Zusammenhangs- und Unterschiedsmaße

Willi Hager

Zusammenhangs- und Unterschiedsmaße ist die allgemeine Bezeichnung für statistische Maße, die zum Ausdruck bringen, wieviel der Variation in den Daten auf systematisierbare Einflüsse wie z.B. experimentelle Behandlungen zurückgeführt werden kann. Diese Maße sind auf der Grundlage statistischer Konzepte definiert für die meisten parametrischen Hypothesen und Tests wie t- und F-Tests sowie deren multivariate Erweiterungen und für Hypothesen über Wahrscheinlichkeiten oder relative Häufigkeiten. Dagegen stehen derartige Maße nur in begrenztem Umfange für Tests über Ranghypothesen oder „ordinale Daten" zur Verfügung. Der empirische Wert dieser Maße ist – mit einer Ausnahme – unabhängig von der Stichprobengröße, und diese Tatsache unterscheidet empirische Werte für Zusammenhangs- und Unterschiedsmaße von empirischen Werten für Teststatistiken wie t, F und χ^2. Zusammenhangs- und Unterschiedsmaße beinhalten also Informationen, die der Signifikanztest oder der statistische Test nicht zu geben vermag.

Beide Arten von Maßen werden häufig als „Effektgrößen" (oder „Effekte") bezeichnet, und zahlreiche Autoren sprechen auch von „Maßen der praktischen Bedeutsamkeit (oder Signifikanz)". Diese Bezeichnung gibt aber zu unangemessenen Assoziationen und Interpretationen der Art Anlaß, daß ein Ergebnis um so bedeutsamer für die Wissenschaft oder für die Praxis ist, je größer der mit ihm verbundene Effekt ist. Aber auch ganz kleine Effekte können von großer Bedeutung für beide Felder sein.

Sofern man die psychologische Forschung konsequent auf die Prüfung psychologischer Hypothesen ausrichtet, kann gesagt werden, daß die meisten dieser psychologischen Hypothesen entweder das Auftreten oder das Ausbleiben eines bestimmten Zusammenhanges oder Unterschiedes vorherzusagen erlauben. Dagegen bezieht sich keine (mir bekannte) psychologische Hypothese auf eine bestimmte Stichprobengröße oder eine bestimmte Irrtumswahrscheinlichkeit, und daher stehen von allen Determinanten des statistischen Tests die Unterschieds- und die Zusammenhangsmaße noch in engster Verbindung zu den psychologischen Inhalten.

Nach diesen einleitenden Vorbemerkungen befaßt sich der vorliegende Beitrag mit der Einteilung und Definition von Zusammenhangs- und Unterschiedsmaßen vor dem Hintergrund der zwei gegenwärtig wichtigsten statistischen Testtheorien (vgl. dazu auch Willmes, in diesem Band). Dabei werden die zahlreichen Maße der Beurteilerübereinstimmung nicht berücksichtigt, weil diese keine Effektmaße im Sinne des Gegenstandes des vorliegenden Beitrages darstellen (vgl. dazu Liebetrau, 1983, S. 31-38). Im einzelnen behandelt werden Effektmaße für Datensätze mit einer unabhängigen und einer abhängigen Variablen (Abschnitt 2), Effektmaße für Datensätze mit mehreren unabhängigen und/oder abhängigen Variablen (Abschnitt 3)

https://doi.org/10.25521/HQM07

und Effektmaße für Datensätze mit kategorialen oder ordinalen Variablen (Abschnitt 4). Der Beitrag schließt mit einer Erörterung der Bedeutung von Zusammenhangs- und Unterschiedsmaßen für die Prüfung psychologischer Hypothesen (Abschnitt 5) und Hinweisen zu weiterführender Literatur (Abschnitt 6).

1 Zusammenhangs- und Unterschiedsmaße in zwei konkurrierenden statistischen Testtheorien

1.1 Theorie des Signifikanztests von R. A. Fisher

In der gegenwärtig noch vorherrschenden Theorie statistischen Schließens, die im Anschluß an die Arbeiten von R. A. Fisher in den zwanziger und den dreißiger Jahren dieses Jahrhunderts (Fisher, 1925, 1935) ihren Siegeszug durch so gut wie alle Bereiche der modernen Wissenschaften antrat, bezieht sich der Signifikanztest auf eine statistische Nullhypothese (H_0), die allgemein besagt, daß kein Unterschied oder kein Zusammenhang zwischen verschiedenen Variablen, meist der oder den unabhängigen und der oder den abhängigen Variablen (UV und AV), besteht. Dieser H_0 steht als Negation eine sog. „Forschungshypothese" gegenüber, die aufgrund der Daten bzw. der Untersuchung „angenommen" werden soll.

In dieser Theorie wird die Frage nach einem geeigneten Unterschieds- oder Zusammenhangsmaß mit der sog. Überschreitungswahrscheinlichkeit, dem „p-Wert", beantwortet. Damit ist die Wahrscheinlichkeit des empirischen Wertes der benutzten Teststatistik oder eines extremeren Wertes unter der Annahme der Gültigkeit der H_0 gemeint. Der empirische Wert der Teststatistik bringt dabei den Abstand der empirischen Daten von der Annahme unter der H_0 zum Ausdruck. Der p-Wert stellt dann eine monoton strikt fallende Funktion der Stichprobengröße und dieses Abstandes dar. Je kleiner er unter sonst gleichen Bedigungen wird, desto größer ist das Vertrauen darin, daß die H_0 „falsch" ist.

Dieses Maß ist als „Effektgröße" vor allem wegen seiner Abhängigkeit vom Stichprobenumfang zwar immer wieder kritisiert worden (vgl. etwa Morrison & Henkel, 1970; Hager & Westermann, 1983, S. 83-87), aber nicht auf dem Boden von Fishers Theorie selbst. Solange man jedoch mit den Begrifflichkeiten und Konzepten dieser Theorie argumentiert, kann kaum ein anderes Maß für den Unterschied oder den Zusammenhang als der p-Wert definiert werden, so kritisierbar dieser Wert auch ist.

1.2 Theorie des statistischen Tests von Neyman und Pearson

Dies verhält sich in der konkurrierenden Theorie statistischer Tests von J. Neyman und E. S. Pearson aus den ausgehenden zwanziger und den dreißiger Jahren anders (Neyman, 1950; vgl. auch Willmes, in diesem Band). Hier steht der tatsächlich mit einem gegebenen Test getesteten H_0 eine statistische Alternativhypothese (H_1) gegenüber, und es ist möglich, die Wahrscheinlichkeit von Resultatsklassen unter der Annahme der Gültigkeit der H_0 wie auch unter der Annahme der Gültigkeit der H_1 zu bestimmen. Man kann daher aus dem Blickwinkel dieser Theorie sagen, daß alle Unterschieds- und alle Zusammenhangsmaße – vereinfacht formuliert – angeben, wie sehr die Daten von der Erwartung unter der H_0 abweichen, wobei diese Angabe

von allen Wahrscheinlichkeitsbetrachtungen unabhängig erfolgen kann. Im folgenden wird ausschließlich auf diese Theorie Bezug genommen, und ferner werden stets gleiche Stichprobenumfänge pro Versuchsbedingung angenommen.

2 Parametrische standardisierte Unterschiedsmaße und Zusammenhangsmaße bei einer UV und einer AV

Üblicherweise werden die (parametrischen) Unterschieds- und die Zusammenhangsmaße als standardisierte Größen definiert, und als Standardisierungsgröße fungiert entweder die Binnenvarianz in den „Populationen" oder die Summe aus dieser Binnenvarianz und mindestens einer systematischen Varianz. Obwohl in beiden Fällen ein „Unterschied" standardisiert wird, will ich nur im ersteren Fall – möglicherweise etwas willkürlich, aber einem verbreiteten Sprachgebrauch folgend – von Unterschiedsmaßen sprechen. Bei der zweiten Art der Standardisierung können die Maße nur Werte zwischen 0 und 1 bzw. -1 und $+1$ annehmen, und man kann sie als (quadrierte) Korrelationskoeffizienten ansehen. Korrelationskoeffizienten stellen Zusammenhangsmaße dar (vgl. auch Hager, 1987).

Im Falle eines Zweigruppenplanes (eine UV mit $J = 2$ Stufen) kann der standardisierte *Unterschied* auf seiten der AV wie folgt definiert werden, wobei μ_1 sowie μ_2 die Erwartungswerte der betrachteten Zufallsvariablen in den beiden zugehörigen „Populationen" bezeichnen und σ_I ihre gemeinsame (Binnen-) Streuung:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma_I}. \tag{1}$$

Die Reihenfolge der Erwartungswerte und damit das Vorzeichen der Differenz wird von der interessierenden statistischen Hypothese festgelegt; im Falle von ungerichteten statistischen Hypothesen (über Erwartungswerte) ist daher $|\delta|$ zu verwenden.

Untersucht man mehr als zwei Versuchsbedingungen ($J > 2$), dann kann man neben den Paarkontrasten, in die stets nur zwei Erwartungswerte wie in (1) eingehen, auch komplexe Kontraste betrachten, die sich auf eine beliebige Anzahl K von Erwartungswerten beziehen. In diesem Fall wird δ_r pro Kontrast ψ_r wie folgt definiert, wobei die c_{jr} die Kontrastkoeffizienten für den Kontrast ψ_r bezeichnen, die nicht alle gleich Null sein dürfen und die sich zu Null aufsummieren müssen ($j = 1, \ldots, J$), und r ($r = 1, \ldots, R$) bezeichnet die Numerierung der Kontraste:

$$\delta_r = \frac{\sum_{j=1}^{J} c_{jr}\mu_j}{\sigma_I}. \tag{2}$$

Bei mehr als $J = 2$ Versuchsbedingungen verwendet man anstelle von Kontrasten zumeist Vergleiche, die aus mehr als einem Kontrast bestehen und die soviele Freiheitsgrade (FG) haben, wie (orthogonale) Kontraste in sie eingehen; die Maximalzahl von orthogonalen Kontrasten in einem Vergleich beträgt bekanntlich $J - 1$. Bei Vergleichen (mit $1 < \mathrm{FG} \leq J - 1$) wird anstelle von Formel (2) für Kontraste die folgende Formel verwendet (vgl. Cohen, 1988, S. 281):

$$\phi^2 = \frac{\sigma_A^2}{\sigma_I^2} = \frac{\sum_{j=1}^{J}(\mu_j - \mu)^2/J}{\sigma_I^2}, \tag{3}$$

wobei σ_A^2 die („Populations"-) Varianz zwischen den Erwartungswerten bezeichnet. Das Maß δ_r kann im übrigen wie folgt in das von Cohen (1988) verwendete Zusammenhangsmaß ϕ_r^2 umgerechnet werden:

$$\phi_r^2 = \frac{\left(\sum_{j=1}^{J} c_{jr}\mu_j\right)^2}{J \cdot \sum_{j=1}^{J} c_j^2 \cdot \sigma_I^2} = \frac{\delta_r^2}{J \cdot \sum_{j=1}^{J} c_{jr}^2}. \tag{4}$$

Bei den *quadrierten Korrelationen* erfolgt die Standardisierung auf die Summe mehrerer („Populations"-) Varianzen, von denen stets eine die Binnenvarianz darstellt (s.o.); sehr häufig wird die Summe aus allen Varianzen gebildet, so daß die Standardisierung dann auf die *totale Varianz* σ_T^2 erfolgt. So kann man das multiple Korrelationsquadrat $\eta_{Y.A}^2$ *im einfaktoriellen Fall* wie folgt definieren, wobei sich unter dem *Allgemeinen Linearen Modell* (vgl. dazu Andres, in diesem Band) die totale Varianz additiv aus den beiden bisher betrachteten Varianzen zusammensetzt:

$$\eta_{Y.A}^2 = \frac{\sigma_A^2}{\sigma_T^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_I^2}. \tag{5}$$

Während bei der vorstehenden Formel eher aus dem Blickwinkel der Varianzanalyse definiert wird, liegt der folgenden alternativen Definition die regressionsanalytische Betrachtungsweise zugrunde. Bei ihr wird das multiple Korrelationsquadrat als Summe quadrierter einfacher Korrelationen ρ_r^2 dargestellt, die sich auf orthogonale Kontraste (oder eine orthogonale Kodierung) beziehen:

$$\eta_{Y.A}^2 = \sum_{r}^{R=J-1} \rho_r^2. \tag{6}$$

Bei nonorthogonalen Kontrasten wird zur Bestimmung von $\eta_{Y.A}^2$ die Summe aus quadrierten semipartiellen Korrelationen (s.u.) gebildet; diese verwendet man z.B. auch bei der Dummy- und bei der Effekt-Kodierung der UV(n) (vgl. dazu Bredenkamp, 1980, sowie Cohen & Cohen, 1983). Hebt man im übrigen bei den einfachen Korrelationen die Standardisierung auf, dann erhält man als unstandardisiertes Zusammenhangsmaß die *Kovarianz*.

Die quadrierten Korrelationen bezeichnet man häufig als Maße der erklärten Varianz. Der Gebrauch des Terminus „erklärt" suggeriert dabei, daß damit eine wissenschaftliche Erklärung geleistet wird. Dies ist jedoch nicht der Fall, denn dazu bedarf es psychologisch-inhaltlicher Theorien. Von daher wäre es günstiger, von *Maßen der systematisierbaren Varianz* zu reden, da diese Varianz auf die Variation der UV oder die Ausprägungen der Prädiktorvariablen zurückgeführt werden kann, sofern Störgrößen etwa durch Randomisierung weitestgehend ausgeschlossen worden sind.

Bei $J = 2$ und bei gleichen Stichprobenumfängen kann nun das Abstandsmaß δ wie folgt in eine quadrierte Produkt-Moment-Korrelation ρ^2 umgerechnet werden:

$$\rho^2 = \frac{\delta^2}{\delta^2 + 4}. \tag{7}$$

Eine allgemeine Umrechnungsformel, die nicht den genannten Beschränkungen unterliegt, ist meines Wissens noch nicht verfügbar. Formel (7) zeigt an, daß *sich Unterschiedsmaße (wie auf die Fehlerstreuung standardisierte Abstände) in den meisten Fällen in Zusammenhangsmaße (wie Korrelationen) überführen lassen – und*

umgekehrt. Dies verwundert natürlich nicht in Anbetracht der Tatsache, daß sich die Varianzanalyse (VA) als Verfahren für Unterschiede als Spezialfall der multiplen Regressionsanalyse (MRA) als allgemeinem Verfahren für Zusammenhänge darstellen läßt (vgl. Bredenkamp, 1980). Dennoch eignet sich der Ansatz der MRA wegen seiner Vielseitigkeit für die Lösung bestimmter Probleme bekanntlich besser als der traditionelle Ansatz der VA. Dies gilt vor allem dann, wenn man simultan mehrere (unabhängige und/oder abhängige) Variablen betrachtet, und/oder dann, wenn man zahlreiche Prädiktoren zu untersuchen hat, die auch miteinander korreliert sein können (s.u.). Demgegenüber zeichnet sich der „Erwartungswertansatz" der VA durch seine einfachere Interpretierbarkeit aus und hat sich trotz seines Charakters eines Spezialfalles der MRA bisher als geeignet bei der Prüfung psychologischer Hypothesen erwiesen (vgl. dazu Hager, 1992a).

Obwohl man bei regressions- und korrelationsanalytischer Betrachtungsweise eher von *Zusammenhangsmaßen* und bei Varianzanalysen eher von *Unterschiedsmaßen* spricht, zeigen die vorstehenden Ausführungen doch, daß beide Arten von Maßen mit jeder der angesprochenen Sichtweisen verbunden werden können.

3 Quadrierte Korrelationen bei mehreren UVn sowie bei mehreren AVn

3.1 Theoretische, „Populations"- oder Definitionsebene

Bei regressionsanalytischer Betrachtungsweise kommt allen *multiplen Korrelationsquadraten* der Status von Zusammenhangsmaßen zu. Sofern man an statistischen Hypothesen über Vergleiche (in varianzanalytischer Terminologie: über Haupteffekte und über Interaktionen) *unabhängig von der Anzahl der Stufen der UV* und auch *unabhängig von der Anzahl der UVn* interessiert ist, sind diese Zusammenhangsmaße den vorher besprochenen Unterschiedsmaßen für (orthogonale und nonorthogonale) Kontraste vorzuziehen. Denn sie sind anwendbar in allen Fällen, in denen eine oder mehrere UVn (Prädiktorvariablen) und eine AV (Kriteriumsvariable) betrachtet werden. Durch die Verwendung von partiellen und semipartiellen Korrelationsquadraten wird es dabei möglich, die gesamten Varianzanteile spezieller UVn oder aber redundante Anteile aus anderen UVn und/oder der AV herauszulösen. Zudem eignen sich diese Zusammenhangsmaße auch für den Fall ungleich großer Stichprobengrößen. Die Definition von partiellen sowie von semipartiellen Korrelationen findet man in einschlägigen Lehrbüchern wie etwa dem von Bortz (1993, Kap. 13) und von Cohen und Cohen (1983).

Dieser Ansatz ist zudem erweiterbar auf die simultane Untersuchung mehrerer Kriteriumsvariablen oder AVn. Obwohl man darüber streiten mag, ob man mehrere Kriteriumsvariablen simultan statistisch analysieren sollte (vgl. Hager, 1992a, S. 93-98, S. 374-381), geschieht dies häufig. In diesem *multivariaten Fall* kann man mit der *kanonischen Korrelation* ρ_c operieren, aus der sich die bisher und die im folgenden betrachteten Produkt-Moment-Korrelationskoeffizienten als *Spezialfälle* herleiten lassen. Liegen dabei J Versuchsgruppen und P AVn vor, so können S $[S = min(J-1, P)]$ kanonische Korrelationen $\rho_{c(s)}$ berechnet werden (vgl. dazu den Beitrag über multivariate Statistik von Andres, in diesem Band). Um aus die-

sen S kanonischen Korrelationen ein einziges (multivariates) Zusammenhangsmaß zu bilden, nämlich die multivariate Erweiterung des univariaten multiplen Korrelationsquadrates (η^2_{mult}), werden in der Literatur zwei Möglichkeiten vorgeschlagen und in Cohen und Cohen (1983, S. 492-494) ausführlicher diskutiert:

$$\eta^2_{mult} = \sum_{s=1}^{S} \rho^2_{c(s)}/S \tag{8}$$

und

$$\eta^2_{mult} = 1 - [1 - \rho^2_{c(1)}][1 - \rho^2_{c(2)}] \ldots [1 - \rho^2_{c(S)}]. \tag{9}$$

Auf die kanonische Korrelation sowie auf andere Zusammenhangsmaße für den multivariaten Fall gehen im einzelnen Bredenkamp (1980, S. 82-85), Cohen (1988; vgl. auch Cohen & Cohen, 1983), Cramér und Nicewander (1979) sowie Wolf (1988) ein.

3.2 Empirische oder Stichprobenebene

Wie werden nun die empirischen Entsprechungen dieser Maße aus den Daten bestimmt? Im Falle von $J = 2$ wird als Unterschiedsmaß d benutzt und wie folgt auf der Grundlage der beiden empirischen Mittelwerte $\bar{y}_1$ und $\bar{y}_2$ und der mittleren empirischen Streuung s_I innerhalb der Versuchsbedingungen berechnet, die im allgemeinen auf den Freiheitsgraden und nicht auf der Anzahl der Versuchspersonen pro Bedingung beruht, obwohl auch diese Variante möglich ist:

$$d = \frac{\bar{y}_1 - \bar{y}_2}{s_I}. \tag{10}$$

Im allgemeinen Fall eines empirischen Kontrastes V_r findet die empirische Entsprechung der Formel (2) Anwendung:

$$d_r = \frac{\sum_{j=1}^{J} c_{jr}\bar{y}_j}{s_I} = \frac{V_r}{s_I}. \tag{11}$$

Die Streuung s_I wird über alle J Versuchsbedingungen bestimmt und beruht daher in einem einfaktoriellen Plan ohne Meßwiederholungen in aller Regel auf $FG = J(n-1)$, wobei n die Anzahl der Werte (Versuchspersonen) in einer Versuchsbedingung bezeichnet. Diese Art der Berechnung eines Unterschiedsmaßes für Kontraste kann auch bei wiederholten Messungen benutzt werden, sofern man stärker an der Vergleichbarkeit der Maße interessiert ist. Dabei wird dann die Tatsache, daß sich ja in aller Regel bei intraindividueller Bedingungsvariation eine relativ zu unabhängigen Stichproben reduzierte Testvarianz ergibt, unbeachtet gelassen; allerdings kann und sollte dieser Umstand gezielt bei der Testplanung (s.u.) berücksichtigt werden. Andere Vorgehensweisen bei der empirischen Bestimmung von Unterschiedsmaßen bei intraindividueller Bedingungsvariation sind natürlich ebenfalls zu rechtfertigen (vgl. dazu Bredenkamp, 1980), so daß in jedem Falle angegeben werden sollte, für welche Variante man sich entschieden hat.

Besonders wichtig ist nun das *multiple Korrelationsquadrat* $R^2_{Y.A}$ als empirisches Pendant zu $\eta^2_{Y.A}$, das bei regressionsanalytischer Sichtweise als die Summe empirischer quadrierter einfacher (oder semipartieller) Korrelationen berechnet werden

kann (s.o.). Bei varianzanalytischer Betrachtungsweise kann sein Wert im einfaktoriellen Fall über die drei Quadratsummen innerhalb, zwischen und total (QSI, QSA und QST) berechnet werden:

$$R^2_{Y.A} = \frac{QSA}{QSA + QSI} = \frac{QSA}{QST}. \tag{12}$$

Wenn man an der Schätzung des $\eta^2_{Y.A}$ in der zugrundeliegenden Population interessiert ist, dann erweist es sich, daß das berechnete multiple Korrelationsquadrat das „wahre" multiple Korrelationsquadrat in der Population überschätzt. Zum Ausgleich dieser Überschätzung sind verschiedene sog. „Schrumpfungskorrekturen" vorgeschlagen worden, deren wichtigste sich u.a. bei Bredenkamp (1980, S. 52), bei Hager und Westermann (1983, S. 164-165) und bei Thompson (1990) finden. Aber diese Schätzung von Populationsmaßen ist bei der Prüfung von psychologischen Hypothesen über aus ihnen abgeleitete statistische Hypothesen (vgl. dazu Bredenkamp, 1972; Hager, 1992a; Hager & Westermann, 1983) weniger bedeutungsvoll als in der herkömmlichen Inferenzstatistik besonders in der Tradition R.A. Fishers, in der es vor allem darauf ankommt, Stichprobenergebnisse auf „Populationen" zu verallgemeinern. Demgegenüber besteht bei der Prüfung psychologischer Hypothesen meines Erachtens kaum Bedarf an derartigen Generalisierungen auf nicht untersuchte Entitäten (vgl. auch Hager & Hasselhorn, 1994, S. 14-18). Allerdings vereinfacht dieses Modell viele Darstellungen, weswegen auch in diesem Beitrag auf seiner Grundlage argumentiert wird. Weitere Details der Berechnung empirischer Werte für die hier behandelten Maße findet man bei Glass, McGaw und Smith (1981) und bei Seifert (1991).

4 Weitere Zusammenhangsmaße

4.1 Produkt-Moment-Korrelationen bei verschiedenen Arten von Variablen

Im folgenden beziehe ich mich in erster Linie auf die empirische Ebene, weil den verschiedenen Korrelationen auf der theoretischen oder Definitionsebene die gleiche Konzeption zugrunde liegt.

a) *Mindestens eine kontinuierliche Variable*: Die bereits erwähnte Produkt-Moment-Korrelation r ist ursprünglich für zwei kontinuierliche Variablen X und Y definiert worden, aber sie läßt sich auch in allen denjenigen Fällen berechnen und interpretieren, in denen diese Voraussetzung nicht gegeben ist.

Ein wichtiger Spezialfall der Produkt-Moment-Korrelation liegt dann vor, wenn die Prädiktorvariable als Kodiervariable nur in den Ausprägungen 0 und 1 auftritt. Im Falle nur zweier Versuchsbedingungen kann dann auf der empirischen Ebene die *punkt-biseriale Korrelation* r_{pbis} berechnet werden, die nichts anderes darstellt als die Produkt-Moment-Korrelation zwischen einer dichotomen (Kodier-) Variablen (X) und einer kontinuierlichen (abhängigen) Variablen (Y). Wenn der Unterschied zwischen den Werten der AV unter zwei verschiedenen Versuchsbedingungen groß ist, dann ist auch die Korrelation (der Zusammenhang) zwischen der kodierten Gruppenzugehörigkeit und der AV groß.

b) *Mindestens eine Rangvariable*: Ein weiterer Spezialfall der Produkt-Moment-Korrelation resultiert dann, wenn beide Variablen Ränge darstellen oder in Ränge transformiert werden. Die *Rangkorrelation von Spearman* (r_S) ist dann die Produkt-Moment-Korrelation zwischen den beiden Rangreihen. Eine mögliche Verallgemeinerung der Rangkorrelation r_S auf $K > 2$ Rangreihen stellt *Kendalls Konkordanzkoeffizient W* dar, der zwar auf einfache Weise in eine mittlere Rangkorrelation $\bar{r}_s$ umgerechnet werden kann, der aber keinen Spezialfall der einfachen oder der multiplen Korrelation darstellt (Marascuilo & McSweeney, 1977, S. 458-466). Weitere Effektmaße für Ranghypothesen findet man in einschlägigen Lehrbüchern der nonparametrischen Statistik wie etwa dem von Marascuilo und McSweeney (1977, Kap. 16) oder von Gibbons (1985, Kap. 7) sowie in Liebetrau (1983, S. 56-85).

Eine Entsprechung der punkt-biserialen Korrelation für Rangdaten (eine Rangvariable und eine dichotome oder Kodiervariable) stellt *Kendalls* τ dar, das allerdings ebenfalls *nicht* als Spezialfall der Produkt-Moment-Korrelation angesehen werden kann (zur Berechnung siehe bspw. Marascuilo und McSweeney, 1977, S. 453-454). Auch nicht als Spezialfall der Produkt-Moment-Korrelation ist eine andere rangmäßige Entsprechung der punkt-biserialen Korrelation, nämlich die *biseriale Rangkorrelation* (vgl. etwa Bortz, 1993, S. 212-213), zu interpretieren, bei der eine Standardisierung auf ihren maximal möglichen Wert erfolgt.

c) *Zwei dichotome Variablen oder Häufigkeiten.* Auch dann, wenn *beide* Wertereihen nur Kodiervariablen darstellen, also nur zwei verschiedene Werte wie 0 und 1 annehmen können, kann die Produkt-Moment-Korrelation berechnet werden. Unter den beschriebenen Bedingungen wird sie als *Phi-Koeffizient* bezeichnet (nicht zu verwechseln mit dem bspw. in Formel (4) verwendeten Maß!), und ihr empirischer Wert kann aus dem Wert für die Testgröße χ^2 berechnet werden (vgl. etwa Bortz, 1993, S. 469-470):

$$\phi = \sqrt{\frac{\chi^2}{N}}. \tag{13}$$

Dieses und weitere Maße der Assoziation in Kontingenztafeln behandeln u.a. Marascuilo und McSweeney (1977, Kap. 8 und 9) und Gibbons (1985, S. 330-342). Der Phi-Koeffizient kann daneben auch als *Spezialfall der kanonischen Korrelation* interpretiert werden (vgl. z.B. Bortz, 1993, S. 596-597), und dies gilt auch für seine Verallgemeinerung auf beliebige Kontingenztafeln. – Eine ausführliche Diskussion verschiedener Korrelationskoeffizienten und Zusammenhangsmaße gibt darüber hinaus Kubinger (1990).

4.2 Wahrscheinlichkeitsdifferenzen

Wenn man die üblichen Verteilungsannahmen nicht treffen will oder wenn die Daten dichotom sind, wird man die psychologische Hypothese häufig nicht über parametrische (Erwartungswert- oder Korrelations-) Hypothesen, sondern über Hypothesen prüfen wollen, die sich auf Wahrscheinlichkeiten π_k oder auf (relative) Häufigkeiten beziehen. Derartige Hypothesen können entweder über die χ^2-Verteilungen (s.o.) oder aber über die exakten Binomialverteilungen getestet werden („Vorzeichen-Test"). Für den letzteren Fall definiert Cohen (1988, S. 147) die Abweichung von der Wahrscheinlichkeit $\pi_0 = .50$ unter der H_0 als Unterschiedsmaß, so daß im Falle

gerichteter Hypothesen resultiert (theoretische Ebene):

$$\gamma = \pi_0 - .50 \quad \text{bzw.} \quad \gamma = .50 - \pi_0. \tag{14}$$

Die empirische Schätzung dieser Wahrscheinlichkeitsdifferenzen erfolgt über relative Häufigkeiten. Testet man die gleiche Art von Hypothesen über die χ^2-Verteilungen, so kann man im häufigsten Fall der Vierfelder-Kontingenztafel die theoretische Entsprechung des Phi-Koeffizienten (s.o.) benutzen, die Cohen (1988, S. 216) ω nennt. Unter Bezugnahme auf die Wahrscheinlichkeiten, auf die sich die dem Test zugrundeliegenden statistischen Hypothesen beziehen, kann dieses Maß auch wie folgt definiert werden:

$$\omega = \sqrt{\sum \frac{(\pi_{1jk} - \pi_{0jk})^2}{\pi_{0jk}}}, \tag{15}$$

wobei die π_{0jk} für die Wahrscheinlichkeit in Feld (j,k) $(j = 1,2;\ k = 1,2)$ unter Gültigkeit der H_0 steht und π_{1jk} für diese Wahrscheinlichkeit unter der Gültigkeit einer spezifischen H_1. Weitere Zusammenhangsmaße für Häufigkeiten stellt Liebetrau (1983, S. 13-31, 39-44) dar.

5 Die Bedeutung von Unterschieds- und Zusammenhangsmaßen für psychologische Hypothesen

Natürlich kann man Unterschieds- und Zusammenhangsmaße auch dann benutzen, wenn man die Testlogik der Testtheorie Fishers bevorzugt, denn die Maße sind ja auch interpretierbar, ohne daß man auf eine Alternativhypothese Bezug nimmt. Und entsprechend findet man gegenwärtig recht häufig die nachträgliche Bestimmung des empirischen Unterschieds oder Zusammenhangs, die immerhin einen Fortschritt gegenüber der völligen Mißachtung dieser Maße darstellt. Aber diese Bestimmung bleibt unverbindlich, solange sie nicht bereits in der Planungsphase durch Maßnahmen der *Testplanung* (Teststärkeanalyse) sozusagen „vorbereitet" worden ist. Denn durch diese soll sichergestellt werden, daß a priori festgelegte Werte der Fehlerwahrscheinlichkeiten eingehalten werden und daß Unterschiede und Zusammenhänge bestimmter Größe (Kriteriumswerte) auch nachgewiesen werden können, wenn es sie in den jeweils angezielten „Populationen" gibt (Hager, 1987; vgl. auch den Beitrag von Buchner, Erdfelder & Faul, in diesem Band).

Obwohl die Beachtung von Zusammenhangs- und Unterschiedsmaßen in einigen Zeitschriften generell gefordert wird, gibt es nach wie vor starke Vorbehalte gegen diese Maße (vgl. etwa die Diskussion zwischen Strack & Rehm, 1984, und Westermann & Hager, 1984). Wenn aber diese Maße von allen Determinanten des statistischen Tests noch am ehesten auf die psychologischen Inhalte bezogen werden können, dann sollte man sie auch systematischer als bisher bei der Entscheidung über die psychologischen Hypothesen berücksichtigen. Der von Westermann und Hager (1982; vgl. auch Hager, 1984, 1992b) unterbreitete Vorschlag sieht vor, zwar die Entscheidung über die statistischen Hypothesen ausschließlich auf dem üblichen Vergleich des empirischen mit dem kritischen Wert der benutzten Teststatistik beruhen zu lassen, aber bei der Entscheidung über das Eintreten oder das Nichteintreten der aus der psychologischen Hypothese abgeleiteten psychologischen Vorhersage

auch den direkten Vergleich des empirischen Effektes mit dem bei der Testplanung festgesetzten Kriteriumswert zu berücksichtigen. Dieser Vergleich sollte aber erst dann durchgeführt werden, wenn die statistischen Tests zu den vorhergesagten Resultaten geführt haben. Diese Entscheidungsstrategie erfordert also keine Änderung der gegenwärtigen Testpraxis, sondern sie erweitert diese nur durch den Einbezug von Informationen, die man den gemeinhin in den Vordergrund gerückten p-Werten nicht direkt entnehmen kann. Aus der Sicht dieser Strategie zeigt der statistische Test, ob ein bestimmter Effekt statistisch signifikant ist, und auf einer nachgeordneten Entscheidungsebene wird dann gefragt, ob dieser *statistisch signifikante Effekt auch von genügender Größe* ist, um mit Blick auf die psychologische Vorhersage und Hypothese von Bedeutung sein zu können (vgl. Hager, 1992b). Andere Entscheidungsstrategien haben beispielsweise Bortz, Österreich und Vogelbusch (1979), Bredenkamp (1972) sowie Witte (1989) vorgeschlagen.

6 Weiterführende Literatur

Wenn man an der Prüfung psychologischer Hypothesen interessiert ist, die über aus ihnen abgeleitete statistische Hypothesen erfolgt, dann kommt den Zusammenhangs- und Unterschiedsmaßen eine besondere Bedeutung zu. Ihre systematische Berücksichtigung erfordert jedoch eine Testplanung. Über diese informieren für verschiedene Tests ausführlich Bredenkamp (1980), Cohen (1988) und Hager (1987); diesen Arbeiten liegt das von Cohen vorgeschlagene Vorgehen zugrunde. Auf einem anderen allgemeinen Ansatz beruht das Buch von Kraemer und Thiemann (1987), die zusätzlich zu den soeben genannten Autoren auch die Testplanung für Tests über verschiedene ordinale Zusammenhangsmaße besprechen. Levin (1975) befaßt sich speziell mit Fragen der Stichprobengröße bei Apriori- und Posthoc-Kontrasten. In seinem Lehrbuch legt Bortz (1993) seiner „Bestimmung optimaler Stichprobenumfänge" einen weiteren allgemeinen Ansatz zugrunde, der auf dem von Cohen (1988) beruht. Fleiss (1981) behandelt die Testplanung ausschließlich für Tests über Häufigkeitsdaten, für die sich ausführliche Tabellen auch bei Bortz et al. (1979) finden. Daneben gibt es verschiedene Programme (z.B. von Erdfelder, Faul & Buchner, 1996; vgl. Buchner, Erdfelder & Faul, in diesem Band), mittels derer die Testplanung zwar in vielen Standardsituationen einfach durchgeführt werden kann, die meines Erachtens aber noch nicht flexibel genug sind, um auch in bei der Hypothesenprüfung oft auftauchenden „Non-Standardsituationen" einsetzbar zu sein.

Literaturverzeichnis

Bortz, J. (1993). *Statistik für Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.

Bortz, J., Österreich, R. & Vogelbusch, W. (1979). Die Ermittlung optimaler Stichprobenumfänge für die Durchführung von Binomial-Tests. *Archiv für Psychologie, 131*, 267–292.

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

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

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

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Cramér, E. M. & Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika, 44*, 43–54.

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

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Gibbons, J. D. (1985). *Nonparametric methods for quantitative analysis* (2nd ed.). Columbus: American Sciences Press.

Glass, G. V., McGaw, B. & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills: Sage.

Hager, W. (1984). Aspekte eines deduktiven Forschungsansatzes in der empirischen Pädagogik: Fragen der Ableitungsvalidität, der Untersuchungsplanung und der Hypothesenbeurteilung. *Zeitschrift für Empirische Pädagogik und Pädagogische Psychologie, 8*, 56–75.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung psychologischer Hypothesen. In G. Lüer (Hrsg.), *Allgemeine experimentelle Psychologie* (S. 43–264). Stuttgart: G. Fischer.

Hager, W. (1992a). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen.* Göttingen: Hogrefe.

Hager, W. (1992b). Eine Strategie zur Entscheidung über psychologische Hypothesen. *Psychologische Rundschau, 43*, 18–29.

Hager, W. & Hasselhorn, M. (1994). Wortnormen: eine Übersicht. In W. Hager & M. Hasselhorn (Hrsg.), *Handbuch deutschsprachiger Wortnormen* (S. 2–34). Göttingen: Hogrefe.

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp & H. Feger (Hrsg.), *Hypothesenprüfung* (= Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 5, S. 24–238). Göttingen: Hogrefe.

Kraemer, H. C. & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park: Sage.

Kubinger, K. D. (1990). Übersicht und Interpretation der verschiedenen Assoziationsmaße. *Psychologische Beiträge, 32*, 290-346.

Levin, J. R. (1975). Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement, 12*, 99–108.

Liebetrau, A. M. (1983). *Measures of association.* Beverly Hills: Sage.

Marascuilo, L. A. & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey: Brooks/Cole.

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

Neyman, J. (1950). *First course in probability and statistics.* New York: Holt, Rinehart and Winston.

Seifert, T. L. (1991). Determining effect sizes in various experimental designs. *Educational and Psychological Measurement, 51*, 341–347.

Strack, F. & Rehm, J. (1984). Theorie testen oder Varianz aufklären? Überlegungen zur Verwendung von Effektgröße als Gütemaß für experimentelle Forschung. *Zeitschrift für Sozialpsychologie*, *15*, 81–85.

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. *Educational and Psychological Measurement*, *50*, 15–31.

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

Westermann, R. & Hager, W. (1984). Zur Verwendung von Effektgrößen in der theorieorientierten Sozialforschung. *Zeitschrift für Sozialpsychologie*, *15*, 159–166.

Witte, E. H. (1989). Die „letzte" Signifikanztestkontroverse und daraus abzuleitende Konsequenzen. *Psychologische Rundschau*, *40*, 76–84.

Wolf, B. (1988). Invariante Test- und Effektmaße sowie approximative Prüfgrößen bei multivariaten parametrischen Analysen. *Empirische Pädagogik*, *2*, 165–197.