Grund scheint mir aber doch zu sein, daß viele Psychologen keine Ausbildung in alternativen Methoden der Datenanalyse bekommen haben. Das Folgende soll nicht als eine solche "Ausbildung" mißverstanden werden. Vielmehr soll der Leser motiviert werden, Methoden anzuwenden, die den individuellen Fragestellungen gerecht werden, Methoden, die die in den Daten tatsächlich vorhandene Information besser repräsentieren können. Genausowenig aber wie Signifikanztesten eine automatische Datenanalyse ermöglicht, bieten diese Methoden oder Methodensammlungen "Kochbuchrezepte".

3 "Error-Bar-Plots" und Konfidenzintervalle

Geoffrey Loftus (1993b) schlägt vor, Signifikanztests mit "Plot-Plus-Error-Bars" (PPE's) zu ersetzen. Was sind PPE's und warum können sie den Signifikanztest ersetzen und darüber hinaus zusätzliche Information liefern? PPE's sind einfache Graphiken, die Mittelwerte und "Error-Bars" für diese Mittelwerte enthalten. Als Error Bars benutzt Loftus meist den Standardfehler. Der Standardfehler ist nichts anderes als eine besondere Variante eines Konfidenzintervalls, dessen exakte Größe von der Art der Stichprobenverteilung des Mittels abhängt. Bei einer Normalverteilung entspricht das Intervall, das durch je einen Standardfehler zu beiden Seiten des Mittels aufgespannt wird, ungefähr einem 67% Konfidenzintervall. Ein Error-Bar-Plot mit je 1.96 Standardfehlern zu beiden Seiten eines normalverteilten Mittels entspricht einem 95% Konfidenzintervall. Ein signifikantes Ergebnis bei einem Test von H_0: μ=0 bei einem zweiseitigen α=.05 wäre gleichbedeutend mit der Tatsache, daß ein 95% Konfidenzintervall für den von uns gefundenen Mittelwert den Wert 0 *nicht* beinhaltet (siehe Huntsberger & Billingsley, 1973, für eine weitergehende Diskussion der Äquivalenz von Konfidenzintervallen und Signifikanztests). Error-Bar-Plots enthalten also im Prinzip die Information, die ein Signifikanztest liefert,

darüber hinaus jedoch auch noch automatisch Mittelwerte und Konfidenzintervalle.

Was sagen uns nun Konfidenzintervalle? Wenn wir unsere Studie sehr oft exakt replizieren, und jedesmal ein 95% Konfidenzintervall berechnen, so werden diese Intervalle in 95% aller Studien das Populationsmittel umschließen, und in 5% aller Studien nicht (siehe Freedman, Pisani, Purves, & Adhikari, 1991, für eine sehr verständliche Diskussion). Dies ist eine Aussage, die zwar etwas informativer, aber ähnlich unbefriedigend ist wie das Ergebnis eines Signifikanztests. Ein weiterer möglicher Nachteil von Konfidenzintervallen, vor allem bei kleineren Stichproben, soll anhand eines Datenbeispiels veranschaulicht werden.

Abbildung 1 zeigt ein PPE, in dem je ein Standardfehler zu beiden Seiten des Mittelwerts abgetragen ist.[3] Die verwendeten Daten stammen aus einer fiktiven "Aufmerksamkeitsstudie" mit "neuropsychologischen Patienten".[4] *Gruppe A* besteht aus 8 Patienten mit Läsionen im Bereich des Frontalhirns. Die "Reaktionszeiten" dieser Patienten in einem bestimmten computergestützten Aufmerksamkeitstest seien (in Millisekunden): 172, 169, 151, 189, 279, 160, 175, und 168. Die entsprechenden Reaktionszeiten bei einer 11 Patienten umfassenden *Gruppe B* mit Läsionen im Stammhirn seien: 194, 172, 213, 203, 180, 203, 203, 195, 182, 198 und 205. Ein t-Test für unabhängige Mittelwertsunterschiede ergibt $t(17)$=-.97, p=.35. Dies könnte zur Schlußfolgerung Anlaß geben, daß kein Unterschied zwischen den beiden Gruppen hinsichtlich der untersuchten Variablen besteht. Betrachtet man die Mittelwerte in den Error-Bar-Plots (Abbildung 1), dann scheinen sich diese auch nicht sehr voneinander zu unterscheiden (183 versus 195). Die Error-Bars selbst sind informativer – sie sind unterschiedlich groß. Das deutet auf eine höhere Varianz in Gruppe A hin. Aus der PPE-Darstellung ist aber nicht ersichtlich, was für diesen Unterschied verantwortlich ist. Diese Information können Verfahren der Explorativen Datenanalyse liefern.

4 Explorative Datenanalyse

Explorative Datenanalyse (EDA) ist, laut Tukey (1977, S. 1), Detektivarbeit – numerische Detektivarbeit, detektivische Zählarbeit oder graphische Detektivarbeit. Es existiert bislang kein allgemeiner Konsens darüber, wo EDA einzuordnen ist, ob sie als eigenständige Sammlung von Statistikverfahren gelten kann oder ein Teil der Deskriptiven Statistik ist. Es gibt auch keine eindeutigen Kriterien dafür, was als EDA-Methode zu betrachten ist und was nicht. Manchmal wird getrennt zwischen EDA, graphischer und robuster Datenanalyse (z.B. Oldenbürger, im Druck). In diesem Beitrag wird der Begriff EDA jedoch in einem sehr umfassenden Sinne gebraucht (siehe auch Polasek, 1988) – im Zweifelsfall werden graphische Analyseverfahren und robuste Datenanalysemethoden der EDA zugerechnet.

Ein wichtiges Merkmal der EDA neben der relativen Einfachheit der verwendeten Verfahren ist das bewußte Einbeziehen von Subjektivität bei der Datenanalyse und -interpretation. Fox and Long (1990) sehen Tukey's Buch (1977) als Ausgangspunkt einer Revolution in der Art und Weise, wie Statistiker über Datenanalyse denken. EDA ist im wesentlichen eine Sammlung von Verfahren zur (semi-) graphischen Beschreibung und Analyse von Daten – das Rüstzeug des "Datendetektivs" zum Auffinden von Mustern, Gesetzmäßigkeiten oder Zusammenhängen (für Übersichten siehe Oldenbürger, im Druck; Smith & Prentice, 1993; Wainer & Thissen, 1993). Dabei sind Überraschungen erwünscht – eine Abbildung ist vor allem dann besonders wertvoll, wenn sie uns zwingt, zu sehen, was wir nie erwartet

[3] Die Graphiken in diesem Beitrag (mit Ausnahme der Stamm & Blatt Diagramme) wurden mit SYGRAPH (Wilkinson, Hill & Vang, 1992) erstellt.

[4] Es wird nicht der Anspruch erhoben, daß die verwendeten Daten repräsentativ für neuropsychologische Patienten sind.

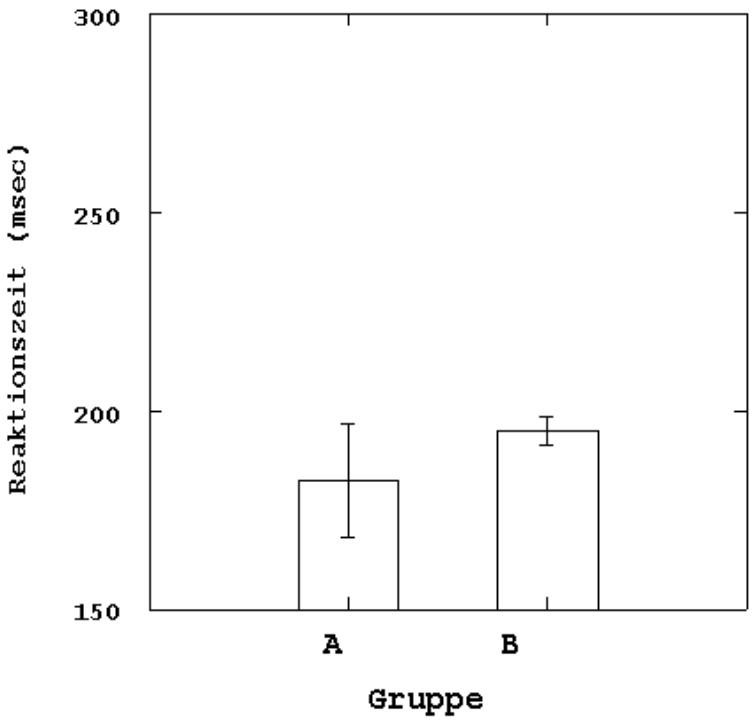


Abbildung 1: PPE (plot plus error-bar) Darstellung des Unterschieds in den Reaktionszeiten (in msec) zwischen Gruppe A und Gruppe B. Die Höhe der Säulen steht für die Mittelwerte und die Länge der "Error-Bars" entspricht jeweils einem Standardfehler zu beiden Seiten der Mittelwerte.

hätten (Tukey, 1977, S. vi).

4.1 Stamm & Blatt Diagramm, Box-Plot

Kehren wir nun zurück zu unserem Vergleich zwischen Gruppe A und Gruppe B. Zwei grundlegende EDA-Verfahren für die Visualisierung von Verteilungen sind "Stamm & Blatt" Diagramm und "Box-Plot". Sehen wir uns zunächst das Stamm & Blatt Diagramm an. Abbildung 2 zeigt einen Sonderfall eines Stamm & Blatt Diagramms, in dem zwei solche Diagramme kombiniert sind, um einen direkten Vergleich von zwei Verteilungen zu ermöglichen. Der "Stamm" in der Mitte des Diagramms enthält den Hundertstelsekunden-Anteil der Reaktionszeiten – "15", der unterste Eintrag im Stamm bedeutet 15 Hundertstelsekunden. Die "Blätter" geben den Rest der Information, den Millisekunden-Anteil. Die schnellste Reaktion in Gruppe A, 151 Millisekunden, wird also repräsentiert als 15 Hundertstelsekunden im Stamm und 1 Millisekunde im Blatt. Man sieht unmittelbar einen Vorteil des Stamm & Blatt Diagramms gegenüber gängigen Histogrammen – das Stamm & Blatt Diagramm konserviert die Rohwerte. Es tut dies in einer sehr übersichtlichen Weise, mit in aufsteigender (oder absteigender) Rangreihe sortierten Zahlen. Rangmaßzahlen wie der Median oder die Quartile, ein wesentlicher Bestandteil von EDA-Prozeduren, können somit leicht gefunden werden. Der Median für Gruppe A ist beispielsweise 170.5 (berechnet als [172+169]/2) und repräsentiert die zentrale Tendenz dieser Verteilung weit besser als der Mittelwert – 183. Bei einigermaßen symmetrischen Verteilungen, wie der für Gruppe B, ist der Unterschied zwischen beiden Maßen in der Regel gering: 198 für den Median versus 195 für den Mittelwert. Ein

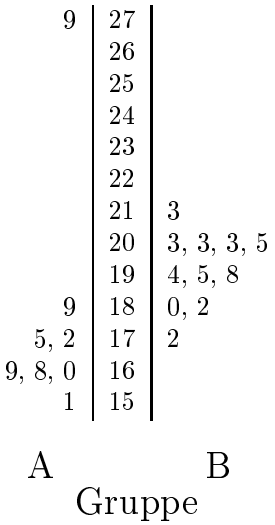
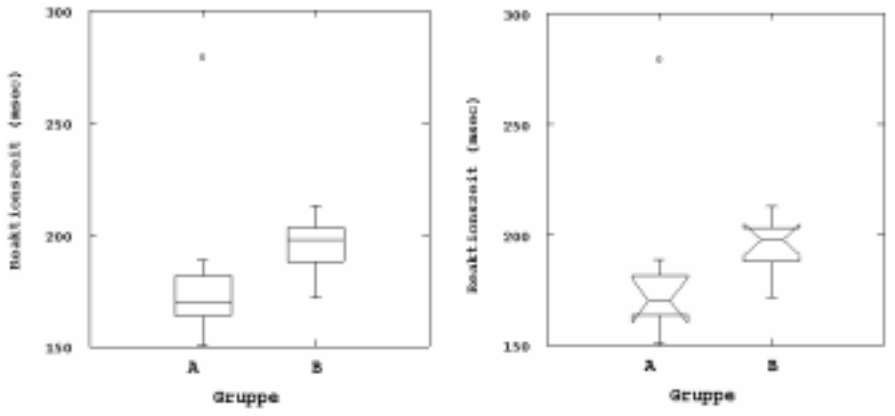


Abbildung 2: Stamm & Blatt Darstellung der Reaktionszeiten (in msec) in Gruppe A und Gruppe B. Der "Stamm" enthält die Hundertstelsekunden und die "Blätter" enthalten den Millisekunden-Anteil für jede Reaktionszeit.

Stamm & Blatt Diagramm kann oft der erste (und manchmal wichtigste) Schritt in der Datenanalyse sein. In unserem Beispiel wird deutlich, daß die Reaktionszeiten der beiden Gruppen sich klar unterscheiden. Diesen klaren Unterschied kann man durch Inspizieren des entsprechenden Error-Bar-Plots oder des p-Werts nicht wahrnehmen. Es wird auch deutlich, warum der Standardfehler für Gruppe A soviel größer ist als der für Gruppe B – der "Ausreißer" in Gruppe A (der Wert 279) beeinflußt Mittelwert und Streuung beträchtlich. Dies wiederum führt zu einem nicht-signifikanten Testergebnis.

Insbesondere bei kleineren Stichproben, in denen Ausreißer oder nichtsymmetrische Verteilungen den Mittelwert stark beeinflussen können, sind Rangmaßzahlen weit weniger verzerrt als auf Mittelung beruhende Maße der zentralen Tendenz einer Verteilung. Box-Plots illustrieren diesen Sachverhalt. Abbildung 3a zeigt die Box-Plots für Gruppe A und Gruppe B. Der Querstrich in der Box markiert jeweils den Median der Verteilung. Die Querstriche an den Enden der Box markieren die "Hinges" oder Quartile der Verteilungen (25% und 75%). Eine Box beinhaltet also (ungefähr) 50% der Werte einer Verteilung. Die Länge dieser Box (Interquartilsabstand) ist völlig unabhängig von extremen Werten, wie z.B. den 279 msec in Gruppe A und ist somit ein resistentes Streuungsmaß. EDA bietet auch eine einfache Methode zur Bestimmung von verschiedenen Klassen von Ausreißern. Ausreißer sind im Box-Plot klar erkennbar, sie liegen außerhalb der kleinen Querstriche ("Whiskers").[5]

[5] In der ursprünglichen Version (Tukey, 1977), die auch heute noch am weitesten verbreitet ist (siehe auch Abbildung 3), werden die Whiskers folgendermaßen bestimmt (für eine theoretische Rechtfertigung siehe z.B. Emerson & Strenio, 1983): Zunächst werden kritische Abstände von den Begrenzungen der Box, sogenannte "inner fences", berechnet, indem man von jeder Begrenzung jeweils 1.5 Interquartilsabstände nach "außen" abträgt. Die zwei Datenpunkte, die jeweils den kritischen Abständen am nächsten sind (auf der Seite, die der Box zugewandt ist) liefern dann die numerischen Werte für die Whiskers. Ein Beispiel – Berechnung des oberen Whiskers für die Gruppe A in Abbildung 3: Die Obergrenze der Box ist 182 (75% Quantil, berechnet als [175+189]/2) und der Interquartilsabstand beträgt 18 (75% Quantil minus 25% Quantil – 182-164). Der kritische obere Punkt ("inner fence") ist somit 209 (182+1.5*18). Nun sucht man den Wert, der (auf der der Box zugewandten Seite) am nächsten an dem kritischen oberen Punkt (209) liegt. Dieser Wert ist in unserem Beispiel 189, und deswegen wird an dieser Stelle auch der kleine "Whisker-Querstrich" eingezeichnet.



(a) (b)

Abbildung 3: Box-Plot-Darstellung der Reaktionszeiten (in msec) in Gruppe A und Gruppe B. Abbildung 3a zeigt die Standardform für Box-Plots. Der Querstrich in der Mitte der Box repräsentiert den Median. Die Box wird begrenzt durch die 25% und 75% Quartile. Ausreißer, wie z.B. der Wert 279 in Gruppe A werden gesondert abgebildet. Abbildung 3b zeigt eine modifizierte Form, die einen "robusten Signifikanztest" ermöglicht. Die Kerben in den Box-Plots entsprechen 95% Konfidenzintervallen. Wenn die Kerben für beide Gruppen sich nicht überlappen, entspricht dies einem signifikanten Testergebnis (bei $\alpha = .05$).

Die Abstände zwischen Median und oberer bzw. unterer Begrenzung der Box geben Aufschluß darüber, ob die Verteilung symmetrisch oder schief ist. In unserem Beispiel wird ersichtlich, daß die Verteilungen beider Gruppen nicht ganz symmetrisch sind – die Verteilung der Werte von Gruppe A ist leicht "linksschief" und die Verteilung der Werte von Gruppe B ist leicht "rechtsschief". Box-Plots können auf verschiedene Weise modifiziert werden (Benjamini, 1988). So kann z.B. Information über die Stichprobengröße in der Breite der Box repräsentiert werden. Abbildung 3b zeigt eine weitere Modifikationsmöglichkeit, (robuste) Konfidenzintervalle, die als Kerben in der Box, mit dem Median als dem Mittelpunkt der Kerbe dargestellt werden (McGill, Tukey & Larson, 1978). Die Länge einer Kerbe in Abbildung 3b repräsentiert jeweils ein 95% Konfidenzintervall. Die Kerben können, wie in diesem Beispiel, auch über die Box hinausgehen. Die Konfidenzintervalle für die beiden Gruppen überlappen sich nicht – dies ist äquivalent mit einem signifikanten Testergebnis. In der Tat ist das Ergebnis eines t-Tests für Mittelwertsunterschiede *ohne* den extremen Wert in Gruppe A $t(16)=4.41$, $p=.0004$. Dies illustriert, wie sehr einzelne extreme Werte, insbesondere bei kleinen Stichproben, parametrische Verfahren beeinflussen können, nicht aber die robusten Verfahren der EDA. Stamm & Blatt Diagramme und Box-Plots sind jedoch nicht speziell nur für kleine Stichproben entwickelt worden, sondern können auch bei relativ großen Stichproben helfen, interessante Informationen gut sichtbar zu machen (siehe Tukey, 1977, für einige Beispiele).

4.2 Weitere EDA-Verfahren

Stamm & Blatt Diagramm und Box-Plots wurden ausführlicher dargestellt, da sie zum einen sehr einfach und zum anderen sehr vielseitig verwendbar sind. Die EDA beinhaltet jedoch eine große und ständig wachsende Anzahl von weiteren Verfahren

Tabelle 1: Ergebnisse für "Gruppe A" in einer hypothetischen Studie. Gezeigt sind Werte von 8 "Patienten" für fünf Variablen.

"Reaktionszeit"	"IQ"	"Genauigkeit"	"Angst"	"Problemlösen"
151	123	50	12	3
160	121	49	11	4
168	115	51	10	4
169	110	50	10	5
172	105	54	11	5
175	103	54	10	5
189	100	55	12	5
279	120	50	20	6

(für detaillierte Beschreibungen siehe neben Tukey, 1977: DuToit, Steyn & Stumpf, 1986; Hoaglin, Mosteller & Tukey, 1983; 1985; Jambu, 1991; Polasek, 1988; Velleman & Hoaglin, 1981). EDA-Verfahren können beliebig erweitert oder ergänzt werden, ja Anwender werden explizit ermutigt, existierende Verfahren weiterzuentwickeln. Bei der Anwendung von EDA-Verfahren geht in der Regel keine Information verloren – sie wird nur in mehrere Komponenten aufgeteilt wie z.B. in *fit* und *residuals* bei der Analyse des Zusammenhangs zweier Variablen oder in *smooth* und *rough* bei der Analyse von Zeitreihen. Diese Aufteilung in jeweils (vorläufige) Modelldaten (fit, smooth) und die Abweichung der empirischen Daten hiervon (residuals, rough) kann Gesetzmäßigkeiten und Zusammenhänge, aber auch charakteristische Abweichungen deutlich sichtbar machen. Im Gegensatz zur herkömmlichen Datenanalyse wird oft ein besonderes Augenmerk auf die individuellen Abweichungen einzelner Datenwerte (z.B. vom Gesamtmedian) gelegt.

Ein weiterer Schwerpunkt der EDA sind multivariate graphische Darstellungen. Zwei Beispiele sollen veranschaulichen, was gemeint ist. Erweitern wir zunächst unsere Beispieldaten für die 8 Patienten der Gruppe A um die Werte aus 4 weiteren Variablen, "IQ", "Genauigkeit", "Angst" und "Problemlösen" (siehe Tabelle 1).

Wenn man Zusammenhänge zwischen mehr als zwei Variablen studieren will, sind einzelne isolierte Streuungsdiagramme oft nicht sehr hilfreich. Eine einfache Kombination individueller Streuungsdiagramme, die "Streuungsdiagramm-Matrix" (scatterplot matrix) vermittelt in solchen Fällen weit mehr Information, da auf einen Blick der Zusammenhang zwischen vielen Variablen sichtbar ist (siehe Cleveland & McGill, 1984, für eine umfassende Diskussion von Streuungsdiagrammen). Abbildung 4 zeigt die Zusammenhänge zwischen den Variablen "Reaktionszeit" (höhere Werte – längere Reaktionszeit), "IQ" (höhere Werte – höherer IQ), "Genauigkeit" (höhere Werte – höhere Genauigkeit) und "Angst" (höhere Werte – größere Angst) für Gruppe A (siehe Tabelle 1).

Sehen wir uns die oberste Reihe der Streuungsdiagramm-Matrix in Abbildung 4 einmal genauer an. Diese Reihe zeigt die Korrelationen zwischen "RT" (Reaktionszeit) und den anderen drei Variablen. Zunächst, in dem Quadrat rechts von "RT", wird deutlich, daß ein verhältnismäßig starker (negativer) linearer Zusammenhang zwischen "RT" (Ordinate) und "IQ" (Abszisse) besteht. Es wird auch ersichtlich, daß ein Patient (der einzelne Kreis rechts oben in dem Quadrat) eine außergewöhnlich lange Reaktionszeit hatte. Gleichzeitig ist der "IQ" Wert dieses Patienten vergleichsweise hoch. Das Ausmaß der Korrelation zwischen "RT" und "IQ" ist deswegen ziemlich niedrig (r=.11), steigt aber drastisch an (zu r=-.93) wenn der Ausreißer von der Analyse ausgenommen wird. Ein ähnliches Bild bietet das dritte Quadrat in der obersten Reihe, das Streuungsdiagramm für "RT" (Ordinate) versus "Genauigkeit" (Abszisse) – abgesehen von dem extremen Wert

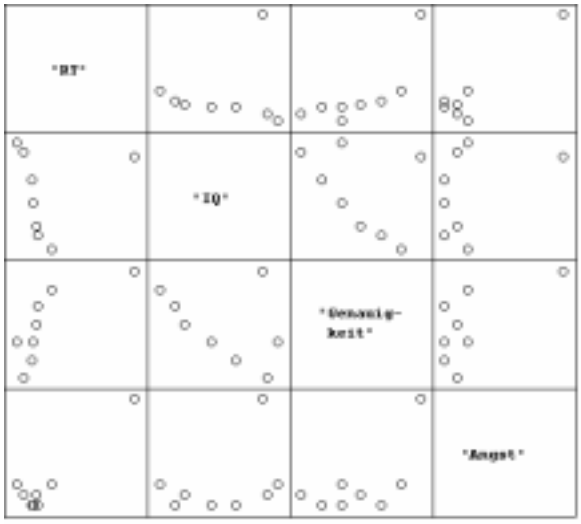


Abbildung 4: Streuungsdiagramm-Matrix (scatterplot matrix), in der gleichzeitig die Zusammenhänge zwischen 4 Variablen, "Reaktionszeit" ("RT"), "IQ", "Genauigkeit" und "Angst" zu sehen sind.

steigt die Genauigkeit mit steigender Reaktionszeit. Das letzte Quadrat in der ersten Zeile zeigt keinen Zusammenhang zwischen "RT" und "Angst". Der entsprechende Korrelationskoeffizient ist jedoch der höchste in der Korrelationsmatrix (r=.93). Entfernt man aber den extremen Wert, so sinkt die Korrelation auf r=-.04. Insgesamt ist ersichtlich, daß, wenn man den Ausreißer entfernt, starke lineare Zusammenhänge zwischen "RT", "IQ" und "Genauigkeit" bestehen, daß aber der Zusammenhang dieser Variablen mit "Angst" verschwindend gering ist. Würde man nur Korrelationskoeffizienten berechnen, käme man auf diametral entgegengesetzte Ergebnisse.

Ist man nicht so sehr an Zusammenhängen zwischen mehreren Variablen über Personen oder Objekte hinweg interessiert, sondern daran, ob und wie sich Personen oder Objekte anhand von mehreren Variablen in Gruppen oder Cluster unterteilen lassen, so hält die EDA auch dafür sehr anschauliche graphische Methoden bereit. Ein Beispiel sind die von Chernoff (1973) eingeführten abstrahierten Gesichter. Jeder Bestandteil eines Gesichts repräsentiert eine Variable, und ein Gesicht repräsentiert die Ausprägungen dieser Variablen für eine Person oder ein Objekt. Würde man nun (aufgrund der Daten in Tabelle 1) nach Subgruppen in Gruppe A suchen, so würde man ähnliche Gesichter zusammengruppieren (siehe Abbildung 5).

Das Ergebnis im Problemlösetest (Tabelle 1, letzte Spalte) wird in Abbildung 5 durch das "Ausmaß des Lächelns" repräsentiert. Patient 8 hat die größte Anzahl von richtigen Lösungen und Patient 1 die geringste. Die Variable "RT" ist durch die Neigung der Augenbrauen und der Augen repräsentiert – Patient 1 hatte die schnellste Reaktionszeit und Patient 8 die langsamste. Das Ergebnis im "Genauigkeitstest" ist durch die Breite der Nasen wiedergegeben – Patient 7 (mit der schmalsten Nase) ist der genaueste. Die Länge der Gesichter zeigt den IQ der Patienten – Patient 1 hat den höchsten Wert und Patient 7 den niedrigsten. Verbleibt noch das Ergebnis des "Angsttests", dargestellt durch die "Haarlänge" – hier hat Patient 8 den höchsten Wert. Für einen genaueren Vergleich, insbesondere wenn die Anzahl der Gesichter größer ist, empfiehlt es sich, diese auszuschneiden und in Gruppen zu sortieren. Aber auch die Anordnung der Gesichter in Abbildung 5 läßt

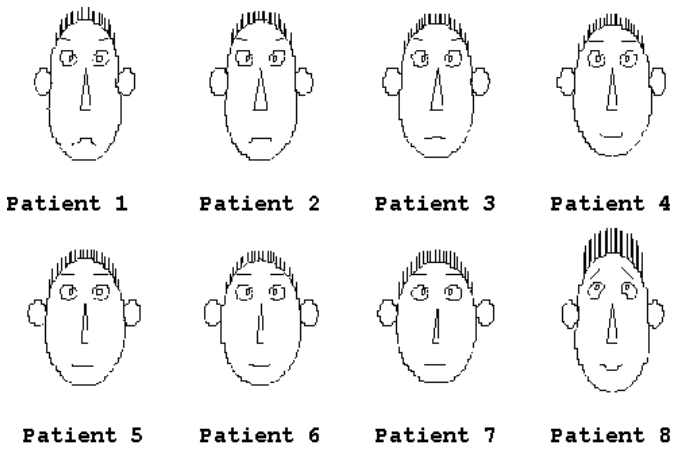


Abbildung 5: Chernoff-Gesichter ermöglichen visuelle Clusterbildung anhand mehrerer Variablen. Ein Gesicht entspricht einem Patienten. Die Bestandteile eines Gesichts repräsentieren verschiedene Variablen und variieren mit den Werten der Variablen.

Gruppierungen erkennen. Zunächst einmal wird deutlich, daß Patient 8 sich stark von allen anderen unterscheidet. Ein zweiter Blick legt nahe, daß seine langsame Reaktionszeit (die Neigung der Augenbrauen) mit seiner erhöhten Angst zu tun haben könnte. Desweiteren könnten die ersten drei Patienten eine Untergruppe bilden. Alle drei haben einen verhältnismäßig hohen IQ, eine schnelle Reaktion und einen eher mäßigen Wert im Genauigkeitstest – im Kontrast zu den Patienten 5, 6 und 7. Die Anzahl der gelösten Probleme legt keine eindeutige Gruppenbildung nahe. Chernoff-Gesichter bieten eine sehr anschauliche Methode für die Darstellung multivariater Zusammenhänge; ihre Nützlichkeit ist allerdings, mehr noch als bei vergleichbaren EDA-Verfahren, von der Variablenzuordnung abhängig.

Trotz ihrer Vielseitigkeit haben auch EDA-Techniken ihre Grenzen. Insbesondere wenn man die in einer Studie gefundenen Ergebnisse hinsichtlich ihrer praktischen Bedeutsamkeit beurteilen will, legt dies oft einen Vergleich mit Ergebnissen aus anderen Studien nahe. Effektgrößen sind hierzu das geeignete Instrumentarium.

5 Effektgrößen

Es existieren mittlerweile viele Möglichkeiten, die Größe eines Effektes anzugeben (für Übersichten siehe Rosenthal, 1993; Rosenthal & Rosnow, 1991; Cohen, 1988). Glücklicherweise sind die meisten dieser Effektgrößen zumindest annäherungsweise ineinander überführbar. Fast alle Maße fallen entweder in die Rubrik "Zusammenhangsmaß oder Maß der erklärten Varianz" oder in die Rubrik " Abstandsmaß" (vgl. Richardson, 1996). Die prinzipielle Äquivalenz dieser beiden Familien von Maßen sei anhand des Korrelationsmaßes r, dem Pearson'schen Korrelationskoeffizient, und des Abstandsmaßes d von Cohen (1962) illustriert (siehe Tatsuoka, 1993, für eine mehr technische Diskussion).

5.1 Äquivalenz von r und d

Ein "Sonnenblumen-Diagramm", ein weiteres EDA-Verfahren, soll illustrieren, warum in vielen Fällen sowohl Abstandsmaße (z.B. d) als auch Korrelationsmaße (z.B. r) gleichwertig verwendbar sind. In Abbildung 6 sind die Ergebnisse der Gruppen A und B in einem fiktiven Problemlösetest (siehe Tabelle 1, letzte Spalte, für die Wer-

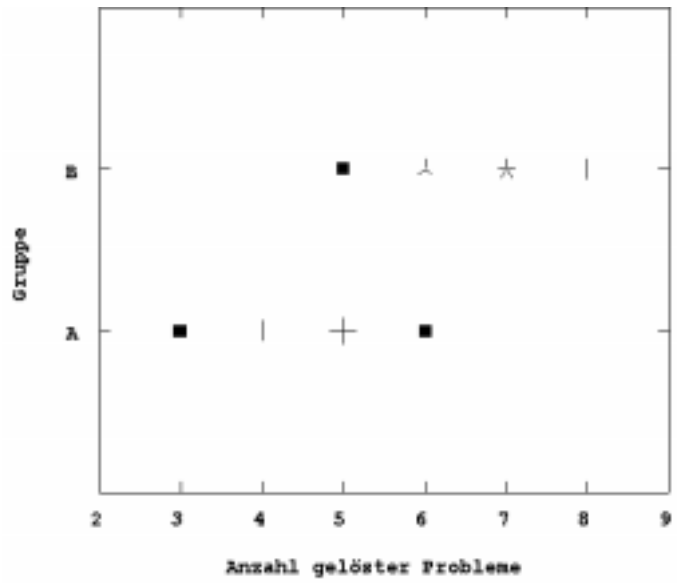


Abbildung 6: Ein "Sonnenblumen-Diagramm" für den Zusammenhang zwischen Gruppenmitgliedschaft (A und B) und der Anzahl der in einem "Problemlösetest" gelösten Aufgaben.

te der Gruppe A) dargestellt. Der Problemlösetest enthält 10 Aufgaben und für jedes Mitglied der Gruppen A und B ist die Zahl der richtigen Lösungen angegeben. In einem gewöhnlichen Streuungsdiagramm würden in diesem Fall viele Werte übereinander gezeichnet werden. Ein Sonnenblumen-Diagramm ermöglicht jedoch auf einfache Weise die Darstellung mehrfach vorkommender Werte, insbesondere bei großen Stichproben (vgl. Cleveland & McGill, 1984). Bei einem Sonnenblumen-Diagramm beginnt man mit einem Punkt für das erste Datum, das in eine bestimmte Kategorie fällt und fügt dann für jedes weitere Datum in dieser Kategorie ein "Blütenblatt" hinzu. Sehen wir uns zunächst die Häufigkeitsverteilung für Gruppe A an (Abbildung 6, untere "Reihe"). Ein Patient hat 3 Lösungen richtig (das kleine Rechteck links unten in Abbildung 6), zwei weitere Patienten haben 4 richtige Lösungen (2 direkt aneinandergefügte Blütenblätter, wiedergegeben als ein längerer vertikaler Strich), vier Patienten haben 5 richtige Lösungen (4 Blütenblätter in Form eines Kreuzes), und ein Patient erreichte 6 Punkte im Problemlösetest. Die obere Reihe von "Sonnenblumen" in Abbildung 6 zeigt die Verteilung der Lösungshäufigkeiten für Gruppe B (ein Patient mit 5, drei Patienten mit 6, fünf Patienten mit 7 und zwei Patienten mit 8 richtigen Lösungen - Werte sind nicht in Tabelle 1 enthalten). Für beide Gruppen lassen sich Mittelwerte und Standardabweichungen berechnen und somit auch ein standardisiertes Abstandsmaß. Aber wo ist die Korrelation? Wenn man die Gruppenzugehörigkeit mit 0 und 1 kodiert, kann man die Korrelation zwischen Gruppenzugehörigkeit und Lösungshäufigkeit berechnen. Das Ausmaß der Korrelation wird generell um so größer, je weniger sich die Werte für die Lösungshäufigkeiten beider Gruppen überlappen. Sie wird 0, wenn diese Werte sich exakt überlappen. Ebenso variiert ein Abstandsmaß mit dem Ausmaß der Überlappung beider Gruppen.

5.2 Berechnung von Effektgrößen

Effektgrößen können auf drei Arten berechnet werden, aus Rohdaten, aus anderen Effektgrößen und aus dem Ergebnis von Signifikanztests (z.B. Friedman, 1968; Rosenthal & Rosnow, 1991). Nehmen wir die Werte in Abbildung 6 als Beispiel. Die Korrelation zwischen Ergebnis im Problemlösetest und Gruppenzugehörigkeit (kodiert als "0" für Gruppe A und als "1" für Gruppe B) ist r=.77. Das Abstandsmaß d (Cohen, 1988, S. 66) wird berechnet als[6]

$$d = \frac{\overline{X}_A - \overline{X}_B}{s_{pooled}}, \text{ wobei } s_{pooled} = \sqrt{\frac{\sum(X_A - \overline{X}_A)^2 + \sum(X_B - \overline{X}_B)^2}{n_A + n_B - 2}}.$$

$\overline{X}_A$ und $\overline{X}_B$ sind die Mittelwerte der beiden Gruppen A und B und n_A und n_B sind die jeweiligen Gruppengrößen. In unserem Beispiel ergeben sich folgende Werte (gerundet): $\overline{X}_A = 4.63$, $\overline{X}_B = 6.73$, $s_{spooled} = 0.91$ und somit d=-2.31.

Das Abstandsmaß d kann nun wieder in ein Korrelationsmaß überführt werden. Dies geschieht mithilfe der Formel

$$r = \sqrt{\frac{d^2 n_A n_B}{d^2 n_A n_B + (n_A + n_B) df}}$$

wobei $df = n_A + n_B - 2$. Das Resultat für unser d=-2.31 ist somit wieder r=.77. Wie kann nun eine Effektgröße aus einem Signifikanztest berechnet werden? Wie schon eingangs erwähnt gilt generell das Gleichungs-Gerüst:

Signifikanztest = Effektgröße x Größe der Studie.

Dieses Gleichungs-Gerüst kann natürlich nicht direkt verwandt werden, sondern hat eher den Status einer "Merkregel". Zwei spezifische Gleichungen, die für die Berechnung von d und r benutzt werden können, sind (Rosenthal & Rosnow, 1991, S. 310):

$$t = d \times \sqrt{\frac{n_A n_B}{n_A + n_B}}$$

und

$$t = \frac{r}{\sqrt{1 - r^2}} \times \sqrt{df},$$

wobei d und der Ausdruck mit den r's jeweils für die Effektgröße stehen und die Größe der Studie jeweils durch eine Funktion von df oder n_A und n_B ausgedrückt ist.

Aufgelöst nach d, bzw r erhalten wir:

$$d = \frac{t\sqrt{n_A + n_B}}{\sqrt{n_A n_B}}$$

und

$$r = \sqrt{\frac{t^2}{t^2 + df}},$$

[6] Die Effektgröße d war von Cohen ursprünglich (für Teststärkeberechnungen) als Populationsmaß definiert worden und bis heute ist der Gebrauch nicht ganz einheitlich. Die hier verwendete Version von d wird manchmal auch as "Hedges's g" bezeichnet (z.B. Rosenthal & Rosnow, 1991, S. 446) und dient zur Schätzung des Populationseffekts. Für sehr kleine Stichproben empfiehlt es sich allerdings, eine Korrekturformel zu verwenden (vgl. Richardson, 1996), da sonst der Populationseffekt überschätzt wird. In diesem Artikel wurde trotz einer verhältnismäßig kleinen Stichprobe die unkorrigierte Version von d verwendet, weil sich verschiedene Zusammenhänge damit leichter und anschaulicher illustrieren lassen.

wobei t für den Wert der t-Statistik mit df Freiheitsgraden steht und n_A und n_B die Größen der beiden Gruppen sind. Der unabhängig berechnete t-Wert ist -4.98 (gerundet). Setzt man diesen t-Wert und die entsprechenden Werte für df (=17), n_A(=8) und n_B(=11) in die beiden Gleichungen ein, so erhält man wieder d=-2.31 und r=.77. Auch diese kurze Übung demonstriert die prinzipielle Äquivalenz von Abstands- und Korrelationsmaßen.

5.3 Interpretation von Effektgrößen

Es gibt keine allgemeingültigen Regeln für die Interpretation von Effektgrößen. In der Regel existiert jedoch eine Forschungstradition, deren Analyse bei der Interpretation helfen kann. Ein Effekt kann dann relativ zu den Effekten, die in dieser Forschungstradition zu finden sind, interpretiert werden. Abhängig von den Fragestellungen, die untersucht werden, kann ein großer Effekt wenig aussagekräftig, und ein kleiner Effekt manchmal schon äußerst wichtig sein (vgl. Rosenthal, 1993). Falls gar keine Anhaltspunkte vorliegen, kann man für eine vorläufige Interpretation die mittlerweile als "Konvention" betrachteten folgenden Werte für d und r verwenden (Cohen, 1992): Als kleine Effekte gelten d=.2 und r=.1, mittlere Effekte sind d=.5 und r=.3, und als große Effekte werden d=.8 und r=.5 betrachtet. Diese Werte waren anfangs nicht empirisch begründet worden, doch die mittleren Effekte scheinen den in verschiedenen Bereichen der Psychologie zu findenden durchschnittlichen Effekten gut zu entsprechen (z.B. Cooper & Findley, 1982; Haase, Waechter & Solomon, 1982; Sedlmeier & Gigerenzer, 1989).

Besonders wichtig ist die Berechnung von Effektgrößen, wenn die H_0 die Operationalisierung der Forschungshypothese ist. Ein nichtsignifikantes Ergebnis sagt in diesem Fall wenig aus, da in solchen Studien die Teststärke oft sehr gering ist (Sedlmeier & Gigerenzer, 1989). Wenn dann auch noch ein substantieller Effekt gefunden wird – schon ein "kleiner" Effekt dürfte in diesem Fall als substantiell gelten – ist es nicht angebracht, das Ergebnis als "Nulleffekt" zu interpretieren.

Wann immer möglich, sollten vor der Interpretation einer Effektgröße die zugrundeliegenden Verteilungen inspiziert werden. Lassen sich starke Asymmetrien oder deutliche Ausreißer erkennen, dann können auch Effektgrößen stark beeinflußt sein. Abhilfen in einem solchen Fall könnten das Nichteinbeziehen von Ausreißern oder eine Transformation der Daten sein. Manchmal werden jedoch auch einfache EDA-Verfahren ausreichen, um solche Daten sinnvoll zu interpretieren.

Bisher haben wir immer über die Analyse der Resultate einzelner Studien gesprochen. Ein konsequenter Schritt von der Effektgrößenberechnung für Einzelstudien hin zur quantitativen Integration einer Reihe von Studien ist die Metaanalyse.

6 Mehr Effektgrößen – Metaanalyse.

Metaanalyse ist ein Sammelname für eine Reihe von Techniken zur quantitativen Integration von Forschungsergebnissen (für Übersichten siehe Bangert-Drowns, 1986; Beelmann & Bliesener, 1994). Ein Beispiel soll demonstrieren, warum die Metaanalyse der herkömmlichen, bei Literaturübersichten oft angewandten "Signifikanz-Zähl" - Methode überlegen ist. Für diese Demonstration habe ich die "psychometrische Metaanalyse"(Hunter & Schmidt, 1990) gewählt, da dieses Verfahren nicht nur Informationen über den Populationseffekt liefert, sondern auch erlaubt, auf elegante Weise Hypothesen über das Zustandekommen der Varianz in den analysierten Studien zu untersuchen.

6.1 Ein Beispiel

Der Input für eine Metaanalyse sind Effektgrößen aus Einzelstudien, aus denen dann, und das ist meist das Hauptanliegen der Metaanalyse, der Populationseffekt geschätzt werden soll. Tabelle 2 zeigt die Effektgrößen (Korrelationen) aus 30 fiktiven Studien zur Wirksamkeit von Aufmerksamkeitstrainings. "Programm X" und "Programm Y" unterscheiden sich nach Ansicht der Experten nur geringfügig und sollen deshalb einer gemeinsamen Analyse unterworfen werden. In jeder der 30 Studien wurde eine Kontrollgruppe mit einer Experimentalgruppe verglichen und die Stichprobengröße pro Gruppe war jeweils n=20.

Zunächst werden in der psychometrischen Metaanalyse die einzelnen Effektgrößen von Artefakten gesäubert, worauf an dieser Stelle nicht eingegangen werden kann (siehe hierzu Hunter & Schmidt, 1990). Danach wird der mit der Stichprobengröße gewichtete durchschnittliche Effekt berechnet: $r_{Mittel} = \sum N_i r_i / \sum N_i$, wobei N_i für die Gesamt-Stichprobengröße in Studie i (konstant N=40 in unserem Beispiel) und r_i für den in Studie i gefundenen Effekt steht. Der gesuchte Mittelwert für unsere 30 Studien ist r_{Mittel}=.36. Beim nächsten Schritt wird klar, warum dieses Verfahren "psychometrische Metaanalyse" genannt wurde. In Analogie zur in der klassischen Testtheorie verwendeten Formel $X_P = T_P + e_P$ – das beobachtete Testergebnis X_P setzt sich zusammen aus dem "wahren Wert" T_P und einem Fehleranteil e_P – verwenden Hunter und Schmidt (1990) die folgende Gleichung: $\sigma_r^2 = \sigma_\rho^2 + \sigma_e^2$ – die Varianz der analysierten Stichproben-Korrelationen setzt sich zusammen aus der Varianz der Populations-Korrelationen und der beim Ziehen von Zufallsstichproben zu erwartenden Fehler-Varianz. Wenn nun die Varianz in den gefundenen Effektgrößen ausschließlich durch den Stichprobenfehler zustande kam, müßte die Varianz der Populations-Korrelationen 0 sein. Dies würde dann heißen, daß die Effektgrößen aus einer einzigen Population (und nicht aus mehreren unterschiedlichen Populationen) stammen. Wie steht es nun in unserem Beispiel damit? Die Fehler-Varianz (σ_e^2) und die Varianz der Stichprobenkorrelationen (σ_r^2) werden folgendermaßen berechnet (Hunter & Schmidt, 1990, S. 108-109 – die Autoren verwenden in ihren Beispielen konsistent das Symbol σ anstelle von s):

$$\sigma_e^2 = \frac{(1 - r_{Mittel}^2)^2}{\overline{N} - 1}$$

und

$$\sigma_r^2 = \frac{\sum [N_i (r_i - r_{Mittel})^2]}{\sum N_i},$$

wobei $\overline{N}$ der Mittelwert aller Stichprobengrößen pro Studie ist. In unserem Beispiel (mit den Werten aus Tabelle 2) ist σ_e^2 =.0194 und die mit der Stichprobengröße pro Studie gewichtete Varianz der Stichproben-Korrelationen ist σ_r^2 =.0246. Die Varianz der Populations-Korrelationen ist somit σ_ρ^2 =.0052. Diese Varianz ist zwar verhältnismäßig klein, schließt aber nicht aus, daß der gefundene durchschnittliche Effekt nicht einen, sondern mehrere Populationseffekte repräsentiert. Der nächste Schritt ist deshalb, nach theoretisch fundierten Moderatorvariablen zu suchen. Eine solche Moderatorvariable in unserem Beispiel ist die Art des angewandten Programms, X oder Y. Die Ergebnisse der entsprechenden Analyse dieser zwei Subgruppen sind, zusammen mit den ursprünglichen Ergebnissen, in Tabelle 3 aufgelistet.

Wenn man für die beiden Programme getrennte Analysen durchführt, werden die Varianzen der Populations-Korrelationen deutlich kleiner (-.0003 für Programm X und .0002 für Programm Y) als die Varianz der Populations-Korrelationen für alle 30 Studien (unterste Zeile in Tabelle 3).[7] Dies deutet darauf hin, daß Programm

[7] Die negative Varianz der Populationskorrelationen für Programm X (σ_ρ = -0.0003) ist zurückführbar auf den Schätzfehler bei der Bestimmung der Varianz der Stichprobenkorrelatio-

Tabelle 2: Ergebnisse aus 30 hypothetischen Studien, in denen Programm X, bzw. Programm Y mit je einer Kontrollgruppe verglichen wurde. Der Treatment-Effekt (Ergebnis für Trainingsgruppe minus Ergebnis für Kontrollgruppe) ist als r wiedergegeben. Die Werte sind Zufallsziehungen aus zwei Stichprobenverteilungen mit den vorgegebenen Mittelwerten ρ=.24 und ρ=.44 für Programm X und Programm Y respektive.

r	Programm	signifikant (α=.05, zweiseitig)
0.41	X	ja
0.63	Y	ja
0.50	Y	ja
0.52	Y	ja
0.43	Y	ja
0.02	X	nein
0.39	X	ja
0.53	X	ja
0.31	Y	nein
0.36	Y	ja
0.43	X	ja
0.15	X	nein
0.33	X	ja
0.33	X	ja
0.32	Y	ja
0.22	X	nein
0.68	Y	ja
0.20	X	nein
0.18	X	nein
0.33	Y	ja
0.62	Y	ja
0.21	Y	nein
0.45	X	ja
0.39	Y	ja
0.44	Y	ja
0.41	X	ja
0.11	X	nein
0.14	X	nein
0.46	Y	ja
0.33	Y	ja

Tabelle 3: Ergebnisse der psychometrischen Metaanalyse (gerundet) aufgeteilt nach "Programm X" versus "Programm Y". Die Ergebnisse für die Gesamtgruppe ("kombiniert") sind zum Vergleich nochmals dargeboten.

	Programm X	Programm Y	kombiniert
r_{Mean}	.29	.44	.36
σ_r^2	.0212	.0169	.0246
σ_e^2	.0215	.0167	.0194
σ_ρ^2	-.0003	.0002	.0052

$\rho = .24$		$\rho = .44$
2	0	
8, 5, 4, 1	1	
2, 0	2	1
9, 3, 3	3	1, 2, 3, 3, 6, 9
5, 3, 1, 1	4	3, 4, 6
3	5	0, 2
	6	2, 3, 8
Programm X		Programm Y

Abbildung 7: Stamm-und-Blatt Darstellung der Ergebnisse aus 30 hypothetischen Studien, in denen Programm X, bzw. Programm Y mit je einer Kontrollgruppe verglichen wurden (jeweils 15 Studien). Der Treatment-Effekt (Ergebnis für Trainingsgruppe minus Ergebnis für Kontrollgruppe) ist als r wiedergegeben. Die Werte sind Zufallsziehungen aus zwei Stichprobenverteilungen mit den Mittelwerten r=.24 und r=.44 für Programm X und Programm Y respektive.

X weniger wirksam ist als Programm Y. In der Tat wurden die Daten aus zwei unterschiedlichen Stichprobenverteilungen für r mit den Mittelwerten ρ=.24 und ρ=.44 für Programm X, bzw Programm Y mittels einer Computersimulation generiert.[8] Abbildung 7, eine Stamm & Blatt Darstellung der Information in Tabelle 2 zeigt (neben der Demonstration, daß ein Stamm & Blatt Diagramm eine viel kompaktere und übersichtlichere Darstellung erlaubt als eine Tabelle) eine für manche vielleicht erstaunliche Tatsache, die man leicht aus dem Blick verliert, wenn man die Ergebnisse von Einzelstudien beurteilt: Die Variation, die alleine aus dem Stichprobenfehler herrührt, ist enorm. Für Programm X variieren die Werte zwischen r=.02 und r=.53 und für Programm Y ergab die Simulation Werte zwischen r=.21 und r=.68. Diese kleine Demonstration verdeutlicht, daß das alleinige Auszählen von Signifikanzen keine zufriedenstellenden Ergebnisse liefern kann. Sie zeigt aber auch den Grund für die Wichtigkeit von Replikationen.

6.2 Probleme der Metaanalyse

Kritiker, aber auch überzeugte Anhänger der Metaanalyse, haben auf eine Reihe von (oft lösbaren) Problemen dieses Ansatzes hingewiesen. Das "Müll rein Müll raus Problem" thematisiert die unterschiedliche Qualität von Studien. Methodisch sehr schwache Studien sollten das Ergebnis einer Metaanalyse weniger stark bestimmen (verzerren) als methodisch "saubere" Studien. In diesem Fall bieten sich zumindest zwei Lösungen an, (i) das Benutzen von Ausschlußkriterien oder (ii) die Einführung einer Moderatorvariablen, anhand derer die Studien nach ihrer methodischen Qualität kodiert werden (was eine getrennte Analyse ermöglicht, falls die Effekte für die Subgruppen sehr unterschiedlich sind). Das "Abhängigkeits-Problem" entsteht, wenn mehrere, nicht aus unabhängigen Stichproben gewonnene Effektgrößen pro Studie in die Analyse eingehen. Vor allem, wenn eine einzelne Studie viele Effektgrößen beisteuert, kann die durchschnittliche Effektgröße, das Hauptergebnis der Metaanalyse, stark verzerrt sein. Die Beschränkung auf eine Effektgröße pro Studie kann manchmal das Problem lösen. Das "Äpfel und Birnen"-Problem kann

nen. Diese geschätzte Varianz wird in der Regel etwas fehlerbehaftet sein, solange die Anzahl der Studien nicht gegen unendlich geht (siehe Hunter & Schmidt, 1990, S. 109-110).

[8] Zunächst wurden zufällig je 15 Stichproben aus einer nichtzentralen t-Verteilung mit 38 *df* gezogen. Die Werte für die Nichtzentralitätsparameter waren 1.5 für "Programm X" und 3 für "Programm Y". Sodann wurden mittels der Formel $r = (t^2/(t^2 + df))^{1/2}$ die t-Werte in Korrelationen transformiert. Alle Berechnungen wurden mit Lisp-Stat (Tierney, 1990) durchgeführt.

sowohl für unabhängige als auch für abhängige Variablen auftreten. Die Analyse einer undifferenzierten Sammlung von Studien zu einem Thema kann zur Nicht-Interpretierbarkeit der resultierenden mittleren Effektgröße führen. Eine Abhilfe bietet die Einteilung in Subgruppen anhand einer oder mehrerer unabhängiger Variablen (z.B. Geschlecht). Gänzlich uninterpretierbar kann das Ergebnis einer Metaanalyse sein, wenn mehrere sehr unterschiedliche abhängige Variablen (z.B. Fremdrating und Lautes-Denken Protokolle) in die Analyse eingehen. In solchen Fällen dürften nach abhängigen Variablen getrennte Analysen der einzige Ausweg sein.

Eine Metaanalyse ist besser als die "Signifikanz-Zähl" Methode, sie ist aber, wie alle hier besprochenen Verfahren, kein automatisches Datenanalyseinstrument. Es gibt sicher Fälle in denen es keinen Sinn macht, eine Metaanalyse durchzuführen. Generell wird das Ergebnis einer Metaanalyse um so befriedigender sein, je präziser die in ihr verfolgte Fragestellung war. Die Bewertung des Ergebnisses wiederum hängt von einer profunden Kenntnis des analysierten inhaltlichen Bereichs ab.

7 Fazit

Ausgehend von einer Kritik des Signifikanztestens wurden verschiedene Alternativen (oder Ergänzungen) dazu vorgeschlagen, insbesondere die Verwendung von Verfahren der EDA und die Berechnung von Effektgrößen. Obwohl ein Teil der hier vorgestellten Verfahren explizit als "explorativ" betitelt ist, heißt das nicht, daß diese Verfahren auf die *Hypothesenfindung* beschränkt bleiben müssen und nicht zur *Hypothesenprüfung* eingesetzt werden können. Eine klare Trennung in einen "Entdeckungskontext", in dem die Hypothesenfindung stattfindet und einen "Begründungskontext", in dem die Hypothese dann geprüft wird, entspricht sowieso nicht der Forschungspraxis (vgl. Gigerenzer, 1991). Die hier vorgestellten Verfahren haben einen berechtigten Platz in allen Stadien psychologischer Forschung (vgl. Erdfelder, 1994). Demgemäß sollten sie auch breiteren Raum in der Statistikausbildung finden. Insbesondere EDA-Verfahren könnten schon im schulischen Stochastikunterricht gewinnbringend eingesetzt werden (Biehler, 1987; Dunkels, 1987).

Dieser Artikel ist ein Versuch, den Leser zur differenzierten Anwendung alternativer Methoden zu motivieren. Die hier vorgestellten Verfahren sollten nicht dazu verführen, sie solange "durchzuprobieren", bis *irgendein* Effekt gefunden ist. Im Zweifelsfall ist es sicher das beste, eine Studie zu replizieren. Es sollte auch klar geworden sein, daß es keine Patentlösungen für die Analyse psychologischer Daten gibt. Der Wunsch nach Patentlösungen ist wohl ein weiterer Grund für die Beliebtheit des Signifikanztest-Rituals (Salsburg, 1985).[9] EDA-Verfahren und auch Effektgrößen kommen diesem Wunsch nicht sehr entgegen. Bei jeder Interpretation von Daten ist explizit subjektives Urteil mit im Spiel. Was angestrebt werden kann, ist ein durch eine geeignete Datenanalyse gewonnener Konsens unter den Experten in einem bestimmten Gebiet, nicht aber das Ersetzen eines Rituals durch ein anderes.

[9] Könnte der Signifikanztest nicht auch so erfolgreich sein, weil im Gegensatz zur Interpretation von EDA-Resultaten oder Effektgrößen Subjektivität keine Rolle spielt, weil er "objektive" Ergebnisse liefert? Tatsächlich ist Signifikanztesten auch mit einer Reihe subjektiver Entscheidungen verbunden (vgl. Berger & Berry, 1988). Zunächst muß ein geeigneter Test ausgewählt werden – muß ich z.B. aufgrund des Skalenniveaus der abhängigen Variablen (z.B. Werte auf einer Rating-Skala) einen parameterfreien Test benutzen oder kann es auch ein gängiges parametrisches Verfahren sein? Sind andere Anwendungsvoraussetzungen wie etwa Varianzengleichheit, Normalverteilung in der Population usw. erfüllt? Wie soll ich mein α und mein β wählen? Und last not least – Wie beurteile ich meinen p-Wert? Die Art und Weise, wie Signifikanztesten manchmal betrieben wird, läßt vergessen, daß auch der Signifikanztest kein automatisiertes Datenanalyse-Instrument ist.

Literatur

[1] Acree, M. C. (1979). Theories of statistical inference in psychological research: A historico-critical study. *Dissertation Abstracts International, 39*, 5073B. (University Microfilms No. 7907000)

[2] Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29*, 189-194.

[3] Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99*, 388-399.

[4] Beelmann, A. & Bliesener, T. (1994). Aktuelle Probleme und Strategien der Meta-analyse. *Psychologische Rundschau, 45*, 211-233.

[5] Benjamini, Y. (1988). Opening the box of a boxplot. *The American Statistician, 42*, 257-262.

[6] Berger, J. O. & Berry, D. A., (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76*, 159-165.

[7] Biehler, R. (1987). Exploratory data analysis and the secondary stochastics curriculum. In R. Davidson & J. Swift (Eds.). *The Proceedings of the Second International Conference on teaching statistics* (S. 79-85). Victoria, B. C.: University of Victoria.

[8] Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials, 3*, 345-353.

[9] Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt/Main: Akademische Verlagsgesellschaft.

[10] Bredenkamp, J., & Feger, H. (1970). Kriterien für die Entscheidung über die Aufnahme empirischer Arbeiten in die Zeitschrift für Sozialpsychologie. *Zeitschrift für Sozialpsychologie, 1*, 43-47.

[11] Butler, D. L., & Neudecker, W. (1989). A comparison of inexpensive statistical packages for microcomputers running MS-DOS. *Behavior Research Methods, Instruments, & Computers, 21*, 113-120.

[12] Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.

[13] Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association, 60*, 361-368.

[14] Cleveland, W. S. & McGill, R. (1984). The many faces of a scatterplot. *Journal of the American Statistical Association, 79*, 807-822.

[15] Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

[16] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

[17] Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

[18] Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

[19] Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*, 997-1003.

[20] Cohen, L. H. (1979). Clinical psychologists' judgment of the scientific merit and clinical relevance of psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 47*, 421-423.

[21] Cooper, H. & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8*, 168-173.

[22] Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta analysis bias. *Professional Psychology: Research and Practice, 17*, 136-137.

[23] Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75-82.

[24] Dunkels, A. (1987). EDA in the primary classroom-graphing and concept formation combined. In R. Davidson & J. Swift (Eds.). *The Proceedings of the Second International Conference on teaching statistics* (S. 79-85). Victoria, B. C.: University of Victoria.

[25] DuToit, S. H. C., Steyn, A. B. W., & Stumpf, R. H. (1986). *Graphical exploratory data analysis*. New York: Springer.

[26] Emerson, J. D. & Strenio, J. (1983). Boxplots and batch comparison. In: D. C. Hoaglin , F. Mosteller & J. W. Tukey (Eds.). *Understanding robust and exploratory data analysis*. New York: Wiley.

[27] Erdfelder, E. (1994).Erzeugung und Verwendung empirischer Daten. In T. Herrmann & W. H. Tack (Hrsg.). *Methodologische Grundlagen der Psychologie* (Enzyklopädie der Psychologie, Themenbereich B, Serie I, Band 1, S.47-97). Göttingen: Hogrefe.

[28] Erdfelder, E., Faul, R., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1-11.

[29] Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, 5, 75-98.

[30] Fox, J. & Long, J. S. (1990). (Eds.). *Modern methods of data analysis*. Newbury Park: Sage.

[31] Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). *Statistics* (2nd ed.). New York: Norton.

[32] Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, 4, 245-251.

[33] Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*, 254-267.

[34] Gigerenzer , G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Ed.), *A handbook for data analysis in the behavioral sciences: Methodological issues* [S. 311-339]. Hillsdale, NJ: Erlbaum.

[35] Gigerenzer, G., & Murray, D. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

[36] Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kr; Krüger, L. (1989). *The empire of chance*. Cambridge: Cambridge University Press.

[37] Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

[38] Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How signifikant is a significant difference? Average effect size of research in Counseling Psychology. *Journal of Counseling Psychology, 29*, 58-65.

[39] Hager, W. & Westermann, R. (1982). Die Elle - 10 Jahre danach. *Zeitschrift für Sozialpsychologie, 13*, 250-252.

[40] Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

[41] Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1985). *Exploring data tables, trends, and shapes*. New York: Wiley.

[42] Hunter, J. E. & Schmidt F. L. (1990). *Methods of meta-analysis*. Newbury Park: Sage.

[43] Huntsberger, D. V. & Billingsley, P. (1973). *Elements of statistical inference* (3rd ed). Boston: Allyn and Bacon.

[44] Jambu, M. (1991). *Exploratory and multivariate data analysis*. Boston: Academic Press.

[45] Kleiter, G. D. (1981). *Bayes Statistik*. Berlin: De Gruyter.

[46] Loftus, G. R. (1993a). Editorial comment. *Memory & Cognition*, *21*, 1-3.

[47] Loftus, G. R. (1993b). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments & Computers, 25*, 250-256.

[48] McGill, R., Tukey, J. W. & Larson, W. A. (1978). Variations of box plots. *The American Statistician, 32*, 12-16.

[49] Meehl, P. E. (1967). Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science, 34*, 103-115.

[50] Meehl, P. E. (1978). Theoretical risks and tabular asteriscs: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

[51] Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*: Chicago: Aldine.

[52] Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

[53] Oldenbürger, H.-A. (im Druck). Exploratorische, graphische und robuste Datenanalyse. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.). *Handbuch Quantitative Methoden*. Weinheim: Psychologie Verlags Union.

[54] Ostmann, A. & Wutke, J. (1994). Statistische Entscheidung. In T. Herrmann & W. H. Tack (Hrsg.). *Methodologische Grundlagen der Psychologie* (Enzyklopädie der Psychologie, Themenbereich B, Serie I, Band 1, S.694-737). Göttingen: Hogrefe.

[55] Polasek, W. (1988). *EDA*. Heidelberg: Springer.

[56] Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers, 28*, 12-22.

[57] Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Ed.), *A handbook for data analysis in the behavioral sciences: Methodological issues* [S. 519-559]. Hillsdale, NJ: Erlbaum.

[58] Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

[59] Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

[60] Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician, 39*, 220-223.

[61] Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 107, 309-316.

[62] Smith, A. F. & Prentice, D. A. (1993). Exploratory data analysis. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Statistical issues* [S. 349-390]. Hillsdale, NJ: Erlbaum.

[63] Tatsuoka, M. (1993). Elements of the general linear model. In: G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Statistical issues* [S. 3-41]. Hillsdale, NJ: Lawrence Erlbaum..

[64] Tierney, L. (1990). *Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics*. New York: Wiley.

[65] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[66] Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 73,* 105-110.

[67] Velleman, P. F. & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston, MA: Duxbury Press.

[68] Wainer, H. & Thissen, D. (1993). Graphical data analysis. In: G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Statistical issues*. Hillsdale, NJ: Lawrence Erlbaum. [S. 391-457].

[69] Westermann, R. & Hager, W. (1982). Entscheidung über statistische und wissenschaftliche Hypothesen: Zur Differenzierung und Systematisierung der Beziehungen. *Zeitschrift für Sozialpsychologie*, *13*, 13-21.

[70] Westermann, R. & Hager, W. (1984). Zur Verwendung von Effektgrößen in der theorie-orientierten Sozialforschung. *Zeitschrift für Sozialpsychologie*, *15*, 159-166.

[71] Wilkinson, L., Hill, M. & Vang, E. (1992). *SYSTAT: Graphics, Version 5.2 Edition.* Evanston, IL: SYSTAT, Inc.

[72] Wottawa, H. (1990). Einige Überlegungen zu (Fehl-) Entwicklungen der psychologischen Methodenlehre. *Psychologische Rundschau*, *41*, 84-107.

Submitted June 13, 1996
Accepted July 30, 1996

