Das Signifikanz-Relevanz-Problem beim statistischen Testen von Hypothesen

Quatember, Andreas

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

DAS SIGNIFIKANZ-RELEVANZ-PROBLEM BEIM STATISTISCHEN TESTEN VON HYPOTHESEN

THE SIGNIFICANCE-RELEVANCE-PROBLEM IN THE CONTEXT OF STATISTICAL TESTING

ANDREAS QUATEMBER

In der empirischen Sozialforschung ist das Signifikanztesten eines der meistverwendeten statistischen Instrumente bei der Suche nach neuen Erkenntnissen. Die ohne Kontextbezug erfolgende Übersetzung der verschiedensten interessierenden Fragestellungen in immer die gleichen statistischen Hypothesen schafft das Problem, dass damit häufig signifikante Resultate erzielt werden, denen es an praktischer Relevanz mangelt. Und dies umso eher, umso genauer das jeweilige Experiment angelegt wird, umso größer also der Stichprobenumfang gewählt wird. In diesem Aufsatz wird auf einen Ausweg für dieses Signifikanz-Relevanz-Problem beim statistischen Hypothesentesten hingewiesen. Dieser besteht aus einer logischen Modifizierung der Signifikanzteststrategie, die uns vom Signifikanz- zum Relevanztest führt. Diese Strategie wird beschrieben und an Beispielen für verschiedene Fragestellungen umgesetzt.

In empirical social research the concept of significance testing is one of the most frequently used statistical instruments in search of new findings. The transformation of various substantive questions into the same statistical hypothesis without consideration of context leads to the problem that for many test results that are statistically significant there is a lack of practical relevance. And the more accurate the experiment, the greater is the problem. This paper shows a way out of this significance-relevance-problem. It consists of a theoretical modification of the strategy used in significance testing, which leads us from tests of significance to tests of relevance. This strategy is shown and applied to various statistical tests.

1 Einleitung

Sehr häufig ist es bei der Suche nach neuen Erkenntnissen in der empirischen Sozialforschung wie auch in anderen Bereichen empirischer Forschung notwendig, eine fundierte Entscheidung über das Zutreffen einer Forschungshypothese zu treffen. Beispiele dafür sind Fragestellungen wie die Nachfolgenden: Hat sich die Quantität der großelterlichen Kinderbetreuung gegenüber einer früheren Erhebung erhöht? Ist der Anteil an von Schülern einer bestimmten Altersklasse gelösten Mathematikaufgaben bei einem standardisierten EU-weiten Test in einem Land kleiner als ein von der Schulbehörde festgelegter Mindestwert und gibt es Unterschiede in den Ergebnissen der beteiligten Länder? Gibt es einen Zusammenhang zwischen der Tageslichtzufuhr am Arbeitsplatz und der Arbeitsleistung von Erwerbspersonen? Hatte die Einführung von Studiengebühren einen Einfluss auf die Prüfungsleistung der Studierenden? usf.

Werden zur Überprüfung dieser Hypothesen Daten auf Basis einer Stichprobe erhoben, dann bedient man sich des Instruments des statistischen Signifikanztests. Für die verschiedensten Fragestellungen sind seit Beginn des 20. Jahrhunderts die dafür geeigneten statistischen Tests entwickelt worden, denen ein- und dieselbe Handlungslogik zu Grunde liegt (siehe für einen kurzen geschichtlichen Abriss etwa: Quatember 2004, und zur Handlungslogik etwa: Quatember 2005: 113ff.): Ausgehend von einer Forschungshypothese werden zwei konkurrierende statistische Hypothesen über einen Parameter θ, die Null- (H_0) und die Einshypothese (H_1) aufgestellt, wobei die letztere zumeist die statistische Übersetzung der Forschungshypothese darstellt. Bei einseitigen Fragestellungen wird der interessierende Parameter nur in eine Richtung überprüft:

$$H_0: \theta \leq \theta_0 \text{ (bzw. } \theta \geq \theta_0) \quad \text{und} \quad H_1: \theta > \theta_0 \text{ (bzw. } \theta < \theta_0)$$

(θ_0 ist ein bestimmter Wert von θ). Bei zweiseitigen Fragestellungen wird gleichzeitig in beide Richtungen getestet:

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

Mit den Daten einer Zufallsstichprobe aus der betreffenden Grundgesamtheit wird im nächsten Schritt eine Teststatistik T berechnet, die hinsichtlich der Hypothesen wesentliche Informationen liefert und deren Stichprobenverteilung bei Gültigkeit der Nullhypothese bekannt ist. Für die darauf basierende Entscheidung über Beibehaltung der Nullhypothese

oder deren Verwerfung zu Gunsten der Einshypothese bezeichnet das Signifikanzniveau α die Wahrscheinlichkeit dafür, sich bei Gültigkeit von H_0 irrtümlich für H_1 zu entscheiden. β gibt die Wahrscheinlichkeit dafür an, sich bei Gültigkeit von H_1 irrtümlich für H_0 zu entscheiden. Die Gegenwahrscheinlichkeit $1-\beta$ wird als die Teststärke bezeichnet.

Mit der Stichprobenverteilung von T bei Gültigkeit von H_0 wird eine Region S, der Ablehnungsbereich von H_0, so aus dem Wertebereich von T bestimmt, dass die vorgegebene Wahrscheinlichkeit α eingehalten wird. Außerdem soll für S gleichzeitig gelten, dass die Wahrscheinlichkeit, dass T bei Gültigkeit von H_1 *nicht* in S liegt, ein Minimum ist. Gilt nun für den konkreten Wert T_0 der Teststatistik T: $T_0 \in S$, dann liegt auf diese Weise ein bei Gültigkeit der Nullhypothese so seltenes Ereignis vor, dass es als signifikant gegen sie sprechend interpretiert wird. Dies hat zur Folge, dass die Nullhypothese verworfen und die Einshypothese bis auf weiteres als gültig akzeptiert wird. Bei $T_0 \notin S$ wird H_0 beibehalten.

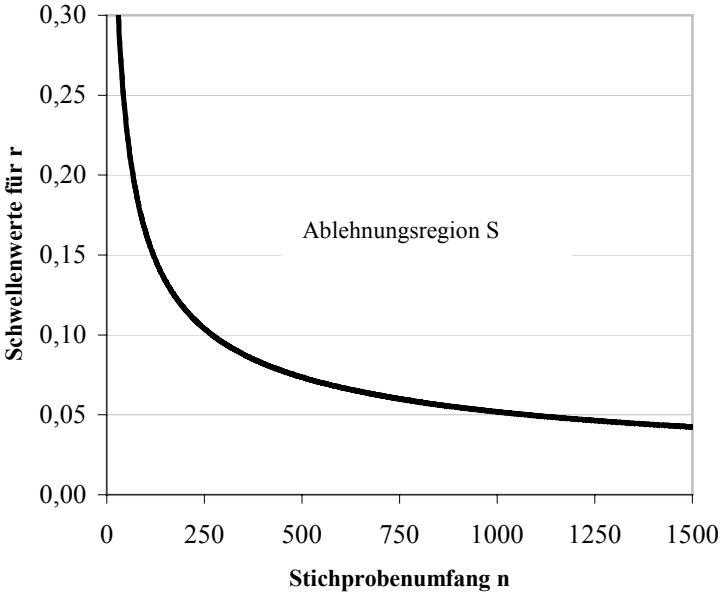
In Statistik-Programmpaketen wird alternativ der zur errechneten Teststatistik T_0 gehörende p-Wert α^* berechnet. Dieser gibt die Wahrscheinlichkeit dafür an, dass bei Gültigkeit der Nullhypothese die Teststatistik T_0 vom Parameterwert θ_0 der Nullhypothese mindestens so weit entfernt liegt wie dies tatsächlich passiert ist. Die Nullhypothese wird hierbei – völlig äquivalent zur oben beschriebenen Vorgehensweise – beibehalten, wenn gilt: $\alpha^* > \alpha$.

2 Das Bedeutsamkeitsproblem statistisch signifikanter Ergebnisse

Ein häufig von den Anwendern als Schwäche der Handlungslogik statistischer Tests interpretiertes Faktum ist der Umstand, dass mit zunehmenden Stichprobenumfängen immer geringer von der Nullhypothese abweichende und somit möglicherweise für die Praxis irrelevante Stichprobenergebnisse signifikant werden. Dies bedeutet, dass eine tatsächlich richtige Einshypothese mit zunehmendem Stichprobenumfang auch mit wachsender Wahrscheinlichkeit erkannt wird, selbst wenn der Parameter θ tatsächlich sehr nahe am Parameterwert bzw. -bereich der Nullhypothese (je nach Fragestellung) liegt.

So liefert z.B. ein einseitiger Test auf gleichsinnigen Zusammenhang zweier metrischer Merkmale (H_0: $\rho \leq 0$ und H_1: $\rho > 0$; ρ … der Korrelationskoeffizient in der Grundgesamtheit) bei einem Stichprobenumfang von n = 1.000 für einen Stichprobenkorrelationskoeffizienten r von 0,06 auf einem Signifikanzniveau α = 0,05 ein signifikantes Testergebnis, das in die Ablehnungsregion S fällt (siehe Abbildung 1).

Abbildung 1 Die Ablehnungsregion S für den Stichprobenkorrelationskoeffizienten r in Abhängigkeit vom Stichprobenumfang beim Testen der Einshypothese H_1: $\rho > 0$ auf einem Signifikanzniveau $\alpha = 0{,}05$



Die Testgröße eines solchen Tests ist bei bivariat normalverteilten Merkmalen

$$z = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

(vgl. etwa: Quatember 2005: 152ff.). Bei großen Stichprobenumfängen gilt: Ist $z \leq z_{1-\alpha}$, das $(1-\alpha)$-Fraktil der Standardnormalverteilung, dann wird die Nullhypothese beibehalten. Aber ist eine Korrelation von 0,06 praktisch relevant? Das heißt, ist es wirklich ein Informationsgewinn, wenn man daraus schließt, dass es auch in der Grundgesamtheit einen gleichsinnigen Zusammenhang zwischen den beiden betrachteten Merkmalen gibt,

wenn dieser offenbar ein sehr schwacher Zusammenhang ist? Einer konkreten Schätzung des einen Merkmals durch das andere mit Hilfe einer Regressionsgeraden würde man z.B. auf Basis eines solch geringen Zusammenhangs nicht vertrauen können.

Ein signifikanter Zusammenhang braucht also, wie Abbildung 1 belegt, bei großen Stichprobenumfängen kein starker Zusammenhang zu sein. Aus diesem Grund werden vereinzelt Maßzahlen vorgeschlagen, welche die praktische Relevanz einer Teststatistik beschreiben sollen. Diese beruhen z.B. bei Korrelationstests auf der Schätzung der durch die unabhängige Variable erklärten Varianz der abhängigen Variablen durch das Bestimmtheitsmaß. So erklärt die unabhängige Variable bei r = 0,06 nicht einmal 0,4 % der Streuung in der abhängigen Variablen, die restlichen mehr als 99,6 % bleiben unerklärt. Demzufolge sollte das Vertrauen in eine Schätzung des Wertes der abhängigen Variablen auf Basis der unabhängigen Variablen äußerst gering sein.

Wenn allerdings solche Schätzungen oder Prognosen nicht benötigt werden, dann hängt es vom jeweiligen Untersuchungsgegenstand ab, ob auch noch so geringe Abweichungen von der Nullhypothese bedeutsam sind. Man denke etwa an einen Test, der den Zusammenhang von Tageslichtzufuhr und Gemütsverfassung und damit Arbeitsleistung von Erwerbspersonen überprüft. Bei dieser Fragestellung sind wohl auch kleine Abweichungen von der Nullhypothese der Unabhängigkeit von Interesse.

Die Signifikanz-Relevanz-Problematik hat gemeinsam mit anderen Ursachen wie der gängigen Publikationspraxis, in der signifikante Ergebnisse statistischer Tests eine deutlich höhere Veröffentlichungschance besitzen als nichtsignifikante, und dem „Alles-mit-Allem-Testen“, dessen Charakteristikum es ist, dass vorhandene Daten mit allen Mitteln, die die statistische Softwarekunst zur Verfügung stellt, ohne Begründung durch interessierende Forschungshypothesen „abgetestet“ werden, zu einer veritablen Krise des Vertrauens in die Signifikanztests geführt, die sich auch in einschlägiger Literatur niederschlägt (vgl. etwa aus unterschiedlichen Bereichen: Begg & Berlin 1988; Bredenkamp 1972; Carver 1993; Kirk 1996; Quatember 1997, 2004; Sahner 1979; Smart 1964; Wilson et al. 1973). Tatsächlich jedoch handelt es sich bei der hier angesprochenen Problematik nicht um eine Schwäche der Methoden des statistischen Testens. Es sind vielmehr die von den Anwendern dieser Methoden festgelegten Hypothesen, die sich oftmals als unbrauchbar erweisen, weil sie inhaltlich nicht das überprüfen, was eigentlich überprüft

werden soll. „Die Entwicklung einer Wissenschaft ausschließlich von der Signifikanz von Ergebnissen abhängig zu machen“ bedeutet deshalb, „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: 28). Soll tatsächlich überprüft werden, ob ein *praktisch bedeutsamer*, gleichsinniger Zusammenhang zwischen zwei Merkmalen besteht und nicht *irgendein* sich von null unterscheidender gleichsinniger Zusammenhang, dann ist die Einshypothese so aufzustellen, dass sie nur jene Parameterwerte enthält, die in der Einschätzung des Anwenders bei der jeweiligen Fragestellung praktisch bedeutsam sind. Nur dann wird die Diskrepanz zwischen Signifikanz und Relevanz von Testergebnissen aufgehoben. Schätzt man z.B. in einem Fall nur „mittelstarke“ Korrelationen größer als 0,2 als praktisch relevant ein, dann ergeben sich für den einseitigen Test auf praktisch relevanten, gleichsinnigen Zusammenhang folgende Hypothesen:

$$H_0: \rho \leq 0{,}2 \quad \text{und} \quad H_1: \rho > 0{,}2$$

Das erfordert dann zwar eine andere Teststrategie (siehe Abschnitt 3), aber auch deren Anwendung folgt der einheitlichen Handlungslogik des Signifikanztestens. In einem anderen Fall aber können tatsächlich alle Korrelationen, die größer als 0 sind, als relevant betrachtet werden (siehe Tageslichtzufuhr und Arbeitsleistung). Dann ergeben sich eben die herkömmlichen Hypothesen

$$H_0: \rho \leq 0 \quad \text{und} \quad H_1: \rho > 0$$

und der oben beschriebene Testverlauf.

Mit zunehmenden Stichprobenumfängen entscheidet man sich beim statistischen Testen von Hypothesen mit zunehmender Wahrscheinlichkeit für die richtige der beiden aufgestellten Hypothesen. Es ist somit von essentieller Bedeutung für die Qualität der gezogenen Schlussfolgerungen, dass die aufgestellten Hypothesen auch das prüfen, was man prüfen möchte, damit dieses Faktum auch tatsächlich das sein kann, was es ist: ein positives Qualitätsmerkmal statistischer Tests. Dies führt uns von unabhängig von der jeweiligen Fragestellung standardisiert ablaufenden Signifikanztests zu kontextbezogenen Relevanztests.

3 Der Relevanztest

Der Relevanzansatz beim Signifikanztesten ermöglicht die Miteinbeziehung der praktischen Bedeutsamkeit durch die Übersetzung der Forschungshypothese in eine kontextbezogene, vorwissengeleitete statistische Hypothese, die ausschließlich den Bereich der im jeweiligen Fall als praktisch bedeutsam eingestuften Parameter θ umfasst (vgl. Hodges & Lehmann 1954: 262). Daraus ergeben sich hinsichtlich der Hypothesenformulierung folgende Modifikationen gegenüber der herkömmlichen Vorgehensweise bei Signifikanztests: Indem wir nun für einseitige Tests mit τ_0 und für zweiseitige Tests mit τ_1 und τ_2 die neu festgelegten Grenzen zwischen den praktisch bedeutsamen und den irrelevanten Parameterwerten bezeichnen, lauten die inhaltlich sinnvollen statistischen Hypothesen für einseitige Fragestellungen:

$$H_0: \theta \leq \tau_0 \text{ (bzw. } \theta \geq \tau_0) \quad \text{und} \quad H_1: \theta > \tau_0 \text{ (bzw. } \theta < \tau_0)$$

Die in den Nullhypothesen angegebenen Parameterbereiche umfassen nun alle praktisch bedeutungslosen Parameterwerte. Für zweiseitige Fragestellungen gilt bei Relevanztests folgende Hypothesenformulierung:

$$H_0: \tau_1 \leq \theta \leq \tau_2 \qquad \text{und} \quad H_1: \theta < \tau_1 \ \vee \ \theta > \tau_2$$

Man sieht, dass sich die Hypothesen beim zweiseitigen Relevanztest für $\tau_1 \neq \theta_0 \neq \tau_2$ (θ_0 sei der Grenzparameterwert beim herkömmlichen Signifikanztest) nicht nur hinsichtlich der damit erfassten Parameterwerte, sondern auch hinsichtlich ihrer Formulierung von jenen der herkömmlichen zweiseitigen Signifikanztests unterscheiden. Bei einseitigen Fragestellungen hingegen wird lediglich die gewohnte Nullhypothesengrenze θ_0 durch die Relevanzgrenze τ_0 ersetzt.

Die Einshypothese eines als Relevanztest konzipierten einseitigen Tests des Korrelationskoeffizienten auf gleichsinnigem Zusammenhang beispielsweise wird – bei Einschätzung von Korrelationskoeffizienten ρ größer als 0,2 als praktisch relevant – demnach $H_1: \rho > 0{,}2$ lauten. Die Nullhypothese lautet dann: $\rho \leq 0{,}2$ (siehe Abschnitt 2).

Wiederum wird mit der Stichprobenverteilung von T bei Gültigkeit der Nullhypothese und dem Signifikanzniveau α jene Region – wir nennen sie nun aber R – bestimmt, die bei Gültigkeit von H_1 gleichzeitig β minimiert. Gilt $T_0 \in R$, dann liegt diesmal jedoch auf

jeden Fall ein signifikantes Testergebnis vor, das auch als praktisch relevant einzustufen ist. Durch eine Erhöhung des Stichprobenumfanges n – also eine Präzisierung des Experiments – werden im Gegensatz zu den unabhängig von der jeweiligen Fragestellung standardisiert ablaufenden Signifikanztests keine unbedeutenden, sondern nur praktisch relevante Testergebnisse signifikant.

Für die einseitige Überprüfung der Hypothese H_1: $\rho > \rho_0$ (bzw. $\rho < \rho_0$), wobei $\rho_0 > 0$ (bzw. < 0) sei, ergibt sich (unter Voraussetzung bivariat normalverteilter Merkmale und $|\rho|$ nicht zu groß) die – von der in Abschnitt 2 für den Test von H_1: $\rho > 0$ angegebenen – abweichende Testgröße

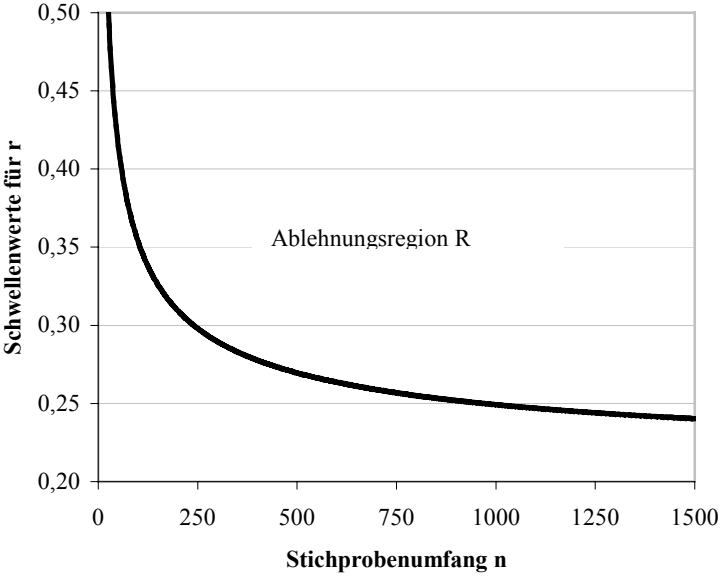
$$z = \frac{\sqrt{n-3}}{2} \cdot \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_0}{1-\rho_0} \right)$$

(vgl. etwa: Bosch 1998: 487f). z ist für großes n wiederum näherungsweise standardnormalverteilt. Die Nullhypothese wird auf dem Signifikanzniveau α beibehalten, wenn gilt: $z \leq z_{1-\alpha}$ (bzw. $z \geq -z_{1-\alpha}$). Abbildung 2 zeigt die daraus für den Test der Hypothesen H_0: $\rho \leq 0{,}2$ und H_1: $\rho > 0{,}2$ in Abhängigkeit vom Stichprobenumfang n auf einem Signifikanzniveau $\alpha = 0{,}05$ errechnete Ablehnungsregion R.

Für den zweiseitigen Test mit den Relevanzgrenzen $\rho_1 = -0{,}2$ und $\rho_2 = +0{,}2$ (in den meisten Fällen wird wie hier gelten: $\rho_1 = -\rho_2$) gilt bei derselben Teststatistik auf dem Signifikanzniveau α die Entscheidungsregel: H_0: $|\rho| \leq 0{,}2$ wird beibehalten, wenn $z \leq z_{1-\alpha/2}$.

Dieser konzeptionelle Schritt vom Signifikanz- zum Relevanztest muss auf alle statistischen Fragestellungen übertragen werden, wenn man an der gängigen Praxis im Zusammenhang mit der Verwendung von Signifikanztests etwas verändern möchte. Dabei gilt es einerseits die Relevanzschwellen, also die Grenzen zu den praktisch bedeutsamen Parametern für die einzelnen Fragestellungen, festzulegen, und andererseits – darauf basierend – die zum Testen dieser Hypothesen geeigneten Vorgehensweisen zu wählen. Zur Festlegung der Relevanzschwellen ist Expertenwissen aus jenem Bereich einzubringen, dem die Fragestellung entstammt, so dass die Hypothesenbildung jeweils kontextbezogen und nicht – unabhängig vom jeweiligen Thema – standardisiert erfolgt.

Abbildung 2 Die Ablehnungsregion R für den Stichprobenkorrelationskoeffizienten r in Abhängigkeit vom Stichprobenumfang beim Testen der Einshypothese H_1: $\rho > 0{,}2$ auf einem Signifikanzniveau $\alpha = 0{,}05$



Im Folgenden werden an weiteren ausgewählten Fragestellungen Relevanzteststrategien vorgestellt und die Auswirkungen dieser Vorgehensweise auf die Testabläufe betrachtet.

4 Einige weitere Beispiele für Relevanzteststrategien

4.1 Relevanztests für Anteilswerte

Es soll beispielsweise überprüft werden, ob der Anteil an Schülern einer bestimmten Altersklasse, die in einem Land eine standardisierte Aufgabe in einem Fach lösen können, durch die Veränderung der fachspezifischen Ausbildung größer als ein vorgegebener Erfahrungswert π (z.B. 0,6) wurde. Auch bei solchen Fragestellungen ist allzu oft die

gängige Signifikanzteststrategie der Prüfung auf *irgendeine* Verbesserung (unabhängig von ihrem Ausmaß) unbefriedigend. So kann sich die Ausbildungsänderung im betroffenen Fach für die Schüler trotz einer einschlägigen Leistungserhöhung nicht lohnen, wenn der Lernerfolg mit vielfältigen „Nebenwirkungen" – wie etwa ein Leistungsverlust in anderen Fächern – erkauft werden muss und die neue Methode beim Test der Wirksamkeit nur geringfügig besser als die alte abschneidet.

In Tests von Hypothesen über einen Anteilswert π, der relativen Häufigkeit einer interessierenden Eigenschaft in der Grundgesamtheit, wirkt sich das Relevanztestkonzept bei einseitigen Fragestellungen nur bei der notwendigen Bestimmung einer Relevanzschranke $\tau_0 = \pi_0$ (z.B. sei $\pi_0 = 0{,}7$ ein im Vergleich zum Erfahrungswert relevant verbesserter Anteil) auf die Formulierung der Hypothesen aus, damit die Nullhypothese alle für diese Fragestellung praktisch bedeutungslosen Anteile umfasst. Eine Nullhypothese H_0: $\pi \leq \pi_0$ ist für große Stichprobenumfänge auf einem Signifikanzniveau α beizubehalten, wenn die relative Häufigkeit p der interessierenden Eigenschaft in der Zufallsstichprobe den Wert

$$\pi_0 + z_{1-\alpha} \cdot \sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}$$

nicht überschreitet. Für H_0: $\pi \geq \pi_0$ gilt dies bei

$$p \geq \pi_0 - z_{1-\alpha} \cdot \sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}$$

Soll also überprüft werden, ob der Anteil der gelösten Aufgaben größer als 0,7 ist, so wird durch

$$0{,}7 + 1{,}65 \cdot \sqrt{\frac{0{,}7 \cdot (1-0{,}7)}{1.000}} = 0{,}724$$

festgelegt, dass unter 1.000 zufällig ausgewählten Probanden mehr als 72,4 % die Aufgabe lösen müssen, damit die Nullhypothese der unbedeutenden Auswirkung der Ausbildungsänderung abgelehnt werden kann.

Bei zweiseitigen Fragestellungen sind zwei Grenzen π_1 und π_2 so festzulegen, dass für H_0: $\pi_1 \leq \pi \leq \pi_2$ wiederum alle als praktisch bedeutungslos eingestuften Parameterwerte umfasst. Die Bestimmung des zweiteiligen Ablehnungsbereiches R auf Basis der in der

Nullhypothese festgelegten Parameterwerte ist für diesen Relevanztest jedoch nicht so einfach wie beim herkömmlichen zweiseitigen Test von H_0: $\pi = \pi_0$ und H_1: $\pi \neq \pi_0$. Die Grenzen der Ablehnregion R sind nur iterativ bestimmbar, sofern π_1 und π_2 nicht weit genug auseinander liegen.

Die in diesem Fall – im Vergleich zur für den einseitigen Test beschriebenen „parameterbasierten" Vorgehensweise über die Kenntnis der Stichprobenverteilung des Anteilschätzers bei Gültigkeit der Nullhypothese – einfachere Möglichkeit zur Testdurchführung ist die approximativ gleichwertige „schätzerbasierte" Teststrategie über die Angabe von Konfidenzintervallen. Bei einem zweiseitigen Test bestimmt man dabei das herkömmliche näherungsweise $(1-\alpha)$-Konfidenzintervall $[\pi_u;\pi_o]$ für den Parameter π (vgl. etwa: Quatember 2005: 115ff.) und behält H_0 bei, wenn es sich mit dem Intervall $[\pi_1;\pi_2]$ der Nullhypothese überschneidet. Ist dies nicht der Fall, dann hat man ein statistisch signifikantes und gleichzeitig praktisch relevantes Testergebnis gefunden.

Sollen die relativen Häufigkeiten π_A und π_B des Auftretens einer interessierenden Eigenschaft aus zwei Grundgesamtheiten A und B (Vergleich der Leistungen von Schülern verschiedener Länder oder zu verschiedenen Zeitpunkten) in einem Zweistichprobentest unabhängiger Stichproben miteinander verglichen werden, dann ist folgende Vorgehensweise eine Relevanzteststrategie: Man legt bei einseitiger Fragestellung für die Differenz von π_A und π_B (mit $\pi_A > \pi_B$) eine Grenzdifferenz δ_0 (z.B. $\delta_0 = 0{,}05$) und bei zweiseitiger Fragestellung zwei Grenzdifferenzen $\tau_1 = \delta_1$ und $\tau_2 = \delta_2$ (mit $\delta_1 < \delta_2$; zumeist wird $\delta_1 = -\delta_2$ gelten, also z.B. $\delta_1 = -0{,}05$ und $\delta_2 = +0{,}05$) als Relevanzgrenzen fest. Damit ergeben sich folgende Hypothesen über die Differenz zweier relativer Häufigkeiten $\delta = \pi_A - \pi_B$ in den Populationen:

$$H_0: \delta \leq \delta_0 \quad \text{und} \quad H_1: \delta > \delta_0$$

bei einseitiger und

$$H_0: \delta_1 \leq \delta \leq \delta_2 \quad \text{und} \quad H_1: \delta < \delta_1 \vee \delta > \delta_2$$

(bei $\delta_1 = -\delta_2$ gilt: H_0: $|\delta| \leq \delta_2$ und H_1: $|\delta| > \delta_2$) bei zweiseitiger Fragestellung.

Mit den Daten zweier unabhängiger Stichproben mit Umfängen n_A und n_B aus den betreffenden Grundgesamtheiten wird die Testgröße $d = p_A - p_B$ (p_A, p_B ... die relativen

Häufigkeiten der interessierenden Eigenschaft in den Stichproben aus A und B) berechnet. Im einseitigen Fall ist H_0 bei großen Stichprobenumfängen auf dem Signifikanzniveau α beizubehalten, wenn gilt (vgl. etwa: Quatember 2005: 137ff.):

$$d \leq \delta_0 + z_{1-\alpha} \cdot \sqrt{\left(\frac{p_A \cdot (1-p_A)}{n_A} + \frac{p_B \cdot (1-p_B)}{n_B} \right)}$$

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

$$[\delta_u; \delta_o] = d \pm z_{1-\alpha/2} \cdot \sqrt{\frac{p_A \cdot (1-p_A)}{n_A} + \frac{p_B \cdot (1-p_B)}{n_B}}$$

und der in der Nullhypothese formulierte Bereich der praktisch irrelevanten Differenzen $[\delta_1;\delta_2]$ einander überschneiden.

4.2 Relevanztests über einen statistischen Zusammenhang zweier nominaler Merkmale

Der statistische Zusammenhang zweier nominaler Merkmale (z.B. soziale Herkunft und Berufstätigkeit) wird im Allgemeinen mit dem χ^2-Test überprüft. Dabei stehen die beiden Hypothesen

$$H_0: \chi^2 = 0 \quad \text{und} \quad H_1: \chi^2 > 0$$

über den Populationsparameter χ^2, der den Zusammenhang mit den Daten einer Kontingenztafel misst, in Konkurrenz. Die standardmäßige Wahl dieser beiden Hypothesen für den Signifikanztest ist wie beim Korrelationstest aus Abschnitt 2 oftmals unbefriedigend, weil bei großen Stichprobenumfängen auch noch so geringe Abweichungen von der Nullhypothese des Fehlens eines statistischen Zusammenhangs mit zunehmender Wahrscheinlichkeit signifikant werden. Sind jedoch solche Zusammenhänge zwischen zwei Merkmalen praktisch irrelevant, was selbstverständlich auch hier nicht immer der Fall sein muss, so sind diese Hypothesen unbrauchbar.

Das Assoziationsmaß V von Cramér mit

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

(N … der Umfang der Grundgesamtheit, r,s ... die Anzahlen der Ausprägungen der beiden Merkmale) ist eine zur Messung der Stärke des Zusammenhangs in der Grundgesamtheit geeignete Kennzahl mit Wertebereich $0 \le V \le 1$. Die oben angeführten standardmäßig formulierten Hypothesen lauten aus der Sicht dieser Kennzahl – wie sich leicht nachvollziehen lässt – ebenfalls:

$$H_0: V = 0 \quad \text{und} \quad H_1: V > 0.$$

Für den Relevanztest des Zusammenhangs ist eine Relevanzschwelle $\tau_0 = V_0$ so zu bestimmen, dass bei der Hypothesenwahl

$$H_0: V \le V_0 \quad \text{und} \quad H_1: V > V_0$$

sämtliche als praktisch unbedeutend eingestuften Parameterwerte für V in die Nullhypothese eingehen.

Ein möglicher Ansatz zur Bestimmung von V_0 führt über die Begründung der Festlegung einer Bedeutsamkeitsschwelle ρ_0 für den Korrelationskoeffizienten ρ zweier metrischer Merkmale (Abschnitt 3). Für 2×2-Tafeln, also Kontingenztafeln mit r = s = 2, gilt nämlich für jede beliebige Kodierung der Ausprägungen der beiden Merkmale:

$$V = |\rho|$$

(vgl. etwa: Hilbert 1998: 95f.). Damit könnte man im Prinzip dieselben Überlegungen wie für ρ auch für V anstellen. Wird darauf basierend ein statistischer Zusammenhang im Ausmaß von $V_0 = 0{,}2$ für 2×2-Tafeln als gerade noch unbedeutend empfunden, so könnte diese Bedeutsamkeitsschwelle verallgemeinernd auch auf allgemeine r×s-Tafeln angewendet werden.

Auf diese Art festgelegte Hypothesen über den Zusammenhang zweier nominaler Merkmale können für nicht allzu kleines V_0 überprüft werden, indem man sich der Bereichschätzung von V bedient (für $V_0 = 0$ wird der Test zum herkömmlichen χ^2-Test).

Mit den Daten einer Zufallsstichprobe ist

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

mit

$$\hat{\chi}^2 = n \cdot \sum_{i,j} \frac{(p_{ij}^o - p_{ij}^e)^2}{p_{ij}^e}$$

(p_{ij}^o ... die in der Stichprobe beobachteten relativen Häufigkeiten der möglichen Merkmalskombinationen, p_{ij}^e ... deren bei Fehlen eines statistischen Zusammenhangs in der Stichprobe zu erwartenden relativen Häufigkeiten) der Schätzer für den Parameter V. Als Schätzer $\widehat{Var}_\infty \hat{V}$ für die asymptotische Varianz $Var_\infty \hat{V}$ von $\hat{V}$ ergibt sich (vgl. etwa: Bishop et al. 1977: 386):

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

mit dem Varianzschätzer für den quadrierten allgemeinen Stichproben-Phikoeffizienten $\hat{\phi}^2 = \frac{\hat{\chi}^2}{n}$:

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

(die p_{ij}'s sind die beobachteten relativen Häufigkeiten aus der Kontingenztafel; p_{i+}, p_{+j} … die beobachteten relativen Randhäufigkeiten). $\hat{V}$ ist asymptotisch normalverteilt mit dem Erwartungswert V und der Varianz $Var_\infty \hat{V}$. Mit den Schätzern $\hat{V}$ und $\widehat{Var}_\infty \hat{V}$ ergibt sich die folgende approximative Untergrenze V_u des einseitigen Konfidenzintervalls zur Sicherheit $1-\alpha$ für V:

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

Gilt $V_u \leq V_0$ aus der Nullhypothese, dann wird diese zum approximativen Signifikanzniveau α beibehalten, sonst zu Gunsten von H_1: $V > V_0$ verworfen.

Beispielsweise habe sich in einer Untersuchung über den statistischen Zusammenhang zwischen sozialer Herkunft (angegeben in 3 sozialen Schichten) und Berufstätigkeit (ja/nein) unter 1.000 zufällig ausgewählten Personen folgende Kontingenztabelle ergeben:

	ja	**nein**	**Summe**
Schicht 1	0,21	0,09	0,30
Schicht 2	0,39	0,11	0,50
Schicht 3	0,10	0,10	0,20
Summe	0,70	0,30	1

Für einen Test von

$$H_0: V \leq 0{,}2 \quad \text{und} \quad H_1: V > 0{,}2$$

errechnet sich mit den Daten dieser Kontingenztabelle eine Kennzahl $\hat{\chi}^2 = 53{,}\dot{3}$, ferner ein Zusammenhang in der Stichprobe von $\hat{V} = 0{,}2309$ und ein Varianzschätzer von

$$\widehat{\mathrm{Var}_\infty}\hat{V} = \frac{1}{4 \cdot 0{,}2309^2} \cdot 0{,}0002309 = 0{,}0010827$$

Somit ist die Untergrenze V_u des einseitigen 0,95-Konfidenzintervalls gegeben durch

$$V_u = 0{,}2309 - 1{,}645 \cdot \sqrt{0{,}0010827} = 0{,}1768$$

und die Nullhypothese wird, da $0{,}1768 \leq 0{,}2$ gilt, zum Signifikanzniveau $\alpha = 0{,}05$ beibehalten. Der gefundene Zusammenhang in der Stichprobe ist nicht groß genug, um daraus auf einen relevanten Zusammenhang in der Grundgesamtheit schließen zu können.

4.3 Relevanztests für Regressionskoeffizienten

Sehr häufig werden in der Sozialforschung zur Beschreibung kausaler Zusammenhänge Regressionsanalysen durchgeführt. Im einfachsten Fall der linearen Einfachregression gilt es die Koeffizienten a und b der Regressionsgeraden

$$y = a + b \cdot x$$

zu schätzen. y und x sind die abhängige bzw. die unabhängige Variable. Die Schätzung von a und b mit den Daten einer Zufallsstichprobe erfolgt mit den Kleinstquadratschätzungen $\hat{b}$, dem Quotienten aus der Stichprobenkovarianz von x und y und der Stichprobenvarianz von x, und $\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$ mit den beiden Stichprobenmittelwerten $\bar{x}$ und $\bar{y}$ der beiden Variablen. Die bei weitem am häufigsten verwendete Hypothesenformulierung ist zweiseitig und lautet:

$$H_0: b = 0 \quad \text{und} \quad H_1: b \neq 0$$

Die Nullhypothese bringt damit zum Ausdruck, dass x doch keinen Einfluss auf y besitzt. Möglicherweise sind jedoch Steigungen b der Regressionsgeraden in der Nähe von null z.B. für Prognosen von y auf Basis von x völlig irrelevant, weil sie sich bei verschiedenen x-Werten nur geringfügig voneinander unterscheiden. Werden Steigungen b in einem Bereich $b_1 \leq b \leq b_2$ für praktisch bedeutungslos erachtet (z.B. $b_1 = -0{,}1$ und $b_2 = +0{,}1$), so ist die Hypothesenformulierung in Hinblick auf die Relevanz der Testergebnisse folgendermaßen zu modifizieren:

$$H_0: b_1 \leq b \leq b_2 \quad \text{und} \quad H_1: b < b_1 \vee b > b_2$$

(bei $b_1 = -b_2$ gilt: $H_0: |b| \leq b_2$ und $H_1: |b| > b_2$). In diesem Fall wird zur Festlegung der Entscheidungsregel wiederum die schätzerbasierte Vorgehensweise verwendet. Das $(1-\alpha)$-Konfidenzintervall $[b_u;b_o]$ für den Regressionskoeffizienten b ist bei bivariat normalverteilten Merkmalen in großen Stichproben gegeben durch

$$[b_u; b_o] = \hat{b} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{\sigma}_b^2}$$

mit dem Schätzer für die Varianz von b

$$\hat{\sigma}_b^2 = \frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

und den zu den jeweiligen x-Werten gehörenden y-Prognosen

$$\hat{y}_i = \hat{a} + \hat{b} \cdot x_i \ ,$$

(siehe etwa: Bosch 1998: 601ff.). Die Nullhypothese wird beibehalten, wenn sich der in ihr enthaltene Bereich $[b_1;b_2]$ und das Konfidenzintervall $[b_u;b_o]$ überschneiden.

Zur Überprüfung von zweiseitigen Hypothesen über den Achsenabschnitt a der Regressionsgeraden bedient man sich für große Stichproben analog des $(1-\alpha)$-Konfidenzintervalls

$$[a_u; a_o] = \hat{a} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{\sigma}_a^2}$$

mit

$$\hat{\sigma}_a^2 = \frac{\left[\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Bei einseitigen Fragestellungen kann man sich auch der (herkömmlich verwendeten) parameterbasierten Vorgangsweise bedienen. Es ist mit $\tau_0 = b_0$ (bzw. a_0) eine Relevanzgrenze für den interessierenden Regressionskoeffizienten festzulegen und damit sind je nach Prüfrichtung die Hypothesen

$$H_0: b \le b_0 \quad \text{und} \quad H_1: b > b_0$$

bzw.

$$H_0: b \ge b_0 \quad \text{und} \quad H_1: b < b_0$$

aufzustellen. Die Teststatistik

$$z = \frac{\hat{b} - b_0}{\hat{\sigma}_b}$$

ist in großen Stichproben unter der Nullhypothese standardnormalverteilt und somit wird H_0 zum Signifikanzniveau α beibehalten, wenn gilt: $z \leq z_{1-\alpha}$ (bzw. $z \geq -z_{1-\alpha}$) (vgl. Bosch 1998: 604).

Für den einseitigen Relevanztest von a ist entsprechend

$$z = \frac{\hat{a} - a_0}{\hat{\sigma}_a}$$

die zu verwendende Teststatistik.

Soll z.B. überprüft werden, ob in der Stichprobenuntersuchung des Zusammenhangs zwischen Tageslichtzufuhr (x) und Arbeitsleistung (y) die Steigung b der Regressionsgeraden den Wert von 0,1 einer früheren Untersuchung noch übersteigt und nicht nur, ob *irgendeine* positive Steigung > 0 vorliegt, so werden mit der Relevanzgrenze $b_0 = 0{,}1$ die Hypothesen

$$H_0: b \leq 0{,}1 \quad \text{und} \quad H_1: b > 0{,}1$$

formuliert. Errechnen sich nun mit den Daten einer Stichprobe vom Umfang n = 1.000 eine Regressionsgerade mit der Gleichung: $y = \hat{a} + \hat{b} \cdot x = 16{,}151 + 0{,}264 \cdot x$ und ein Schätzer $\hat{\sigma}_b^2 = 0{,}0021$, so beträgt die Teststatistik

$$z = \frac{0{,}264 - 0{,}1}{0{,}046} = 3{,}565$$

Somit ist wegen 3,565 > 1,645 die Nullhypothese auf dem Signifikanzniveau $\alpha = 0{,}05$ zu verwerfen. Die Steigung ist signifikant größer als 0,1 und damit in diesem Fall auch praktisch bedeutsam.

4.4 Einfache Varianzanalyse auf relevante Mittelwertsunterschiede

Der üblichen Vorgehensweise bei der einfachen Varianzanalyse liegt die Fragestellung zu Grunde, ob sich k Gruppen ($k \geq 2$) hinsichtlich eines interessierenden Merkmals x unterscheiden. Dies wird in die statistische Hypothese übersetzt, dass sich mindestens zwei der k Gruppenmittelwerte unterscheiden. Anders formuliert: Es ist zu überprüfen, ob die Streuung der Gruppenmittelwerte größer als null ist. Eine diesbezügliche Fragestellung wäre beispielsweise, ob sich die Ergebnisse bei Leistungstests von Schülern in k verschiedenen Ländern der EU unterscheiden.

Im herkömmlichen Test (wir begnügen uns mit der Darstellung der Idee für gleiche Stichprobenumfänge m in den Gruppen) wird dazu – unter der Voraussetzung der Normalverteilung des Merkmals x und derselben Varianz σ^2 von x in jeder der k Gruppen – die Zufallsvariable

$$\hat{F} = \frac{\frac{1}{k-1} \cdot \sum_{i=1}^{k} m \cdot (\bar{x}_i - \bar{x})^2}{\frac{1}{n-k} \cdot \sum_{i=1}^{k} \sum_{j=1}^{m} (x_{ij} - \bar{x}_i)^2}$$

($n = m \cdot k$ ... Gesamtstichprobenumfang, $\bar{x}$ ... Gesamtstichprobenmittelwert, $\bar{x}_i$ ... Stichprobenmittelwert der i-ten Gruppe ($i = 1,2,...k$), x_{ij} ... Merkmalsausprägung von x bei der j-ten Erhebungseinheit der i-ten Gruppe) als Teststatistik verwendet. Die Mittelwertshypothesen

$$H_0: \mu_i = \mu \text{ für alle i } (i = 1,2,...k) \quad \text{und} \quad H_1: \mu_i \neq \mu \text{ für mindestens ein i}$$

lassen sich mit dem zu $\hat{F}$ gehörenden Parameter F folgendermaßen formulieren:

$$H_0: F = 0 \quad \text{und} \quad H_1: F > 0$$

Bei Gültigkeit von H_0 ist die Teststatistik $\hat{F}$ zentral F-verteilt mit k–1 bzw. n–k Freiheitsgraden. H_0 wird beibehalten, wenn $\hat{F} \leq F_{k-1;n-k;1-\alpha}$ ist, dem $(1-\alpha)$-Fraktil einer solchen F-Verteilung (vgl. etwa: Bosch 1998: 495ff.).

Doch natürlich ist auch hier nicht jeder Mittelwertsunterschied praktisch relevant. Cohen (1969) schlägt vor, zur Bestimmung des Ausmaßes der Abweichung von der in der Nullhypothese formulierten Behauptung die Größe

$$f = \frac{\sigma_\mu}{\sigma}$$

(σ_μ ... die Standardabweichung der Gruppenmittelwerte) zu verwenden (vgl. Cohen 1969: 267). Die Hypothesen über die Gruppenmittelwerte lauten für den herkömmlichen Signifikanztest auch mit diesem Parameter

$$H_0: f = 0 \text{ und } H_1: f > 0$$

In die Größe f kann die Vorstellung des Anwenders der Varianzanalyse über praktisch irrelevante bzw. relevante Mittelwertsunterschiede als Expertenwissen durch die Bestimmung einer gerade noch als praktisch unbedeutend empfundenen Standardabweichung der Mittelwerte $\sigma_\mu \geq 0$ einfließen. Für die Standardabweichung σ sind Informationen aus früheren Untersuchungen oder abermals begründete Schätzungen zu verwenden.

Für den Relevanztest auf Mittelwertsunterschiede in k unabhängigen Gruppen mit den Hypothesen

$$H_0: f \leq f_0 \text{ und } H_1: f > f_0$$

sei damit die Bedeutsamkeitsschwelle $\tau_0 = f_0$ bestimmt.

Eine andere Möglichkeit der Bestimmung der Bedeutsamkeitsschwelle f_0 ist die der Verwendung einer Konvention darüber, welche Effektgrößen f als klein, welche als mittel und welche als groß einzustufen sind. Überprüft ein Experimentator dann etwa seine Vermutung eines mittleren Effektes, dann kann er auf diese Kategorisierungen zurückgreifen und so f_0 für die Überprüfung von relevanten Unterschieden festlegen. Cohen (1969) etwa spricht bei $f = 0{,}1$ von einem kleinen, bei $f = 0{,}25$ von einem mittleren und bei $f = 0{,}4$ von einem großen Effekt (vgl. Cohen 1969: 277ff.).

Der Einbau der Relevanzstrategie in die einfache Varianzanalyse führt jedenfalls dazu, dass für $f_0 > 0$ die Teststatistik $\hat{F}$ nun asymptotisch nichtzentral F-verteilt ist mit dem Nichtzentralitätsparameter $n \cdot f_0^2$ und k–1 bzw. n–k Freiheitsgraden (vgl. Cohen 1969: 404f.).

Die Bestimmung des (1–α)-Fraktils $F_{n \cdot f_0^2, k-1, n-k, 1-\alpha}$ dieser Verteilung aus Tabellen (z.B. unter: *http://calculators.stat.ucla.edu/cdf* oder für nicht zu kleine m und f_0 durch Verwendung der Transformation (vgl. Laubscher 1960: 1111)

$$z = \frac{\sqrt{(2 \cdot (n-k)-1) \cdot \frac{k-1}{n-k} \cdot F} - \sqrt{2 \cdot (k-1+n \cdot f_0^2) - \frac{k-1+2 \cdot n \cdot f_0^2}{k-1+n \cdot f_0^2}}}{\sqrt{\frac{k-1}{n-k} \cdot F + \frac{k-1+2 \cdot n \cdot f_0^2}{k-1+n \cdot f_0^2}}} ,$$

wobei z asymptotisch standardnormalverteilt ist und sich somit für $z = z_{1-\alpha}$ mit α, dem Signifikanzniveau, durch Umformung nach der Variablen F das gesuchte (1–α)-Fraktil $F_{n \cdot f_0^2, k-1, n-k, 1-\alpha}$ näherungsweise bestimmen lässt) ermöglicht die Berechnung der asymptotischen Ablehnungsregion R der Nullhypothese des Relevanztests. Gilt $\hat{F} \leq F_{n \cdot f_0^2, k-1, n-k, 1-\alpha}$, so ist H_0 beizubehalten.

Z.B. ergibt sich für den Leistungstest von Schülern in k = 4 Staaten mit m = 1.000 und somit n = 4.000 bei der Überprüfung eines relevanten „kleinen Effekts“ mit den Hypothesen

$$H_0: f \leq 0{,}1 \quad \text{und} \quad H_1: f > 0{,}1$$

ein Wert $F_{n \cdot f_0^2 = 40, k-1=3, n-k=3.996, 1-\alpha=0{,}95} = 21{,}963$ (im Vergleich dazu gilt beim herkömmlichen Test auf *irgendeinen* Effekt $f_0 = 0$: $F_{k-1=3, n-k=3.996, 1-\alpha=0{,}95} = 2{,}607$). Ist die in der Stichprobe errechnete Kennzahl $\hat{F}$ größer als diese Schranke, so wird auf die Einshypothese übergegangen und man hat ein signifikantes und relevantes Testergebnis gefunden.

5 Zusammenfassung und praktische Umsetzung

Die vorgestellte Relevanzteststrategie versetzt uns in die Lage, das Signifikanz-Relevanz-Problem in der Anwendung der herkömmlichen Signifikanztests innerhalb des Signifikanztestkonzepts zu lösen. Dabei wird für die Verwendung der statistischen Methoden der Einsatz von Expertenwissen in jenem Bereich gefordert, dem die Fragestellung entstammt, damit es in die Bestimmung der für den jeweiligen Test nötigen Relevanzschwellen einfließen kann. Auf diese Art erhöht man die Wahrscheinlichkeit dafür, dass inhaltlich uninteressante Effekte nichtsignifikant bleiben.

Es stellt sich heraus, dass für die Relevanzteststrategie häufig die schätzerbasierte Vorgehensweise beim statistischen Testen mit der Bestimmung ein- oder zweiseitiger Konfidenzintervalle einfacher ist als die herkömmliche Vorgehensweise über die Stichprobenverteilung der Teststatistik bei Gültigkeit der Nullhypothese und die Bestimmung von p-Werten. Die Relevanztests für Anteile bei zweiseitiger Fragestellung und auf statistischem Zusammenhang zweier nominaler Merkmale sind Beispiele dafür.

Dies ist von besonderer Bedeutung bei Verwendung von Statistik-Programmpaketen. Bei bestimmten Fragestellungen können bei manchen Paketen nur die vom jeweiligen Inhalt unabhängig formulierten Standardhypothesen mit der parameterbasierten Vorgehenswiese getestet werden (z.B. beim Chiquadrattest). Daraus würde für die Anwendbarkeit der Relevanzteststrategie auch in solchen Fällen die Notwendigkeit der Ausarbeitung neuer Programme resultieren, die die Festlegung der Bedeutsamkeitsschwellen durch den Anwender ermöglichen und die daraus abgeleiteten Änderungen im Testverlauf der verschiedenen statistischen Tests implementieren. Auch der dadurch entstehende zusätzliche Aufwand könnte jedoch in keiner Weise gegen den Unsinn des Testens inhaltlich ungeeigneter Hypothesen aufgewogen werden.

Die schätzerbasierten Vorgehensweisen lassen sich jedoch als Alternative ohne zusätzlichen Aufwand überall dort sofort umsetzen, wo solche Intervalle für interessierende Parameter sowieso standardmäßig mit ausgegeben werden. Der Anwender der Relevanzteststrategie hat dann im Entscheidungsschritt nur mehr die Ergebnisse dieser Intervalle mit der in der – natürlich vorab formulierten – Nullhypothese enthaltenen Menge aller praktisch irrelevanten Parameter zu vergleichen. Haben diese Mengen nichts gemein, dann hat man auf einem Signifikanzniveau α ein statistisch signifikantes *und* praktisch relevantes Testergebnis gefunden.

Literatur

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

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

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation*. 2. Auflage. Berlin: Springer Verlag.

Bosch, K. (1998). *Statistik-Taschenbuch*. 2. Auflage. München: Oldenbourg Verlag.

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

Carver, R. P. (1993). The Case Against Statistical Significance Testing, Revisted. *Journal of Experimental Education, 61* (4), 287-292.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Hilbert, A. (1998). *Zur Theorie der Korrelationsmaße*. Lohmar: Josef Eul Verlag.

Hodges, J. L. & Lehmann, E. L. (1954). Testing the Approximate Validity of Statistical Hypotheses. *Journal of the Royal Statistical Society, B, 16,* 261-268.

Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement, 56* (5), 746-759.

Laubscher, N. F. (1960). Normalizing the noncentral t and F distributions. *Annals of Mathematical Statistics, 31,* 1105-1112.

Quatember, A. (1997). Die Veränderung der sozial-, wirtschaftwissenschaftlichen und medizinischen Forschung durch die Verwendung statistischer Programmpakete: Bestandsaufnahme und Verbesserungsvorschläge. In W. Bandilla & F. Faulbaum (Ed.), *SoftStat '97. Advances in Statistical Software 6* (pp. 309-316). Stuttgart: Lucius & Lucius.

Quatember, A. (2004). Der statistische Signifikanztest in der Krise. *IFAS Research Paper Series*. 2004-08. Website: www.ifas.jku.at (unter „Forschung" und „Research Reports").

Quatember, A. (2005). *Statistik ohne Angst vor Formeln*. München: Pearson Verlag.

Sahner, H. (1979). Veröffentlichte empirische Sozialforschung: Eine Kumulation von Artefakten? *Zeitschrift für Soziologie, 8*(3), 267-278.

Smart, R.G. (1964). The Importance of Negative Results in Psychological Research. *The Canadian Psychologist, 5a* (4), 225-232.

Wilson, F. D., Smoke, G. L. & Martin, J. D. (1973). The Replication Problem in Sociology: A Report and a Suggestion. *Sociological Inquiry, 43* (2)*,* 141-149.

Korrespondenzadresse

Ass.-Prof. Dr. Andreas Quatember
IFAS - Institut für Angewandte Statistik
Johannes Kepler Universität Linz
Altenbergerstraße 69
A-4040 Linz/Österreich
email: andreas.quatember@jku.at
Internet: www.ifas.jku.at