

Edgar Erdfelder, Axel Buchner, Franz Faul und Martin Brandt

## GPOWER: Teststärkeanalysen leicht gemacht

### Zusammenfassung

Statistische Tests können mehr Schaden anrichten als Nutzen bringen, wenn die Stärke (*Power*) eines Tests in Anwendungen unberücksichtigt bleibt. GPOWER<sup>1</sup> ist ein Computerprogramm, das die Durchführung von Teststärkeanalysen für ein sehr breites Spektrum gebräuchlicher statistischer Tests erlaubt. Allerdings ist das Programm in der bisherigen Form nur bei Standardtests leicht anwendbar; Tests, die nicht zu dieser Klasse zählen, erfordern Zusatzkenntnisse, die in diesem Beitrag für eine Reihe von *z*-Tests und *F*-Tests vermittelt werden.

### Die Notwendigkeit der Teststärkekontrolle

Bei statistischen Entscheidungen zwischen einer Nullhypothese ( $H_0$ ) und einer Alternativhypothese ( $H_1$ ) können zwei Arten von Fehler auftreten: Zum einen kann  $H_0$  gelten, aber es wird fälschlich zugunsten von  $H_1$  entschieden (Fehler 1. Art oder  $\alpha$ -Fehler), zum anderen kann  $H_1$  gelten, aber es wird fälschlich zugunsten von  $H_0$  entschieden (Fehler 2. Art oder  $\beta$ -Fehler). Lange Zeit hat in der Psychologie der auf Sir Ronald Aylmer Fisher zurückgehende Nullhypothesen-Signifikanztest dominiert, der ausschließlich die  $\alpha$ -Fehlerwahrscheinlichkeit kontrolliert und  $\beta$  sowie dessen Komplement  $1 - \beta$ , die so genannte Stärke oder *Power* des statistischen Tests, unberücksichtigt lässt. Dies hat zur Konsequenz, dass Nullhypothesen zwar mit kontrollierten Fehlerraten verworfen, aber nie angenommen werden können. Bredenkamp (1969, 1972) hat gezeigt, dass dieser Sachverhalt insbesondere dann fatale Konsequenzen hat, wenn wissenschaftliche Hypo-

---

<sup>1</sup> GPOWER ist ein Freeware-Computerprogramm. Es darf nicht zu kommerziellen Zwecken weiterverbreitet werden. Die jeweils aktuellen Versionen kann man kostenlos für verschiedene Plattformen beziehen über die URL <http://www.psych.uni-duesseldorf.de/aap/projects/gpower/>

thesen mittels statistischer Tests geprüft werden. Dies ist in der psychologischen Forschung nach wie vor der Regelfall. Meistens implizieren psychologische Hypothesen statistische Alternativhypothesen, etwa in Form einer Erwartungswertdifferenz zwischen einer Experimental- und einer Kontrollgruppe ( $H_1: \mu_1 > \mu_2$ ). Ist dies der Fall, so können wir die Alternativhypothese zwar bei signifikantem Testresultat als bewährt betrachten, nicht aber zurückweisen. Dies ist im Rahmen einer deduktivistischen Methodologie, die den Erkenntnisfortschritt an die Elimination falscher Hypothese koppelt (Popper, 1934/1982), völlig unakzeptabel. Keineswegs unproblematischer ist der seltenere Fall, dass die psychologische Hypothese eine Nullhypothese impliziert (z. B.  $H_0: \mu_1 = \mu_2$ ). Eine solche Hypothese kann im Rahmen des Fisherschen Signifikanztests keine fundierten Bewährungsurteile erfahren, sodass unklar bleibt, woran man ihren Erfolg eigentlich messen soll.

Der Verzicht auf Teststärkekontrollen hat äußerst bedenkliche Konsequenzen für die Publikationspraxis psychologischer Fachzeitschriften: Signifikante Befunde, die eine  $H_1$  stützen, lassen sich vergleichsweise leicht publizieren, nicht signifikante Ergebnisse dagegen nur schwer oder gar nicht. Die Folge ist eine systematische Verzerrung der Befundlage zugunsten von  $H_1$ -konformen Hypothesen und gegen  $H_0$ -konforme Hypothesen (Bredenkamp, 1972, 1980). Folglich ist zu erwarten, dass sich unter den publizierten Befunden ein erheblicher Anteil von  $\alpha$ -Fehlern verbirgt, die auch nicht aufgedeckt werden, weil gescheiterte Replikationen mit nicht signifikanten Befunden nur minimale Chancen haben, publiziert zu werden. Eine weitere, eng damit zusammenhängende Konsequenz ist, dass gelegentlich vermeintliche Effektunterschiede für verschiedene abhängige Variable publiziert und psychologisch interpretiert werden, hinter denen möglicherweise gar keine realen Effektunterschiede stecken. Buchner und Wippich (2000) sowie Buchner und Brandt (2003) haben beispielsweise gezeigt, dass die in der Forschung zum impliziten Gedächtnis gerne inhaltlich interpretierten einfachen Dissoziationen zwischen expliziten und impliziten Gedächtnistests auch dadurch zustande kommen können, dass statistische Tests auf Unterschiede zwischen Bedingungen auf der Basis von typischen expliziten Gedächtnistests deutlich teststärker sind als Unterschiedstests auf der Basis von impliziten Gedächtnistests. Wenn also eine bestimmte experimentelle Manipulation Effekte auf Maße aus expliziten, nicht aber auf Maße aus impliziten Gedächtnistests zeigt, dann muss das nicht daran liegen, dass die interessierende experimentelle Manipulation nur jene Strukturen oder Prozesse betrifft, die einem expliziten Gedächtnis zugeordnet werden. Der Grund kann auch sein, dass der Unterschiedstest

auf der Basis der Daten aus dem impliziten Gedächtnistest wegen dessen geringerer Reliabilität viel zu wenig Teststärke besaß.

Zusammenfassend bleibt festzuhalten, dass Signifikanztests, welche Teststärkegesichtspunkte ignorieren, Gefahr laufen, »... worse than nothing« zu sein (J. Neyman, zitiert nach Gigerenzer u. Murray, 1987).

### Wie lassen sich Teststärkekontrollen realisieren?

Zwei Einwände haben der routinemäßigen Anwendung von Teststärkeanalysen bei Signifikanztests in der Vergangenheit im Weg gestanden: der Einwand, dass Teststärkekontrollen oft unmöglich seien (z. B. Wottawa, 1981, S. 157) sowie der häufig unter der Hand geäußerte Einwand, dass Teststärkekontrollen zu aufwändig und damit letztlich unpraktikabel seien. Der erste Einwand stützt sich auf das Argument, dass oft unklar sei, welche Punkthypothese eigentlich als  $H_1$  fungiere beziehungsweise welche Effektgröße für die zugrunde liegende Population zu veranschlagen sei. Die Festlegung einer geeigneten Effektgröße kann jedoch im Regelfall unter Erwägung der praktischen Bedeutung, die eine entsprechende Abweichung von der  $H_0$  hat, auf rein apriorischer Basis erfolgen, gegebenenfalls unter Berücksichtigung der Effektstärkekonventionen von Cohen (1988) sowie der Ergebnisse verwandter Studien, die zuvor durchgeführt wurden. Die Effektgrößenspezifizierung verlangt lediglich eine Entscheidung darüber, ab welcher Effektgröße die  $H_0$  bei kontrollierter Fehlerrate  $\beta$  zurückgewiesen werden soll oder – anders ausgedrückt – bis zu welcher Effektgröße man eine Abweichung von der  $H_0$  für noch vernachlässigbar hält.

Gewarnt werden muss vor der so genannten retrospektiven Teststärkeanalyse, bei der die Effektstärke aus denselben Stichprobendaten geschätzt wird, für die auch die Teststärkeanalyse durchgeführt werden soll. Fehlerhaft ist dieses Vorgehen, das beispielsweise im Programmsystem SPSS als »Observed Power« (beobachtete Teststärke) bezeichnet wird, in erster Linie deshalb, weil die Effektstärke in der Stichprobe je nach Stichprobenumfang erheblich von der wahren Effektstärke in der zugrunde liegenden Population abweichen kann und im Regelfall noch nicht einmal ein erwartungstreuer Schätzer der Populationseffektstärke ist. Besonders bei kleineren Stichprobenumfängen können die Ergebnisse also sehr irreführend sein. Aber selbst wenn man von dieser Problematik absieht, stellt sich die Frage, warum ausschließlich die vorliegenden Daten darüber entscheiden sollten, wie groß die zu entdeckende Effektstärke sein muss. Die so bestimmte Ef-

fektstärke und die daraus errechnete Teststärke wären dann ja völlig losgelöst von den Kosten- und Nutzenaspekten, die mit statistischen Entscheidungen verbunden sind. Derartige Teststärkekontrollen verlieren unseres Erachtens ihren Sinn.

Ist der zweite Einwand gegen Teststärkekontrollen berechtigt? Ob der für Teststärkeanalysen notwendige Aufwand im Regelfall wesentlich über dem Aufwand bei konventionellen Fisherschen Signifikanztests liegt, kann auch für die Vergangenheit mit Fug und Recht bezweifelt werden. Bredenkamp (1969) hat bereits vor 35 Jahren eine leicht anwendbare approximative Formel zur Berechnung des minimalen Stichprobenumfangs bei Mittelwertvergleichen publiziert, mit dem eine gewünschte Teststärke bei vorgegebener Effektstärke garantiert werden kann. Bredenkamps (1969) Approximationsformel funktioniert trotz ihrer einfachen Anwendbarkeit »amazingly well even for very small sample sizes«, wie ihr Kupper und Hafner (1989, S. 101) im *American Statistician* in einem vergleichenden Übersichtsreferat zu verschiedenen Stichprobenumfangsformeln bescheinigt haben,<sup>2</sup> und sie deckt außerdem den dominierenden Anwendungsfall statistischer Tests in der Psychologie ab (Zwei-Gruppen-*t*-Test). Trotzdem sind Anwendungen dieser Formel wie auch später entwickelter Techniken der Teststärkekontrolle (z. B. Cohen, 1988) in der psychologischen Forschung rar geblieben.

Dass wir uns vor etwa zehn Jahren trotz dieser wenig motivierenden Sachlage entschlossen haben, mit dem Programm GPOWER ein weiteres Werkzeug der Teststärkeanalyse zu entwickeln (Erdfelder et al., 1996), liegt in der Einsicht begründet, dass sich viele Anwendungsprobleme von Teststärkeanalysen mit einem PC-basierten Computerprogramm optimal lösen lassen. Vor allem drei Gründe sprechen für GPOWER. Erstens kann sich die Notwendigkeit sehr genauer Powerberechnungen ergeben, insbesondere wenn  $\alpha$  oder  $\beta$  bei multiplen Tests zum gleichen Datensatz adjustiert werden müssen (Hager, 1992; Westermann u. Hager, 1986). GPOWER kann dies problemlos leisten, während Approximationsformeln oder entsprechende Tabellenwerke (z. B. Cohen, 1988) zu ungenau sind. Zweitens lässt sich das Praktikabilitätsargument nur mit einem sehr allgemeinen und leicht handhabbaren Instrument überzeugend entkräften, das – wie GPOWER – Teststärkeanalysen für eine breite Palette gebräuchlicher statistischer Tests auf den verbreiteten Computerplattformen realisieren kann.

---

<sup>2</sup> Die Vollständigkeit gebietet es, an dieser Stelle hinzuzufügen, dass Kupper und Hafner (1989) die Formel von Bredenkamp (1969) zwar darstellen und kommentieren, ihn aber nicht als Urheber zitieren, sondern stattdessen vier englischsprachige Arbeiten, die in den Jahren 1983 bis 1988 publiziert wurden.

Drittens ist der Standardfall bisheriger Teststärkeanalysen, die so genannte A-priori-Poweranalyse (Cohen, 1988), die den erforderlichen Stichprobenumfang zur Realisierung einer gewünschten Teststärke  $1-\beta$  bei vorgegebenem  $\alpha$ -Niveau vor der Durchführung einer Untersuchung berechnet, zwar der Idealtypus der Teststärkekontrolle, keineswegs aber die einzig sinnvolle Möglichkeit, Teststärkeüberlegungen in statistische Entscheidungen oder in die Bewertung statistischer Entscheidungen anderer Forschergruppen einfließen zu lassen. Je nach Sachlage können noch vier weitere Formen der Teststärkeanalyse wertvolle Dienste leisten, die sich nur mithilfe von Computerprogrammen sinnvoll realisieren und kombinieren lassen: Post-hoc-Analysen (Cohen, 1988), Kompromissanalysen (Erdfelder et al., 1996), Sensitivitätsanalysen (z. B. Welsch et al., 2000, Kap. 5.4) und Kriteriumsanalysen. A-priori-, Post-hoc- und Kompromissanalysen sind bereits als Auswahloptionen in GPOWER 2.0 implementiert, während die zuletzt genannten Analysen bislang nur durch iterative Anwendung von Post-hoc-Analysen realisiert werden können. Für Version 3.0 von GPOWER sind Sensitivitäts- und Kriteriumsanalysen aber ebenfalls als Auswahloptionen vorgesehen.

### Formen der Teststärkeanalyse

Die fünf Formen der Teststärkeanalyse decken unterschiedliche Anwendungsfälle ab. *A-priori-Analysen* ermöglichen die Planung von Studien und Experimenten, die vorgegebene Güteeigenschaften der statistischen Entscheidungen erfüllen, das heißt  $\alpha$ ,  $\beta$  und die erwartete Stärke des Effekts in der Population stehen fest und gesucht wird die dafür notwendige Stichprobengröße  $N$ . *Post-hoc-Analysen* liefern als Ergebnis die Teststärke  $1-\beta$  beziehungsweise die Wahrscheinlichkeit eines  $\beta$ -Fehlers zu einem gegebenen  $N$ , dem gewählten  $\alpha$  und einer bestimmten Effektstärke. Meist werden diese Analysen eingesetzt, um zu prüfen, mit welcher Fehlerwahrscheinlichkeit eine Entscheidung zugunsten einer  $H_0$  gefallen ist.

*Kompromiss-Analysen* können dann sinnvoll sein, wenn das verfügbare  $N$  gemessen am idealen  $N$  zu klein ist, etwa bei klinischen Studien zu einem seltenen Krankheitsbild. Um der Alternativhypothese eine faire Chance zu geben, wird ein Kompromiss zwischen  $\alpha$  und der Teststärke  $1-\beta$  angestrebt. Dazu wird zunächst entschieden, wie wichtig  $\beta$  im Vergleich zu  $\alpha$  sein soll. Anschließend bestimmt man auf der Basis dieser Gewichtung und des verfügbaren  $N$  nicht nur  $\beta$ , sondern simultan auch  $\alpha$  und den damit

verbundenen Wert der Teststatistik. Ein weiterer interessanter Anwendungsfall für Kompromiss-Analysen ist bei Modellanpassungstests gegeben, wo das verfügbare  $N$  bisweilen »zu groß« sein kann, sodass bereits triviale Abweichungen der vom Modell implizierten von der empirischen Datenstruktur zu einer Zurückweisung des Modells führen würden. In diesem Fall legt man über die Effektgröße fest, welche Abweichung man tolerieren möchte. Auf dieser Basis und aufgrund des vorliegenden  $N$  und der gewünschten Gewichtung von  $\beta$  relativ zu  $\alpha$  liefert die Analyse dann die Werte für  $\beta$  sowie für  $\alpha$  und den damit verbundenen Wert der Teststatistik, bei dessen Überschreiten das Modell zurückgewiesen wird.

*Sensitivitäts-Analysen* liefern eine Antwort auf die Frage, wie groß die Stärke eines Effekts sein muss, damit man diesen mit einem bestimmten begrenzten  $N$  und mit bestimmten idealen Fehlerwahrscheinlichkeiten  $\alpha$  und  $\beta$  entdecken kann. Das Ergebnis einer solchen Analyse kann durchaus sein, dass man einen bestimmten interessierenden Effekt mit den gegebenen Ressourcen und den vorliegenden Ansprüchen an die Fehlerwahrscheinlichkeiten gar nicht entdecken kann, weil man annehmen muss, dass der interessierende Effekt deutlich kleiner ist als der unter den gegebenen Voraussetzungen entdeckbare Effekt. Wie beurteilt man, ob ein interessierender Effekt »zu klein« ist? Wie üblich kommen hierfür entweder Effektstärkekonzventionen oder die Ergebnisse verwandter Studien infrage. Beispielsweise müsste in einer Zwei-Gruppen- $t$ -Test-Situation bei  $\alpha = \beta = .05$  und einem maximal verfügbaren  $N = 30$  ( $n_1 = n_2 = 15$ ) der entdeckbare Gruppenunterschied eine Effektstärke von  $d = 1.23$  besitzen. Wenn wir in Anlehnung an Cohen (1988) für unsere Fragestellung höchstens »mittlere« Effekte, das heißt Effekte der Größe  $d = 0.5$  in Bezug auf den Gruppenunterschied erwarten, ist klar, dass unsere begrenzten Ressourcen und unsere Ansprüche an die Fehlerwahrscheinlichkeiten zu einer Testsituation führen würden, die nicht sensitiv genug ist, um einen solchen Effekt zu entdecken. In diesem Fall kann man sich zum einen entscheiden, eine solche Untersuchung nicht durchzuführen. Das spart immerhin Zeit und Geld. Alternativ kann man mithilfe einer Kompromiss-Analyse prüfen, ob sich unter den gegebenen Ressourcenbegrenzungen statt der idealen immer noch akzeptable Wahrscheinlichkeiten für  $\alpha$ - und  $\beta$ -Fehler finden lassen.

*Kriteriumsanalysen* schließlich liefern den kritischen Wert der jeweiligen Teststatistik (und damit auch das  $\alpha$ -Fehlerniveau) zu vorgegebenem Stichprobenumfang sowie gegebener Effektstärke und Teststärke. Diese Art der Analyse ist insbesondere bei der Durchführung so genannter Äquivalenztests hilfreich. Äquivalenztests sind vor allem in der biometrischen Pharmaforschung gebräuchlich, in der man häufig vor dem Problem steht, die

Effektivitätsäquivalenz eines neuen, kostengünstigen Medikaments mit einem etablierten, teuren Medikament nachzuweisen. Ganz ähnliche Probleme können sich aber auch in der Psychologie ergeben, wenn man etwa mit einem Experimental-Kontrollgruppenvergleich zeigen will, dass eine kostengünstige Kurzformpsychotherapie genauso effektiv wie eine teure etablierte Langzeittherapie wirkt (Rogers et al., 1993). Der Grundgedanke hierbei ist, die Null- und die Alternativhypothese eines gewöhnlichen Zwei-Gruppen- $t$ -Tests gewissermaßen gegeneinander auszutauschen:  $H_0$  besagt, dass die beiden Gruppen nicht äquivalent sind, sodass die mittlere Effektivität der etablierten Behandlung ( $\mu_e$ ) die mittlere Effektivität der neuen Behandlung ( $\mu_n$ ) mindestens um einen klein gewählten Grenzwert  $\Delta$  überschreitet ( $H_0: \mu_e - \mu_n \geq \Delta$ ). Die Alternativhypothese der Äquivalenz behauptet hingegen, dass die Effektivität der etablierten Behandlung gar nicht oder allenfalls nur minimal über die der neuen Behandlung dominiert ( $H_1: \mu_e - \mu_n < \Delta$ ). Diese Hypothesen lassen sich mit einem modifizierten einseitigen  $t$ -Test für zwei unabhängige Gruppen gegeneinander testen. Wenn die Untersuchung bereits durchgeführt wurde und die Stichprobenumfänge in den beiden Gruppen bereits feststehen, stellt sich die Frage nach dem kritischen Wert und der Entscheidungsregel, mit der zwischen  $H_0$  und  $H_1$  bei kontrollierten Fehlerwahrscheinlichkeiten zu entscheiden ist. Kriteriumsanalysen helfen hier weiter. Die Verteilung unter  $H_0$  ist in diesem Fall eine nonzentrale  $t$ -Verteilung zur Effektgröße  $d = \Delta/\sigma$ , wobei  $\sigma$  die (ggf. erwartungstreu zu schätzende) Populationsstreuung in den beiden Gruppen ist. Will man den Äquivalenztest bei  $\alpha = .05$  durchführen, ist in GPOWER als geforderte »power«  $1 - \alpha = .95$  vorzugeben, womit wunschgemäß sichergestellt ist, dass 95 Prozent der nonzentralen Verteilung zur Punkthypothese  $\mu_e - \mu_n = \Delta$  oberhalb vom kritischen Wert und 5 Prozent unterhalb vom kritischen Wert liegen. Berechnet man nun eine Kriteriumsanalyse, so erhält man mit  $\alpha$  den Anteil der zentralen  $t$ -Verteilung rechts vom Entscheidungskriterium. Die Teststärke des Äquivalenztests, also die Wahrscheinlichkeit, bei faktisch vorliegender Äquivalenz ( $\mu_e - \mu_n = 0$ ) auch korrekt zugunsten der Äquivalenzhypothese zu entscheiden, ist folglich gerade das Komplement des in der Kriteriumsanalyse errechneten  $\alpha$ -Werts.

Kriteriumsanalysen ermöglichen die Durchführung von Äquivalenztests für prinzipiell alle Testfamilien, nicht nur für Zwei-Gruppen- $t$ -Tests. Zu beachten ist lediglich, dass die Größe, die in GPOWER als *Power* bezeichnet wird, bei Äquivalenztests gerade  $1 - \alpha$  entspricht, während umgekehrt die Größe, die GPOWER als *alpha* bezeichnet, bei Äquivalenztests gerade der  $\beta$ -Fehlerwahrscheinlichkeit entspricht.

## Spezielle Teststärkeanalysen mit GPOWER

Das Programm GPOWER ist als allgemeines Programm für Teststärkeanalysen konzipiert worden (Erdfelder et al., 1996). Tatsächlich kann das Programm die im letzten Abschnitt beschriebenen Formen der Teststärkeanalyse für alle statistischen Tests durchführen, deren Prüfgrößen unter der Null- und der Alternativhypothese einer Normalverteilung, einer (nonzentralen)  $\chi^2$ -Verteilung, einer (nonzentralen)  $t$ -Verteilung oder einer (nonzentralen)  $F$ -Verteilung folgen. Fast alle in den Sozial- und Verhaltenswissenschaften geläufigen Tests erfüllen diese Bedingung.<sup>3</sup> Die konkrete Handhabung des Programms ist allerdings nur für einige häufig vorkommende Standardtests einfach. Diese Fälle sind von Erdfelder et al. (1996) sowie Buchner et al. (1996) an konkreten Beispielen erläutert worden. Im Folgenden soll deshalb ausschließlich auf einige spezielle Testprobleme eingegangen werden, deren Lösung mit GPOWER Zusatzwissen verlangt, das bislang in der Literatur nicht dargestellt wurde.

### Spezielle Tests mit normalverteilten Teststatistiken

In diesem Abschnitt werden Teststärkeanalysen für Proportions- und Korrelationstests dargestellt. Der Vorzeichentest (*sign test*) ist als Spezialfall eines Proportionstests ebenfalls eingeschlossen.

#### Tests zu einer Proportion

Nehmen wir an, in einer Stichprobe des Umfangs  $N$  sei die Proportion eines bestimmten Merkmals erhoben worden, das heißt die relative Häufigkeit  $x/N$ , mit der eine dichotome Variable (z. B. Erfolg vs. Misserfolg einer Psychotherapie) eine bestimmte Ausprägung (z. B. Erfolg) annimmt. Zugrunde gelegt wird ein Binomialverteilungsmodell, das heißt, die  $N$  Beobachtungen seien unabhängig aus derselben Grundgesamtheit gezogen worden. Getestet werden soll beispielsweise die einseitige Nullhypothese, dass die Proportion höchstens den Wert  $p_0$  annimmt. Unter der Alternativhypothese wird entsprechend ein größerer Wert  $p_1$  für die Proportion angenom-

---

<sup>3</sup> Für einige Tests gilt dies allerdings nur asymptotisch, sodass die Powerberechnungen bei kleinem Stichprobenumfang als Approximationen zu werten sind.

men. So kann etwa von Interesse sein, ob die Erfolgsquote einer neuen Psychotherapieform den Wert  $p_0 = .60$  einer etablierten Therapie nicht überschreitet ( $H_0: p \leq p_0$ ) oder aber größer ist ( $H_1: p > p_0$ ). Als zu entdeckende substantielle Verbesserung werde z. B.  $p_1 = .80$  festgelegt.

Unter  $H_0$  ist dann bekanntlich die Prüfstatistik

$$z = \frac{x - N \cdot p_0}{\sqrt{N \cdot p_0 \cdot (1 - p_0)}}$$

asymptotisch standardnormal verteilt (z. B. Hays, 1972, S. 305). Gilt  $H_1$ , so ist die gleiche Statistik ebenfalls normal verteilt, allerdings mit der Streuung

$$\sigma_1 = \frac{\sqrt{N \cdot p_1 \cdot (1 - p_1)}}{\sqrt{N \cdot p_0 \cdot (1 - p_0)}}$$

und dem Erwartungswert

$$\mu_1 = \frac{N \cdot (p_1 - p_0)}{\sqrt{N \cdot p_0 \cdot (1 - p_0)}} = \sqrt{N} \cdot \frac{p_1 - p_0}{\sqrt{p_0 \cdot (1 - p_0)}}$$

Post-hoc-Teststärkeanalysen für diesen Fall lassen sich mit GPOWER mittels der Prozedur »Other  $t$ -Tests« durchführen. Da die nonzentrale  $t$ -Verteilung mit Nonzentralitätsparameter  $\delta$  für  $df \rightarrow \infty$  in eine Normalverteilung mit Mittelwert  $\delta$  und Streuung 1 übergeht, setzt man zunächst  $df = 32000$ , das heißt den maximal von GPOWER zugelassenen Eingabewert, und wählt einen einseitigen Test (»one-tailed«). Der Nonzentralitätsparameter  $\delta$  entspricht in der Prozedur »Other  $t$ -Tests« dem Produkt der Parameter  $\sqrt{N}$  und  $f$  (= Effektstärke). Man wählt also  $N$  und  $f$  in GPOWER so, dass das Produkt von  $\sqrt{N}$  und  $f$  gerade  $\mu_1$  entspricht, dem Erwartungswert der Teststatistik unter  $H_1$ . Für  $N$  ist folglich der Stichprobenumfang einzusetzen und für die Effektstärke

$$f = \frac{p_1 - p_0}{\sqrt{p_0 \cdot (1 - p_0)}}$$

Ein letztes zu lösendes Problem betrifft die Tatsache, dass GPOWER bei »Other  $t$ -tests« mit  $df = 32000$  von einer Streuung  $\sigma = 1$  unter  $H_1$  ausgeht, während im gegebenen Fall die tatsächliche Streuung – wie oben ausgeführt –  $\sigma_1$  beträgt. Dieses Problem löst man dadurch, dass man als Signifikanzniveau  $\alpha$  nicht das faktisch gewünschte Niveau eingibt, sondern ein adjustiertes Niveau  $\alpha'$ , welches so gewählt wird, dass der kritische Wert der  $z$ -Statistik zum Niveau  $\alpha'$ , abgekürzt  $z(\alpha')$ , zum entsprechenden kritischen Wert für das nominelle  $\alpha$ -Niveau,  $z(\alpha)$ , wie folgt in Beziehung steht:

$$z(\alpha') = z(\alpha)/\sigma_1.$$

Diese Adjustierung garantiert, dass die korrekte Teststärke berechnet wird. Da es sich um einen asymptotischen Test handelt, ist natürlich vorauszusetzen, dass der Stichprobenumfang nicht zu klein ist. Die Übertragung auf den Fall zweiseitiger Tests liegt auf der Hand: Hier sind zwei Ablehnungsbereiche für  $H_0$  zu betrachten statt eines Bereichs, was durch Anwählen der Option »two-tailed« geschieht.

Man beachte, dass die geschilderte Prozedur auch asymptotische Poweranalysen für den Vorzeichentest (*sign test*) korrekt durchführt, weil dieser nichts anderes ist als ein Proportionstest mit der Nullhypothese  $H_0: p = .50$  (vgl. Hays, 1972, S. 193–197).

Betrachten wir hierfür ein konkretes Rechenbeispiel. Es seien  $N = 100$  Fälle beobachtet worden. Die Erfolgsquote unter  $H_0$  betrage  $p_0 = .50$ , unter  $H_1$  dagegen  $p_1 = .70$ . Wie groß ist die Power des einseitigen Vorzeichentests bei  $\alpha = .05$ ? Wie gesagt ist für  $df$  die Obergrenze 32000 und für  $N$  der Wert 100 im Menüpunkt »Other *t*-Tests«, »Post hoc Analysis«, festzulegen. Setzt man die genannten  $p_0$ - und  $p_1$ -Werte in die oben genannte  $f$ -Formel ein, ergibt sich der für die Effektgröße einzugebende Wert  $f = .33806$ . Für alpha ist in GPOWER 2.0 aus den genannten Gründen aber nicht .05 zu wählen. Der zu  $\alpha = .05$  (einseitig) gehörige  $z$ -Wert von 1.649 ist zunächst durch  $\sigma_1$  zu dividieren, was den adjustierten  $z$ -Wert 1.7947 ergibt. Der zugehörige adjustierte  $\alpha$ -Wert (einseitig) ist etwa .0364. Setzt man ihn für  $\alpha$  in GPOWER ein und klickt auf »Calculate«, so erhält man die erfreuliche Powerschätzung von  $1 - \beta = .9437$ .

### Tests zu einer Korrelation

Werden in einer Stichprobe der Größe  $N$  zwei quantitative Variablen  $X$  und  $Y$  (z. B. die Konzentrationsfähigkeit und der IQ der Probanden) gemessen, die in der Population bivariat normal verteilt sind, können Hypothesen über die Korrelation der beiden Variablen getestet werden. Häufig steht man vor der Situation, dass das Ausmaß des linearen Zusammenhangs getestet werden soll. So könnte aufgrund einer Theorie behaupten werden, dass die Konzentrationsfähigkeit und die Intelligenz in der Population mit .50 korrelieren. In diesem Fall soll die  $H_0: \rho_{XY} = \rho_0 = .50$  getestet werden. Die Differenz zwischen der gemäß der Fisherschen  $r$ -nach- $z$ -Formel transformierten Korrelation  $r_{XY}$  in der Stichprobe und der  $z$ -transformierten  $H_0$ -Korrelation in der zugrunde liegenden Population,

$$0.5 \cdot \ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right) - 0.5 \cdot \ln\left(\frac{1+\rho_0}{1-\rho_0}\right),$$

ist unter der  $H_0$  approximativ normal verteilt mit dem Erwartungswert Null und Varianz  $1/(N-3)$  (Hays, 1972, S. 662–663). Unter einer bestimmten Alternativhypothese ( $H_1: \rho_{XY} = \rho_1$ ) ist diese Differenz ebenfalls normal verteilt mit der gleichen Varianz, allerdings mit dem Erwartungswert

$$\mu_1 = 0.5 \cdot \ln\left(\frac{1+\rho_1}{1-\rho_1}\right) - 0.5 \cdot \ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = q_2',$$

welcher genau dem Effektstärkemaß  $q_2'$  nach Cohen entspricht (Cohen, 1988). Multipliziert man die  $z$ -Differenzen mit dem Kehrwert der Streuung  $1/\sqrt{1/(N-3)}$ , erhält man standardnormal verteilte Variablen. Unter der  $H_0$  ist der Erwartungswert weiterhin Null, während er unter der  $H_1$  nun  $q_2'/\sqrt{1/(N-3)} = q_2' \cdot \sqrt{N-3}$  ist. Die Teststärkeberechnung für diesen Fall ist mit GPOWER nun recht einfach. Zunächst sind in der Prozedur »Other  $t$ -Tests« wieder 32000 Freiheitsgrade zu wählen, sodass die  $t$ -Verteilung eine Normalverteilung approximiert. Da der Nonzentralitätsparameter  $\delta$  in dieser Prozedur als Produkt von Effektstärke  $f$  und  $\sqrt{N}$  errechnet wird, erhält man den gewünschten Wert  $\delta = q_2' \cdot \sqrt{N-3}$ , wenn man für die Effektstärke den Wert  $q_2'$  wie oben definiert und als Stichprobengröße den Wert  $(N-3)$  eingibt. Man wählt zusätzlich noch das gewünschte Signifikanzniveau  $\alpha$  und – je nach Fragestellung – einen einseitigen (»one-tailed«) oder zweiseitigen (»two-tailed«) Test, um die korrekte Power des Tests zu erhalten. Wie schon im Fall des im letzten Abschnitt geschilderten Proportionstests ist allerdings zu beachten, dass es sich um einen asymptotischen Test handelt, der nur bei nicht zu kleinem  $N$  hinreichend genau ist.

Nehmen wir an, die Nullhypothese einer Korrelation von  $\rho_0 = .5$  soll gegen die Alternativhypothese einer Korrelation von  $\rho_1 = .6$  bei  $\alpha = .05$  einseitig getestet werden. Es wurden  $N = 103$  Messwertpaare untersucht, sodass die Varianz der  $z$ -transformierten Korrelation  $1/100$  und die Streuung somit  $0.10$  beträgt. Das Effektmaß  $q_2'$  nach Cohen ist, wie man einer Fisher  $r$ -nach- $z$ -Tabelle entnehmen kann (z. B. Hays, 1972), in diesem Fall  $q_2' = .6931 - .5493 = .1430$ . Man errechnet folglich mit GPOWER 2.0 die Power, indem man »Other  $t$ -Tests«, »Post-hoc analysis« mit  $df = 32000$  auswählt, das gewünschte  $\alpha = .05$  und »one-tailed« spezifiziert, für die Effektstärke den Wert  $f = q_2' = .1430$  und für  $N$  das Inverse der Varianz (100) angibt. Man erhält dann durch Klicken auf »Calculate« die ernüchternde Powerbestimmung von nur  $1 - \beta = .4149$ .

### Vergleiche von zwei unabhängigen Korrelationen

Nehmen wir an, dass in zwei unabhängigen Stichproben jeweils  $N$  Werte zweier quantitativer Variablen  $X$  und  $Y$ , die in beiden zugrunde liegenden Population jeweils bivariat normal verteilt sind, gemessen wurden. In diesem Fall kann die Hypothese geprüft werden, dass die beiden Populationskorrelationen identisch sind ( $H_0: \rho_{1(XY)} = \rho_{2(XY)}$ ). Beispielsweise könnte man daran interessiert sein, ob ein Persönlichkeitsmerkmal in zwei klinischen Populationen gleich mit einer Kriteriumsvariablen korreliert. Da die  $z$ -Transformation einer Korrelation nach Fisher (s. Hays, 1972, S. 662–663) normal verteilt ist mit der Varianz  $1/(N-3)$ , ist die Teststatistik

$$z_1 - z_2 = 0.5 \cdot \ln\left(\frac{1+r_{1(XY)}}{1-r_{1(XY)}}\right) - 0.5 \cdot \ln\left(\frac{1+r_{2(XY)}}{1-r_{2(XY)}}\right)$$

unter der  $H_0$  ebenfalls normal verteilt mit dem Erwartungswert Null und der Varianz  $2/(N-3)$ . Unter der  $H_1: \rho_{1(XY)} \neq \rho_{2(XY)}$  ist die Statistik ebenfalls normal verteilt, allerdings mit dem Erwartungswert

$$\mu_1 = 0.5 \cdot \ln\left(\frac{1+\rho_{1(XY)}}{1-\rho_{1(XY)}}\right) - 0.5 \cdot \ln\left(\frac{1+\rho_{2(XY)}}{1-\rho_{2(XY)}}\right) = q$$

und der Varianz  $2/(N-3)$ . Man beachte, dass  $\mu_1$  gerade dem Effektstärkemaß  $q$  von Cohen (1988) entspricht.

Multipliziert man die  $z$ -Differenzen mit  $1/\sqrt{2/(N-3)}$ , erhält man standardnormal verteilte Variablen. Unter der  $H_0$  ist der Erwartungswert weiterhin Null, während er unter der  $H_1$  nun  $q/\sqrt{2/(N-3)} = q \cdot \sqrt{(N-3)/2}$  ist. Die Teststärkeberechnung für diesen Fall ist mit GPOWER nun problemlos möglich. Zunächst sind in der Prozedur »Other  $t$ -Tests« erneut 32000 Freiheitsgrade zu wählen, sodass die  $t$ -Verteilung eine Normalverteilung approximiert. Der Nonzentralitätsparameter  $\delta$  entspricht – wie bereits aus den letzten beiden Abschnitten bekannt – dem Produkt aus der Effektstärke  $f$  und  $\sqrt{N}$ . Um  $\delta = q \cdot \sqrt{(N-3)/2}$  zu erhalten, wählt man also für die Effektstärke  $f$  den Wert  $q$  wie oben definiert und als Stichprobengröße den Wert  $(N-3)/2$ . Ansonsten verfährt man analog zum Test für eine Korrelation.

Für den Fall, dass die untersuchten Korrelationen auf ungleich großen Stichproben beruhen ( $N_1 \neq N_2$ ) kann  $N$  über

$$N' = \frac{2(N_1-3)(N_2-3)}{N_1 + N_2} + 3$$

bestimmt werden (s. Cohen, 1988).

Auch hier soll ein konkretes Beispiel das Vorgehen verdeutlichen. Bei  $\alpha = .01$  soll die Nullhypothese gleicher Korrelationen in zwei Populationen, aus denen jeweils 103 Versuchspersonen zufällig entnommen wurden, gegen die Alternativhypothese getestet werden, dass die Korrelation in der ersten Population .70, in der zweiten Population dagegen .90 ist. Fisher  $r$ -nach- $z$ -Tabellen (Hays, 1972) kann man entnehmen, dass hierfür  $q = 1.472 - .8673 = .6047$  das Effektmaß bildet. Man wählt also wie in den vorherigen Beispielen in GPOWER 2.0 »Other  $t$ -Tests«,  $df = 32000$ , »Post-hoc analysis«, nunmehr aber  $\alpha = .01$  und »two-tailed«. Für die Effektgröße ist  $f = .6047$  anzugeben und für  $N = (103 - 3)/2 = 50$ . Durch Klicken auf Calculate erhält man dann als Powerberechnung den erfreulich hohen Wert von  $1 - \beta = .9554$ .

## Spezielle Tests mit $F$ -verteilten Teststatistiken

### Multivariate Varianzanalysen

Bredenkamp und Erdfelder (1985) empfehlen als Teststatistik bei multivariaten Varianzanalysen das Pillai-Bartlett- $V$  als multivariates Testkriterium. Bei Gültigkeit der  $H_0$  gilt, dass

$$F = \frac{\frac{V(h)/s(h)}{df_1}}{\frac{1-V(h)/s(h)}{df_2}}$$

approximativ  $F(df_1, df_2)$ -verteilt ist. Hierbei repräsentiert  $V(h)$  das Pillai-Bartlett- $V$  für die zu testende Hypothese und  $s(h)$  das Minimum aus  $(p, n(h))$ , wobei  $p$  die Anzahl der abhängigen Variablen und  $n(h)$  die Anzahl der Prädiktoren für den zu testenden Effekt bezeichnet. Für die Freiheitsgrade gilt

$$df_1 = p \cdot n(h)$$

und

$$df_2 = s(h) \cdot (N - k - p + s(h)),$$

wobei  $N$  die Gesamtzahl aller Versuchspersonen in allen  $k$  Gruppen des Ver-

suchsplans repräsentiert. Der Term  $V(h)/s(h)$  variiert zwischen 0 und 1 und kann als multivariates  $R^2$  oder  $\eta^2$  betrachtet werden. Ein einfaches Maß für die standardisierte multivariate Populationseffektgröße erhält man nach

$$f^2_{\text{mult}} = \frac{V(h)/s(h)}{1 - V(h)/s(h)} = \frac{V(h)}{s(h) - V(h)}.$$

Vergleiche mit Veröffentlichungen exakter Teststärketabellen (Pillai u. Jayachandran, 1967; Stevens, 1980) zeigen, dass mit GPOWER berechnete Teststärken sehr nahe an den exakten Teststärken liegen, wenn man annimmt, dass bei Gültigkeit der  $H_1$  die Pillai-Bartlett- $V$ -Statistik approximativ  $F(df_1, df_2, \lambda)$ -verteilt ist mit den oben genannten Freiheitsgraden und dem Nonzentralitätsparameter

$$\lambda = s(h) \cdot N \cdot f^2_{\text{mult}}.$$

Bei der Durchführung einer globalen MANOVA zur Prüfung der Frage, ob alle verwendeten Prädiktoren Varianz in den zwei oder mehr abhängigen Variablen aufklären, gilt  $n(h) = k - 1$ . Für spezielle MANOVAs zur Prüfung des Effekts einer Untermenge der Prädiktoren gilt  $n(h) =$  Anzahl der zu testenden Prädiktoren. Für einen Haupteffekt  $A$  mit  $a$  Stufen gilt also  $n(h) = a - 1$ , für einen Haupteffekt  $B$  mit  $b$  Stufen gilt  $n(h) = b - 1$  und für die  $A \times B$ -Interaktion gilt  $n(h) = (a - 1) \cdot (b - 1)$ .

Post-hoc-Teststärkeanalysen für diese Fälle lassen sich mit der Prozedur »Other  $F$ -Tests« in GPOWER durchführen. In dieser Prozedur wird der Nonzentralitätsparameter  $\lambda$  als Produkt aus Effektstärke  $f^2$  und der Stichprobengröße  $N$  berechnet. Um das gewünschte  $\lambda = s(h) \cdot N \cdot f^2_{\text{mult}}$  zu erhalten, wählt man für  $f^2$  den Wert  $f^2_{\text{mult}}$  und als Stichprobenumfang  $s(h) \cdot N$ . Als Nenner- und Zählerfreiheitsgrade sind die oben angegebenen einzusetzen.

### Multivariate Varianzanalysen für Messwiederholungspläne

Varianzanalysen für Messwiederholungspläne sind stets multivariate Varianzanalysen, bei denen die Stufen der messwiederholten Variablen als verschiedene abhängige Variable betrachtet werden (O'Brian u. Kaiser, 1985). Nehmen wir an, der im vorangegangenen Abschnitt erwähnte Faktor  $B$  repräsentiere keinen Gruppenfaktor, sondern eine messwiederholte Variable,

Faktor  $A$  jedoch weiterhin verschiedene Gruppen. Dann ergeben sich  $p = b - 1$  abhängige Variablen, weil die  $b$  Stufen der Variablen  $B$  mit geeigneten orthogonalen Kontrastvariablen in  $b - 1$  abhängige Variablen rekodiert werden.

Für den Test des Haupteffekts  $A$  gilt

$$df_1 = a - 1$$

und

$$df_2 = N - a,$$

wobei  $N$  die Anzahl der Versuchspersonen repräsentiert. Der Nonzentralitätsparameter ist definiert als

$$\lambda = N \cdot \left( \frac{m}{1 + (m-1) \cdot \rho} \right) \cdot f^2,$$

wobei  $N$  die Gesamtzahl aller Versuchspersonen,  $m$  die Stufen der messwiederholten Variablen,  $f^2$  den Effektgrößenindex für den univariaten ANOVA-Fall (Effektvarianz/Fehlervarianz, vgl. Cohen, 1988) und  $\rho$  die Populationskorrelation zwischen den Stufen der messwiederholten Variablen repräsentiert. Letzteres mag überraschen angesichts der Tatsache, dass es bei  $m$  Stufen einer messwiederholten Variablen  $(m^2 - m)/2$  Korrelationen zwischen diesen gibt. Nach Stevens (1996) ist es jedoch vertretbar, für diesen Fall eine Art »durchschnittlicher« Korrelation zwischen den Stufen der messwiederholten Variable anzunehmen. Für  $m = 1$  (also ohne Messwiederholung) reduziert sich die Berechnung des Nonzentralitätsparameters auf  $\lambda = N \cdot f^2$ , den Nonzentralitätsparameter für den univariaten ANOVA-Fall.

Für den  $F$ -Test für die messwiederholte Variable  $B$  ergeben sich die Freiheitsgrade nach

$$df_1 = b - 1$$

und

$$df_2 = s(h) \cdot (N - k - p + s(h)),$$

wobei  $N$  die Anzahl der Versuchspersonen und  $k$  die Anzahl der Gruppen des Versuchsplans – hier also die Stufen der Variablen  $A$  – repräsentiert.

Der Nonzentralitätsparameter ergibt sich approximativ nach

$$\lambda = N \cdot m \cdot \frac{f^2}{1-\rho}$$

Der  $F$ -Test für die  $A \times B$ -Interaktion zwischen der Gruppiervariablen  $A$  und der messwiederholten Variablen  $B$  schließlich hat folgende Freiheitsgrade:

$$df_1 = (a - 1) \cdot (b - 1),$$

$$df_2 = (N - a) \cdot (b - 1),$$

Der Nonzentralitätsparameter ist wie beim Haupteffekt der messwiederholten Variablen  $B$  definiert als

$$\lambda = N \cdot m \cdot \frac{f^2}{1-\rho}$$

Die Vorgehensweise bei der Durchführung einer Post-hoc-Teststärkeanalyse mithilfe von GPOWER ist analog zu der im vorigen Abschnitt beschriebenen.

### Varianzanalysen für Messwiederholungspläne – der so genannte »univariate« Fall

Ein Spezialfall der multivariaten Varianzanalysen für Versuchspläne mit Messwiederholungen ist die so genannte »univariate Varianzanalyse für Messwiederholungen« – eigentlich eine Fehlbenennung, denn Messwiederholungs-Varianzanalysen sind grundsätzlich multivariat. Was hier jedoch gemeint ist, ist eine multivariate Varianzanalyse für Messwiederholungspläne, bei der die restriktive und selten realistische Annahme gemacht wird, dass die Varianzen aller Stufen der messwiederholten Variablen homogen und die Korrelationen zwischen allen Stufen der messwiederholten Variablen identisch sind. Dies wird oft als die Sphäritätsannahme bezeichnet. Es gibt selten einleuchtende Gründe, warum man diese Annahme machen sollte, doch in einigen speziellen Situationen kann sie vielleicht akzeptabel sein: Wenn etwa die Reihenfolge der Messwiederholungen über die Versuchspersonen hinweg randomisiert wurde oder wenn mehr Stufen der abhängigen Variablen als Versuchspersonen vorliegen. Im zuletzt genannten Fall ist das multivariate Modell nicht identifiziert. Dann profitiert man davon, dass die erwähnten Restriktionen auf den Varianzen und Kovarianzen dazu führen, dass weniger Parameter geschätzt werden müssen und dadurch zusätzliche Freiheitsgrade zur Verfügung stehen. Entsprechend erge-

durch zusätzliche Freiheitsgrade zur Verfügung stehen. Entsprechend ergeben sich für unser bereits oben verwandtes Design die Freiheitsgrade für den  $F$ -Test für die messwiederholte Variable  $B$  nach

$$df_1 = b - 1$$

und

$$df_2 = (N - a) \cdot (b - 1)$$

wobei  $N$  wieder die Anzahl der Versuchspersonen bezeichnet. Der Nonzentralitätsparameter ergibt sich nach

$$\lambda = N \cdot m \cdot \frac{f^2}{1 - \rho}$$

Hier müssen wir nicht den Umweg über eine gedachte »durchschnittliche« Populationskorrelation  $\rho$  gehen, denn mit der Sphäritätsannahme postulieren wir ja gerade, dass es nur eines einzigen Werts bedarf, um sämtliche Interkorrelationen zwischen den Stufen der messwiederholten Variablen zu repräsentieren. Ist diese Annahme erfüllt, so ist der so genannte univariate Zugang teststärker als die im letzten Abschnitt beschriebene multivariate Analyse von Messwiederholungsplänen.

Ist die Sphäritätsannahme verletzt, muss eine Korrekturgröße  $\varepsilon$  berücksichtigt werden, für die zwei mögliche Schätzer mit leicht unterschiedlichen Eigenschaften existieren: Greenhouse-Geisser- $\hat{\varepsilon}$  und Huynh-Feldt- $\hat{\varepsilon}$ . Zur Korrektur werden bei der Entscheidung über die statistische Signifikanz eines Testergebnisses die Zähler- und Nennerfreiheitsgrade des  $F$ -Tests mit dem gewählten Schätzer für  $\varepsilon$  multipliziert und dann die Signifikanz auf der Basis des  $F$ -Werts für die reduzierten Werte  $df'_1 = df_1 \cdot \hat{\varepsilon}$  und  $df'_2 = df_2 \cdot \hat{\varepsilon}$  bewertet.

Zur Approximation der Teststärke des  $\varepsilon$ -korrigierten  $F$ -Tests haben Muller und Barton (1989) ein analoges Verfahren vorgeschlagen. Dabei wird zusätzlich auch der Nonzentralitätsparameter  $\lambda$  mit dem  $\varepsilon$ -Schätzer multipliziert und die Teststärkeanalyse mit  $df_1$ ,  $df_2$  und  $\lambda' = \lambda \cdot \hat{\varepsilon}$  durchgeführt.

Für den Gruppierfaktor ändert sich gegenüber der multivariaten Varianzanalyse für messwiederholte Pläne dagegen nichts, sodass sich eine Teststärkeanalyse einfach an dem oben beschriebenen Vorgehen orientieren kann.

Die konkrete Vorgehensweise bei der Durchführung einer Post-hoc-Teststärkeanalyse mithilfe von GPOWER ist analog zu der oben beschriebenen.

## Schlussfolgerungen

Teststärkeüberlegungen sind bei der Bewertung statistischer Entscheidungen, besser aber schon bei der Planung von Untersuchungen unumgänglich. Wie wir zu zeigen versucht haben, bieten verschiedene Formen von Teststärkeanalysen für eine breite Palette statistischer Tests ein sehr flexibles Instrumentarium sowohl zur rationalen Planung von Untersuchungen wie auch zur inferenzstatistischen Analyse oder Neubewertung bereits durchgeführter Experimente und Studien. Mit leicht zu handhabenden Werkzeugen wie GPOWER lassen sich diese Analysen problemlos umsetzen, sodass der in der Vergangenheit häufig geäußerte Einwand, Teststärkeanalysen seien aufwändig und daher unpraktikabel, nicht mehr überzeugen kann.

## Literatur

- Bredenkamp, J. (1969). Über die Anwendung von Signifikanztests bei theorie-testenden Experimenten. *Psychologische Beiträge*, 11, 275–285.
- Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung*. Frankfurt a. M.: Akademische Verlagsgesellschaft.
- Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente*. Darmstadt: Steinkopff.
- Bredenkamp, J., Erdfelder, E. (1985). Multivariate Varianzanalyse nach dem V-Kriterium. *Psychologische Beiträge*, 27, 127–154.
- Buchner, A., Brandt, M. (2003). Further evidence for systematic reliability differences between explicit and implicit memory tests. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 56A, 193–209.
- Buchner, A., Erdfelder, E., Faul, F. (1996). Teststärkeanalysen. In E. Erdfelder, T. Meiser, R. Mausfeld u. G. Rudinger (Hg.), *Handbuch Quantitative Methoden* (S. 123–136). Weinheim: Psychologie Verlags Union.
- Buchner, A., Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, 40, 227–259.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, NJ: Erlbaum.
- Erdfelder, E., Faul, F., Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers*, 28, 1–11.
- Gigerenzer, G., Murray, D.J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
- Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen*. Göttingen: Hogrefe.
- Hays, W.L. (1972). *Statistics for the social sciences* (2. Aufl.). London: Holt, Rinehart & Winston.
- Kupper, L.L., Hafner, K.B. (1989). How appropriate are popular sample size formulas? *The American Statistician*, 43, 101–105.

- Muller, K.E., Barton, C.N. (1989). Approximate power for repeated measures ANOVA lacking sphericity. *Journal of the American Statistical Association*, 84, 549–555.
- O'Brien, R.G., Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, 97, 316–333.
- Pillai, K.C.S., Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika*, 54, 195–210.
- Popper, K.R. (1934/1982). *Logik der Forschung* (7. Aufl.). Tübingen: J.C.B. Mohr (im Original publiziert 1934).
- Rogers, J.L., Howard, K.I., Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113, 553–565.
- Stevens, J.P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin*, 88, 728–737.
- Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3. Aufl.). Mahwah, NJ: Erlbaum.
- Welsch, W., Heunecke, O., Kuhlmann, H. (2000). *Auswertung geodätischer Überwachungsmessungen*. Heidelberg: Herbert Wichmann Verlag.
- Westermann, R., Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics*, 11, 117–146.
- Wottawa, H. (1981). *Psychologische Methodenlehre* (2. Aufl.). München: Juventa.