Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2004-08

Der statistische Signifikanztest in der Krise
-
Bestandsaufnahme einer Vertrauenskrise und Vorschläge zu ihrer Überwindung

Andreas Quatember

Dezember 2004

Der statistische Signifikanztest in der Krise - Bestandsaufnahme einer Vertrauenskrise und Vorschläge zu ihrer Überwindung

Zusammenfassung: Signifikanztests statistischer Hypothesen gehören zu den am meisten verwendeten statistischen Methoden. Aus ihrer praktischen Anwendung erwachsen durch Umstände wie der gängigen Publikationspraxis mit der hauptsächlichen Berücksichtigung signifikanter Testergebnisse und dem – der Handlungslogik der Signifikanztestkonzeption völlig zuwiderlaufenden – „forschungshypothesenfreien Alles-mit-allem-Testen“ massive Zweifel an der Gültigkeit der mit diesen Anwendungen gewonnen Erkenntnisse. Zudem führt die absolut kontextunabhängige Übersetzung der verschiedensten wissenschaftlichen Fragestellungen in immer dieselben statistischen Hypothesen zu einer Fülle signifikanter, aber praktisch völlig irrelevanter Testergebnisse. Dieses Faktum lässt sich dadurch erklären, dass fast nie wirklich das getestet wird, was man eigentlich testen möchte.
In diesem Aufsatz wird ausgehend von einer ausführlichen Bestandsaufnahme der Praxis des Signifikanztestens in der empirischen psychologischen Forschung ein theoretisch fundierter Ausweg aus dieser umfassenden Vertrauenskrise dargelegt. Dieser beinhaltet auch den konzeptionellen Schritt vom Signifikanz- zum Relevanztest.

Abstract: Significance Testing of statistical hypotheses is one of the most used concepts of statistics these days. From the practical use, there arise serious problems, that are based on factors like the current publication practice of publicating almost only significant results and the method of “testing everything-with everything” without any research hypothesis, which infringes the logic of significance tests. These problems let us doubt strongly, that the knowledge, that results from these significance tests, is valid. Besides, the totally context-independent transformation of the various scientific questions in always the same statistical hypotheses leads to many test results, that are statistically significant, but practically completely irrelevant. This fact is very simply explained by the circumstance, that there almost never is inspected the hypotheses, that is wanted to be.
In this paper a theoretically sound way out of this crisis of confidence is shown based on a detailed stock-taking of the practice of significance testing in empirical psychological research. This way out comprises the conceptional step from tests of significance to tests of relevance.

Schlüsselwörter: Signifikanztest, Relevanztest, Publikationspraxis, Forschungshypothesenfreies Testen

1 Einleitung

Der wissenschaftliche Erkenntnisfortschritt basiert häufig auf der Entscheidung über Gültigkeit oder Nichtgültigkeit einer aufgestellten Theorie. In der Psychologie bedient man sich wie in vielen anderen Bereichen zu diesem Zweck seit Jahrzehnten des Instruments des Signifikanztestens statistischer Hypothesen. Dies in den letzten Jahren insofern umso mehr, als durch die ständige Weiterentwicklung von Statistik-Programmpaketen wie SAS oder SPSS oder die Implementierung eines umfangreichen Statistikmoduls in das Tabellenkalkulationsprogramm EXCEL die Verfügbarkeit dieser statistischen Methoden für jedermann, also auch für echte statistische Laien, wesentlich erleichtert wurde.
Verschiedene Umstände haben zu einer in zahllosen kritischen Publikationen dokumentierten Erschütterung des Vertrauens der Anwender in die Qualität der mit diesen Methoden der schließenden Statistik gewonnenen Erkenntnisse geführt. Dies lässt es notwendig erscheinen, die Art der Anwendung dieser Methoden der schließenden Statistik vom Gesichtspunkt der damit verbundenen Erwartungen einer gründlichen Überprüfung zu unterziehen.
Nach einem kurzen geschichtlichen Abriss der Überprüfung statistischer Hypothesen und einer sich daraus ergebenden Darstellung der Handlungslogik des statistischen Testens wird in dieser „Bestandsaufnahme“

- die Publikationspraxis geschildert, in der hauptsächlich signifikante Ergebnisse „zählen“,
- das forschungshypothesenfreie Testen beschrieben, in dem ohne Zugrundelegung von eigentlichen Fragestellungen die Daten nach allen Regeln der statistischen Softwarekunst „ausgequetscht“ werden, und
- auf das Problem der oftmals fehlenden praktischen Relevanz statistisch signifikanter Resultate eingegangen.

Von der Beschreibung der gegenwärtigen empirischen Forschungspraxis ausgehend werden für die sich daraus in Hinblick auf die Funktionstauglichkeit der Signifikanztestkonzeption ergebenden Probleme theoretisch fundierte Lösungsvorschläge unterbreitet, deren Ziel es ist, dem statistischen Signifikanztest bei der Suche nach neuen Erkenntnissen in der Praxis wieder mit jener Kompetenz auszustatten, die dieses Verfahren theoretisch auszeichnet.

1.1 Ein kurzer geschichtlicher Abriss zur Entwicklung der Handlungslogik des statistischen Signifikanztests

Die Theorie der zu den verschiedensten Fragestellungen entwickelten Signifikanztests ist praktisch zur Gänze ein Kind des 20. Jahrhunderts. Die dieser Theorie zugrunde liegende Logik wurde aber schon wesentlich früher angewandt. Ein Beispiel einer solchen frühen Anwendung ist im Bereich der Qualitätsprüfung die im 13. Jahrhundert in England eingeführte Gewichtsprüfung der in der königlichen Münzstätte geprägten Gold- und Silbermünzen (vgl. Stigler, 1977).
Der schottische Arzt John Arbuthnott überprüfte zu Beginn des 18. Jahrhunderts durch die Aufzeichnungen über die Geschlechtsverteilung der Geburten in London in 82 auf-

einanderfolgenden Jahren die Hypothese (*griechisch:* Hypothese = Unterstellung), dass gleich viele Mädchen wie Knaben geboren werden und kam zu dem Schluss, dass dies nicht der Fall sein kann (vgl. Arbuthnott, 1710). Sein Argument war folgendes: Wären die Wahrscheinlichkeiten für eine Knaben- bzw. eine Mädchengeburt gleich, dann würden die zur Verfügung stehenden Beobachtungen, dass Jahr für Jahr die Zahl der Knabengeburten jene der Mädchengeburten überwog, so unwahrscheinlich sein, dass aus den Beobachtungen auf die Ungültigkeit der Hypothese zu schließen ist. Diesen Umstand interpretierte er als Nachweis der Existenz Gottes!
1885 beschrieb F.Y. Edgeworth erstmalig ein objektives Verfahren für das Testen der Angemessenheit einer statistischen Hypothese durch Festlegung bestimmter Kriterien, die zur Ablehnung dieser Theorie führen (vgl. Kotz, Johnson, 1988, S.468). Ergebnisse, die diese Kriterien erfüllen, die also als Zeichen gegen die Theorie zu werten sind, bezeichnete er als signifikant (*lateinisch:* signum facere = ein Zeichen machen).
Die Entwicklung der modernen Theorie der Untersuchung statistischer Hypothesen, der sogenannten Testtheorie, begann jedoch erst mit dem englischen Mathematiker Karl Pearson. Pearson entwickelte die erste theoretische Grundlage für einen statistischen Test als er sich am Ende des 19. Jahrhunderts unabhängig von praktischen Anwendungen mit dem theoretischen Problem der Überprüfung der Anpassung beobachteter Verteilungen an theoretische beschäftigte. Das Ergebnis war im Jahre 1900 die Theorie für den Chiquadrat-Anpassungstest (vgl. Pearson, 1900). Pearsons Idee war, dass bei guter Übereinstimmung von Empirie und Theorie die über r Klassen des Wertebereiches eines Merkmals x berechnete Teststatistik Chiquadrat χ^2 mit

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(h_i^o - h_i^e\right)^2}{h_i^e} \qquad (1)$$

(h_i^o ... in der Stichprobe beobachtete Häufigkeit der i-ten Klasse von x; h_i^e ... theoretische Häufigkeit der i-ten Klasse bei Zutreffen der erwarteten Verteilungsform) einen kleinen Wert annehmen müsste. Ein großer Wert sei dann als Hinweis für eine schlechte Übereinstimmung zu deuten. Aus der Kenntnis der Verteilungsfunktion F der Zufallsvariablen χ^2 nach (1) bei Gültigkeit der theoretischen Verteilung, überließ Pearson es dem Anwender seines Tests mittels des sogenannten p-Wertes der Teststatistik χ^2, definiert als

$$p = 1 - F(\chi^2)$$

darüber zu befinden, ob die beobachtete Verteilung hinreichend genau mit der theoretischen übereinstimmt. Der p-Wert der Teststatistik gibt also die Wahrscheinlichkeit dafür an, dass bei Zutreffen der angenommenen Verteilungsform die Teststatistik einen höheren Wert als den Errechneten anzeigt. Pearsons eigene diesbezügliche Schlüsse lassen sich aus seinen Anmerkungen zu verschiedenen diesbezüglichen Ergebnissen ableiten: So interpretierte er einen p-Wert von 0,28 als Hinweis für eine „ziemlich gute Übereinstimmung“ und einen solchen von 0,1 noch als „nicht sehr unwahrscheinlich(es)“ Resultat, während er ein p von 0,01 als ein – hinsichtlich des Zutreffens der vermuteten Verteilungsform – „sehr unwahrscheinliches Resultat“ (Cowles, Davis, 1982, S.556) bezeichnete. Bei p-Werten um 0,1 schienen sich bei ihm also erste Zweifel über das Zutreffen der theoretischen Verteilung einzustellen und bei solchen von 0,01

schien er offenbar schon davon überzeugt zu sein, dass die Anpassung an die angenommene Verteilung unzureichend sei.
William S. Gosset, der seine wissenschaftlichen Arbeiten unter dem Pseudonym Student veröffentlichte, war von 1899 an in der Guiness-Brauerei zu Dublin tätig. Seine Aufgaben umfassten die Aufsicht über die dortigen Qualitätsprüfungen, wofür er aus der Notwendigkeit kleiner Stichprobenumfänge die sich daraus ergebende Prüfverteilung entwickelte. Für den so entstandenen t-Test von Mittelwerten legte Gosset nun fest, dass eine Abweichung der Teststatistik von der „Nullhypothese“ im Ausmaß des dreifachen wahrscheinlichen Fehlers des Mittelwertes – ein Begriff, der vom deutschen Astronomen und Mathematiker Friedrich W. Bessel 1818 erstmals verwendet wurde – als signifikant gelten sollte (vgl. ebd., S.556). Dieses von ihm verwendete Abstandsmaß gibt die Differenz zwischen dem oberen Quartil und dem Mittelwert einer normalverteilten Zufallsvariablen an, so dass die halbe Verteilungsfläche innerhalb des einfachen wahrscheinlichen Fehlers um den Mittelwert angesiedelt ist. Übertragen auf die heutzutage gebräuchliche Abstandsmessung mittels eines Vielfachen der Standardabweichung der Zufallsvariablen, entspricht der einfache wahrscheinliche Fehler etwa dem 0,674-fachen dieser Standardabweichung. Gossets dreifacher wahrscheinlicher Fehlers entspricht als „Signifikanzgrenze“ somit ca. dem 2,023-fachen der Standardabweichung. Die durch diese Grenze festgelegte Schranke für „auffällige“ p-Werte liegt bei zweiseitigen Tests bei etwa 0,043. p-Werte, die diese Grenze unterschritten, markierten nach Gosset also signifikante Teststatistiken.
Ronald Aylmer Fisher entwickelte in den Zwanziger Jahren neben der für die Schätztheorie richtungsweisenden Maximum-Likelihood-Methode zum Auffinden von geeigneten Schätzern weitere Testmethoden wie den F- oder den exakten Test, die Varianz-, die Kovarianz- und die Diskriminanzanalyse (vgl. Kotz, Johnson, 1988, S.469). Er präferierte die Verwendung der Standardabweichung an Stelle des wahrscheinlichen Fehlers. Als Mindestabstand zur Identifizierung signifikanter Abweichungen von der Ausgangshypothese legte Fisher – hier offenbar Gosset folgend – das Zweifache der Standardabweichung fest.
Die zunehmende Verwendung statistischer Testmethoden in den Sozialwissenschaften (und die zunehmende Verwendung von nichtnormalen Prüfverteilungen) ließen es in Hinblick auf ein breiteres Verständnis der zugrunde gelegten Handlungslogik als notwendig erachten, die Grenze zwischen nichtsignifikanten und signifikanten Testergebnissen durch das sogenannte Signifikanzniveau α, anzugeben. Dieses gibt die bedingte Wahrscheinlichkeit des Überschreitens dieser Grenze bei Zutreffen der als „Nullhypothese“ bezeichneten Hypothese an (vgl. Berkson, 1942, S.325; *englisch:* to nullify = entkräften; das ist jene Hypothese, die man für null und nichtig erklären möchte). α besitzt für die zweifache Standardabweichung bei zweiseitigen Fragestellungen konkret einen Wert von ca. 0,046 und dürfte von Fisher zur weiteren Vereinfachung auf 0,05 aufgerundet worden sein. Der Wert $\alpha = 0{,}05$ hat sich in der Folge in vielen Bereichen als konventionelles Signifikanzniveau etabliert.

1.2 Die „klassischen“ Vorgehensweisen beim statistischen Signifikanztesten

Die „klassische“ Signifikanztestkonzeption nach Fisher folgt nachstehender Handlungslogik:
Am Beginn steht eine wissenschaftliche Hypothese, die Forschungshypothese, die sich einem Substanzwissenschaftler – z.B. durch langjährige Beobachtungen – als Vermutung aufdrängt und für deren Zutreffen auch eine erklärende Theorie angeboten werden kann. Diese Forschungshypothese ist im nächsten Schritt in eine statistische Hypothese zu übersetzen, welche hinsichtlich eines oder mehrerer Merkmale eine Aussage über einen oder mehrere Parameter enthält, welche die zu diesen Merkmalen gehörende Verteilung. Diese Hypothese als zu überprüfende Aussage zumeist zur statistischen Eins- oder Alternativhypothese H_1, während die dazu komplementäre Aussage zur Nullhypothese H_0 des Tests wird. An Letzterer wird im Rahmen des Signifikanztests so lange festgehalten bis auf Basis der Daten einer zufällig aus der betreffenden Grundgesamtheit gezogenen Stichprobe von Erhebungseinheiten ernsthafte Zweifel an ihrer Gültigkeit entstehen. Dies bedeutet, dass die beiden Hypothesen nicht gleichberechtigt sind. Die Testphilosophie ist konservativ. Die Nullhypothese ist – in völliger Analogie zur sogenannten Unschuldsvermutung im Strafrecht – bevorzugt und die Teststrategie besteht darin, Indizien gegen ihre Gültigkeit zu suchen. Diese Hinweise sind schließlich hinsichtlich ihrer Vereinbarkeit mit der Nullhypothese zu bewerten und auf Basis dieser Bewertung ist eine Entscheidung über Beibehaltung der Nullhypothese oder deren Ablehnung zu Gunsten der Einshypothese zu treffen.
Dabei werden beim Testen eines Parameters – bezeichnen wir ihn allgemein mit θ – folgende Hypothesenformulierungen unterschieden:
Signifikanztests der Art

$$H_0: \theta \leq \theta_0 \quad \text{und} \quad H_1: \theta > \theta_0$$

bzw.

$$H_0: \theta \geq \theta_0 \quad \text{und} \quad H_1: \theta < \theta_0$$

behandeln einseitige, gerichtete Fragestellungen. Beide Hypothesen sind hierbei sogenannte zusammengesetzte Hypothesen. Sie bestehen also aus mehr als einem Wert aus dem möglichen Wertebereich Ω des gesuchten Parameters θ.
Ein Test der Hypothesen

$$H_0: \theta = \theta_0 \quad \text{und} \quad H_1: \theta \neq \theta_0$$

betrifft eine zweiseitige, ungerichtete Fragestellung. Bei zweiseitigen Fragestellungen ist die Einshypothese zusammengesetzt, während die Nullhypothese eine sogenannte einfache, nur eine Ausprägung aus Ω umfassende Hypothese ist.
Wie man sieht haben H_0 und H_1 in jedem Fall eine Zerlegung des gesamten möglichen Wertebereichs Ω des gesuchten Parameters zu sein, so dass genau eine der beiden Hypothesen wahr ist.
Im nächsten Schritt ist der zur Überprüfung der aufgestellten Hypothesen geeignete statistische Test zu bestimmen. Mit der Wahl des adäquaten Tests wird auch jene Zufallsvariable festgelegt, mit deren Hilfe die Entscheidung zwischen den beiden Hypothesen getroffen werden kann. Es ist dies die Teststatistik T.

Für die weitere Vorgangsweise beim Signifikanztesten stehen zwei Alternativen zur Verfügung: Für die „parameterorientierte“ Vorgangsweise muss die Zufallsstichproben-Verteilung der Teststatistik unter der Nullhypothese bekannt sein. Damit lässt sich nach Vorgabe des Signifikanzniveaus α des Tests jener Wertebereich E_1 für T festlegen, der diejenigen Realisationen von T enthält, die für den Fall der Gültigkeit der Nullhypothese nur mit Wahrscheinlichkeit α auftreten, und gleichzeitig die Wahrscheinlichkeit β, dass T nicht in E_1 zu liegen kommt, obwohl H_1 richtig ist, minimiert. In E_1 liegende und somit am deutlichsten gegen die Nullhypothese sprechende Teststatistiken werden als Indiz gegen die Gültigkeit von H_0 gewertet. E_1 ist deshalb der Ablehnungsbereich der Nullhypothese.
An dieser Stelle ist aus der interessierenden Grundgesamtheit eine Zufallsstichprobe zu ziehen und mit den damit erhaltenen Daten die Teststatistik T zu berechnen. Dies ist der zeitlich und meist auch finanziell aufwendigste Schritt im Ablauf eines statistischen Tests. Die Bedeutung der Ziehung einer Zufallsstichprobe für qualitativ hochstehende Rückschlüsse auf Zutreffen oder Nichtzutreffen statistischer Hypothesen wird von den Anwendern zumeist unterschätzt. Die zur Bestimmung des Ablehnungsbereiches E_1 notwendige Kenntnis der Stichprobenverteilung der vom jeweiligen Signifikanztest verwendeten Teststatistik T bei Gültigkeit der Nullhypothese liegt nur bei Zufallsstichproben vor. Willkürliche Auswahlen, bei denen befragt wird, wer sich dazu bereit erklärt – dazu gehören etwa Befragungen via Internet –, und bewusste Auswahlen wie Quotenauswahlen, lassen die Bestimmung der Stichprobenverteilung von T nicht zu (vgl. hierzu: Quatember, 2001, S. 33f und 93ff)!
Nach der Berechnung der Teststatistik T wird festgestellt, ob T in der Menge E_1 zu liegen kommt oder nicht. Teststatistiken T, die in diesen Ablehnungsbereich fallen, nennt man signifikant auf dem Niveau α. In einem solchen Fall wird die Nullhypothese zu Gunsten der Einshypothese verworfen. Eine häufig von Anwendern verwendete äquivalente Entscheidungsregel bei der parameterorientierten Vorgangsweise des Signifikanztestens besteht in der Berechnung des zur berechneten Teststatistik gehörenden p-Wertes (nach unserer Definition in Abschnitt 1.1) und der Entscheidung gegen die Beibehaltung von H_0, wenn $p \leq \alpha$ ist, was nichts anderes bedeutet, als dass T in E_1 zu liegen gekommen ist.
Eine Alternative zu dieser parameterorientierten Vorgangsweise mit asymptotisch gleichen Eigenschaften basiert auf die Bestimmung von Bereichschätzern für die in den Hypothesen verwendeten Parameter. Bereichschätzer überdecken in $(1-\alpha)\cdot 100$ % der Fälle den gesuchten Parameter.
Bei zweiseitigen Fragestellungen entscheidet man sich demzufolge bei der „schätzerbasierten“ Vorgangsweise auf dem Signifikanzniveau α „aus Mangel an Beweisen“ für die Beibehaltung von H_0, wenn der Parameterwert θ_0 aus der Nullhypothese Element des mit den Daten einer Zufallsstichprobe errechneten $(1-\alpha)$-Konfidenzintervalls $[\theta_u;\theta_o]$ für den Parameter θ ist: $\theta_0 \in [\theta_u;\theta_o]$. Andernfalls entscheidet man sich für die Akzeptierung von H_1.
Bei einseitigen Tests der Art

$H_0: \theta \leq \theta_0$ und $H_1: \theta > \theta_0$

berechnet man lediglich die untere Vertrauensschranke θ_u zur Sicherheit $1-\alpha$ und bleibt bei H_0, wenn $\theta_0 \in [\theta_u; \infty]$ ist. Andernfalls hat man offenbar ein starkes Indiz gegen H_0 gefunden, das eine Entscheidung für H_1 nahelegt.
Bei einseitigen Tests der Art

$H_0: \theta \geq \theta_0$ und $H_1: \theta < \theta_0$

berechnet man analog dazu natürlich die obere Vertrauensschranke θ_o zur Sicherheit $1-\alpha$ und entscheidet für die Beibehaltung von H_0, wenn gilt: $\theta_0 \in [-\infty; \theta_o]$.

Diese Vorgangsweise unterscheidet sich von der parameterorientierten dadurch, dass nun die Schätzer aus der Stichprobe und nicht die Stichprobenverteilung der Teststatistik unter dem Parameter der Nullhypothese die Entscheidungsregeln vorgeben. Beide Alternativen weisen jedoch asymptotisch gleiche Fehlerwahrscheinlichkeiten auf.

1.3 Zur Interpretation der Testergebnisse

Welche Vorgangsweise man auch wählt – die parameterorientierte oder die schätzerbasierte –, bei der Entscheidung über das Zutreffen von Hypothesen auf Basis von Stichproben sind solche Hypothesen natürlich weder verifizier- noch falsifizierbar, da kein Stichprobenergebnis einer statistischen Hypothese wirklich widersprechen kann. Tatsächlich verifizier- bzw. falsifizierbar wären die Hypothesen nur durch die Bestimmung der gesuchten Parameter.
Bereits im Jahr 1934 brachte Karl Popper seine kritische Haltung zu dieser Induktionslogik zum Ausdruck: „Der häufigste Fehler ist zweifellos der, dass den *Wahrscheinlichkeitshypothesen*, also den hypothetischen Häufigkeitsansätzen, *Hypothesenwahrscheinlichkeit* zugeschrieben wird. Man versteht diesen Fehlschluss wohl am besten, wenn man sich daran erinnert, dass die Wahrscheinlichkeitshypothesen ihrer logischen Form nach, also ohne Berücksichtigung unserer methodologischen Falsifizierbarkeitsforderung, weder verifizierbar noch falsifizierbar sind: Verifizierbar sind sie nicht, weil sie allgemeine Sätze sind, und sie sind nicht streng falsifizierbar, weil sie nie in logischem Widerspruch zu irgendwelchen Basissätzen stehen können. Sie sind also [nach Reichenbach] „völlig unentscheidbar“. Nun können sie sich aber, wie wir gezeigt haben, *besser oder schlechter „bestätigen“*, d.h., sie können mit anerkannten Basissätzen besser oder schlechter übereinstimmen. Das ist die Situation, an welche die Wahrscheinlichkeitslogik anknüpft: In Anlehnung an die klassisch-induktionslogische Symmetrie zwischen Verifizierbarkeit und Falsifizierbarkeit glaubt man, den unentscheidbaren Wahrscheinlichkeitsaussagen abgestufte Geltungswerte zuschreiben zu können, „stetige Wahrscheinlichkeitsstufen, deren unerreichbare Grenzen nach oben und unten Wahrheit und Falschheit sind“ [Reichenbach]. Nach unserer Auffassung sind jedoch die Wahrscheinlichkeitsaussagen, wenn man sich nicht entschließt, sie durch Einführung einer methodologischen Regel falsifizierbar zu machen, eben wegen ihrer völligen Unentscheidbarkeit *metaphysisch*. Die Folge ihrer Nichtfalsifizierbarkeit ist dann nicht, dass sie sich etwa „besser“ oder „schlechter“ oder auch „mittelgut“ bewähren können, sondern sie können sich dann *überhaupt nicht empirisch bewähren*; denn da sie nichts verbieten, also mit jedem Basissatz vereinbar sind, so könnte ja *jeder beliebige* (und beliebig komplexe) einschlägige Basissatz als „Bewährung“ angesprochen werden“ (Popper, 1989, S.207f, Hervorhebungen wie dort).

In der Konzeption der Signifikanztests wird nicht die Forschungshypothese des Substanzwissenschaftlers in die statistische Nullhypothese übersetzt. Diese Funktion übernimmt die ihr widersprechende Hypothese, an der so lange festgehalten wird bis massive Zweifel ihrer Gültigkeit entgegenstehen. Die von Popper geforderten „methodologischen Regeln“ werden aufgestellt und damit bereits vor dem Experiment konkret festgelegt, unter welchen Bedingungen man die Nullhypothese zu verwerfen bereit ist, um auf die Einshypothese, d.i. die Forschungshypothese, überzugehen. Diese Verwerfungskriterien werden durch die Wahl des Signifikanzniveaus α so festgelegt, dass „unbegründete und voreilige Schlussfolgerungen zugunsten der Forschungshypothese erheblich erschwert werden“ (Bortz, Döring, 1995, S.28).
Man sieht, dass gerade weil statistische Hypothesen nicht falsifizierbar sind, die Nullhypothese nach der Logik des Signifikanztestens solange als zutreffend zu gelten hat, bis Stichprobenergebnisse dies (massiv) in Zweifel ziehen. Ein nichtsignifikantes Testergebnis, das zur Beibehaltung der Nullhypothese führt, hat somit die Bedeutung eines die Forschungshypothese falsifizierenden Testergebnisses. Die aus der Forschungshypothese abgeleitete Theorie sollte dann nicht weiter verfolgt werden, was nach Popper ebenso wie der (vorläufige) Akzeptierung einer neuen Theorie als Erweiterung unseres Wissens zu werten ist, verhindert dies doch wissenschaftliche Fehlentwicklungen: „Und wenn wir auch sagen können, daß es nur die Theorie und nicht das Experiment ist, nur die Idee und nicht die Beobachtung, die der Wissenschaftsentwicklung immer wieder den Weg zu neuen Erkenntnissen weist, so ist es doch immer wieder das Experiment, das uns davor bewahrt, unfruchtbare Wege zu verfolgen, das uns hilft, ausgefahrene Geleise zu verlassen und uns vor die Aufgabe stellt, neue zu finden“ (Popper, 1989, S.214).
Ein signifikantes Testergebnis aber liefert die Grundlage für die Ablehnung der Nullhypothese, was einer Bewährung der Forschungshypothese entspricht. Diese gilt deshalb natürlich nicht im geringsten als verifiziert, woraus sich ebenso selbstverständlich im Popperschen Sinne die Verpflichtung des redlichen Wissenschafters ergibt, durch die anhaltende Suche nach Fakten, die seiner Theorie widersprechen – etwa durch Überprüfung in weiteren Studien –, diese erneut in Frage zu stellen.
Die Handlungslogik des Signifikanztests hält sich m.E. in jenem Maße an den Popperschen Falsifikationismus als es die Konzeption nur möglich macht. Unbeantwortet bleibt natürlich die Frage, „wieviele Anwender (statistischer Tests; Anm.d.Verf.) sich überhaupt dem von Popper begründeten Falsifikationismus verpflichtet fühlen“ (Ostmann, Wutke, 1994, S.731), denn es besteht der Verdacht, dass „nicht wenige Anwender ... wohl eigentlich „implizite Verifikationisten“ (sind), die mit jeder „Signifikanz“ den Fundus gesicherten Wissens erweitern wollen“ (ebd.).
Im Übrigen kann sich die Einschätzung auch nichtsignifikanter Ergebnisse als Beitrag zum Erkenntnisfortschritt nur dann positiv in den verschiedenen Wissenschaften manifestieren, wenn die Publikationspraxis in den Disziplinen auch einen so definierten Erkenntnisfortschritt zulässt.

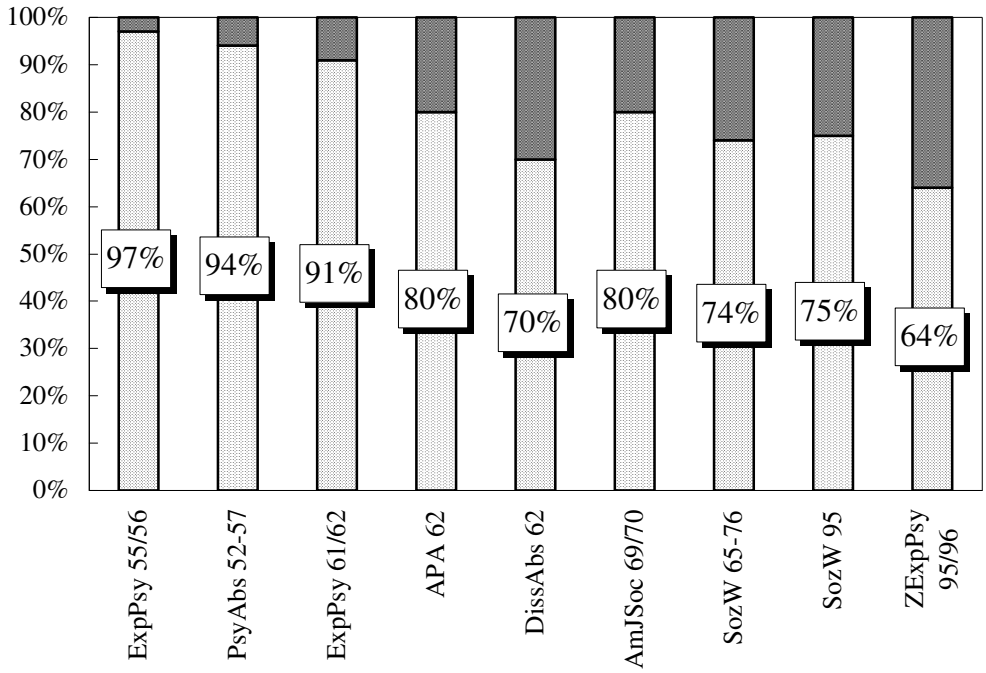
2 Die Publikationspraxis und ihre Auswirkungen

Die in der empirischen psychologischen Forschung und auch anderen Bereichen wie der klinischen Forschung oder der empirischen Sozialforschung im Zusammenhang mit Ergebnissen von Signifikanztests gängige Publikationspraxis ist ein Umstand, der das Zutreffen der im letzten Abschnitt beschriebenen Interpretationen von damit erzielten Forschungsergebnissen anzweifeln lässt. Nichtsignifikante Testergebnisse werden offenbar für uninteressant gehalten und besitzen daher eine geringere Veröffentlichungschance als signifikante Ergebnisse. In einem anderen Umfeld – man denke etwa an von der Industrie finanzierte Studien zur Umweltverschmutzung – könnten umgekehrt hauptsächlich nichtsignifikante Teststatistiken an die Öffentlichkeit gelangen und signifikante Ergebnisse „verschwinden“. Das Problem dabei ist: „Was nicht berichtet wird, existiert nicht, oder etwas vorsichtiger: Seine Chancen, zu einem Teil der von den Zeitgenossen wahrgenommenen Wirklichkeit zu werden, sind minimal“ (Noelle-Neumann, 1980, S.216). Die Publikationskultur ist dabei identisch mit der gesamten Medienkultur, „das ist die Auswahl von Welt, wie sie die Medien anbieten, und soweit die Welt außer Reichweite, außer Sicht eines Menschen liegt, ist es meist die einzige Ansicht der Welt, die er besitzt“ (ebd., S.216).

Die Konsequenz einer solchen Publikationspraxis, die eben nur eine bestimmte „Auswahl von Welt“ und nicht die Welt selbst anbietet, ist, dass Untersuchungen, die nicht zu dieser Auswahl gehören, von anderen Forschern solange wiederholt werden bis sie (möglicherweise irrtümlich) ein signifikantes Ergebnis hervorbringen. Wird dieses publiziert, weil man nirgends nachlesen kann, dass zum betreffenden Thema auch einige nichtsignifikante Untersuchungsergebnisse vorliegen, dann wird dieser Irrtum Bestandteil der des „Wissens“! Auf diese Ergebnisse basierend werden dann „Klienten psychotherapiert, Organisationen umstrukturiert, Erziehungsstile vermittelt, Medikamente eingeführt, Maßnahmen im wirtschaftlichen Bereich getroffen, Züchtungen von Nutzpflanzen vorgenommen etc.“ (Witte, 1980, S.58).

Verschiedene in der Vergangenheit durchgeführte Erhebungen zur Verteilung der Signifikanz bzw. Nichtsignifikanz der Testergebnisse in veröffentlichten Artikeln in Periodika der Psychologie und Soziologie belegen gemeinsam mit einer für einen Vortrag des Autors im Jahr 1997 durchgeführten Erhebung in vier Zeitschriften aus diesen Bereichen die Behauptung, dass vorwiegend signifikante Testergebnisse publiziert werden (Abbildung 1).

Abbildung 1:
Die Verteilung der mit Signifikanztestergebnissen veröffentlichten Artikeln hinsichtlich der Signifikanz bzw. Nichtsignifikanz der Testergebnisse (der Anteil von Artikel mit mehrheitlich signifikanten Ergebnissen ist hell gepunktet, jener mit überwiegend nicht-signifikanten Ergebnissen dunkel gepunktet dargestellt).



ExpPsy 55/56: Vollerhebung von N = 294 Artikeln mit Signifikanztests aus 4 zufällig ausgewählten psychologischen Zeitschriften der Jahrgänge 1955 bzw. 1956: *Journal of Experimental Psychology* (1955), *Journal of Comparative and Physiological Psychology* (1956), *Journal Journal of Clinical Psychology* (1955) und *Journal of Social Psychology* (1955) (Sterling, 1959, S.31f).

PsyAbs 52-57: Zufallsauswahl von n = 100 Artikeln mit Signifikanztests aus den *Psychological Abstracts* der Jahrgänge 1952-1957 (Sterling, 1959, S.32 Fußnote).

ExpPsy 61/62: Vollerhebung von N = 309 Artikeln mit Signifikanztests der Jahrgänge 1961 bzw. 1962 derselben Periodika, die Sterling (1959) für seine Untersuchung ausgewählt hat: *Journal of Experimental Psychology* (1962), *Journal of Comparative and Physiological Psychology* (1962), *Journal Journal of Clinical Psychology* (1961) und *Journal of Social Psychology* (1962) (Smart, 1964, S.226f).

APA 62: Zufallsauswahl von n = 83 Abstracts mit Signifikanztests aus den Referaten des Kongresses der American Psychological Association im Jahr 1962 (Smart, 1964, S.226f).

DissAbs 62: Vollerhebung der N = 86 Dissertationen aus 3 zufällig ausgewählten Ausgaben der *Dissertation Abstracts* des Jahres 1962 (Smart, 1964, S.227).

AmJSoc 69/70: Vollerhebung von N = 76 Artikeln mit Signifikanztests der Ausgaben von Juli 1969 - Juni 1970 der Zeitschriften: *American Journal of Sociology*, *The American Sociological Review* und *Social Forces* (Wilson et al., 1973, S.143).

SozW 65-76: Vollerhebung der N = 94 Artikel mit Signifikanztests von Jahrgängen zwischen 1965 und 1976 folgender Zeitschriften: *Soziale Welt* (1965-1976), *Kölner Zeitschrift für Soziologie und Sozialpsychologie* (1965-1976) und *Zeitschrift für Soziologie* (1972-1976) (Sahner, 1979, S.267ff).

SozW 95: Vollerhebung der N = 16 Artikel mit Signifikanztests des Jahrganges 1995 derselben Periodika, die Sahner (1979) für seine Untersuchung ausgewählt hat: *Soziale Welt*, *Kölner Zeitschrift für Soziologie und Sozialpsychologie* und *Zeitschrift für Soziologie*.
ZExpPsy 95/96: Vollerhebung von N = 47 Artikeln mit Signifikanztests in den Jahrgängen 1995 und 1996 der *Zeitschrift für Experimentelle Psychologie*.

Diese Verteilung könnte nun theoretisch auch damit erklärt werden, dass lediglich gut begründete Forschungshypothesen einem statistischen Test unterzogen werden und deshalb die „Signifikanzquote" so hoch ist. Doch bereits Sterling (1959), in dessen Untersuchungen (siehe Abbildung 1) einmal 286 von 294 (97 %) und ein andermal 94 von 100 Beiträgen eine Mehrheit an signifikanten Testergebnissen enthielten, merkt dazu an, dass nichtsignifikante Ergebnisse von Signifikanztests „mit Sicherheit ... mit geringerer Häufigkeit in der Literatur vorkommen als es im Laboratorium vernünftigerweise erwartet werden kann – auch wenn angenommen wird, daß alle Experimentatoren bei der Auswahl ihrer Hypothesen außerordentlich geschickt sind" (ebd., S.34). Diese Vermutung wird durch den Umstand unterstrichen, dass unter den (unveröffentlichten) Dissertationen und Kongressvorträgen aus dem Bereich der Psychologie die Anteile mit überwiegend signifikanten Testergebnissen geringer sind als unter den in psychologischen Zeitschriften veröffentlichten Artikeln (siehe dazu APA 62 und DissAbs 62 in Abbildung 1). Smart (1964) erhob zusätzlich 37 Dissertationen auf Basis einer Zufallsauswahl aus den *Dissertation Abstracts* des Jahrganges 1956: Laut den *Psychological Abstracts* wurden bis zum Jahr 1963 die Ergebnisse von 12 der 23 Dissertationen mit überwiegend signifikanten Resultaten, aber nur 2 der 14 mit überwiegend nichtsignifikanten Resultaten in Artikelform publiziert (vgl. ebd., S.227f).
In den Jahrgängen 1995 und 1996 einer psychologischen und im Jahrgang 1995 mehrerer soziologischer deutschsprachiger Zeitschriften finden sich ebenfalls klare Mehrheiten an „signifikanten Artikeln", woraus man schließen darf, dass diese Praxis auch heute noch die in den angesprochenen Disziplinen übliche ist.
Auf Grund geringer Aussicht auf wissenschaftliche Anerkennung werden bereits veröffentlichte Untersuchungen auch kaum wiederholt oder solche Wiederholungen kaum publiziert. Sterling (1959) fand in seiner bereits erwähnten Untersuchung unter 362 Aufsätzen aus dem Bereich der Psychologie keinen einzigen Aufsatz, der eine Replikation einer bereits veröffentlichten Untersuchung enthielt (ebd., S.31) und Wilson et al. (1973) unter 76 Aufsätzen aus dem Bereich der Soziologie lediglich vier (ebd., S.143)! Dies beschreibt eine Forschungspraxis, die nicht darauf abzielt, eigene oder die Forschungsergebnisse anderer in Zweifel zu ziehen. Dabei könnten gerade die Ergebnisse der Replikationen solcher Untersuchungen einen weiteren Beitrag bei der Wahrheitsfindung leisten, so dass solche im Popperschen Sinne geradezu eine Verpflichtung für die an der Wahrheit interessierte Forschungsgemeinschaft darstellen müssten.
Eine Idee zur Vermeidung der Publikation solcher Fehlentscheidungen wäre folglich, sich darum zu bemühen, die zum betreffenden Forschungsgegenstand durchgeführten Untersuchungen vollständig zu erfassen und die Ergebnisse der unabhängigen Einzeluntersuchungen zu aggregieren. Dies ist die Aufgabe der sogenannten metaanalytischen Techniken (vgl. etwa: Bortz, Döring, 1995, S.589ff). Unerlässlich für eine ernstzunehmende Anwendbarkeit dieser Methoden ist die Zugänglichkeit aller zu

einem Thema durchgeführten Erhebungen. Diese Voraussetzung ist angesichts der vorherrschenden Publikationspraxis, in der nichtsignifikante Ergebnisse von statistischen Untersuchungen in aller Regel verschwinden, im Allgemeinen aber nicht gegeben. In der Medizin gibt es in einigen speziellen Forschungsbereichen wie z.B. in der Krebsforschung eine verpflichtende Registrierung von Studien bei einzuhaltenden Qualitätsmaßstäben (vgl. Begg, Berlin, 1988, S.439). Solche Register könnten für die Metaanalyse der Forschungsergebnisse von Nutzen sein. Neben diesem praktischen Problem ist die Schwierigkeit der Vereinheitlichung der Ergebnisse diverser Studienergebnisse, die mit verschiedenen Stichprobenverfahren, Erhebungs- oder Messmethoden usw. gewonnen wurden, mit der sich die Metaanalyse auseinandersetzt (vgl. etwa: Sedlmeier, 1996, S.63), geradezu vernachlässigbar.
Somit kann m.E. der gängigen Bewertung ausschließlich signifikanter Testergebnisse als Erkenntnisfortschritte nur mit der Forderung nach einem diesbezüglichen Paradigmenwechsel begegnet werden. Auch nichtsignifikante Testergebnisse sind als Erfolge auf dem Weg zur Wahrheit zu werten und dienen vor allem der umfassenden Einschätzung weiterer Forschungsresultate zu denselben oder ähnlichen Fragestellungen. Damit könnte sich die empirische Forschung von der einseitigen Jagd nach Signifikanzen wieder zu einer nach allen Seiten offenen Suche nach Erkenntnisgewinnen weiter- oder – besser gesagt – zurückentwickeln.

3 Das „Alles-mit-Allem-Testen“ und seine Auswirkungen

Gemeinsam mit „jener üblen Forschungspraxis, bei der nur „Signifikanzen“ zählen“ (Ostmann, Wutke, 1994, S.722) forcierte der massive Einsatz statistischer Programmpakete die oben geschilderten Konsequenzen in geradezu beängstigendem Ausmaß. Diese Programmpakete ersetzen „die überaus komplizierte und sensible Handlungslogik des Hypothesentestens“ durch die „schlichte Automatik der Berechnung von p-Werten“ (Diepgen, 1994, S.11), auf Basis derer die Entscheidungen getroffen werden. Dies hat zur Folge, dass der nach Veröffentlichungen strebende Wissenschaftler die ihm zur Verfügung stehenden Daten nach allen Regeln der statistischen Softwarekunst „ausquetschen“ kann, d.h., dass er Alles mit Allem testet, ohne dass im Einzelnen dahinter eine begründete Forschungshypothese steht (vgl. Quatember, 1997).
Während Sahner (1979) bei seiner Erhebung (siehe Abbildung 1) von Artikeln, in denen Signifikanztests verwendet wurden, in soziologischen Zeitschriften der Jahrgänge 1965 bis 1976 im Durchschnitt 51,6 (!) berichtete Signifikanztests pro Artikel feststellte (vgl. ebd., S.278), lag dieser Mittelwert bei denselben Zeitschriften im Jahr 1995 bei 114,9 (Median: 80,5). In den Jahrgängen 1995 und 1996 der Zeitschrift für Experimentelle Psychologie wurde im Mittel Resultate von 37,8 Signifikanztests pro Artikel berichtet (Median: 44).
Sahner merkt 1979 an: „Es ist zu erwarten, daß die Relationen zwischen signifikanten und nicht signifikanten Ergebnissen zukünftig sich verändern werden und zwar dergestalt, daß zunehmend mehr nichtsignifikante Ergebnisse ausgezählt werden können. Dies dürfte ein Ergebnis der zunehmenden Verwendung der EDV sein. Während früher (mein Vorurteil) nur die vermutlich signifikanten Beziehungen errechnet und referiert

worden sind, werden heute alle möglichen Tests durchgeführt“ (ebd., S.271, Fußnote). Diese Prognose ist m.E. voll eingetroffen.
Eine solche Handlungsweise war vor Einführung von Programmpaketen in dem Ausmaß, wie sie heute betrieben wird, vom Aufwand her gar nicht möglich. Es wird somit eine Vorgangsweise gewählt, die im krassen Widerspruch zur klassischen Handlungslogik von Signifikanztests steht und als „forschungshypothesenfreies Testen“ tituliert werden kann.
Das dieser Vorgangsweise eigene Abwarten der Anwender darauf, welche aus der Unmenge berechneter Teststatistiken signifikant werden, birgt indes nicht geringe Gefahren. Denn „der Witz ist, *daß wir stets etwas Besonderes finden, wenn wir nicht nach etwas Bestimmten suchen*. Irgendwelche Muster entstehen letztlich immer.“ Und „interessant sind sie nur, wenn eine Theorie sie vorhergesagt hat. Deshalb gehört es zum Standard wissenschaftlicher Studien, daß *erst* das Untersuchungsziel und die Hypothese angegeben werden müssen und *dann* die Daten erhoben werden. Wer aber nach *irgendwelchen Mustern* in Datensammlungen sucht und *anschließend* seine Theorien bildet, schießt sozusagen auf die weiße Scheibe und malt danach die Kreise um das Einschußloch“ (von Randow, 1994, S.94; Hervorhebungen wie dort). Die strukturierte, theoretisch fundierte Durchführung dieser Vorgangsweise bei riesigen Datenmengen ist gegenwärtig unter dem Begriff „Datamining“ Arbeitsgegenstand vor allem der Programmierer statistischer Programmpakete.
Betrachten wir zur Verdeutlichung des Problems doch einmal folgendes „Experiment“: Es befinden sich 400 Studierende in einem Hörsaal. Diese sollen unabhängig voneinander die Ausgänge von sechs aufeinanderfolgenden Münzwürfen vorhersagen. Die Wahrscheinlichkeit dafür, dass dies einem bestimmten Studierenden, der keine wahrsagerischen Qualitäten aufweist, vollständig gelingt, beträgt

$$\left(\frac{1}{2}\right)^6 = 0{,}015625 \ .$$

Dies ist ein so seltenes Ereignis, dass man nach der Logik des Signifikanztestens bei dessen Eintreffen von einem (auf dem 5%-Niveau) signifikanten Testergebnis sprechen kann. Man hätte somit die Nullhypothese, dass die Vorhersagewahrscheinlichkeit π einer Versuchsperson bei einem Münzwurf ½ beträgt, zu verwerfen. Dies natürlich dann umso mehr, wenn die betreffende Person schon vorher unter dem Verdacht stand, hellseherisch veranlagt zu sein, und dieser Umstand den Anlass zur Überprüfung dieser Hypothese lieferte.
Prüft man jedoch anstelle einiger weniger „einschlägig verdächtiger“ Studierender gleich 400 forschungshypothesenfrei, dann werden, sofern alle unabhängig voneinander raten, durchschnittlich

$$400 \cdot \left(\frac{1}{2}\right)^6 = 6{,}25$$

alle Ausgänge richtig vorhersagen. Dies fordert die statistische Theorie. Der Experimentator aber geht nach der verqueren Logik des forschungshypothesenfreien Testens mit den Personen, die alle Ausgänge richtig vorhersagten, und nur mit diesen an die (staunende) Öffentlichkeit, welche die näheren Umstände des Experiments nicht erfährt, ohne eine Theorie für die Fähigkeiten der Versuchspersonen anbieten zu können.

Es macht jedoch einen großen Unterschied für die qualitative Einschätzung der Untersuchungsergebnisse, ob wenige, aber begründete oder eine Unzahl unbegründeter Tests durchgeführt worden sind. Oder ist der werte Leser dieses Aufsatzes schon einmal auf die Idee gekommen, den Lottogewinner vom letzten Wochenende für parapsychologisch veranlagt zu halten?
Ein anderes lehrreiches Beispiel: „Nehmen wir an, ein Möchtegern-Berater druckt ein Firmenemblem auf extrafeines Briefpapier und verschickt 32 000 Briefe an potentielle Kapitalanleger. Die Briefe berichten vom hochentwickelten Computermodell seiner Firma, von seiner reichen Erfahrung in Finanzangelegenheiten und seinen guten Geschäftsbeziehungen. In 16 000 der Briefe wird die Vorhersage gemacht, daß die betreffende Aktie steigen wird, in den anderen 16 000 wird vorausgesagt, daß sie fällt. Es ist nun völlig gleichgültig, ob die Aktie steigt oder fällt; jedenfalls wird denjenigen, die eine korrekte „Vorhersage“ erhalten haben, ein weiterer Brief zugesandt. 8 000 von ihnen wird für die darauf folgende Woche vorhergesagt, daß die Aktie steigt, den anderen 8 000 das Gegenteil. Damit haben naturgemäß 8 000 Leute eine weitere korrekte Vorhersage erhalten.
Dieses Spiel wird noch ein paar Mal wiederholt, bis schließlich 500 Personen sechs richtige Voraussagen erhalten haben. Diesen 500 Leuten wird nun das Angebot unterbreitet, daß sie auch in der siebten Woche diese wertvolle Information erhalten könnten − wenn sie 500 Dollar bezahlen. Wenn alle 500 Adressaten darauf eingehen, verdient der Berater bei diesem >Geschäft< 250 000 Dollar“ (Paulos, 1993, S.83f).
Die Kenntnis ausschließlich eines bestimmten Teils und nicht der gesamten Versuchsanordnung verändert die Beurteilung der Qualität der Ergebnisse jener Tests, die „ein Signum lieferten“, offenbar entscheidend. Die forschungshypothesenfreie Handlungslogik verstößt gravierend gegen die klassische Handlungslogik, derer man sich bei Verwendung von Signifikanztests zu unterwerfen hat, wenn man sich auf ihre theoretischen Eigenschaften berufen möchte. Die – wenn überhaupt – nachträglich auf Basis signifikanter (und somit publizierbarer; siehe Abschnitt 2) Testergebnisse formulierten Theorien zur Erklärung dieser Ergebnisse hatten nie die Chance, innerhalb des Signifikanztestkonzepts widerlegt zu werden! Ein beträchtlicher Teil des so erzeugten „Wissens“ ist falsch.
Die Anpassung der für den einzelnen Signifikanztest konzipierten Theorie an diese von den Anwendern gewählte Strategie mittels Einführung eines gemeinsamen α-Fehlers α_{ges} für die Gesamtheit aller k Tests wird als α-Adjustierung bezeichnet (vgl. etwa: Stelzl, 1982, S.117ff). α_{ges} stellt darin in Abänderung der Wahrscheinlichkeit α, bei einem Test bei tatsächlichem Zutreffen der Nullhypothese eine Fehlentscheidung zu treffen, jene nun zu kontrollierende Wahrscheinlichkeit dar, bei Gültigkeit aller Nullhypothesen dies nur mindestens einmal zu tun.
Werden mit den Daten zu Merkmalen derselben Versuchspersonen mehrere statistische Tests durchgeführt, so sind die Testergebnisse nicht statistisch unabhängig voneinander. Beschreibe x die Anzahl der Fehlentscheidungen bei Gültigkeit sämtlicher Nullhypothesen. Es gilt dann für den Erwartungswert E von x bei k durchgeführten Tests:

$$E\,x = \sum_{i=1}^{k} \alpha = \alpha_{ges}$$

und somit:

$$\alpha = \frac{\alpha_{ges}}{k} \ .$$

Der Nachteil der α-Adjustierung ist offensichtlich, dass sich durch die Verringerung des α-Fehlers für die Einzeltests (wenn man nun α_{ges} mit 0,05 festlegt) gleichzeitig auch deren Teststärke $1-\beta$ (massiv) verringert. Dies bedeutet, dass bei kleinen Stichprobenumfängen nur mehr wesentlich größere Abweichungen von der Nullhypothese im Vergleich zur herkömmlichen Vorgangsweise erkannt werden können. Beim oben beschriebenen „Experiment" mit 400 Studierenden würde beispielsweise die Wahrscheinlichkeit α für ein irrtümlich signifikantes Testergebnis bei Gültigkeit der Nullhypothese für $\alpha_{ges} = 0{,}05$ wegen der hier vorliegenden Unabhängigkeit der einzelnen Tests mit

$$\alpha = 1 - \sqrt[400]{0{,}95} \approx 0{,}00013$$

festgelegt. Da beim Stichprobenumfang von n = 6 Würfen die Wahrscheinlichkeit für sechs richtige Voraussagen größer als diese Wahrscheinlichkeit ist, kann in diesem Experiment nach α-Adjustierung sogar gar kein Testergebnis signifikant sein!

Dies ist für das Auffinden von Abweichungen von der Nullhypothese natürlich genauso unbefriedigend wie das multiple Alles-mit-Allem-Testen ohne α-Adjustierung. Diese bewirkt eigentlich auch keine Änderung in der Handlungslogik des forschungshypothesenfreien Testens: die Theorien werden erst nach dem Test entworfen und können daher mit ihm nicht geprüft worden sein. Auf diese Art und Weise behält der Volksmund ganz und gar recht, wenn es heißt, dass man „mit Statistik alles beweisen kann".

Angesichts der durch den Einsatz von statistischen Programmpaketen erst möglich gewordenen und von der in Abschnitt 2 geschilderten Publikationspraxis auf der Jagd nach Signifikanzen noch geförderten Methode des forschungshypothesenfreien Testens ist von den Anwendern der statistischen Signifikanztests verstärkt wissenschaftliche Redlichkeit einzufordern, damit sie ihr Fach wegen der geschilderten Mängel dieser Vorgangsweise nicht mit „Scheinwissen" überschwemmen. Zu dieser Redlichkeit gehört die selbstverständliche Unterwerfung unter die klassische Handlungslogik und der ausschließlichen Überprüfung vorab formulierter Forschungshypothesen. Wie in Abschnitt 1.2 gezeigt wurde, garantiert nur dies die volle Qualität der Signifikanztestkonzeption. Die Ergebnisse multiplem „Alles-mit-Allem-Testens" dürfen keinesfalls schon als Forschungsergebnisse behandelt werden, wie dies häufig der Fall ist.

Und dennoch können sie einen nützlichen Beitrag auf dem Weg zu neuen Erkenntnissen darstellen, wenn sie, sofern sie durch eine sinnvolle erklärende Theorie unterlegt werden können, als Forschungshypothesen in einer neuen Untersuchung überprüft werden. Sie dürfen eben bestenfalls das Nachdenken des Substanzwissenschaftlers zur Auffindung interessanter Fragestellungen unterstützen. Die konsequente und ausnahmslose Einhaltung dieser Forderung würde zu einer Verringerung der Anzahl durchgeführter statistischer Tests führen und damit auch zu einer Verringerung von Fehlentscheidungen. Eine geringere Anzahl auch noch fachlich begründeter Tests produziert ganz einfach eine geringere Anzahl von Irrtümern.

4 Die Hypothesenformulierung und ihre Auswirkungen

4.1 Statistisch signifikant oder praktisch bedeutsam?

Ein Faktum des Signifikanztestens ist der Umstand, dass mit zunehmendem Stichprobenumfang n jeder auch noch so gering von dem in der Nullhypothese formulierten Wertebereich abweichende Parameter θ erkannt wird. Dies heißt, dass die sogenannte Gütefunktion des Tests, die die Wahrscheinlichkeit für die Annahme der Nullhypothese in Abhängigkeit vom wahren Parameterwert θ misst, bei konstantem Signifikanzniveau α mit wachsendem n in jenem Bereich der H_1, der an die H_0 grenzt, einen immer steileren Verlauf nimmt. Es nimmt somit für jedes $\theta \in H_1$ die Wahrscheinlichkeit $1-\beta$ dafür zu, dies auch zu erkennen. $1-\beta$ wird als Teststärke eines Signifikanztests bezeichnet.
Dieses positive Qualitätsmerkmal funktionstauglicher Hypothesentests hat für ein- bzw. zweiseitige Fragestellungen etwas unterschiedliche Konsequenzen:
Bei einseitigen Fragestellungen liegt für einen Stichprobenumfang n die Gütefunktion des Tests im Bereich der zu H_0 gehörigen Parameterwerte immer unterhalb der zu einem geringeren Stichprobenumfang gehörenden Gütefunktion desselben Tests. Wir erkennen demnach für $\theta \neq \theta_0$ die Wahrheit, also das tatsächliche Vorliegen von H_0 oder H_1, bei wachsendem Stichprobenumfang mit zunehmender Sicherheit. Für $\theta = \theta_0$ besitzt diese Wahrscheinlichkeit natürlich konstant den durch die Festlegung des Signifikanzniveaus α fixierten Wert $1-\alpha$.
Besitzt der Parameter θ, über welchen eine Hypothese mittels eines Signifikanztests zu überprüfen ist, einen (zumindest theoretisch) stetigen Wertebereich Ω, wie dies im Allgemeinen der Fall ist, so ist die Wahrscheinlichkeit dafür, dass dieser Parameter einen ganz bestimmten Wert aus Ω annimmt, (praktisch) gleich null. Wenn der Parameter θ aber gar nicht exakt den Wert θ_0 aufweisen kann, dann bedeutet die mit wachsenden Stichprobenumfängen zunehmende Sicherheit für eine richtige Entscheidung für zweiseitige Fragestellungen gleichzeitig, dass die Wahrscheinlichkeit für eine Entscheidung auf H_1 gegen eins konvergiert:

$$\lim_{n\to\infty} \Pr(T \in E_1) = 1$$

Ein signifikantes Ergebnis ist bei einer Testgröße mit einem stetigen Wertebereich bei dieser Fragestellung mit zunehmendem Stichprobenumfang also garantiert und zwar einfach deshalb, weil dann die Nullhypothese einer zweiseitigen Fragestellung gar nicht wahr sein kann. Die einzige Erklärung für eine Entscheidung zu Gunsten der Nullhypothese ist dann tatsächlich das Auftreten eines β-Fehlers wegen eines für eine richtige Entscheidung zu gering gewählten Stichprobenumfanges. Dies ist jener Umstand, der von Anwendern der Signifikanztests als Schwäche der Konzeption empfunden wird. Tatsächlich jedoch unterstreicht dies die Qualität der Signifikanztestkonzeption, weil das lediglich bedeutet, dass mit zunehmendem Stichprobenumfang, also mit Erhöhung des Aufwandes in finanzieller und zeitlicher Hinsicht, die Wahrheit, und weicht diese auch noch so gering von der in der Nullhypothese formulierten Behauptung ab, mit immer größerer Wahrscheinlichkeit ans Licht kommt.
Von einem in seinem Fachbereich kundigen Experimentator muss auf Basis seines Expertenwissens gefordert werden können, dass er seinen Forschungshypothesen häufiger,

als dies derzeit der Fall ist, eine Richtung gibt. Die große Zahl ungerichteter, zweiseitiger Fragestellungen, die in der empirischen Forschung überprüft werden, ist tatsächlich nichts als ein Ergebnis des forschungshypothesenfreien Testens. Eine zweiseitige Hypothese soll aber nur dann formuliert werden, wenn auch tatsächlich eine Veränderung des interessierenden Parameters in beliebiger Richtung überprüft werden soll. So darf sich – in einem Beispiel aus der statistischen Qualitätskontrolle – nach einer Reparatur an einer Abfüllmaschine der Mittelwert eines Füllgewichts im Vergleich zum Normgewicht weder verringert noch erhöht haben. Um dies zu testen, ist offenbar eine zweiseitige Fragestellung angebracht. Einem in seinem Bereich kundigen Forscher muss aber abverlangt werden können, z.B. in seiner zu prüfenden Forschungshypothese über den Zusammenhang zweier Merkmale die Richtung der Korrelation als Einshypothese vorzugeben und nicht mangels einer diesen Zusammenhang stützenden Theorie zweiseitig testen zu müssen.

Ein statistisch signifikantes Ergebnis ist allerdings nicht automatisch auch praktisch bedeutsam, wie dies von Anwendern immer wieder irrtümlich angenommen wird. Diese Fehleinschätzung kommt vereinzelt sogar dadurch zum Ausdruck, dass die beiden Begriffe als Synonyme verwendet werden. Wenn man aber etwa Hypothesen über den Mittelwert μ eines Merkmals x z.B. in der Art H_0: $\mu = \mu_0$ und H_1: $\mu \neq \mu_0$ formuliert, so überprüft man eben, ob der Mittelwert von x sich von μ_0 unterscheidet. Man testet auf diese Weise natürlich nicht gleichzeitig, ob dieser Mittelwert auch noch praktisch bedeutsam von μ_0 abweicht. „Die Entwicklung einer Wissenschaft ausschließlich von signifikanten Ergebnissen abhängig zu machen“ bedeutet demzufolge, „daß Theorieentwicklungen weiter verfolgt werden, die auf minimalen, wenngleich statistisch signifikanten Effekten beruhen, deren Erklärungswert für reale Sachverhalte eigentlich zu vernachlässigen ist“ (Bortz, Döring, 1995, S.28).

Tatsächlich ist es aber – wie deutlich gemacht werden soll – nicht die statistische Methode des Signifikanztestens, die hier eine Schwäche aufweist. **Es sind vielmehr die von den Anwendern dieser Methoden formulierten Hypothesen, die häufig unbrauchbar sind, weil sie offenbar nicht das überprüfen, was der Anwender überprüfen möchte.** Wenn man nicht nur einfach überprüfen möchte, ob der Mittelwert von x von μ_0 abweicht, sondern ob er dies in einem praktisch bedeutsamen Ausmaß tut, dann muss die Einshypothese anders formuliert werden, als dies herkömmlich der Fall ist. Und zwar so, dass sie nicht alle Parameterwerte enthält, die nicht μ_0 sind, sondern lediglich jene, die in der Einschätzung des Anwenders praktisch bedeutsam sind. Und dies führt uns von der Konzeption des Signifikanz- zu jener des Relevanztests.

In der Literatur wurden verschiedene Lösungsansätze für die Relevanzproblematik vorgeschlagen:

Mit dem Argument der besseren Einschätzbarkeit der Untersuchungsergebnisse durch den Konsumenten schlagen andere Autoren vor, als Ergebnisse der Analyse von Stichprobendaten standardmäßig lediglich die betreffenden Konfidenzintervalle anzugeben (vgl. etwa: Brandstätter, 1999). Tatsächlich besteht keinerlei Notwendigkeit, sich in der empirischen Forschung in jedem Fall einer Art von Signifikanztestritual zu unterwerfen, wenn die eigentliche Fragestellung nur eine Auskunft über die Größenordnung eines interessierenden Parameters verlangt und nicht auf eine Entscheidung zwischen zwei konkurrierenden Hypothesen hinausläuft. Oft jedoch besteht das Wesen des wissenschaft-

lichen Fortschritts jedoch in der fundierten Überprüfung einer neuen Theorie. In diesen Fällen darf es m.E. weder den Anwendern der statistischen Methoden noch den Konsumenten der Forschungsergebnisse überlassen sein, aus den Ergebnissen von Bereichschätzungen ihre persönlichen Schlüsse hinsichtlich der Hypothesen zu ziehen. Es gilt unter allen Umständen zu vermeiden, dass bei gleichen Forschungsergebnissen der eine Anwender so, der andere anders entscheiden kann. Die Entscheidungen dürfen keine „Geschmacksache" werden. Deshalb ist auch die einheitliche Handlungslogik der Signifikanztests mit vorab klar formulierten Hypothesen und der genauen Festlegung jener Kriterien, die für eine Abkehr von H_0 zu Gunsten von H_1 erfüllt sein müssen, im Popperschen Sinne ein wesentliches Merkmal einer qualitativ hochstehenden Anwendung von Signifikanztests bei der Suche nach neuen Erkenntnissen.

Ein deshalb m.E. problemgerechterer Ansatz, die praktische Bedeutsamkeit der Teststatistik bei den statistischen Signifikanztests mit zu berücksichtigen, ist der Vorschlag, sie bereits bei der Formulierung der Hypothesen in den Test einzubauen. Eine Möglichkeit dazu wäre etwa, die Mindest-„Effektgröße", also jenen Abstand von dem in einer herkömmlichen Nullhypothese festgelegten Parameterwert θ_0 zu bestimmen, der die Grenze θ_1 zur praktischen Bedeutsamkeit markiert. Diese Grenze wird zur Einshypothese eines Hypothesentests, der der Wert θ_0 als Nullhypothese gegenübergestellt wird (vgl. etwa: Bortz, Döring, 1995, S.566). Beim einem einseitigen Korrelationstest würden in diesem Ansatz etwa – bei Einschätzung eines Korrelationskoeffizienten ρ von der Größenordnung $\rho \geq 0{,}4$ als praktisch bedeutsam – die Hypothesen H_0: $\rho = 0$ (oder $\rho \leq 0$) und H_1: $\rho = 0{,}4$ (oder $\rho \geq 0{,}4$) aufgestellt werden. Die Bestimmung der praktisch bedeutsamen Werte von ρ ist natürlich in gewissem Sinne willkürlich. Sie kann aber doch daraus abgeleitet werden, dass ein Erklärungswert des linearen Zusammenhangs von $\rho^2 \cdot 100\,\%$ hinsichtlich der Varianz als ausreichend empfunden wird. In dem von uns gewählten Beispiel wäre dies (bei $\rho = 0{,}4$) ein Erklärungswert von 16 %. Eine Hilfe kann für die Festlegung dieser Schwellen leistet jene Literatur, die sich mit Effektgrößen für verschiedene Tests als Ergänzung zum Signifikanztesten beschäftigt (vgl. hierzu etwa: Cohen, 1969). Aus den dort angeführten Klassifikationen lassen sich Bedeutsamkeitsschwellen für verschiedene Teststatistiken ableiten. Jedenfalls ist in diesem Ansatz Expertenwissen aus dem der Fragestellung zugeordneten Bereich bei der Bestimmung dieser Schwellen einzubringen.

Die Anwendung dieses (Neyman-Pearsonschen) Ansatzes hat in fast allen in der Praxis relevanten Fragestellungen einen entscheidenden Nachteil: Er wird der Realität nicht gerecht. Wenn die beiden Hypothesen keine Zerlegung des zulässigen Parameterraums bilden, dann gibt es eine von null verschiedene Wahrscheinlichkeit dafür, dass keine der beiden Hypothesen zutrifft. Im Falle der Korrelation in obigem Beispiel ist die Summe der Wahrscheinlichkeiten für das Zutreffen von H_0 oder H_1 bei den zusammengesetzten Hypothesen $\rho \leq 0$ und $\rho \geq 0{,}4$ ungleich eins. Bei zwei einfachen Hypothesen wie $\rho = 0$ und $\rho = 0{,}4$ ist sie bei stetigen Teststatistiken sogar gleich null, die Entscheidung für eine der beiden Hypothesen ist dann sicher falsch, denn der Wert des gesuchten Parameters entspricht mit Sicherheit nicht jenem aus der Null- und auch nicht jenem aus der Einshypothese! Wir riskieren nach Festlegung der bedingten Wahrscheinlichkeiten α bzw. β für Fehlentscheidungen bei Gültigkeit einer der beiden Hypothesen mit diesem Ansatz einen weiteren Fehler, der nicht kontrolliert wird. Das ist der Fehler γ, dass wir

uns für eine der aufgestellten Hypothesen entscheiden, obwohl keine der beiden richtig ist.
Auf der Suche nach wissenschaftlichen Erkenntnissen ist eine solche Vorgangsweise völlig inakzeptabel. Es besteht nämlich keinerlei begründbarer Anlass, auf Basis eines – wie auch immer gearteten – Testresultats zu behaupten, dass eine der beiden Hypothesen zutrifft. Die eine konkrete Hypothese bringt nach der verwendeten Logik das Testresultat nur mit größerer Wahrscheinlichkeit hervor als die andere.
Die Grundidee dieser Vorgangsweise wird in der empirischen Forschung verwendet, um jenen Mindeststichprobenumfang berechnen zu können, der für herkömmliche Signifikanztests nach Festlegung von θ_0 jenen Mindeststichprobenumfang berechenbar macht, der einen bestimmten bedeutsamen „Effekt", dokumentiert durch die Festlegung des praktisch bedeutsamen Parameterwertes θ_1, mit einer vorab festgelegten Wahrscheinlichkeit $1-\beta$ entdeckt. Im Übrigen jedoch ist diese Vorgangsweise eine herkömmliche Anwendung der Signifikanztestkonzeption. Denn getestet wird selbstverständlich nach wie vor, ob eine Nullhypothese z.B. der Form $\theta \leq \theta_0$ beibehalten werden kann oder nicht. Ein signifikantes Testergebnis ist dann noch lange nicht praktisch bedeutsam. Die Signifikanzschwelle liegt dann für $1-\beta > 0{,}5$ zwischen θ_0 und θ_1, so dass signifikante Testergebnisse auftreten können, die bedauerlicherweise unter der bei der Bestimmung des notwendigen Stichprobenumfangs festgelegten Bedeutsamkeitsschwelle liegen und somit die Bedeutsamkeitsproblematik erneut auftritt.

4.2 Die Relevanztestkonzeption

4.2.1 Die Formulierung von Relevanzhypothesen

Es steht fest, dass die Hypothesen der Signifikanztests von den Anwendern häufig in der Hinsicht unbrauchbar formuliert werden bzw. die in den Statistikprogrammpaketen standardmäßig vorgegebenen Hypothesen oft ungeeignet sind, dass nicht das überprüft wird, was der Anwender eigentlich überprüfen möchte. Dies ist natürlich nicht der Signifikanztestkonzeption anzulasten. Diese funktioniert bei korrekter Anwendung in der im Abschnitt 1.2 beschriebenen Art und Weise. Der folgende Ansatz zur Lösung der Relevanzproblematik beim statistischen Signifikanztesten basiert auf einer Idee von Hodges und Lehmann (1954) und ermöglicht die Miteinbeziehung der praktischen Bedeutsamkeit durch die Übersetzung der Forschungshypothese in eine kontextbezogene statistische Hypothese, die ausschließlich den Bereich der als praktisch bedeutsam eingestuften Parameter umfasst.
Gegenüber den herkömmlichen Signifikanztests ergeben sich aus dieser Idee für Relevanztests modifizierte Hypothesenformulierungen:
Man legt für einseitige Tests mit τ_0 und für zweiseitige Tests mit τ_1 und τ_2 die Grenzen zwischen den praktisch bedeutsamen und den unbedeutenden Parameterwerten so fest, dass folgende Hypothesen entstehen:
Einseitige Fragestellungen sind

$$H_0: \theta \leq \tau_0 \quad \text{und} \quad H_1: \theta > \tau_0$$

bzw.

$$H_0: \theta \geq \tau_0 \quad \text{und} \quad H_1: \theta < \tau_0 ,$$

wobei die in den Nullhypothesen angegebenen Parameterbereiche nun alle praktisch bedeutungslosen Parameterwerte umfassen.
Bei zweiseitigen Fragestellungen gilt nun:

$H_0: \tau_1 \le \theta \le \tau_2$ und $H_1: \theta < \tau_1 \vee \theta > \tau_2$

(mit $\tau_1 \le \tau_2$) und auch hierbei umfasst H_0 wieder die praktisch irrelevanten Parameterwerte.
Wir lehnen demnach mit der in Abschnitt 1.2 beschriebenen Handlungslogik H_0 wegen der nunmehr gewählten Hypothesenformulierung nur mehr dann ab, wenn sich statistisch signifikant eine praktisch bedeutsame Abweichung von H_0 feststellen lässt! Nur wenn bei einseitigen Tests die Bedeutsamkeitsschwelle τ_0 und die Grenze θ_0 beim herkömmlichen Signifikanztest bzw. bei zweiseitigen Tests die Schwellen τ_1, τ_2 und der Parameterwert θ_0 identisch sind, wenn also die Parameterwerte in der herkömmlichen Signifikanztestkonzeption in einem bestimmten Kontext den Bedeutsamkeitsschwellen entsprechen, gehen die Relevanztests in die herkömmlichen Signifikanztests über.
Für die „parameterorientierte“ Vorgangsweise muss die Zufallsstichproben-Verteilung der Teststatistik unter einer nun möglicherweise anderen Nullhypothese als beim herkömmlichen Signifikanztest bekannt sein. Damit lässt sich aber wieder analog zur Vorgangsweise in Abschnitt 1.2 jener Wertebereich E_1 für die Teststatistik T festlegen, der diejenigen Realisationen von T enthält, die als starkes Indiz gegen die Nullhypothese gewertet werden. Wird ein solches T realisiert, dann haben wir ein signifikantes und diesmal auch praktisch relevantes Testergebnis gefunden.
Bei der alternativen, schätzerbasierten Vorgangsweise entscheidet man sich bei zweiseitigen Fragestellungen auf dem Signifikanzniveau α für die Beibehaltung von H_0, wenn sich der Parameterbereich aus der Nullhypothese $[\tau_1;\tau_2]$ und jener des $(1-\alpha)$-Konfidenzintervalls $[\theta_u;\theta_o]$ überlappen: $\{\theta \,|\, \tau_1 \le \theta \le \tau_2\} \cap \{\theta \,|\, \theta_u \le \theta \le \theta_o\} \neq \varnothing$. Ist der Durchschnitt der beiden Mengen jedoch die leere Menge $\varnothing$, dann wird das Ergebnis der Teststatistik T als starkes Indiz gegen die Nullhypothese gewertet und man hat damit ein signifikantes und auch gleichzeitig relevantes Ergebnis vorliegen.
Bei einseitigen Tests der Art

$H_0: \theta \le \tau_0$ und $H_1: \theta > \tau_0$

berechnet man lediglich die untere Vertrauensschranke θ_u zur Sicherheit $1-\alpha$ und bleibt bei H_0, wenn $\tau_0 \in [\theta_u;\infty]$ ist.

Bei einseitigen Tests der Art

$H_0: \theta \ge \tau_0$ und $H_1: \theta < \tau_0$

berechnet man analog dazu natürlich die obere Vertrauensschranke θ_o zur Sicherheit $1-\alpha$ und entscheidet für die Beibehaltung von H_0, wenn gilt: $\tau_0 \in [-\infty;\theta_o]$.

Zur rezeptartigen Ausformulierung der konkreten Teststrategien in Hinblick auf die ganze Vielfalt der statistischen Fragestellungen sind jeweils die Grenzen zu den praktisch bedeutsamen Effektgrößen festzulegen. Dies muss in Form von Konventionen in der jeweiligen Substanzwissenschaft vorgenommen werden, um diese Entscheidung – in aller Regel – nicht dem einzelnen Anwender zu überlassen. Dazu wird es erst notwendig sein, dort, wo dies nicht schon längst der Fall ist, zuerst Erfahrungen mit der Relevanztestphilosophie zu sammeln, ehe man mit diesen Bedeutsamkeitsschwellen umzugehen lernt. Eine wertvolle Starthilfe kann dabei die bereits erwähnte Literatur leisten, die sich

mit Effektgrößen für verschiedene Tests beschäftigt. Auf jeden Fall ist hier Expertenwissen aus dem jeweiligen Anwendungsgebiet in ganz anderem Ausmaß als bei den herkömmlichen Signifikanztests einzubringen, um die kontextbezogen geeigneten Relevanzschranken bestimmen zu können.
In den folgenden Abschnitten werden Relevanzteststrategien für einige statistische Fragestellungen vorgestellt.

4.2.2 Einstichproben-Mittelwerts-Relevanztests

Auch bei solchen Fragestellungen ist allzu oft die gängige Signifikanzteststrategie unbefriedigend: So reicht es doch wohl in der Tat nicht, dass z.B. eine Gruppe von Lernenden, die mit einer neuen Methode unterrichtet wurde, im Schnitt signifikant bessere Resultate als die Kontrollgruppe liefert, um den Aufwand, der mit der Einführung einer neuen Methode entstehen würde, zu rechtfertigen. Wenn sich der durchschnittliche Lernerfolg im Vergleich zur herkömmlichen Methode nur geringfügig unterscheidet, ist die praktische Bedeutsamkeit des Testergebnisses fraglich.
Im Rahmen der Qualitätskontrolle wird bei Anwendung der herkömmlichen Signifikanztestlogik bei genügend großen Stichprobenumfängen eine Nullhypothese über einen Mittelwert (von Schraubenlängen beispielsweise) auch dann verworfen, wenn die aufgetretene Abweichung von der Norm aufgrund ihrer Geringfügigkeit gar nicht korrigiert werden kann (oder zumindest: nicht korrigiert werden braucht). Auch in der medizinischen Forschung kann sich die Einführung eines neuen Medikamentes für die betroffenen Patienten trotz der Erhöhung des Anteils an geheilten Patienten nicht lohnen, wenn der Heilerfolg gegenüber dem herkömmlichen Medikament mit zusätzlichen Nebenwirkungen „erkauft“ werden muss und das neue Medikament beim Test nur geringfügig besser als das alte abschneidet. In solchen Fällen ist die Einbeziehung der praktischen Bedeutsamkeit schlichtweg notwendig.
Das Relevanzkonzept wirkt sich bei solchen Mittelwertstests in einseitigen Fragestellungen nur durch die notwendige Bestimmung einer Relevanzschranke μ_0 auf die Formulierung der Hypothesen aus, damit die Nullhypothese $\mu \le \mu_0$ bzw. $\mu \ge \mu_0$ alle für diese Fragestellung praktisch bedeutungslosen Mittelwerte umfasst. In den asymptotischen Ablehnregionen

$$E_1 = \left\{ \bar{x} \,\middle|\, \bar{x} > \mu_0 + z_{1-\alpha} \cdot \frac{s}{\sqrt{n}} \right\} \quad \text{für } H_0\text{: } \mu \le \mu_0$$

bzw.

$$E_1 = \left\{ \bar{x} \,\middle|\, \bar{x} < \mu_0 - z_{1-\alpha} \cdot \frac{s}{\sqrt{n}} \right\} \quad \text{für } H_0\text{: } \mu \ge \mu_0$$

($\bar{x}$ ... der Stichprobenmittelwert des interessierenden Merkmals x, s ... die Stichprobenstandardabweichung von x) ist μ_0 nun als Bedeutsamkeitsschwelle möglicherweise anders festzulegen als beim herkömmlichen Signifikanztest.
Bei zweiseitigen Fragestellungen sind zwei Grenzen μ_1 und μ_2 für H_0: $\mu_1 \le \mu \le \mu_2$ so festzulegen, dass H_0 wiederum alle als praktisch bedeutungslos eingestuften Parameterwerte umfasst. Die Bestimmung des Ablehnungsbereiches E_1 ist mit dieser Hypothesen-

wahl jedoch nicht so einfach wie beim herkömmlichen zweiseitigen Test von H_0: $\mu = \mu_0$ gegen H_1: $\mu \neq \mu_0$.
Der asymptotische Ablehnbereich E_1 ist nun folgendermaßen definiert:

$$E_1 = \left\{ \bar{x} \middle| \bar{x} > \mu_o \vee \bar{x} < \mu_u \right\}$$

mit

$$\mu_o = \mu_2 + z_{1-\gamma} \cdot \frac{s}{\sqrt{n}}$$

und

$$\mu_u = \mu_1 - z_{1-\gamma} \cdot \frac{s}{\sqrt{n}} \ ,$$

wobei das Fraktil $z_{1-\gamma}$ der Standardnormalverteilung so zu wählen ist, dass

$$Pr(\bar{x} \in E_1 \middle| \mu_1 \leq \mu \leq \mu_2) \leq \alpha$$

mit $Pr(\bar{x} \in E_1) = \alpha$ nur für $\mu = \mu_1$ und $\mu = \mu_2$. Deshalb gilt für die Wahrscheinlichkeit γ, dass sie abhängig vom Abstand zwischen μ_1 und μ_2 zwischen $\alpha/2$ und α liegt und umso mehr zu α geht, umso größer der Abstand und auch je größer der Stichprobenumfang wird. Für $\mu_1 = \mu_2$ gilt: $\gamma = \alpha/2$. Für $\mu_1 - \mu_2 \to \infty$ gilt: $\gamma = \alpha$.
Die für den jeweiligen Test zutreffenden Grenzen μ_u und μ_o sind unter diesen Bedingungen lediglich iterativ bestimmbar.
Die alternative, schätzerbasierte Vorgangsweise zum Testen von Hypothesen lässt sich jedoch bei beiden Fragestellungen sehr einfach anwenden. Im Falle einseitiger Hypothesen sind ausgehend vom Stichprobenergebnis $\bar{x}$ die untere bzw. obere Vertrauensschranke μ_u bzw. μ_o zur Sicherheit $1-\alpha$ zu bestimmen und damit auf Beibehaltung oder Ablehnung von H_0: $\mu \leq \mu_0$ bzw. $\mu \geq \mu_0$ zu schließen. Diese Vertrauensschranken sind näherungsweise bestimmt durch

$$\begin{matrix} \mu_u \\ \mu_o \end{matrix} = \bar{x} \mp z_{1-\alpha} \cdot \frac{s}{\sqrt{n}}$$

Gilt im ersteren Fall einseitiger Fragestellungen (d.i. H_0: $\mu \leq \mu_0$) $\mu_0 \notin [\mu_u; \infty]$ und im zweiteren (d.i. H_0: $\mu \geq \mu_0$) $\mu_0 \notin [-\infty; \mu_o]$, dann wird H_0 auf dem Niveau α zu Gunsten der Einshypothese H_1: $\mu > \mu_0$ bzw. $\mu < \mu_0$ aufgegeben.
Bei einem zweiseitigen Test mit H_0: $\mu_1 \leq \mu \leq \mu_2$ bestimmt man nun einfach das $(1-\alpha)$-Konfidenzintervall $[\mu_u;\mu_o]$ für den Parameter μ und entscheidet auf dem Niveau α für die Beibehaltung von H_0, wenn die Mengen $\left\{ \mu \middle| \mu_1 \leq \mu \leq \mu_2 \right\}$ und $\left\{ \mu \middle| \mu_u \leq \mu \leq \mu_o \right\}$ mit

$$\begin{matrix} \mu_o \\ \mu_u \end{matrix} = \bar{x} \pm z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

nicht elementfremd sind. Ansonsten geht man auf H_1: $\mu < \mu_1 \vee \mu \leq \mu_2$ über.

4.2.3 Zweistichproben-Mittelwerts-Relevanztests

Will man die Mittelwerte μ_A und μ_B aus zwei Grundgesamtheiten A und B miteinander vergleichen, so ist oft nicht jede auch noch so geringe, von null verschiedene Differenz von Bedeutung. Für den Test auf relevanten Mittelwertsunterschied legt man bei einsei-

tiger Fragestellung eine Grenzdifferenz $\delta_0 = \mu_A - \mu_B$ (mit $\mu_A > \mu_B$) und bei zweiseitiger Fragestellung zwei Grenzdifferenzen δ_1 und δ_2 (mit $\delta_1 < \delta_2$) als Relevanzgrenzen fest und formuliert damit die Hypothesen für die Mittelwertsdifferenz δ in den Populationen A und B folgendermaßen:

$H_0: \delta \le \delta_0$ und $H_1: \delta > \delta_0$

bei einseitiger und

$H_0: \delta_1 \le \delta \le \delta_2$ und $H_1: \delta < \delta_1 \vee \delta > \delta_2$

bei zweiseitiger Fragestellung.

Mit den Stichproben mit den Umfängen n_A und n_B aus den betreffenden Grundgesamtheiten berechnet man als Testgröße die Stichprobenmittelwertsdifferenz $d = \bar{x}_A - \bar{x}_B$. Im einseitigen Fall ist für große Stichprobenumfänge H_0 auf dem Niveau α bei schätzerbasierter Relevanzteststrategie zu verwerfen, wenn gilt:

$$\delta_0 < d - z_{1-\alpha} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \equiv \delta_u$$

(s_A^2, s_B^2 ... die Stichprobenvarianzen von x in den beiden Stichproben aus A und B). Die Differenz rechts des Kleinerzeichens stellt dabei die untere Vertauensschranke δ_u für die Differenz δ zur Sicherheit $1-\alpha$ dar.

Im zweiseitigen Fall wird für große Stichprobenumfänge die Nullhypothese auf dem Niveau α verworfen, wenn das $(1-\alpha)$-Konfidenzintervall $[\delta_u;\delta_o]$ mit

$$\begin{matrix}\delta_o \\ \delta_u\end{matrix} = d \pm z_{1-\alpha} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

kein Element mit dem in der Nullhypothese formulierten Bereich der praktisch irrelevanten Differenzen δ gemeinsam hat. Eine solche Stichprobenmittelwertsdifferenz d wird dann als starkes Indiz gegen H_0 gewertet und man hat ein statistisch signifikantes Ergebnis erhalten, das auch noch praktisch bedeutsam auf dem Niveau α ist.

Für $\delta_0 = 0$ bei einseitigen bzw. $\delta_1 = \delta_2 = 0$ bei zweiseitigen Fragestellungen geht die Relevanzteststrategie in die herkömmliche Signifikanzteststrategie bei schätzerbasierter Vorgangsweise des Testens über.

4.2.4 Korrelations-Relevanztests

In der Praxis werden Korrelationen ρ zwischen zwei normalverteilten Merkmalen fest ausschließlich daraufhin geprüft, ob sie von null verschieden sind. Die Hypothesen lauten deswegen zumeist:

$H_0: \rho \le 0$ und $H_1: \rho > 0$

oder

$H_0: \rho \ge 0$ und $H_1: \rho < 0$

bei einseitiger Fragenformulierung und

$H_0: \rho = 0$ und $H_1: \rho \ne 0$

bei zweiseitiger. Es ist bei diesem Test üblich, als Testgröße

$$t = \sqrt{(n-2) \cdot \frac{r^2}{1-r^2}} \qquad (2)$$

zu verwenden. Die für die parameterorientierte Vorgangsweise bei diesen Tests nötige Stichprobenverteilung von t ist eine t-Verteilung mit n−2 Freiheitsgraden. t ist somit für große Stichprobenumfänge n näherungsweise standardnormalverteilt. Und beim ersten der drei möglichen Tests hat man genau dann ein genügend großes Indiz gegen H_0 gefunden, um sie auf dem Niveau α abzulehnen, wenn der Stichprobenkorrelationskoeffizient r so groß ist, dass $t > z_{1-\alpha}$ wird ($z_{1-\alpha}$ ... das $(1-\alpha)$-Fraktil der Standardnormalverteilung). Bei der zweiten der oben angeführten Fragestellungen ist diese Entscheidung zu treffen, wenn r so klein ist (ein großer negativer Wert), dass $t < -z_{1-\alpha}$ wird. Schließlich gilt für den zweiseitigen Test der Korrelation, dass H_0 auf diesem Niveau abgelehnt wird, wenn der Betrag von r so groß ist, dass gilt: $|t| > z_{1-\alpha/2}$.

Doch auch hier gilt natürlich, dass eine sehr gering von null abweichende Korrelation möglicherweise praktisch völlig irrelevant ist. Aus (2) geht jedoch hervor, dass jede auch noch so geringe Stichprobenkorrelation r statistisch zu einem signifikanten Testergebnis t führt, wenn nur der Stichprobenumfang ausreicht. Will man aber so geringe Abweichungen von der Korrelation null gar nicht feststellen, so ist die Einshypothese, die man überprüfen möchte, mit den oben angegebenen Standardhypothesen nicht korrekt formuliert worden.

Unter der Normalverteilungsvoraussetzung sind für Relevanztests bei einseitigen Fragestellungen folgende mögliche Hypothesenpaare zu formulieren:

$$H_0: \rho \le \rho_0 \quad \text{und} \quad H_1: \rho > \rho_0$$

bzw.

$$H_0: \rho \ge \rho_0 \quad \text{und} \quad H_1: \rho < \rho_0$$

So könnte die Einshypothese eines als Relevanztest konzipierten einseitigen Tests des Korrelationskoeffizienten etwa lauten, dass die Korrelation ρ zwischen zwei Merkmalen größer als $\rho_0 = 0{,}4$ ist ($H_1: \rho > 0{,}4$). Die Nullhypothese lautet dann: $\rho \le 0{,}4$. Hier legt z.B. $\theta_0 = \rho_0 = 0{,}4$ die Relevanzgrenze für den Parameter ρ fest.

Bei zweiseitiger Fragestellung ergeben sich die Hypothesen

$$H_0: \rho_1 \le \rho \le \rho_2 \quad \text{und} \quad H_1: \rho < \rho_1 \vee \rho > \rho_2$$

mit $\rho_1 < \rho_2$.

Bei $\rho_0 = 0$ im einseitigen und $\rho_1 = \rho_2 = 0$ im zweiseitigen Fall ist der Relevanztest der herkömmliche Signifikanztest. Es ist tatsächlich vorstellbar, dass jeder auch noch so geringer Zusammenhang zwischen zwei Merkmalen in einem bestimmten Kontext von praktischer Bedeutsamkeit ist. Bei $\rho_0 \neq 0$ bedient man sich bei parameterorientierter Vorgangsweise folgender Teststrategie:

Für normalverteilte Merkmale gilt bei $\rho_0 \neq 0$ und nicht zu groß, dass die Testgröße

$$z = \frac{\sqrt{n-3}}{2} \cdot \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_0}{1-\rho_0} \right) \tag{3}$$

für große n näherungsweise standardnormalverteilt ist („Fisher-Transformation“). Die Stichprobenkorrelation r ist demnach bei einer einseitigen Fragestellung vom Typ H_0: $\rho \le \rho_0$ und H_1: $\rho > \rho_0$ auf dem Niveau α signifikant größer als die Bedeutsamkeitsschwelle ρ_0, wenn sie so viel größer als ρ_0 ist, dass $z > z_{1-\alpha}$ ist.

Im zweiten Fall mit H_0: $\rho \le 0{,}4$ wird analog diese Hypothese erst verworfen, wenn r so viel kleiner als ρ_0 ist, dass $z < -z_{1-\alpha}$ ist.

Das Problem der parameterorientierten Vorgangsweise beim zweiseitigen Korrelationstest entspricht dem auch beim Mittelwertstest aufgetretenen. Eine Lösung ist wieder, die asymptotisch gleich gute schätzerbasierte Vorgangsweise zu wählen: Aus der standardnormalverteilten Testgröße z nach (3) erhält man, indem man z durch das α-Fraktil auf der rechten Seite durch $z_{1-\alpha/2}$ bzw. $-z_{1-\alpha/2}$ ersetzt, die Grenzen ρ_u und ρ_o des $(1-\alpha)$-Konfidenzintervalls für ρ:

$$\begin{matrix}\rho_o \\ \rho_u\end{matrix} = \frac{\exp\left(\ln\frac{1+r}{1-r} \pm \frac{2\cdot z_{1-\alpha/2}}{\sqrt{n-3}}\right)-1}{\exp\left(\ln\frac{1+r}{1-r} \pm \frac{2\cdot z_{1-\alpha/2}}{\sqrt{n-3}}\right)+1} \qquad (4)$$

Gilt dann, dass sich die beiden Bereiche $[\rho_1;\rho_2]$ und $[\rho_u;\rho_o]$ nicht überschneiden, dann hat man mit r ein starkes Indiz gegen H_0 gefunden. H_0 wird somit auf dem Niveau α abgelehnt und H_1 angenommen. Bis auf weiteres gilt dann, dass eine praktisch bedeutsame Korrelation existiert.

Diese schätzerbasierte Vorgangsweise darf natürlich auch für einseitige Tests gewählt werden. Man bestimmt dazu nur die passenden einseitigen Vertrauensschranken für ρ aus (4), indem man dort $z_{1-\alpha/2}$ durch $z_{1-\alpha}$ ersetzt.

Liegen zwei ordinale Merkmale vor, dann lässt sich der Zusammenhang zwischen den beiden Merkmalen mittels des Spearmanschen Rangkorrelationskoeffizienten ρ_S beurteilen. Dazu sind die Merkmalsausprägungen beider Merkmale in Rangreihen zu transformieren und dann die Korrelation dieser Rangreihen zu bestimmen.

Für die Hypothesen eines Relevanztest auf bedeutsame Abhängigkeit zweier Rangmerkmale gelten dieselben Überlegungen wie für diese Tests zweier metrischer Merkmale. In die Nullhypothesen sind wiederum alle praktisch bedeutungslosen Werte aus dem Wertebereich von ρ_S aufzunehmen.

Für die parameterorientierte Vorgangweise wird die Stichprobenverteilung des Spearmanschen Korrelationskoeffizienten für $\rho_S \neq 0$ (vgl. für kleine Stichproben etwa: Henze, 1979) benötigt. Als einfachere Vorgangsweise bietet sich die schätzerbasierte an. Asymptotisch gelten für beide Vorgangsweisen dieselben Entscheidungsregeln wie für den Korrelationskoeffizienten ρ bei metrischen Merkmalen. So ist bei der schätzerbasierten Vorgangsweise bei genügend großen Stichprobenumfängen lediglich r in (4) durch den Stichprobenrangkorrelationskoeffizienten r_S zu ersetzen (vgl. etwa: Kraemer, 1974, S. 114).

4.2.5 Der Chiquadrattest auf relevante Abhängigkeit

Auch beim Testen auf Abhängigkeit unter nominalen Merkmalen tritt natürlich das Relevanzproblem statistisch signifikanter Ergebnisse auf. Mit wachsendem Stichprobenumfang werden auch noch so geringe Unterschiede in den bedingten Verteilungen eines Merkmals bei fester Ausprägung des anderen mit zunehmender Sicherheit erkannt und bei der Wahl zwischen den Hypothesen

$H_0: \chi^2 = 0$ und $H_1: \chi^2 > 0$

konvergiert die Wahrscheinlichkeit einer Entscheidung für H_1 gegen eins. Sind geringe Zusammenhänge zwischen zwei Merkmalen aber praktisch irrelevant, so sind die Hypothesen unbrauchbar formuliert worden.
Mit dem Assoziationsmaß V von Cramer mit

$$V = \sqrt[+]{\frac{\chi^2}{n \cdot (\min(r,s)-1)}}$$

(r,s ... die Anzahlen der Ausprägungen der beiden Merkmale) wird der mit der Kennzahl χ^2 gemessene Zusammenhang zweier nominaler Merkmale normiert: $0 \le V \le 1$. Für den herkömmlichen χ^2-Test auf Abhängigkeit lauten die Hypothesen aus der Sicht von V somit:

H_0: $V = 0$ und H_1: $V > 0$.

Für den Relevanztest des Zusammenhangs wäre eine Bedeutsamkeitsschwelle V_0 so zu festzulegen, dass bei Formulierung der Hypothesen

H_0: $V \le V_0$ und H_1: $V > V_0$

der für V in H_0 beschriebene Wertebereich alle als praktisch unbedeutend eingestuften Parameterwerte für V umfasst.
Die Schwierigkeit der beschriebenen Vorgangsweise liegt natürlich in der Bestimmung der Bedeutsamkeitsschwelle V_0. Ein Ansatz ist die in Abschnitt IV.1 angegebene Möglichkeit der Begründung zur Festlegung einer Bedeutsamkeitsschwelle ρ_0 für den Korrelationskoeffizienten ρ durch den mit dem Bestimmtheitsmaß errechneten Erklärungswert des linearen Zusammenhangs. Für 2×2-Tafeln, also Kontingenztafeln mit r,s = 2 gilt für jede beliebige Kodierung der Ausprägungen der beiden Merkmale:

$$V = |\rho|$$

(vgl. etwa: Hilbert, 1998, S.95f). Damit ließen sich bei solchen Tafeln auch für V dieselben Überlegungen anstellen. Wird dann ein Zusammenhang im Ausmaß von V = 0,4 auch für 2×2-Tafeln als gerade noch unbedeutend empfunden, so legen wir dies auch auf allgemeine r×s-Tafeln um.
Bei der einfacher handhabbaren schätzerbasierten Vorgangsweise des Testens statistischer Hypothesen legt man für diesen einseitigen Test eine approximative untere Vertrauensschranke V_u durch

$$V_u = \hat{V} - z_{1-\alpha} \cdot \sqrt{\widehat{Var}_\infty \hat{V}}$$

fest. Darin ist $\hat{V}$ das Assoziationsmaß V in der Stichprobe und $\widehat{Var}_\infty \hat{V}$ der Schätzer für die asymptotische Varianz dieses Stichprobenergebnisses mit

$$\widehat{Var}_\infty \hat{V} = \frac{1}{4 \cdot (\min(r,s)-1) \cdot V^2} \cdot \widehat{Var}_\infty (\hat{\phi}^2)$$

(vgl. etwa: Bishop et al., 1977, S.386). Der Varianzschätzer $\widehat{Var}_\infty (\hat{\phi}^2)$ für das Quadrat des Stichproben-Phikoeffizienten $\hat{\phi}^2$ einer 2×2-Tafel ist gegeben durch

$$\widehat{\mathrm{Var}}_{\infty}(\hat{\phi}^2) = \frac{1}{n} \cdot \left\{ 4 \cdot \sum_{i,j} \frac{p_{ij}^3}{p_{i+}^2 \cdot p_{+j}^2} - 3 \cdot \sum_{i} \frac{1}{p_{i+}} \cdot \left(\sum_{j} \frac{p_{ij}^2}{p_{i+} \cdot p_{+j}} \right)^2 - \right.$$

$$-3 \cdot \sum_{j} \frac{1}{p_{+j}} \cdot \left(\sum_{i} \frac{p_{ij}^2}{p_{i+} \cdot p_{+j}} \right)^2 +$$

$$\left. +2 \cdot \sum_{i,j} \left(\frac{p_{ij}}{p_{i+} \cdot p_{+j}} \cdot \left(\sum_{k} \frac{p_{kj}^2}{p_{k+} \cdot p_{+j}} \right) \cdot \left(\sum_{\ell} \frac{p_{i\ell}^2}{p_{i+} \cdot p_{+\ell}} \right) \right) \right\}$$

(alle indizierten p's sind in der Stichprobe beobachtete relative Häufigkeiten).
Gilt $V_0 \in [V_u; 1]$, dann hat man mit dem Stichprobenergebnis zu geringe Indizien gegen die Nullhypothese eines praktisch bedeutungslosen Zusammenhangs gefunden. Ist dies nicht der Fall, so wird H_0 auf dem Niveau α verworfen und statt dessen auf H_1 übergegangen. Das Stichprobenergebnis ist mithin signifikant und praktisch bedeutsam.

5. Zusammenfassung

Ausgehend von der Darstellung der klassischen Vorgangsweisen beim Signifikanztesten statistischer Hypothesen und den damit verbundenen Interpretationsmöglichkeiten statistischer Testergebnisse wurden in diesem Aufsatz verschiedene Gründe aus der Anwendung dieser Methoden der schießenden Statistik angesprochen, die insgesamt zu einer Vertrauenskrise der Anwender in die Funktionstauglichkeit dieser Methoden geführt haben. Eine der Ursachen ist die Publikationspraxis in den meisten empirischen Fachzeitschriften. Es wurde belegt, dass Ergebnisse von Signifikanztests eine wesentliche höhere Veröffentlichungschance besitzen, wenn es hauptsächlich signifikante Resultate sind. Daraus ergibt sich, dass, sofern die nichtsignifikanten Untersuchungsergebnisse nicht dokumentiert werden, diese niemals Teil der Erkenntnisse eines Faches werden und ein signifikantes Forschungsergebnis für eine bestimmte Fragestellung somit nicht in die Reihe aller Untersuchungsergebnisse zu diesem Untersuchungsgegenstand gestellt werden kann. Auf den Punkt gebracht ist die logische Konsequenz daraus, dass nichtsignifikante Tests von weiteren Forschern solange wiederholt werden bis sich (möglicherweise irrtümlich) ein signifikantes Ergebnis einstellt. Erst dieses kann Bestandteil des Wissens eines Faches werden. Nur eine Veränderung der bisherigen Publikationspraxis durch die Erkenntnis, dass auch nichtsignifikante Resultate begründeter Forschungshypothesen das Wissen über den Forschungsgegenstand bereichern und dass desgleichen auch Replikationen früherer Untersuchungen tun, kann hier eine nötige Veränderung herstellen.
Eine weitere die Praxis der Anwendung beschreibende Verhaltensweise beim Signifikanztesten ist das forschungshypothesenfreie „Alles-mit-Allem-Testen". Sein Charakteristikum ist, dass vorhandene Daten mit allen Mitteln, die die statistische Softwarekunst zur Verfügung stellt, „ausgequetscht" werden, ohne dass dies im Einzelnen eine Forschungshypothese begründet. Es zeigt sich, dass diese Vorgangsweise die Anzahl von Fehlentscheidungen bei den Testergebnissen erhöht. Die erst *nach* signifikanten Testresultaten formulierten Forschungshypothesen wurden tatsächlich niemals überprüft.

Diese Vorgangsweise, die erst durch die leichte Handhabbarkeit statistischer Programmpakete ermöglicht wurde, widerspricht der klassischen Handlungslogik beim Testen von Hypothesen, die alleine aber Garant für qualitativ hochwertige Ergebnisse bei der Entscheidung auf Beibehaltung oder Ablehnung von Hypothesen sind. Trotzdem könnten solcherarts produzierte Testresultate einen Beitrag für die empirische Forschung leisten, wenn sie als das interpretiert werden, was sie tatsächlich sind: Hinweise auf mögliche interessante Fragestellungen, die, sofern sie mit einer erklärenden Theorie unterlegt werden können, einer neuerlichen (im eigentlichen Sinn: erstmaligen) Überprüfung unterzogen werden müssen.
Eine sich direkt aus der Signifikanztestlogik ergebende praktische Ursache für die beobachtete Vertauenskrise ist ferner, dass diese Tests natürlich genau diejenigen Hypothesen testen, die vom Anwender (oder seinem Statistik-Programmpaket) vorgegeben werden. Sind diese nicht die Hypothesen, die der Anwender wirklich prüfen möchte, so sind sie für die betreffende Fragestellung offenbar falsch formuliert worden. So führt eine dem Untersuchungsgegenstand nicht angepasste Hypothesenformulierung sehr häufig dazu, dass praktisch bedeutungslose „Signifikanzen“ erzeugt werden und dieser Umstand irrtümlich als Schwäche dieser statistischen Verfahren ausgelegt wird. Tatsächlich aber ergibt sich aus diesem Problem die Notwendigkeit der Formulierung von Relevanzhypothesen. Dies entspricht einer Abkehr von „standardmäßig“ angewendeten Signifikanztests zu kontextbezogenen Relevanztests, die bei der Hypothesenformulierung das fachspezifische Expertenwissen des Anwenders umsetzen. Dies wirkt sich auf die Teststrategien in unterschiedlichen Ausmaßen aus. Den Ergebnissen der Anwendung von Relevanztests ist jedoch gemeinsam, dass statistisch signifikante Resultate automatisch auch praktisch bedeutsam sind.
Die Umsetzung all dieser Vorschläge kann zu einer neuen Praxis der Anwendung dieser statistischen Vorgangsweise führen, deren Ergebnisse jenem qualitativem Standard entsprechen, der das Vertrauen in diese bei korrekter Anwendung funktionierenden Verfahren wiederherstellt. Damit würde die Methode wieder zu dem gemacht werden, was sie sein kann: Ein wesentliches Verfahren der empirischen Forschung bei der Suche nach neuen Erkenntnissen.

Literatur

[1] Arbuthnott, J. (1710). An argument for Devine Providence, taken from the constant regularity observ’d in the births of both sexes. In: Kendall, M., Plackett, R.L. (Hrsg.) (1977). *Studies in the History of Statistics and Probability*. Volume II. Charles Griffin & Company Limited. London. 30-34.

[2] Begg, C.B., Berlin, J.A. (1988). Publication Bias: a Problem in Interpreting Medical Data. *Journal of the Royal Statistical Society*. Series A. Volume 151. Part 3. 419-463.

[3] Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*. Volume 37. 325-335.

[4] Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press. Cambridge.

[5] Bortz, J., Döring, N. (1995). *Forschungsmethoden und Evaluation*. 2. Auflage. Springer Verlag. Berlin.

[6] Brandstätter, E. (1999). Konfidenzintervalle als Alternative zu Signifikanztests. Methods of Psychological Research Online (www.pabst-publishers.de/mpr/). Volume 4. No. 2.
[7] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press. New York.
[8] Cowles, M., Davis, C. (1982). On the Origins of the .05 Level of Statistical Significance. *American Psychologist*. Volume 37. No. 5. S. 553-558.
[9] Diepgen, R. (1994). Inferenzstatistische Sprachspiele in den Humanwissenschaften: Eine kleine Fallstudie. *Stochastik in der Schule*. 14(1), S. 10-22.
[10] Henze, F.H.-H. (1979). The Exact Noncentral Distributions of Spearman's r and Other Related Correlation Coefficients. *Journal of the American Statistical Association*. Volume 74. Number 366. S. 459-465.
[11] Hilbert, A. (1998). *Zur Theorie der Korrelationsmaße*. Josef Eul Verlag. Lohmar.
[12] Hodges, J.L., Lehmann, E.L. (1954). Testing the Approximate Validity of Statistical Hypotheses. *Journal of the Royal Statistical Societ*y. Series B. Volume 16. S.261-268.
[13] Kotz, S., Johnson, N.L. (eds.) (1988). *Enzyclopedia of Statistical Sciences*. Volume 8. John Wiley & Sons.
[14] Kraemer, H.C. (1974). The Non-Null Distribution of the Spearman Rank Correlation Coefficient. *Journal of the American Statistical Association*. Volume 69. Number 345. S. 114-117.
[15] Noelle-Neumann, E. (1980). *Die Schweigespirale*. R. Riper & Co. Verlag. München.
[16] Ostmann, A., Wutke, J. (1994). *Statistische Entscheidung*. In: Herrmann, T., Tack, W.H. (Hrsg.). Methodologische Grundlagen der Psychologie. Hogrefe Verlag für Psychologie. Göttingen.
[17] Paulos, J.A. (1993). *Zahlenblind*. Wilhelm Heyne Verlag. München.
[18] Pearson, K. (1900). On the Criterion that a Given System of Derivations from the Probable in the Case of a Correlated System of Variables is Such that it Can reasonably be Supposed to Have Arisen from Random Sampling. In: Kotz, S., Johnson, N.L. (eds.) (1992). *Breakthroughs in Statistics*. Volume II. Springer-Verlag. New York. S. 11-28.
[19] Popper, K. (1989). *Logik der Forschung*. Neunte, verbesserte Auflage. Mohr. Tübingen.
[20] Quatember, A. (1997). Die Veränderung der sozial-, wirtschaftwissenschaftlichen und medizinischen Forschung durch die Verwendung statistischer Programmpakete: Bestandsaufnahme und Verbesserungsvorschläge. In: Bandilla, W., Faulbaum, F. (Hrsg.) (1997). *SoftStat '97. Advances in Statistical Software 6*. Lucius & Lucius. Stuttgart. S. 309-316.
[21] Quatember, A. (2001). *Die Quotenverfahren – Stichprobentheorie und -praxis*. Shaker Verlag. Aachen.
[22] Randow, G. von (1994). *Das Ziegenproblem*. Rowohlt. Reinbek bei Hamburg.
[23] Sahner, H. (1979). Veröffentlichte empirische Sozialforschung: Eine Kumulation von Artefakten? *Zeitschrift für Soziologie*. 8(3). 267-278.

[24] Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. Methods of Psychological Research Online (www.pabst-publishers.de/mpr/). Volume 1. No. 1. 45-68.
[25] Smart, R.G. (1964). The Importance of Negative Results in Psychological Research. *The Canadian Psychologist*. Vol. 5a. No. 4. S. 225-232.
[26] Stelzl, I. (1982). *Fehler und Fallen in der Statistik*. Verlag Hans Huber. Bern.
[27] Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance -or vice versa. *Journal of the American Statistical Association*, 54, 30-34.
[28] Stigler, S.M. (1977). Eight Centuries of Sampling Inspection: The Trial of the Pyx. *Journal of the American Statistical Association*. Volume 72. S. 439-500.
[29] Wilson, F.D., Smoke, G.L., Martin, J.D. (1973). The Replication Problem in Sociology: A Report and a Suggestion. *Sociological Inquiry*. 43 (2). S. 141-149.
[30] Witte, E. H. (1980). *Signifikanztest und statistische Inferenz*. Ferdinand Enke Verlag. Stuttgart.

Ass. Prof. Dr. Andreas Quatember
IFAS - Institut für Angewandte Statistik
Johannes Kepler Universität Linz
Altenbergerstraße 69
A-4040 LINZ
Österreich
E-Mail: andreas.quatember@jku.at
Tel.: +43 (732) 2468-8267
Fax: +43 (732) 2468-9846
Internet: www.ifas.jku.at