Teststärkeanalysen

Axel Buchner, Edgar Erdfelder und Franz Faul

Die Teststärke (*power*) eines statistischen Tests ist die Wahrscheinlichkeit, sich bei Anwendung dieses Tests für eine der Nullhypothese (H_0) entgegengestellte einfache Alternativhypothese (H_1) zu entscheiden, wenn diese in der zugrundeliegenden Population tatsächlich gültig ist. Ist die Alternativhypothese keine einfache (Punkt-) Hypothese, sondern – wie häufig in der Psychologie – eine zusammengesetzte Hypothese, so ist für jeden Parameterwert, der in den Bereich der zusammengesetzten Alternativhypothese fällt, eine Teststärke definiert (vgl. Willmes, in diesem Band).

Offenbar muß die Teststärke bekannt sein, um einzuschätzen, welche Chance einer bestimmten Alternativhypothese bei einem statistischen Test gegeben wird oder gegeben wurde. Dieses Wissen liefern uns Teststärkeanalysen. Zugleich erfahren wir dadurch etwas über die komplementäre Wahrscheinlichkeit, daß fälschlich zugunsten von H_0 entschieden wird, obwohl eine bestimmte H_1 gilt („Fehler zweiter Art" oder „β-Fehler"). Bezeichnet man die Wahrscheinlichkeit eines Fehlers zweiter Art bei Gültigkeit einer bestimmten Alternativhypothese mit β, so ist die Teststärke[1] $1 - \beta$.

Obwohl in der psychologischen Forschung routinemäßig mit statistischen Tests gearbeitet wird, finden sich Teststärkeanalysen sehr selten. An Kritik dieses wegen der Bedeutung des Teststärkekonzepts unbefriedigenden Zustandes hat es ebensowenig gefehlt wie an prinzipiell praktikablen Verbesserungsvorschlägen (Bredenkamp, 1972, 1980; Cohen, 1988; Erdfelder, Faul & Buchner, 1996; Gigerenzer, 1993; Hager, 1987; Sedlmeier & Gigerenzer, 1989). Bislang sind jedoch keine spürbaren Veränderungen zu verzeichnen. Ziel dieses Kapitels ist deshalb, durch Verweis auf einfach handhabbare Hilfsmittel und eine Reihe von konkreten Anwendungsbeispielen zur Verbreitung von Teststärkeanalysen in der psychologischen Forschung beizutragen.

Wenn Teststärkeanalysen durchgeführt werden, dann sind zwei Varianten besonders häufig, nämlich Apriori- und Posthoc-Teststärkeanalysen. Diese und noch eine dritte wichtige Variante – nämlich Kompromiß-Teststärkeanalysen – wollen wir hier behandeln. Kurz charakterisiert sind Apriori-Analysen *idealistisch*, Posthoc-Analysen werden oft *kritisch* gebraucht, und Kompromiß-Analysen können als *pragmatisch* bezeichnet werden. Bei Apriori-Analysen sucht man *vor* einer Untersuchung die notwendige Stichprobengröße N, die bei Gültigkeit einer bestimmten H_1 und festgelegter α-Fehlerwahrscheinlichkeit eine gewünschte (hohe) Teststärke $1 - \beta$ garantiert.[2] Posthoc-Analysen werden dagegen erst durchgeführt, wenn die Unter-

[1] Man beachte, daß die Teststärke im Beitrag von Willmes (in diesem Band) mit β statt mit $1 - \beta$ bezeichnet wird. Die von uns gewählte Notation ist in der psychologischen Methodenlehre gebräuchlicher, die von Willmes in der mathematischen Statistik.

[2] α bezeichnet wie üblich die Wahrscheinlichkeit des „Fehlers erster Art", d.h. sich bei Gültigkeit von H_0 fälschlich für die H_1 zu entscheiden.

https://doi.org/10.25521/HQM10

suchung bereits abgeschlossen ist (N also festliegt) und kritisch nachgefragt wird, welche Bewährungschance eine spezifische H_1 bei diesem N und dem gewählten α überhaupt hatte. Kompromiß-Teststärkeanalysen schließlich sind insofern pragmatisch, als sie der Tatsache Rechnung tragen, daß unsere Ressourcen oft nicht für das ideale N ausreichen. Wir müssen also mit einem suboptimalen N leben, wollen aber dennoch bestimmten Alternativhypothesen eine Chance geben und suchen folglich nach einem rationalen Kompromiß zwischen einem möglichst kleinen α und einer möglichst großen Teststärke $1 - \beta$. Dazu überlegen wir, welches Gewicht q wir der Fehlerwahrscheinlichkeit β im Vergleich zu α geben wollen ($q = \beta/\alpha$) und errechnen dann β sowie α und den damit verbundenen kritischen Wert der jeweiligen Teststatistik als Funktion von N, q und der Größe des durch die H_1 definierten Effekts (Erdfelder, 1984). Bisweilen kommt es auch vor, daß wir nicht ein zu kleines, sondern ein zu großes N haben, etwa bei großangelegten Feldstudien zur Validierung von Tests. Auch in diesem Fall sind Kompromiß-Teststärkeanalysen angebracht.

Die Teststärke eines statistischen Tests ist genau dann eindeutig festgelegt, wenn einerseits die nonzentrale Verteilung einer Teststatistik und andererseits der vom α-Niveau abhängende kritische Wert (beim zweiseitigen Test: die kritischen Werte) der Teststatistik bekannt sind. Die nonzentrale Verteilung ist die Verteilung der Teststatistik, die bei Gültigkeit einer bestimmten H_1 für einen bestimmten Stichprobenumfang resultiert. Wir werden uns in diesem Beitrag auf Teststärkeanalysen für die wichtigsten statistischen Tests in der psychologischen Forschung konzentrieren, nämlich t-, F- und χ^2-Tests. Die nonzentralen Verteilungen der entsprechenden Teststatistiken sind im Falle der Gültigkeit der Verteilungsvoraussetzungen entweder exakt oder approximativ nonzentrale t-, F- und χ^2-Verteilungen, die durch die Freiheitsgrade (df) und einen sogenannten Nonzentralitätsparameter eindeutig festgelegt sind (zur Definition und Genese dieser Verteilungen vgl. Johnson & Kotz, 1970, und Fußnote 2 im Kapitel zum Allgemeinen Linearen Modell von Andres, in diesem Band). Das zentrale Problem bei der Durchführung von Teststärkeanalysen für die genannten Tests besteht darin, die Freiheitsgrade des Tests sowie den Nonzentralitätsparameter, der zu einer bestimmten Alternativhypothese und einem bestimmten Stichprobenumfang gehört, korrekt zu bestimmen. Hat man diese beiden Angaben, so kann man bei Benutzung entsprechender Tabellenwerke (z.B. Cohen, 1988; Odeh & Fox, 1991) oder Computerprogramme (vgl. die Übersicht bei Goldstein, 1989) ablesen bzw. errechnen, wie groß die Wahrscheinlichkeit ist, daß ein Wert aus der betreffenden nonzentralen t-, F- oder χ^2-Verteilung größer oder gleich dem kritischen Wert ist, der durch das α-Niveau festgelegt ist. Das ist aber gerade die Teststärke $1 - \beta$ des Tests relativ zur gewählten Alternativhypothese.

Wir werden im folgenden Beispiele für alle drei genannten Tests besprechen und dabei immer deutlich machen, wie die jeweilige Alternativhypothese und der Stichprobenumfang mit dem Nonzentralitätsparameter der nonzentralen t-, F- oder χ^2-Verteilung zusammenhängen. Zugleich werden wir konkret angeben, wie man die Berechnung von Stichprobenumfängen (bei Apriori-Teststärkeanalysen), Teststärken (bei Posthoc-Teststärkeanalysen) und α- sowie β-Werten (bei Kompromiß-Teststärkeanalysen) mit dem Programm GPOWER einfach und bequem durchführen kann. Dieses Programm existiert in zwei bezüglich der verwendeten numerischen Verfahren äquivalenten und hinsichtlich der Benutzeroberfläche ähnlichen Versionen. Eine

Version läuft auf allen IBM-kompatiblen Computern (Faul & Erdfelder, 1992), die andere auf Macintosh-Computern (Buchner, Faul & Erdfelder, 1992).[3]

Alternativhypothesen können bei Benutzung von GPOWER auf zweierlei Weise spezifiziert werden. Eine Möglichkeit ist die Eingabe von Parametern des zugrundeliegenden statistischen Modells (z.B. von Populationsmittelwerten oder -varianzen), die einer einfachen Alternativhypothese entsprechen. Eine zweite Möglichkeit ist die Eingabe standardisierter Effektgrößen, wie sie durch Cohen (1988) definiert wurden. Diese Effektgrößen sind Funktionen der Parameter des statistischen Modells, die als standardisierte Maße für den „Abstand" einer bestimmten H_1 von einer H_0 interpretiert werden können (vgl. Hager, in diesem Band). Ein bestimmter Wert eines solchen Effektmaßes charakterisiert somit immer eine Klasse von einfachen Alternativhypothesen, die gleich stark von der H_0 abweichen, in dem Sinne, daß für alle diese einfachen Alternativhypothesen die gleiche Teststärke resultiert.

Häufig fällt es mangels präziser Theorien und einschlägiger Voruntersuchungen schwer, eine bestimmte einfache Alternativhypothese als für Teststärkeberechnungen relevant auszuzeichnen. Deshalb liegt es nahe, von standardisierten Effektgrößen auszugehen. Cohen (1988) hat Vorschläge unterbreitet, welche Effektgrößenwerte als „klein", „mittel" oder „groß" gelten können. Ist die Auszeichnung einer bestimmten einfachen Alternativhypothese nicht möglich, so empfehlen wir die Orientierung an den Cohenschen Effektgrößenkonventionen, wobei von der konkreten Fragestellung abhängig gemacht werden sollte, ob „kleine", „mittlere" oder „große" Effekte durch den Test aufgedeckt werden sollen. Impliziert eine psychologische Hypothese beispielsweise eine H_0, so mag man bereits an der Entdeckung kleiner Abweichungen von H_0 interessiert sein. Geht es dagegen um den Effektivitätsnachweis für eine teure Therapie, so wird man vielleicht „mittlere" oder gar „starke" Effekte der Therapie fordern und hinnehmen, daß die Teststärke für kleine Effekte gering ausfällt.

Die Cohenschen Konventionen sind nützliche Orientierungspunkte, können aber bei unkritischer Verwendung insofern in die Irre führen, als sie suggerieren, daß eine bestimmte Effektgröße in verschiedenen Anwendungskontexten immer die gleiche psychologische Bedeutung hat. Gegen derartige Fehleinschätzungen kann man sich schützen, indem man keinen Effektgrößenwert festlegt, ohne ihn zunächst in exemplarische Parameter des zugrundeliegenden Modells „rückübersetzt" zu haben. So erhält man eine Vorstellung davon, was eine bestimmte Effektgröße – z.B. in termini von Populationsmittelwerten und -varianzen – bedeuten kann.

1 Teststärkeanalysen für t-Tests

1.1 t-Tests für unabhängige Stichproben

In einer inzwischen klassischen und häufig zitierten Studie haben Warrington und Weiskrantz (1970) die Erinnerungsleistungen amnestischer Patienten mit denen einer

[3]Beide Versionen können kostenlos mittels *anonymous ftp* über das *Internet* vom *ftp-Server* der Universität Trier (ftp.uni-trier.de, Verzeichnis pub/pc/msdos, Datei gpower2d.exe (deutsche Version) oder gpower2i.exe (englische Version) bzw. pub/mac/local, Datei gpower21.sit) kopiert werden. Alternativ sind beide Versionen auch über das *World Wide Web* zugänglich (http://www.psychologie.uni-trier.de:8000/projects/gpower.html). Einzelheiten hierzu wie auch zu den Algorithmen und zur Handhabung der Versionen findet man in Erdfelder et al. (1996).

Kontrollgruppe verglichen und dabei neben üblichen direkten Gedächtnisprüfverfahren wie Rekognition auch indirekte Prüfverfahren wie Wortstammergänzung verwendet. Indirekte Prüfverfahren sollen Erfahrungsnachwirkungen ohne eine ausdrückliche Erinnerungsinstruktion erfassen. Das interessante Ergebnis war, daß Amnestiker beispielsweise im Rekognitionsmaß signifikant schlechter abschnitten als die Kontrollgruppe (Mittelwerte 8 vs. 13), nicht aber bei der Wortstammergänzungsaufgabe (Mittelwerte 14.5 vs. 16).

Kann man aufgrund dieser Ergebnisse begründet behaupten, daß Amnestiker in indirekten Gedächtnismaßen – zumindest im Hinblick auf Wortstammergänzung – genauso gut wie gesunde Kontrollpersonen sind? An den oben erwähnten Untersuchungsergebnissen fällt zunächst auf, daß zumindest auf deskriptiver Ebene sehr wohl ein Unterschied zwischen Amnestikern und der Kontrollgruppe in der Wortstammergänzung existiert. Bedenkt man nun, daß die Stichprobe aus lediglich 4 Amnestikern und 8 Personen in der Kontrollgruppe bestand, so muß vermutet werden, daß die Teststärke des berichteten t-Tests für unabhängige Stichproben nicht besonders groß gewesen sein kann. Zusätzlich wird sie dadurch verringert, daß die beiden Gruppen ungleich besetzt waren. Für den Nonzentralitätsparameter δ der nonzentralen t-Verteilung (Johnson & Kotz, 1970, Kap. 31) gilt nämlich[4]

$$\delta = d \cdot \sqrt{\frac{n_1 \cdot n_2}{N}}\,, \tag{1}$$

wobei n_1 und n_2 die Stichprobengrößen der beiden Gruppen sind, $N = n_1 + n_2$ und $d = |\mu_1 - \mu_2|/\sigma$ ist. d ist der von Cohen (1988) verwendete Effektgrößenindex für t-Tests für unabhängige Stichproben. μ_1 und μ_2 sind die Populationsmittelwerte für die beiden Gruppen, deren Differenz an σ, der gemeinsamen Standardabweichung in der Population, standardisiert wird. Die H_0 des einseitigen t-Tests ist $\mu_2 - \mu_1 \leq 0$. In (1) sieht man: Je ungleicher die Gruppen bei gleichem Gesamtstichprobenumfang N für ein bestimmtes d besetzt sind, desto kleiner wird δ und damit die Teststärke.

Wie groß war – bezogen auf die Wortstammergänzungsaufgabe – die Teststärke bei Warrington und Weiskrantz (1970), wenn die zugrundeliegenden Populationsmittelwerte 14.5 für Amnestiker und 16 für die Kontrollpopulation betrugen? Nehmen wir an, die Streuung des Leistungsmaßes, die für die beiden Stichproben leider ebensowenig berichtet wird wie der empirische t-Wert, sei in den zugrundeliegenden Populationen jeweils 3. Wenn die Annahme vernünftig ist, daß die Leistungsdaten normalverteilt sind, bedeutet dies, daß 95% der Gedächtnisleistungen in einem Intervall von $2\sigma = 6$ um den Gruppenmittelwert liegen sollten. Wir wählen in GPOWER ≪Post hoc≫ als Typ der Analyse und Teststärkeanalysen für ≪t-Test (means)≫. Da eine einseitige Fragestellung vorliegt, ist der t-Test ≪one-tailed≫. Dann ermitteln wir mit ≪Calc d≫ $d = (16 - 14.5)/3 = 0.5$ als Effektgröße, was „mittleren" Effekten nach Cohen entspricht. Wie groß war die Chance, diesen Effekt beim benutzten Niveau $\alpha = .05$ zu entdecken? Wir spezifizieren in GPOWER $\alpha = .05, n_1 = 4$ und $n_2 = 8$. Das Ergebnis ist ernüchternd. Die Teststärke dieses Tests beträgt gerade einmal 0.1887. GPOWER liefert zusätzlich den zum α-Niveau gehörenden kritischen

[4]Gleichung (1) reduziert sich für $n_1 = n_2$ zu $\delta = (d/2) \cdot \sqrt{N} = f \cdot \sqrt{N}$ für die t-Test-Situation, wobei f die standardisierte Effektgröße ist, die von Cohen (1988) auch im Zusammenhang mit F-Tests verwendet wird.

t-Wert $t(10) = 1.8125$ und den aus den Stichprobengrößen und d resultierenden Nonzentralitätsparameter $\delta = 0.8165$ (vgl. Gleichung (1)).

Fazit: Der von Warrington und Weiskrantz durchgeführte Test hatte kaum eine Chance, kleine oder mittlere Defizite der Amnestiker in der Wortstammergänzungsaufgabe aufzudecken. Erst ab einer Populationsmittelwertsdifferenz von 6.5085 zugunsten der Kontrollpopulation ($d = 2.1695$) wäre die Teststärke .95 und damit $\alpha = \beta$. Will man einen Effekt der Größe $d = 0.5$ bei $n_1 = 4$ und $n_2 = 8$ mit gleichen α- und β-Risiken entdecken ($q = 1$), so führt eine Kompromiß-Teststärkeanalyse mittels der ≪Compromise≫ Option von GPOWER zu der Empfehlung, $\alpha = \beta = .3422$ zu wählen (kritischer Wert: $t(10) = 0.4186$). Obwohl dieses Signifikanzniveau unter den gegebenen Umständen unseres Erachtens noch das kleinste Übel wäre, wird doch deutlich, daß dieser statistische Test kaum besser als ein Münzwurf ist, den man über Akzeptanz oder Verwerfung von H_0 entscheiden läßt.

1.2 t-Tests für abhängige Stichproben

Im Anschluß an die klassische Arbeit von Gesell und Thompson (1929) ist eine Reihe von Experimenten mit monozygoten Zwillingspaaren durchgeführt worden, bei denen jeweils ein zufällig ausgewähltes Zwillingskind ein bestimmtes motorisches Training erhielt, während das andere Kind des Paares keine Förderung bekam (zusammenfassend: McGraw, 1946). Auf diese Weise konnte experimentell untersucht werden, ob bestimmte Entwicklungsprozesse (z.B. Laufen lernen oder die Blase kontrollieren) ausschließlich reifungsbedingt sind oder ob Umwelteinflüsse die Entwicklung beschleunigen bzw. verlangsamen.

Wenn wir eine statistisch aussagekräftige Replikation der klassischen Experimente planen, die wir mit dem t-Test für abhängige Stichproben auswerten wollen, und wenn wir 20 Zwillingspaare untersuchen können, welche statistischen Fehlerrisiken wählen wir dann vernünftigerweise?

Wir bezeichnen das Alter, in dem das nicht geförderte bzw. das geförderte Kind eines Zwillingspaares eine bestimmte motorische Leistung beherrscht, mit X bzw. Y. Die H_0 des (einseitigen) t-Tests für abhängige Stichproben ist dann $\mu_{x-y} = \mu_x - \mu_y \leq 0$, wobei μ_{x-y} den Populationsmittelwert der Altersdifferenzen pro Paar bezeichnet. Der Effektgrößenindex d_z ist für diesen Fall definiert als

$$d_z = \frac{|\mu_{x-y}|}{\sigma_{x-y}} = \frac{|\mu_{x-y}|}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2cov_{xy}}} = \frac{|\mu_{x-y}|}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y}}\,, \tag{2}$$

wobei σ_{x-y} die Streuung der $(X - Y)$-Differenzen, cov_{xy} die Kovariation und ρ die (positive) Korrelation zwischen den X- und Y-Werten in der Population unter H_1 ist. Je größer diese Korrelation, desto größer ist bei sonst gleichen Bedingungen der Effektgrößenindex d_z. Bei Gültigkeit der H_1 ist die Referenzverteilung für unsere Teststatistik eine nonzentrale t-Verteilung mit $N-1$ Freiheitsgraden (N steht hier für die Anzahl der Zwillings- bzw. Meßwert*paare*) und dem Nonzentralitätsparameter

$$\delta = \frac{|\mu_{x-y}|}{\sigma_{x-y}} \cdot \sqrt{N} = d_z \cdot \sqrt{N}\,. \tag{3}$$

Nehmen wir nun für unser Beispiel an, daß ein durchschnittlicher Entwicklungsunterschied von 2 Monaten bezüglich einer bestimmten motorischen Leistung entdeckt

werden soll. Aus Voruntersuchungen sei bekannt, daß die Streuung der Altersdifferenzen bezüglich dieses Merkmals 4 Monate betrage. Demzufolge ist die zu entdeckende Effektgröße $d_z = 2/4 = 0.5$. Bei der Teststärkeanalyse mit GPOWER ist nun zu beachten, daß nicht die Option ≪t-Test (means)≫ angewählt werden darf, da diese für *unabhängige* Stichproben konzipiert ist und daher automatisch von $N - 2$ Freiheitsgraden ausgeht. Beim t-Test für abhängige Stichproben betragen die Freiheitsgrade aber, wie gesagt, $N - 1$, so daß wir ≪Other t-Tests≫ wählen, wo wir die Freiheitsgrade unabhängig von N wählen können. Hier wird nun nicht d, sondern f als Effektgrößenindex verlangt, was aber kein Problem ist, da der Nonzentralitätsparameter gemäß der Formel

$$\delta = f \cdot \sqrt{N} \tag{4}$$

errechnet wird. Wenn man für f also d_z gemäß Gleichung (2) eingibt (im Beispiel also 0.5), so erhält man den richtigen Nonzentralitätsparameter.

Da wir nur auf 20 Zwillingspaare zurückgreifen können, entschließen wir uns zu einer Kompromiß-Teststärkeanalyse. Beide Fehlerwahrscheinlichkeiten β und α sollen gleich groß sein, so daß $q = \beta/\alpha = 1$. Nun können wir mit GPOWER die Teststärkeanalyse durchführen. Wir wählen ≪Compromise≫ als Typ der Analyse. Der Test ist wieder einseitig, wir wählen also ≪one-tailed≫. Als ≪effect size f≫ spezifizieren wir, wie gesagt, $d_z = 0.5$. Bei ≪N≫ geben wir die Anzahl der Meßwert- bzw. Zwillingspaare, also 20, ein. Als ≪Beta/alpha ratio≫ ist $q = \beta/\alpha = 1$ zu spezifizieren. Schließlich hat unser t-Test für abhängige Stichproben $N - 1 = 19$ Freiheitsgrade. GPOWER berechnet als Nonzentralitätsparameter $\delta = 2.2361$ und schlägt uns vor, $\alpha = \beta = .1357$ zu wählen, so daß die Teststärke $1-\beta = .8643$ beträgt. Die Teststatistik muß den kritischen Wert $t(19) = 1.1328$ überschreiten, wenn H_0 (d.h. hier: die Reifungshypothese) verworfen werden soll. Dieses Ergebnis ist etwas erfreulicher als das im letzten Abschnitt erzielte. Wem es noch nicht erfreulich genug ist, der muß die finanziellen und zeitlichen Ressourcen z.B. soweit erhöhen, daß 45 Zwillingspaare untersucht werden können. Wie eine Posthoc-Analyse mit ≪Other t-Tests≫ und den Eingaben $\alpha = .05, f = .5, N = 45$ und $df = 44$ zeigt, ist erst bei dieser Stichprobengröße eine Teststärke von $1 - \beta = .9512$ garantiert.

Unser Beispiel bezog sich auf Zwillingspaare. Analog verfahren wir aber natürlich auch in anderen Fällen „abhängiger" Messungen, etwa dann, wenn X und Y Meßwiederholungen an denselben Vpn zugrunde liegen.

1.3 t-Tests für Korrelationen

Berry und Broadbent (1984) untersuchten die Beziehung zwischen der Leistung bei der Steuerung einer kleinen Computersimulation und der Fähigkeit, eine bestimmte Form von Fragen über das simulierte System zu beantworten. Als überraschend, aber wichtig wurde bewertet, daß etwa in Experiment 1 die Korrelation zwischen beiden Aufgabenteilen negativ war (zwischen -.25 und -.30): Je besser man die Simulation steuern kann, desto weniger kann man darüber auf andere Weise Auskunft geben; so schien es jedenfalls. Allerdings war die negative Korrelation nicht statistisch signifikant, wofür die Autoren die geringe Stichprobengröße ($N = 12$) verantwortlich machten. Wir können uns nun fragen, wie groß in diesem Experiment die Chance war, Korrelationen einer bestimmten Größe überhaupt aufzudecken.

Betrachten wir zunächst die Definition des Nonzentralitätsparameters δ bei t-Tests für Korrelationen zwischen zwei Merkmalen:

$$\delta = \sqrt{\frac{\rho^2}{1-\rho^2}} \cdot \sqrt{N} \,, \tag{5}$$

wobei ρ die Populationskorrelation nach der H_1 und N die Stichprobengröße – hier also die Anzahl der Meßwertpaare – ist. Cohen (1988) definiert Korrelationen von $\rho = .30$ als „mittelgroße" Effekte. Welche Chance hätte man in einem Experiment wie dem von Berry und Broadbent (1984), einen Effekt in dieser Größenordnung zu entdecken? In GPOWER wählen wir zunächst ≪Post hoc≫ als Typ der Analyse und Teststärkeanalysen für ≪t-Test (correlations)≫. Der t-Test ist ≪one-tailed≫, da wir die $H_0 : \rho \geq 0$ gegen die $H_1 : \rho < 0$ testen wollen. Als Effektgröße können wir direkt $r = .30$ eingeben. Die Stichprobengröße beträgt $N = 12$, und $\alpha = .05$. Der Nonzentralitätsparameter ist $\delta = 1.0894$, der t-Wert der Analyse müßte $t(10) = -1.8125$ unterschreiten. Die Teststärke beträgt lediglich $1 - \beta = .2648$, was Berry und Broadbents (1984) Erklärung für die Nichtsignifikanz plausibel erscheinen läßt.

Wie groß sollte eine Stichprobe sein, damit man mit einer Teststärke von 0.95 mittelgroße Effekte entdecken kann? Wir wählen ≪A priori≫ als Typ der Analyse, geben .95 als Teststärke sowie .30 als Effektgröße an und erhalten eine erforderliche Stichprobengröße von $N = 111$. Der kritische Wert der Teststatistik liegt dann bei $t(109) = -1.6590$, der Nonzentralitätsparameter beträgt $\delta = 3.3133$.

2 Teststärkeanalysen für F-Tests

Für F-Tests im Rahmen von Regressions- und Varianzanalysen (vgl. das Kapitel zum Allgemeinen Linearen Modell von Andres, in diesem Band) verwenden wir als Effektgrößenindex Cohens f bzw. f^2. Zunächst gehen wir auf Varianzanalysen ein, dann auf Regressionsanalysen.

2.1 Varianzanalysen mit fixierten Effekten

Der Bezug von f^2 zum Nonzentralitätsparameter λ der nonzentralen F-Verteilung (Johnson & Kotz, 1970, Kap. 30) ist gegeben durch

$$\lambda = f^2 \cdot n \cdot k = f^2 \cdot N \,, \tag{6}$$

wobei n die Anzahl der Vpn in jeder der k Gruppen des Designs ist. Der Effektgrößenindex f selbst ist

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \,, \tag{7}$$

wobei η^2 der Anteil der gesamten Populationsvarianz ist, der durch die in der H_1 spezifizierten Gruppenunterschiede in der Population aufgeklärt wird. Im Falle ungleicher Gruppengrößen n_j gilt für den Effektgrößenindex f

$$f = \frac{\sqrt{\frac{\sum_{j=1}^{k} n_j \cdot (\mu_j - \bar{\mu})^2}{N}}}{\sigma} \,. \tag{8}$$

In Gleichung (8) bezeichnet n_j die Anzahl der Versuchspersonen, μ_j den Populationsmittelwert in Gruppe j, $\bar{\mu} = (\sum_{j=1}^{k} n_j \cdot \mu_j)/N$ das gewichtete Mittel der k Populationsmittelwerte, N die Gesamtstichprobengröße und σ die in allen Gruppen gleiche Streuung in der Population.

Einfaktorielle Designs

Am einfachsten sind Teststärkeanalysen für einfaktorielle ANOVAs. Ein gutes Beispiel liefert die Arbeit von Schmitt, Hoser und Schwenkmezger (1991), in der untersucht wurde, wie stark verschiedene Facetten von Ärgerausdruck auftreten als Funktion der Verantwortlichkeit eines fremden Schädigers für den angerichteten Schaden. Die Verantwortlichkeit wurde in sechs Stufen variiert. Nehmen wir an, wir wollten die Untersuchung von Schmitt et al. (1991) replizieren. Unsere H_0 ist, daß sich die sechs Gruppen in ihrem Ärgerzustand nicht unterscheiden werden. Als Schätzer für die Größe des zu entdeckenden Populationseffekts nehmen wir den empirischen Effekt in der vorliegenden Untersuchung, den wir mit GPOWER leicht berechnen können. Nach einem Klick auf ≪Calc f≫ können wir einfach die einzelnen Gruppengrößen (20), die empirischen Mittelwerte der gemessenen Ärgerausdruckswerte (nämlich 15.3, 18.3, 20.5, 22.7, 23.3, und 24.8; siehe Schmitt et al., 1991, S. 641) sowie deren durchschnittliche Streuung ($\approx$ 6.5; M. Schmitt, persönliche Mitteilung, April 1995) eintragen und f nach Gleichung (8) berechnen. Wir erhalten $f = 0.4963$. Wieviele Personen benötigen wir für die Replikation, wenn wir $\alpha = \beta = .05$ für akzeptabel halten? In GPOWER wählen wir ≪A priori≫ als Typ der Analyse und ≪F-Test (ANOVA)≫ als Test mit der Option ≪Global≫.[5] Wir setzen $\alpha = .05, 1 - \beta = .95, f = 0.4963$ und die Anzahl der Gruppen auf 6. Der Nonzentralitätsparameter beträgt $\lambda = 22.1682$, der kritische F-Wert liegt bei $F(5, 84) = 2.3231$. Wir benötigen insgesamt $N = 90$ Personen – 15 in jeder Gruppe – für die Replikation.

Mehrfaktorielle Designs

Beispielhafte Teststärkeanalysen für komplexere Designs und verschiedene varianzanalytische Modelle liefert Koele (1982). Bleiben wir zunächst beim Modell fixierter Effekte, das auch der Untersuchung von Schmitt et al. (1991) zugrunde lag. Koele (1982) orientiert sich an einem $A \times B$-Design. Faktor A hat $k_A = 3$, Faktor B hat $k_B = 4$ Stufen. Wie sieht die Teststärke für die beiden Haupteffekte und die Interaktion aus? Im Prinzip können wir uns am Vorgehen bei einfaktoriellen Designs orientieren. Veränderungen ergeben sich etwa dadurch, daß die Anzahl der Nennerfreiheitsgrade durch die weiteren Faktoren geringer wird (nämlich $N - k_A \cdot k_B$). Um Koeles (1982) Beispiele nachvollziehen zu können, wählen wir in GPOWER zunächst ≪Post hoc≫ als Typ der Analyse, ≪F-Test (ANOVA)≫ als Typ der Teststärkeanalysen – nun allerdings mit der Option ≪Special≫ – und spezifizieren $\alpha = .05$. Als Effektgröße setzt Koele (1982) $f^2 = 0.05$ voraus; wir spezifizieren

[5] Globale Tests beziehen sich auf die globale Nullhypothese, daß keinerlei Mittelwertsunterschiede zwischen den Gruppen des Designs vorliegen. Spezielle Tests beziehen sich auf Nullhypothesen über bestimmte Teilmengen linearer Kontraste, z.B. Haupteffekte, Wechselwirkungen und Trendkomponenten (vgl. Erdfelder et al., 1996).

in GPOWER also $f = \sqrt{0.05} = 0.2236$. Alternativ hätten wir f nach Gleichung (7) bzw. (8) berechnen können. Beide Varianten stehen uns unter ≪Calc f≫ zur Verfügung, wobei wir berücksichtigen müssen, daß η^2 hier als *partielles* η^2 zu interpretieren ist. In jeder der 12 Zellen des Designs finden sich 10 Beobachtungen, wir spezifizieren also für die Gesamtstichprobe $N = 120$. Die Anzahl der Zellen ist unter ≪Groups≫ einzugeben. Um unter diesen Bedingungen die Teststärke für den Faktor A berechnen zu können, benötigen wir nur noch die Zählerfreiheitsgrade für diesen Faktor, nämlich $k_A - 1 = 2$. Diesen Wert geben wir bei ≪Numerator DF≫ ein. Der Nonzentralitätsparameter beträgt $\lambda = 5.9996$, der kritische F-Wert liegt bei $F(2, 108) = 3.0804$. Die Teststärke beträgt 0.5714. Völlig analog berechnen wir für den vierstufigen Faktor B mit $k_B - 1 = 3$ Zählerfreiheitsgraden eine Teststärke von 0.5020; der kritische F-Wert liegt bei $F(3, 108) = 2.6887$. Das ist schon nicht berauschend, doch die Teststärke wird noch geringer, wenn wir die Interaktion der Faktoren A und B betrachten. Hier haben wir bei $(k_A - 1) \cdot (k_B - 1) = 6$ Zählerfreiheitsgraden eine Teststärke von lediglich 0.3806 und einen kritischen F-Wert von $F(6, 108) = 2.1837$. Geringe Abweichungen zwischen diesen Werten und denen, die Koele (1982) berichtet, gehen darauf zurück, daß Koele Approximationen, GPOWER dagegen „präzise" Routinen zur Berechnung der relevanten Verteilungen verwendet (siehe Erdfelder et al., 1996). Größere Differenzen können auftreten im Vergleich zu dem Verfahren, das Cohen (1988) bei der Verwendung seiner Teststärketabellen für spezielle Effekte und Interaktionen in komplexen Designs vorschlägt. Wie sich zeigen läßt, ist Cohens Verfahren fehlerbehaftet (siehe Bradley, Russell & Reeve, in press; Erdfelder et al., 1996; Koele, 1982, Fußnote 1).

Mittelwertkontraste

Bisweilen interessiert die Frage, ob ein Effekt auf einen ganz bestimmten Mittelwertkontrast zurückgeht, beispielsweise einen linearen Trend. Mittelwertkontraste werden durch geeignete Kodierungen in Kontrastvariablen realisiert. Bei Trendanalysen sind dies orthogonale polynomiale Kontraste. Die H_0 ist in diesem Fall, daß der betrachtete Mittelwertkontrast in der Population keine Varianz in der abhängigen Variablen aufklärt, oder anders ausgedrückt, daß die partielle Populationskorrelation R_p zwischen der abhängigen Variablen und der Kontrastvariablen für den betrachteten Kontrast gleich 0 ist. Entsprechend ergibt sich der Effektgrößenindex f nach

$$f = \sqrt{\frac{R_p^2}{1 - R_p^2}} \,. \tag{9}$$

Nehmen wir an, wir erwarten für den linearen Trend bei Faktor B des im letzten Abschnitt besprochenen mehrfaktoriellen Designs einen Effekt der Größe $f = 0.1$, weil wir davon ausgehen, daß das partielle Korrelationsquadrat zwischen der Kontrastvariablen für den linearen Trend und der abhängigen Variablen $R_p^2 = .01$ beträgt. Der lineare Trend hat einen Zählerfreiheitsgrad, weil wir nur die Korrelation zwischen einer Kontrastvariablen und der abhängigen Variablen betrachten. Der Nonzentralitätsparameter beträgt $\lambda = 1.2$, der kritische F-Wert liegt bei $F(1, 108) = 3.9290$. Wie sich zeigt, ist die Teststärke in diesem Fall mit .1922 äußerst gering.

2.2 Regressionsanalysen

Für F-Tests im Zusammenhang mit Regressionsanalysen ist bei der Definition des Nonzentralitätsparameters nur das zu berücksichtigen, was in den vorangegangenen Abschnitten schon für F-Tests bei Varianzanalysen dargelegt wurde. Allerdings verwenden wir hier mit Cohen (1988) nicht f, sondern f^2 als Effektgrößenindex, wobei $f^2 = PV_s/PV_e$, PV_s der durch die Prädiktoren erklärte systematische Varianzanteil und PV_e der Anteil an Fehlervarianz in der Population ist. Alternativ ist es im Kontext von multiplen Regressionsanalysen auch sinnvoll, f^2 zu schreiben als

$$f^2 = \frac{R^2_{Y \cdot A}}{1 - R^2_{Y \cdot A}} , \tag{10}$$

wobei $R^2_{Y \cdot A}$ das multiple Korrelationsquadrat einer Menge von Prädiktoren A mit einem Kriterium Y in der Population repräsentiert. Gleichung (10) weist eine offenkundige Verwandtschaft mit Gleichung (7) auf, was schlicht daran liegt, daß die einfaktorielle Varianzanalyse als ein Spezialfall der multiplen Regression mit kategorialen Prädiktoren aufgefaßt werden kann.

Betrachten wir ein konkretes Beispiel. Mallinckrodt und Bennett (1992) haben die Effekte von Arbeitslosigkeit und deren Begleitumstände auf die psychische Gesundheit untersucht. Dabei interessierte, ob und welche Beziehungen zwischen sechs verschiedenen Formen sozialer Unterstützung und berichteten Depressionssymptomen von Arbeitslosen bestehen. Dies ist die klassische Fragestellung für eine multiple Regressionsanalyse. Wir prüfen die multiple Korrelation einer Menge von Prädiktoren A mit einem Kriterium Y, wobei der durch die Prädiktormenge A bei Y aufgeklärte Varianzanteil $R^2_{Y \cdot A}$ interessiert. Die H_0 besagt, daß $R^2_{Y \cdot A}$ in der Population Null ist. Mallinckrodt und Bennett (1992) fanden tatsächlich Zusammenhänge der sechs Prädiktoren mit Depressionssymptomen: Je geringer die soziale Unterstützung, desto eher waren Depressionssymptome als Folge der Arbeitslosigkeit zu konstatieren.

Wir planen eine Replikation dieser Untersuchung unter deutschen Verhältnissen. Nehmen wir an, wir erwarten ein multiples Korrelationsquadrat von mindestens .35 zwischen den verschiedenen Formen sozialer Unterstützung und den Depressionssymptomen, weil es sich erst ab dieser Größenordnung lohnt, über die systematische Einbeziehung dieser Variablen in Betreuungsprogramme für arbeitslose Menschen nachzudenken. Wir sind außerdem bereit, ein α-Fehlerrisiko von .10 in Kauf zu nehmen. Die β-Fehlerwahrscheinlichkeit sollte aber mit .01 recht gering sein, denn wir wollen möglichst keine effektiven Einflußgrößen übersehen.

Wie groß muß unsere Stichprobe sein? In GPOWER wählen wir ≪A priori≫ als Typ der Analyse und ≪F-Test (MCR)≫ als Test mit der Option ≪Global≫. Wir setzen $\alpha = .10, 1 - \beta = .99$ und die Anzahl der Prädiktoren auf 6 (die sechs verschiedenen Formen sozialer Unterstützung). Mit ≪Calc f2≫ berechnen wir, daß bei einem multiplen Korrelationsquadrat zwischen den sechs Prädiktoren und dem Kriterium von $R^2_{Y \cdot A} = .35$ der Effektgrößenindex $f^2 = 0.5385$ beträgt. Unter diesen Bedingungen beträgt der Nonzentralitätsparameter $\lambda = 28.002$, der kritische F-Wert liegt bei $F(6, 45) = 1.9094$, und wir benötigen 52 Personen für unsere Replikation.

Nach Mallinckrodt und Bennett (1992) sollen vor allem emotionale Bewältigungsressourcen effektiv sein. Unter den sechs Facetten sozialer Unterstützung, die

wir in unserer multiplen Regressionsanalyse berücksichtigt haben, qualifizieren sich zwei am ehesten für diese Kategorie: das Vorhandensein emotionaler Bindungen und die Wertschätzung, die einem von anderen entgegengebracht wird. Wie sieht in unserer Replikation die Teststärke für diese Untermenge an Prädiktoren aus?

Formal testen wir hierbei die Korrelationen von zwei Mengen von Prädiktoren A und B mit einem Kriterium Y. Hierbei bezeichnet $R^2_{Y \cdot A,B}$ den durch A und B zusammen aufgeklärten Varianzanteil, $R^2_{Y \cdot A}$ den durch A alleine aufgeklärten Varianzanteil und $R^2_{Y \cdot (B \cdot A)} = R^2_{Y \cdot A,B} - R^2_{Y \cdot A}$ den Varianzanteil, den die Prädiktorenmenge B *zusätzlich* zur Menge A aufklärt. Die H_0 ist, daß die Prädiktorenmenge B in der Population keine zusätzliche Varianz aufklärt, also $R^2_{Y \cdot (B \cdot A)} = 0$. Entsprechend ist der Effektgrößenindex f^2 nun definiert als

$$f^2 = \frac{R^2_{Y \cdot A,B} - R^2_{Y \cdot A}}{1 - R^2_{Y \cdot A,B}} = \frac{R^2_{Y \cdot (B \cdot A)}}{1 - R^2_{Y \cdot A,B}} \,. \tag{11}$$

Nehmen wir an, wir erwarten, daß $R^2_{Y \cdot (B \cdot A)} = R^2_{Y \cdot A,B} - R^2_{Y \cdot A} = .35 - .10 = .25$, dann ist $f^2 = .25/.65 = 0.3846$. Alternativ können wir auch das partielle Korrelationsquadrat $R^2_{Y B \cdot A}$ betrachten, d.h. den Varianzanteil, der von der vom A-Einfluß bereinigten Prädiktormenge B an der Varianz aufgeklärt wird, die *nicht* von Menge A aufgeklärt wird:

$$R^2_{Y B \cdot A} = \frac{R^2_{Y \cdot (B \cdot A)}}{1 - R^2_{Y \cdot A}} \,. \tag{12}$$

Für diesen Fall erhalten wir f^2 völlig parallel zu (10) nach

$$f^2 = \frac{R^2_{Y B \cdot A}}{1 - R^2_{Y B \cdot A}} \,. \tag{13}$$

Wegen der Parallelität zu (10) ist Gleichung (13) auch parallel zum oben eingeführten Effektgrößenindex $f^2 = PV_s/PV_e$, nur daß hier PV_s der *durch die Prädiktormenge* B erklärte systematische Varianzanteil und PV_e die Residualvarianz ist. Für welche Art der Bestimmung von f^2 man sich entscheidet, hängt davon ab, über welche der involvierten R^2-Ausdrücke man Informationen besitzt. In GPOWER ist wegen der Konsistenz zum globalen F-Test die zuletzt diskutierte Form der Bestimmung des Effektgrößenindex implementiert.

Wie berechnen wir nun mit GPOWER die Teststärke für diesen Fall? Wir wählen ≪Post hoc≫ als Typ der Analyse und ≪F-Test (MCR)≫ als Test, allerdings mit der Option ≪Special≫. Wir setzen $\alpha = .10, N = 52$ und – wie zuvor – die Anzahl der Prädiktoren auf 6 (die sechs verschiedenen Formen sozialer Unterstützung). Diese Angabe wird von GPOWER bei der Berechnung der Nennerfreiheitsgrade berücksichtigt: df_n = N– Prädiktoren -1. Als Effektgröße ermitteln wir mit ≪Calc f2≫ ein $f^2 = 0.3846$, da $R^2_{Y B \cdot A} = .25/(1 - .10) = 0.2\overline{7}$. Als Zählerfreiheitsgrade des Tests geben wir die Anzahl der Prädiktoren an, deren zusätzlicher Effekt getestet werden soll, in unserem Falle also 2. Der Nonzentralitätsparameter beträgt $\lambda = 20.004$. Wir erhalten unter diesen Bedingungen eine Teststärke von $(1 - \beta) = 0.9912$, was recht akzeptabel ist. Der kritische F-Wert liegt bei $F(2, 45) = 2.4245$.

3 Teststärkeanalysen für χ^2-Tests

Zwei Arten von χ^2-Tests sind in der Psychologie gebräuchlich (vgl. Cohen, 1988): Kontingenztests, bei denen es um Abweichungen (H_1) vom Modell der totalen stochastischen Unabhängigkeit (H_0) zweier oder mehrerer kategorialer Variablen geht, und Modellanpassungstests (*goodness-of-fit tests*). Die Referenzverteilung zur Teststärkeberechnung ist in beiden Fällen die nonzentrale χ^2-Verteilung (Johnson & Kotz, 1970, Kap. 28), deren Nonzentralitätsparameter

$$\lambda = w^2 \cdot N \tag{14}$$

multiplikativ vom Stichprobenumfang N und dem quadrierten Effektstärkeindex

$$w = \sqrt{\sum_{i=1}^{m} \frac{(p_{1i} - p_{0i})^2}{p_{0i}}} \tag{15}$$

abhängt. Hierbei bezeichnet m die Gesamtzahl der Beobachtungskategorien, p_{0i} die Wahrscheinlichkeit für Kategorie i unter H_0 und p_{1i} die Wahrscheinlichkeit für Kategorie i unter H_1 (vgl. auch Hager, in diesem Band).

Betrachten wir ein Beispiel für Kontingenztests. Nehmen wir an, eine Therapieform X habe eine recht hohe Erfolgsquote von $p_x = .88$, aber den Nachteil, sehr teuer zu sein. Eine neue Therapieform Y wird vorgeschlagen, die erheblich kostengünstiger ist. Sie soll aber nur dann zum Einsatz kommen, wenn sie keine (bedeutsam) schlechteren Erfolgsquoten hat. Wir prüfen daher die $H_0 : p_y \geq p_x$ mit einem *einseitigen* χ^2-Kontingenztest für 2×2-Kontingenztafeln: Die Zeilenvariable ist die Therapieform (X vs. Y), die Spaltenvariable der Erfolg bzw. Mißerfolg der Therapie. Jeweils die Hälfte der Stichprobe weisen wir X bzw. Y zu. Wir wollen relativ sicher sein, einen Nachteil der Therapieform Y zu entdecken, wenn er existiert, und legen daher die Teststärke auf $1 - \beta = .95$ fest. Im Gegensatz dazu akzeptieren wir ein Risiko von $\alpha = .20$, die Therapie Y fälschlicherweise als schlechter zu verwerfen. Wir wollen nur dann von einem bedeutsamen Nachteil von Y sprechen, wenn die Erfolgsquote mindestens .09 unter der von X liegt. Wie groß muß unter diesen Bedingungen der Gesamtstichprobenumfang N sein?

Zur Beantwortung dieser Frage wählen wir in GPOWER ≪A priori≫ als Typ der Analyse und ≪Chi-square Test≫ als Typ des Tests. Da wir eine einseitige Fragestellung haben, wählen wir nicht $\alpha = .20$, sondern $\alpha = .40$, da GPOWER von ungerichteten χ^2-Tests ausgeht. Als gewünschte Teststärke geben wir .95 ein. Den Effektgrößenindex w, der zu der einfachen Alternativhypothese $p_x = .88$, $p_y = .88 - .09 = .79$ gehört, können wir in GPOWER leicht mit ≪Calc w≫ berechnen. Da jeweils 50% der Stichprobe Behandlung X bzw. Y bekommen, sind die Wahrscheinlichkeiten für die Zellen der Kontingenztafel unter H_1 gerade $.5 \cdot .880 = .440$ und $.5 \cdot .120 = .060$ für Erfolg bzw. Mißerfolg bei Therapie X und analog $.5 \cdot .79 = .395$ sowie $.5 \cdot .21 = .105$ für Erfolg bzw. Mißerfolg bei Therapie Y. Die H_0, die stochastische Unabhängigkeit von Therapie und Ergebnis bei gleichen Randsummen behauptet, impliziert gleiche Erfolgswahrscheinlichkeiten ($.5 \cdot .835 = .4175$) und Mißerfolgswahrscheinlichkeiten ($.5 \cdot .165 = .0825$) für beide Therapien. Setzt man diese H_0- und H_1-Wahrscheinlichkeiten mittels ≪Calc w≫ in Gleichung (15) ein, so

resultiert $w = .1212$. Nun müssen wir nur noch $df = 1$ für den Vierfelder-χ^2-Test eingeben. Der Nonzentralitätsparameter beträgt $\lambda = 6.1696$. Die Apriori-Analyse ergibt ein erforderliches N von 420. Als kritischer χ^2-Wert ist $\chi^2_{(1)} = .7083$ zu wählen. Überschreitet die χ^2-Statistik diesen Wert *und* ist die Stichprobenerfolgsquote für Therapie Y kleiner als die für Therapie X, dann nehmen wir die H_1 an, d.h. die neue Therapie Y wird zurückgewiesen. Andernfalls können wir unsere H_0 beibehalten, so daß die kostengünstigere Therapie Y angewandt werden kann. Alle diese Angaben gelten lediglich approximativ, da die exakte Verteilung der χ^2-Statistik nur asymptotisch (für $N \rightarrow \infty$) eine χ^2-Verteilung ist. Bei $N = 420$ ist die Abweichung von der asymptotischen Verteilung jedoch vernachlässigbar.[6]

4 Weiterführende Literatur

Dieses Kapitel beschränkte sich auf Teststärkeanalysen für einfache und häufig benutzte statistische Tests. Zur weiteren Vertiefung dieser Punkte sei auf Cohen (1988) verwiesen. Denjenigen, die sich für Teststärkeanalysen im Rahmen von Meßwiederholungsdesigns und approximative Teststärkeanalysen für F-Tests im Rahmen multivariater Varianzanalysen interessieren, empfehlen wir O'Brien und Muller (1993) sowie Muller, LaVange, Ramey und Ramey (1992). Die entsprechenden Teststärkeanalysen können ebenfalls mit GPOWER durchgeführt werden, wobei allerdings die ≪Other F-Tests≫ Option zu wählen und Sorge zu tragen ist, daß der Nonzentralitätsparameter korrekt spezifiziert wird. Für F-Tests im Rahmen von Zufallseffekt-ANOVAs und gemischten ANOVA-Modellen empfehlen wir Koele (1982).

Abschließend soll noch darauf hingewiesen werden, daß es eine ökonomische Alternative zur Apriori-Festlegung von Stichprobenumfängen im Rahmen der Neyman-Pearson-Statistik gibt, die eine Kontrolle von α- und β-Fehlerrisiken bei durchschnittlich geringeren Stichprobengrößen ermöglicht: Sequentielles Testen. Diepgen (in diesem Band) gibt eine Einführung in diesen Zugang zur statistischen Inferenz und nennt weiterführende Literatur. Sequentielles Testen sollte immer dann als Option in Erwägung gezogen werden, wenn parallel zur Datenerhebung eine sequentielle Analyse der Daten leicht durchgeführt werden kann. Dies ist z.B. bei computergestützten ($N = 1$)-Experimenten der Fall, bei denen die Daten der Versuchsperson direkt vom Computer erfaßt und ausgewertet werden können.

Literaturverzeichnis

Berry, D. C. & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology, 36A*, 209–231.

[6]Eine Monte-Carlo-Studie mit 10000 Zufallsstichproben des Umfangs $N = 420$ aus der o.g. H_1-Population ergab, daß die faktische (empirisch geschätzte) *power* bei $1 - \hat{\beta} = .9514$ liegt, wenn jede Überschreitung des kritischen Wertes .7083 zu einer H_0-Verwerfung führt (hierfür sind die Teststärkeberechnungen ausgelegt), und bei $1 - \hat{\beta} = .9509$, wenn nur die Überschreitungen zählen, bei denen zusätzlich die Stichprobenerfolgsquote von Therapie X größer war. Die Teststärkeangaben behalten also auch bei Verwendung der o.g. „einseitigen" Entscheidungsregel ihre Gültigkeit.

Bradley, D. R., Russell, R. L. & Reeve, C. P. (in press). Statistical power in complex experimental designs. *Behavior Research Methods, Instruments, & Computers.*

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt am Main: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Buchner, A., Faul, F. & Erdfelder, E. (1992). *GPOWER: A priori, post-hoc, and compromise power analyses for the Macintosh* [Computer-Programm]. Bonn: Psychologisches Institut der Universität.

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

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des β-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle. *Zeitschrift für Sozialpsychologie, 15,* 18–32.

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

Faul, F. & Erdfelder, E. (1992). *GPOWER: A priori, post-hoc, and compromise power analyses for MS-DOS* [Computer-Programm]. Bonn: Psychologisches Institut der Universität.

Gesell, A. & Thompson, H. (1929). Learning and growth in identical infant twins: An experimental study of the method of co-twin control. *Genetic Psychology Monographs, 6,* 1–124.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues (S. 311–339).* Hillsdale: Erlbaum.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *American Statistician, 43,* 253–260.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Ed.), *Allgemeine Experimentelle Psychologie* (S. 43–264). Stuttgart: Fischer.

Johnson, N. L. & Kotz, S. (1970). *Distributions in statistics. Continuous univariate distributions - 2.* New York: Wiley.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin, 92,* 513–516.

Mallinckrodt, B. & Bennett, J. (1992). Social support and the impact of job loss in dislocated blue-collar workers. *Journal of Counseling Psychology, 39,* 482–489.

McGraw, M. B. (1946). Maturation of behavior. In L. Carmichael (Ed.), *Manual of child psychology* (S. 332–369). New York: Wiley.

Muller, K. E., LaVange, L. M., Ramey, S. L. & Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association, 87,* 1209–1226.

O'Brien, R. G. & Muller, K. E. (1993). Unified power analysis for t-tests through multivariate hypotheses. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (S. 297–344). New York: Dekker.

Odeh, R. E. & Fox, M. (1991). *Sample size choice. Charts for experiments with linear models* (2nd. ed.). New York: Dekker.

Schmitt, M., Hoser, K. & Schwenkmezger, P. (1991). Schadensverantwortlichkeit und Ärger. *Zeitschrift für Experimentelle und Angewandte Psychologie, 38,* 634–647.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Warrington, E. K. & Weiskrantz, L. (1970). Amnesic syndrome: Consolidation or retrieval? *Nature, 228,* 628–630.