Validitätsprobleme bei der statistischen Modellbildung mit kleinen Stichproben in der Lebenslaufforschung

Kelle, Udo; Prein, Gerald

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

VALIDITÄTSPROPLEME BEI DER STATISTISCHEN MODELLBILDUNG MIT KLEINEN STICHPROBEN IN DER LEBENSLAUFFORSCHUNG

Dr. Udo Kelle & Dr. Gerald Prein

In leicht gekürzter Form erschienen als:

Universität Bremen
Sonderforschungsbereich 186
"Statuspassagen und Risikolagen im Lebensverlauf"
Bereich Methoden, Statistik und EDV
Postfach 330 440
28334 Bremen
Tel.: 0421/218-4168
0421/218-4169
Fax: 0421/218-4153
email: ukelle@sfb186.uni-bremen.de
gprein@sfb186.uni-bremen.de

Die neuere Soziologie des Lebenslaufs, deren Forschungsschwerpunkte auf den Schnittstellen von Lebenslauf- und Biografieforschung, Arbeitsmarkts- und Berufssoziologie, Sozialpolitikforschung und Medizinsoziologie liegen (vgl. HEINZ 1991, 1993)[1], eröffnet auch für die Gerontologie neue sozialwissenschaftliche Perspektiven, weil dort die Gestaltung von Lebensläufen in modernen Industriegesellschaften sowie die an den "Statuspassagen" der Biografie auftretenden Risiken und ihre wohlfahrtsstaatliche Bearbeitung im Mittelpunkt des Forschungsinteresses stehe. Gegenstand von Untersuchungen zu Erwerbs- und Familienbiographien von Frauen im Rentenalter (BORN, KRÜGER 1993), der kulturellen Veränderung von Altersbildern (GÖCKENJAN 1993 a,b), Rehabilitation bei gesundheitsbedingter Berufsunfähigkeit (BEHRENS 1992, 1993a,b) bilden nämlich psychosoziale Aspekte physiologischer und sozialer Alterungsprozesse und deren Verarbeitung durch die Individuen, ebenso wie makro-sozietäre Determinanten und Auswirkungen dieser Prozesse.

Diese neuartige sozialwissenschaftliche Blickweise auf den Lebenslauf bedurfte allerdings auch neuartiger methodischer Instrumente, wie sie bspw. Panel- und Längsschnittdesigns (KASPRZYK et al. 1989) sowie Verfahren der Analyse von Ereignissen im Zeitverlauf (*"event history analysis"*) darstellen. Neuartige methodische Werkzeuge bringen allerdings auch neuartige methodische Probleme und *"threats for validity"* (COOK, CAMPBELL 1979) mit sich. Validitätsbedrohungen meßtheoretischer Art werden ausführlich von Kemnitz (in diesem Band) diskutiert. Wir wollen uns dahingegen hier auf eine bestimmte Art inferenzstastischer Probleme i.e.S., d.h. auf Probleme der *statististical conclusion validity* (COOK, CAMPBELL 1979, S. 39 ff), konzentrieren, wie sie im im Kontext der Lebensverlaufsforschung sehr häufig auftreten: Probleme *kleiner Fallzahlen* und *geringer Zellenbesetzungen*.

- geringe Stichprobenumfänge ergeben sich sehr häufig bei der Erhebung von Längsschnittdaten in der Lebensverlaufsforschung etwa durch *panel attrition*;

- Lebenslaufforschung erfordert vielfach komplexe theoretische Modelle, die auf der Ebene der statistischen Modellierung eine *Vielzahl erklärender Variablen* und Interaktionseffekte berücksichtigen. Komplexe Modelle führen leicht – auch bei *prima facie* großen oder sehr großen Stichproben – zu geringen Besetzungen in einzelnen Zellen.

Insbesondere bei der Erforschung von Strukturen der Primärversorgung, Rehabilition und Prävention im Kontext der Lebenslaufforschung werden Probleme geringer Zellenbesetzung in besonderer Weise virulent, denn:

---

[1] Die Grundlage der folgenden Ausführungen bilden empirische Studien am Sonderforschungsbereich 186 der Deutschen Forschungsgemeinschaft "*Statuspassagen und Risikolagen im Lebensverlauf*", in dessen Methoden- und Statistikabteilung die Verfasser dieses Beitrages tätig sind.

- in den hier verwendeten quasi-experimentellen und ex-post-facto experimentellen Designs lassen sich kleine Fallzahlen oftmals nicht vermeiden. Beispiele lassen sich hierfür in der medizinsoziologischen Forschung leicht finden: man denke an die Studie von Uta Gerhardt zu Rehahabilitationsverläufen und sozial-ökonomischen Coping bei chronisch niereninsuffizienten Patienten (GERHARDT 1986). Die Studie wurde, um relevante sozialstrukturelle Faktoren zu kontrollieren, regional begrenzt, so daß sich eine Fallzahl von 60 Familien ergab. Auch Untersuchungen, die nur im Kontext bestimmter Institutionen durchgeführt werden können, müssen sich oft auf kleine Fallzahlen beschränken. Oft kann hier von Seiten der Forscher auf die Größe der Stichprobe keinerlei Einfluß genommen werden. Als Beispiel mag hierfür die jüngst von Behrens und Dohrenburg an einer Rehabilitationsklinik durchgeführte Verbleibsstudie über berufliche Rehabilitanden dienen, wo der reha-ärtzliche Rat für einen Tätigkeitswechsel als Zugangskriterium für den Stichprobenzugang diente (BEHRENS 1993b): Die Untersuchungsstichprobe umfasste schließlich 27 Fälle.

- Oftmals sind vielfältige Wechselwirkungen zwischen biologisch-medizinischen, psychologischen und sozialen Einflußgrößen zu berücksichtigen, die zur Konstruktion äußerst komplexer Modelle und damit zu geringen Zellenbesetzungen führen kann.

Welche besonderen inferenzstatistischen Probleme verbinden sich nun mit kleinen Fallzahlen und geringen Zellenbesetzungen?

Ziel von inferenzstatistischen Teststrategien ist es, *rational begründete Entscheidungen* über die Ablehnung oder Beibehaltung von Hypothesen zu treffen, d.h. zu entscheiden, ob ein Tatbestand, der im Rahmen eines Experiments oder auf der Grundlage einer Stichprobenerhebung festgestellt wurde, über den Kontext der Untersuchungssituation sowie der untersuchten Gruppe hinaus verallgemeinert werden kann. Folgendes inferenzstatistisches Vorgehen kommt hierbei i.d.R. – auf der Basis der in den Sozialwissenschaften üblicherweise eingesetzten Testtheorie von Neyman und Pearson (NEYMAN, PEARSON 1967) – zum Einsatz:

1. Auf der Grundlage theoretischer Annahmen zum Gegenstandsbereich sowie meßtheoretischer Überlegungen zum Skalenniveau der untersuchten Variablen wird eine Teststatistik (etwa: Mittelwerte, Varianzen, der χ^2-Wert ...) für die jeweilige Stichprobe berechnet.

2. Sodann wird die Wahrscheinlichkeitsverteilung dieser Teststatistik unter der Annahme ermittelt, daß der beobachtete Zusammenhang im Experiment oder in der untersuchten Stichprobe nur zufällig zustandegekommen ist, d.h. *nicht* auf einem systematischen Effekt beruht. Diese Annahme wird als Nullhypothese (H_0) bezeichnet.

3. Sinkt die Wahrscheinlichkeit, den empirisch festgestellten Effekt unter Annahme der Nullhypothese zu beobachten, unter ein vorab festgelegtes sog. α-Niveau, wird die Nullhypothese verworfen und – bis zum Auftauchen neuer Gegenevidenz – die Alternativhypothese (H_1) beibehalten.

Beim traditionellen Vorgehen wird also nur die Wahrscheinlichkeit für ein konkretes Ergebnis unter der Annahme der Nullhypothese ermittelt und damit auch nur eine Entscheidung darüber, ob die Nullhypothese begründet verworfen werden kann. Normalerweise ist damit keine Aussage darüber verbunden, wie wahrscheinlich die jeweiligen Ergebnisse unter der Annahme einer jeweils spezifischen Alternativhypothese sind. Der übliche Signifikanztest sichert ein bestimmtes Ergebnis also nur gegen den Zufallsfehler, indem er überprüft, wie wahrscheinlich das empirische Ergebnis unter Zufallsbedingungen erzeugt werden kann. Dieser Fehler wird gemeinhin als der α-Fehler oder Fehler erster Art bezeichnet.

Hiermit entsteht jedoch eine weitere Fehlerquelle: es ist durchaus möglich, die Nullhypothese unzutreffenderweise beizubehalten. Dieser Fehler zweiter Art (oder β-Fehler) läßt sich allerdings nur dann bestimmen, wenn eine Alternativhypothese als Punkthypothese spezifiziert wird. In diesem Fall läßt sich die Wahrscheinlichkeitsverteilung der Teststatistik unter der Annahme der Alternativhypothese berechnen. Hierbei lassen sich vier Fälle unterscheiden:

1. Im unproblematischen Fall fällt jedes Ergebnis, das unter der Annahme von H_1 sehr wahrscheinlich ist, in den Ablehnungsbereich von H_0. In diesem Fall ist die Beibehaltung von H_0 mit einem relativ geringen Fehler zweiter Art behaftet.

2. Ein problematischer Fall tritt dann auf, wenn die unter H_1 und H_0 erwarteten Stichprobenstatistiken sehr nah beieinander liegen. Werte, die in den "Annahmebereich" von H_0 fallen, sind auch unter der Annahme daß H_1 gilt, sehr wahrscheinlich. Das β-Fehlerrisiko ist dementsprechend sehr hoch.

3. Derselbe Fall tritt ein, wenn die untersuchte Stichprobe relativ klein ist, da die Varianz der Prüfverteilung mit sinkendem Stichprobenumfang anwächst. Auch hier überschneiden sich beide Verteilungen in einem solchen Maße, daß die Ablehnung von H_1 mit einem unakzeptabel hohen Fehlerrisiko zweiter Art belastet ist. Bei der χ^2-Verteilung tritt dieser Fall natürlich auch dann auf, wenn die Zahl der Freiheitsgrade wächst.

4. Auch die Senkung des α–Niveaus (d.h. die Annahme eines 1% oder 0,1% Niveaus) führt zum selben Resultat.

Das bedeutet: Ein Fehler zweiter Art tritt vor allem dann sehr wahrscheinlich auf, wenn

1. sehr konservativ getestet wird (d.h. nur sehr oder hoch signifikante Ergebnisse akzeptiert werden),

2. die unter H_0 und H_1 erwarteten Werte relativ zur Stichprobengröße sich nur wenig voneinander unterscheiden,

3. die Stichprobe relativ klein bzw. die Anzahl der Freiheitsgrade relativ hoch ist, d.h. wenn der geprüfte Zusammenhang relativ komplex ist.

Dies führt zu dem bekannten Effekt, daß relativ kleine und deswegen oft relativ unbedeutsame Unterschiede bei großen Stichproben relativ häufig statistisch signifikant werden oder umgekehrt: daß deutliche Effekte und starke Unterschiede bei kleinen Stichproben relativ selten statistisch signifikant werden. Oder anders ausgedrückt, mit sinkendem Stichprobenumfang sinkt die Wahrscheinlichkeit einen empirisch bedeutsamen Effekt inferenzstatistisch begründet nachzuweisen, gesetzt den Fall, man verwendet die übliche Teststrategie.

Welche Folgen dieser Umstand insbesondere für die statistische Untersuchung von Lebensläufen haben kann, soll hier exemplarisch anhand einer Studie über Berufsbiografien älterer Frauen (KRÜGER, BORN, KELLE 1989; BORN 1993; ERZBERGER 1993) aufgezeigt werden. Die Untersuchungspopulation bildete eine Kohorte von Frauen in zwei ausgewählten Regionen, die kurz nach Kriegsende ihre Ausbildung in einem von fünf Ausbildungsberufen (Friseurinnen, Verkäuferinnen, Schneiderinnen, Kaufmännische Angestellte und Kinderpflegerinnen) abgeschlossen hatten. Eines der Untersuchungsziele war die Bestimmung erklärender Faktoren für den Ablauf weiblicher Berufsbiographien. Hierzu wurden Erwerbsbiografien auf der Basis zahlreicher Merkmale einer Clusteranalyse unterzogen, wobei eine Vier-Clusterlösung die besten Resultate (i.S. hoher Varianzaufklärung) erbrachte. Die folgenden vier “erwerbsbiographischen Cluster” konnten dabei identifiziert werden:

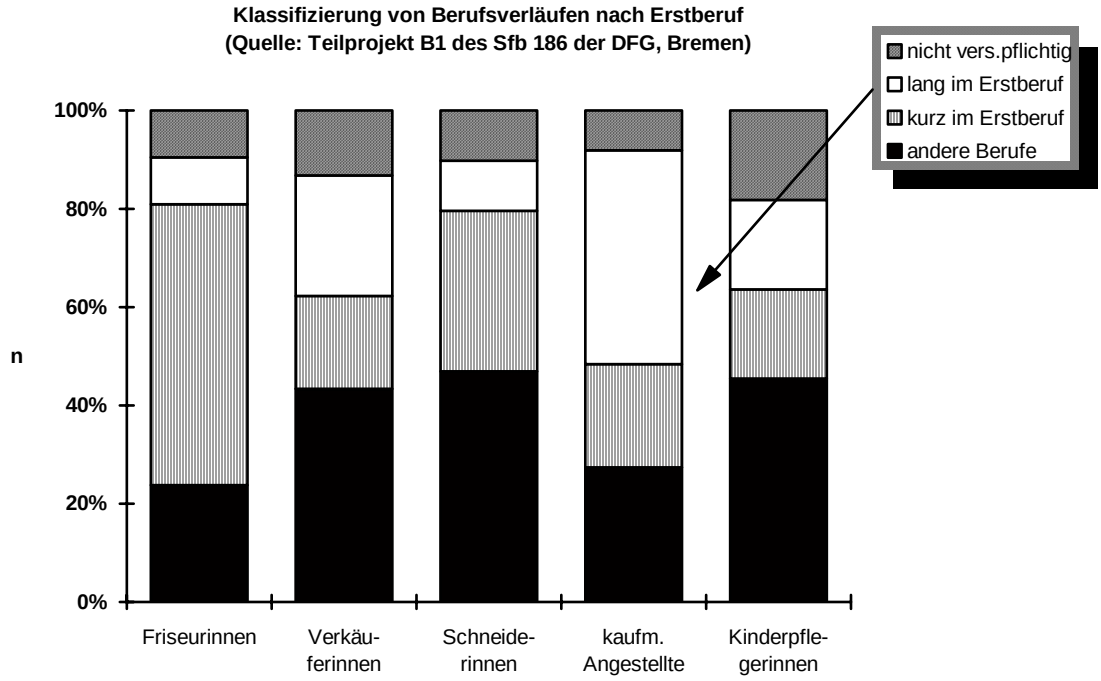
1. Frauen, die nur eine kurze Zeit im erlernten Beruf tätig gewesen sind, und dann zumeist als Hausfrau oder in **geringfügigen Beschäftigungsverhältnissen unterhalb der Sozialversicherungspflichtgrenze** tätig gewesen waren,

2. Frauen, die im Verlauf ihrer **gesamten Erwerbsbiografie überwiegend im gelernten Beruf** tätig gewesen waren,

3. Frauen, die eine **kurze Zeit im gelernten Beruf** gearbeitet hatten (durchschnittlich etwa 7 Jahre) um anschließend ganz aus dem Erwerbsleben auszuscheiden,

4. Frauen, die die **meiste Zeit in anderen Berufen** als ihrem gelernten Erstberuf tätig gewesen waren.

Einer klassischen These der Berufssoziologie zufolge wird das Erwerbsverhalten von Frauen vor allem durch den Erwerbsstatus des Mannes beeinflußt, so daß bspw. Frauen von Arbeitern aufgrund ökonomischer Zwänge eine höhere Erwerbsbeteiligung aufweisen als Frauen von Angestellten oder Beamten. Dahingegen konnte aufgrund des erhobenen Datenmaterials festgestellt werden, daß die Zugehörigkeit zu einem der "erwerbsbiographischen Cluster" in hohem Maße abhängig vom gelernten Erstberuf war. So zeigte sich etwa, daß weibliche kaufmännische Angestellte wesentlich öfter zur Gruppe derjenigen Frauen gehörten, welche fast ihr gesamtes Leben im gelernten Erstberuf tätig waren als bspw. Schneiderinnen, während der Anteil der Friseurinnen in der Gruppe derjenigen Frauen, die nach durchschnittlich 7 Jahren ihren Beruf verließen, um danach vorwiegend Familienarbeit zu leisten, deutlich höher war als der Anteil anderer Berufsgruppen. Die folgende Grafik zeigt die Aufteilung der vier erwerbsbiografischen Cluster auf die fünf untersuchten Berufe



Es lagen also nunmehr zwei konkurrierende Modelle zur Erklärung weiblicher Erwerbsbeteiligung im Lebenslauf vor:

- Gemäß dem *konventionellen Ansatz* ist die Dauer weiblicher Erwerbstätigkeit im Lebenslauf im wesentlichen abhängig vom **Berufsstatus des Mannes** und der damit verbundenen **ökonomischen Situation der Familie**.

- Aufgrund des in der erwähnten Studie *erhobenen Datenmaterials* konnte davon ausgegangen werden, daß der **erlernte Erstberuf der Frau** einen zusätzlichen maßgeblichen Einfluß besitzt.

Um diese beiden konkurrierenden Modelle zu beurteilen, wurde in einer weiteren Befragung zusätzlich die Erwerbsbiographie der Ehemänner der anfangs befragten Frauen erhoben. Eine hohe *panel attrition* führte dabei zu hohen Ausfällen (ein Effekt, der dadurch verschlimmert wurde, daß die Adressen des Ausgangssamples aufgrund überzogener administrativer Datenschutzmaßnahmen nicht mehr im vollen Umfang verfügbar waren), so daß nur noch eine Fallzahl von 74 Ehepaaren realisiert wurde. Die folgenden Ergebnisse wurden von uns im Rahmen einer querschnittsorientierten Zwischenauswertung ermittelt. Ihre Darstellung soll an dieser Stelle nicht der Diskussion berufssoziologischer Theorien dienen, sondern der Exemplifizierung der eingangs dargestellten methodologischen Probleme, die sich aus kleinen Fallzahlen ergeben.

Für die Untersuchungsstichprobe wurde der überwiegende Erwerbsstatus des Mannes, der gelernte Erstberuf der Frau, sowie die Zugehörigkeit der Ehefrau zu einem der vier erwerbsbiographischem Cluster ermittelt. Aufgrund der kleinen Fallzahl wurden die Variablen stark vereinfacht. Beim Erstberuf der Frau wurde zwischen kaufmännischen Angestellten und anderen Berufen unterschieden, beim Beruf des Mannes zwischen Arbeitern, Angestellten bzw. Beamten sowie Selbständigen, bei der Berufsbiographie der Frau, ob sie jenem erwerbsbiographischen Cluster von Frauen angehörte, die die überwiegende Zeit ihres Lebens in ihrem gelernten Beruf tätig war, oder nicht.

Beruf des Mannes ⇒	**Arbeiter**		**Angestellter / Beamter**		**Selbständiger**	
Erstberuf ⇒ Erw.tätigkeit i. Erstberuf ⇓	kaufm. Angestellte	sonstige	kaufm. Angestellte	sonstige	kaufm. Angestellte	sonstige
lang	2	3	3	2	0	0
kurz	1	10	13	28	3	6

Quelle: Erhebung des Teilprojektes B1 des Sonderforschungsbereichs 186 der DFG in Bremen

Die beiden konkurrierenden Kausalmodelle über die Determinanten weiblicher Erwerbsbiographien konnten nun in Form der folgenden Hypothesen operationalisiert werden:

- Dem traditionellen Ansatz zufolge müssen Frauen von Arbeitern öfter jenem erwerbsbiographischen Cluster bzw. jener Gruppe von Frauen angehören, die lange Zeit im erlernten Erstberuf tätig waren, *und das unabhängig von ihrem gelernten Erstberuf.* Dies wollen wir als die *Nullhypothese* bezeichnen.

- Die *Alternativhypothese* schließt an die bereits dargestellten Ergebnisse der ersten Studie an, derzufolge wir davon ausgehen müssen, daß Frauen, die einen kaufmännischen Beruf gelernt haben, häufiger als andere jenem erwerbsbiographischem Cluster bzw. jener Gruppe von Frauen angehören, die lange Zeit in ihrem gelernten Erstberuf tätig waren, *und das unabhängig vom Berufsstatus des Ehemannes.*

Um diese beiden konkurrierenden Hypothesen zu überprüfen, wurden multivariate Verfahren zur Analyse mehrdimensionaler Kontigenztafeln (log-lineare Modelle) benutzt. In einem log-linearen Modell mit drei Variablen lassen sich Nullhypothese und Alternativhypothese folgendermaßen spezifizieren:

H_0: Es besteht kein Zusammenhang zwischen gelernten Erstberuf und dem Ausmaß weiblicher Erwerbstätigkeit im Lebenslauf, evtl. Unterschiede im Datenmaterial können allein durch den Berufsstatus des Mannes erklärt werden (bedingte Unabhängigkeit).

H_1: Unterschiede zwischen weiblichen Erwerbsverläufen sind einerseits auf den Berufsstatus des Mannes und andererseits auf den gelernten Erstberuf der Frau zurückzuführen.

In den Daten zeigte sich nun ein signifikanter Zusammenhang zwischen dem männlichen Erwerbsstatus und der Clusterzugehörigkeit der Frau, wohingegen eine Interaktion zwischen dem Erwerbsstatus des Mannes und dem Ausbildungsberuf der Frau nicht festgestellt werden konnte. Problematischer war hingegen der Interaktionseffekt zwischen dem Erstberuf der Frau und deren Clusterzugehörigkeit, (d.h. der Effekt, der zuvor in dem umfangreicheren Datensatz mit nur weiblichen Erwerbsverläufen hatte identifiziert werden können): Bei diesem Effekt waren weder die Effektparameter gegen Null gesichert, noch war der Partial-χ^2 für diesen Effekt auf dem 5%-Niveau signifikant.

Strategien zur Schätzung des Fehlerrisikos zweiter Art[2]

Eine traditionelle Teststrategie, bei der die Entscheidung für oder oder gegen ein bestimmtes Modell allein aufgrund der Höhe des α-Fehlers getroffen wird, würde zur Beibehaltung des konventionellen Modells und zur Verwerfung der Hypothese vom Einfluß des Erstberufs führen. Da in diesem Fall sowohl der Umfang der Stichprobe relativ klein als auch der zu testende Effekt relativ (d.h. im Vergleich zum Stichprobenumfang) schwach ausgeprägt war, mußten wir jedoch bei der Verwerfung von H_1 mit einem relativ hohen Fehler zweiter Art rechnen.

Strategien zur Schätzung des Fehlers zweiter Art sind zwar in der sozial- und verhaltenswissenschaftlichen Literatur bereits seit Ende der siebziger Jahre bekannt (vgl. COHEN 1988; WITTE 1980; AGRESTI 1990; HAGENAARS 1990; eine Übersicht bei PREIN, KLUGE, KELLE 1993), werden jedoch in der Praxis fast nie angewendet. Die bei der Modellbildung mit kategorialen Daten anzuwendenden Verfahren beruhen auf der Anwendung der bekannten χ^2-Teststatistiken. Diese sind bei ausreichender Stichprobengröße asymptotisch χ^2-verteilt unter der Annahme der Nullhypothese und – unter der Annahme der Alternativhypothese – nichtzentral χ^2-verteilt mit den Nichtzentralitätsparameter l, der für die Likelihood-Statistik wie folgt bestimmt wird:

$$\lambda = 2n\sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij}\ln\frac{\pi_{ij}}{\pi_{ij}(M)} \qquad (1)^3$$

Aus der Stichprobengröße, dem gewählten α-Niveau sowie diesem Nichzentralitätsparameter läßt sich die Wahrscheinlichkeit eines β-Fehlers bestimmen als Wahrscheinlichkeit, daß ein nichtzentraler χ^2 mit n Freiheitsgraden und einem Nichtzentralitätsparameter l kleiner oder gleich dem χ^2-Wert für n Freiheitsgrade bei einem spezifizierten α-Niveau ist:

$$p(\text{Fehler 2. Art}) = p[X^2{}_{\nu\lambda} \le \chi^2{}_{\nu}(\alpha)] \qquad (2)$$

In dem von uns gezeigten Beispiel betrug die Wahrscheinlichkeit für einen Fehler zweiter Art asymptotisch berechnet .82.

Allerdings ist die Anwendung dieses Verfahren bei kleinen und schief verteilten Stichproben, die zu kleinen erwarteten Zellenbesetzungen führen, mit schwerwiegenden Problemen

---

[2] Die folgenden teststatistischen Überlegungen und die dargestellten Verfahren zur Bestimmung des β-Fehlerrisikos sind detailliert und ausführlich dargestellt in PREIN, KELLE 1994 sowie in PREIN, KLUGE, KELLE 1993.

[3] In dieser Formel kennzeichnet n den Stichprobenumfang, r die Anzahl der Zeilen, c die Anzahl der Spalten, π_{ij} die wahre Zellenwahrscheinlichkeit in Zelle ij, $\pi_{ij}(M)$ die Zellenwahrscheinlichkeit in Zelle ij gemäß Modell M.

behaftet. In diesen Fällen gilt die Anwendung asymptotischer Approximationen nicht als *arte legis*, da bei der Bestimmung der Parameter der Prüfverteilung erhebliche Schätzfehler auftreten können. Wann die Wahrscheinlichkeit für einen Schätzfehler inakzeptabel hoch ist, wird in der Regel nur aufgrund eingebürgerter "Daumenregeln" wie "Cochran´s Kriterium" entschieden, wonach keine erwartete Häufigkeit unter eins und höchstens ein Fünftel der erwarteten Häufigkeiten unter fünf liegen sollte. Eine wesentlich bessere Lösung für dieses Problem bietet die Anwendung von Verfahren zur Benutzung von exakten Prüfverteilungen oder von Monte-Carlo-Verfahren zur Schätzung der Parameter der exakten Prüfverteilung, wie sie in den letzten Jahren für den Bereich der Biostatistik (MEHTA, PATEL 1983) entwickelt wurden.

Solche Verfahren zur Bestimmung von Prüfverteilungen bei kleinen Stichproben werden jedoch bislang nur im Rahmen konventioneller Teststrategien eingesetzt. Da, wie bereits dargestellt, gerade hier die Gefahr einer ungerechtfertigten Zurückweisung der Alternativhypothese sehr hoch ist, kann deren Benutzung nicht unbedingt angeraten werden: Das Risiko für einen Fehler zweiter Art, d.h. das Risiko, daß eine auf Basis des Datenmaterials relativ plausible Alternativhypothese zurückgewiesen werden muß, weil die Fallzahl zu klein ist, ist hier relativ hoch.

Eine solche Ausgangslage erfordert die Anwendung von exakten Verfahren oder von Monte-Carlo-Methoden zur Schätzung der Fehlerrisikos, ein Vorgehen, daß bislang in der sozialwissenschaftlichen Forschungspraxis u. W. noch nie angewendet wurde. Aus diesem Grunde mußten wir, da entsprechende Algorithmen in der Literatur nicht beschrieben werden, einen Algorithmus zur Monte-Carlo-Simulation einfacher, hierarchischer log-linearer Modell (mit direkten Schätzern) entwickeln[4]. Die Anwendung dieses Verfahrens auf unser Beispiel zeigt tatsächlich, daß die üblichen asymptotischen Approximationen den Fehler zweiter Art deutlich überschätzen. Trotzdem würde in diesem Fall bei der Zurückweisung der Alternativhypothese mit einer Wahrscheinlichkeit von .68 ein β-Fehler begangen.

Schlußfolgerungen für einen rationalen Einsatz von Signifikanztests

Bei kleinen Stichproben kann eine konventionelle Teststrategie, bei der nur das Fehlerrisiko erster Art ermittelt wird, vielfach nicht als Grundlage für eine rationale

[4] Der Algorithmus ist abgedruckt in PREIN, KLUGE, KELLE 1993.

Entscheidungsstrategie dienen, da mit sinkendem Stichprobenumfang die Wahrscheinlichkeit steigt, daß Alternativhypothesen ungerechtfertigterweise zurückgewiesen werden. Eine rationale Entscheidungsstrategie muß als zusätzliche Information das Risiko des Fehlers zweiter Art berücksichtigen.

Während bei einer konventionellen Teststrategie zwei Entscheidungen möglich sind, nämlich die Verwerfung oder Beibehaltung der Nullhypothese, muß bei dem von uns vorgeschlagenen Vorgehen mit vier Möglichkeiten gerechnet werden:

	$X^2 > \chi^2(\alpha)$	$X^2 \leq \chi^2(\alpha)$
p(Fehler 2. Art) gering	*(1)* *Verwerfe H_0, akzeptiere H_1*	*(2)* *Behalte H_0 bei, verwerfe H_1*
p(Fehler 2. Art) hoch	*(3)* *Verwerfe H_0, akzeptiere H_1 nicht*	*(4)* *Verwerfe H_0 nicht, suche nach Evidenz oder Gegenevidenz für H_1*

In den Fällen (1) und (2) kann wie gewohnt verfahren werden: dann, wenn das Fehlerrisiko zweiter Art gering ist, führt eine Unterschreitung des kritischen α-Niveaus zur Zurückweisung der Nullhypothese, eine Überschreitung zu deren Beibehaltung. In diesen beiden Fällen ist eine Entscheidung über die Verwerfung der Null- oder der Alternativhypothese auf der Basis des vorliegenden Datenmaterials problemlos möglich.

Die Fälle (3) und (4) beziehen sich dahingegen auf jene problematischen Situationen, die bei der Anwendung konventioneller Teststrategien nicht aufgedeckt werden können:

Fall (3): Liegen die in der Stichprobe ermittelten Kennwerte sowohl hinsichtlich der Nullhypothese als auch der Alternativhypothese unterhalb der kritischen Werte, müssen sowohl Alternativ- als auch Nullhypothese abgelehnt werden. Der in der Stichprobe beobachtete Zusammenhang ist dann sowohl unter Annahme der Nullhypothese als auch unter der Annahme der Alternativhypothese gleichermaßen unwahrscheinlich, jedenfalls unwahrscheinlicher, als wir es zu akzeptieren bereit sind. (Bei der multivariaten Modellbildung könnte dies etwa der Fall sein bei einem stark simplifizierten Modell).

Fall (4): Liegen die ermittelten Prüfstatistiken in beiden Fällen oberhalb der kritischen Marke für den Fehler erster und zweiter Art, so kann weder Null- noch Alternativhypothese verworfen werden. Das beobachtete Ergebnis ist sowohl wahrscheinlich unter der Annahme des von uns postulierten theoretischen Zusammenhangs, als auch dann, wenn ein zufälliger Effekt vorausgesetzt wird.

In den beiden letztgenannten Fällen ist auf der Basis des vorliegenden Datenmaterials keine rational begründete Entscheidung über die Geltung oder Zurückweisung einer statistischen Hypothese möglich; vielmehr zeigt ein solches Ergebnis, daß die verwendete Datenbasis zu schmal ist, um eine statistisch begründete Aussage zu treffen. Bei einer konventionellen Inferenzstrategie müßte jedoch im Fall (4) die Alternativhypothese zurückgewiesen werden, unabhängig davon, wie wahrscheinlich sie angesichts des vorliegenden Datenmaterials erscheint. Eine erweiterte Teststrategie zeigt jedoch, daß es in diesem Fall sinnvoll ist, die Entscheidung aufzuschieben, und sich um weitere empirische Evidenz bemühen (d.h. bspw. weitere zusätzliche Daten zu sammeln) Die von uns vorgeschlagene Entscheidungsstrategie soll also dazu dienen, jene Fälle zu identifizieren, in denen ein konservatives Vorgehen bei der Testung von Hypothesen zu schwerwiegenden Fehlentscheidungen führen kann. Eine Teststrategie, die die Abschätzung des Fehlerrisikos zweiter Art mit einbezieht, kann demgegenüber helfen, jene Fälle aufzufinden, in denen das empirische Material als Basis für eine rationale Entscheidung für oder gegen eine getestete Hypothese nicht ausreichend ist.

Ein wissenschaftstheoretisches Addendum:

Der Signifikanztest ist fester Bestandteil der “Folklore der Forschung” und wird oft intuitiv als Grundlage entweder einer dogmatisch verifikationistischen oder dogmatisch falsifikationistischen Forschungslogik misinterpretiert:

- Im Rahmen einer verifikationistischen Forschungslogik wird ein “erfolgreicher” Signifikanztest fälschlich als “Beweis” für eine Alternativhypothese betrachtet. Dies ist schon deswegen falsch, da ein klassischer Signifikanztest Wahrscheinlichkeiten nur unter der Annahme der Nullhypothese zu bestimmen hilft und hieraus keine Aussage über die Wahrscheinlichkeit eines empirischen Ergebnisses unter der Annahme einer spezifizierten Alternativhypothese logisch ableitbar ist. Methodologisch stellt ein Signifikanztest den Versuch dar, eine bestimmte Art von (Zufalls)fehlern auszuschließen, er kann allerdings nicht dazu dienen, eine bestimmte Hypothese zu “belegen”.

- Da die Beibehaltung der Nullhypothese immer mit dem Risiko eines Fehlers zweiter Art verbunden ist, kann ein Signifikanztest auch nicht zur strengen “Falsifikation” einer Hypothese im Sinne einer dogmatischen Falsifikationslogik genutzt werden. Auch Karl Popper, der von einem naiven und dogmatischen Falsifikationismus im Gegensatz zu vielen seiner Epigonen weit entfernt war (vgl. LAKATOS 1982), hat i.ü. darauf aufmerksam

gemacht, das Falsifikationen immer nur auf der Basis konventioneller Übereinkünfte der Forschergemeinschaft und unproblematisierter Hintergrundannahmen gelten (POPPER 1989, S.69ff.). Diese Übereinkünfte und diese Annahmen können jedoch, wie in der Wissenschaftsgeschichte vielfach geschehen, problematisiert und damit der Überprüfungsvorgang von neuem in Gang gesetzt werden.

Signifikanztests können also weder Beweise noch Widerlegungen von Hypothesen im strengen Sinne darstellen, sie liefern nur, wenn alle verfügbaren Informationen ausgeschöpft worden sind und die Datenbasis korrekt erhoben wurde, gerechtfertigte Gründe für rationale Zusicherungen, sie dienen als *"Gebrauchswerkzeug beschränkter Intelligenzen zur Gewinnung nicht der wirklich und wahrhaftig bestmöglichen (...) sondern der besterreichbaren Antwort — der besten Antwort, der wir uns unter vorhandenen Bedingungen versichern können"* RESCHER 1987, S.21) Die konventionelle Vorgehensweise bei der Anwendung des Signifikanztests stellt jedoch in zahlreichen Fällen ein ungenügendes Gebrauchswerkzeug dar, welches den Erkenntnisfortschritt mehr behindert als ihn fördert. Wir plädieren deshalb dafür, bei der Analyse kleiner Stichproben die konventionelle Teststrategie so zu erweitern, daß der Erkenntnisgewinn nicht durch ein übermäßig konservatives und rigoroses Testen behindert und der Bedeutung statistischer Signifikanz bei kleinen Stichproben Rechnung getragen wird.

LITERATUR:

AGRESTI, ALAN (1990): *Categorical Data Analysis*. New York: John Wiley & Sons.

BEHRENS, JOHANN (1992): In Würde alt werden im Handwerk. In: IKK-BUNDESVERBAND (Hg.): *Fachtagung innovative Perspektiven der Gesundheitsförderung im Zeichen der weiteren Gesundheitsreformpolitik.* Bonn.

BEHRENS, JOHANN (1993a): Strategien der Gesetzlichen Kranken- und Rentenversicherung bei begrenzter Tätigkeitsdauer, Befragungs- und Diskussionsergebnisse. In: FRIEDRICH-EBERT-STIFTUNG (Hg.): *Betrieblicher Gesundheitsschutz auf dem Prüfstand.* Bonn.

BEHRENS, JOHANN (1993b): Erste Ergebnisse einer Rehabilitationsverlaufsstudie. Vortrag gehalten auf der Herbsttagung der Sektion Medizinsoziologie der DGS in Bremen.

BORN, CLAUDIA (1993): Abhängigkeiten zwischen ehepartnerlichen Erwerbsverläufen in der BRD – Dilemmata und Dissonanzen zwischen Struktur und Norm. In: BORN, CLAUDIA, KRÜGER, HELGA (Hg.) (1993): *Erwerbsverläufe von Ehepartnern und die Modernisierung weiblicher Lebensläufe.* Weinheim: Deutscher Studien Verlag. S.71-88.

BORN, CLAUDIA, KRÜGER, HELGA (Hg.) (1993): *Erwerbsverläufe von Ehepartnern und die Modernisierung weiblicher Lebensläufe.* Weinheim: Deutscher Studien Verlag.

COHEN, JACOB (1988): *Statistical Power Analysis for the Behavioral Sciences* (1969). Hillsdale, N.J: Lawrence Erlbaum Ass.

COOK, THOMAS D.; CAMPBELL, DONALD T. (1979): *Quasi-Experimentation.* Design & Analysis Issues for Field Settings. Boston: Houghton Mifflin Company

ERZBERGER, CHRISTIAN (1993): *Erwerbsarbeit im Eheleben. Männlicher und weiblicher Erwerbsverlauf zwischen Dependenz und Unabhängigkeit.* Bremen: Arbeitspapier Nr. 16 des Sfb 186.

GERHARDT, UTA (1986): *Patientenkarrieren*. Ein medizinsoziologische Studie. Frankfurt/M.: Suhrkamp

GÖCKENJAN, GERD (1993a): Old Age caught between State and Medical Profession Interests . In: *Dynamis, Acta Hispanica*, Vol 13.

GÖCKENJAN, GERD (1993b): Hilfebedürftigkeit als Rahmung der Statuspassage ins hohe Alter. In: LEISERING, LUTZ u.a.: *Moderne Lebensläufe im Wandel.* Weinheim: Deutscher Studien Verlag.

HAGENAARS, JACQUES (1990): *Categorical Longitudinal Data.* Log-Linear, Panel, Trend, and Cohort Analysis. Newbury Park; London; New Delhi: Sage

HEINZ, WALTER (Hg.) (1991): *The Life Course and Social Change: Comparative Perspectives.* Weinheim: Deutscher Studien Verlag

HEINZ, WALTER (Hg.) (1993): *Institutions and Gatekeeping in the Life Course.* Weinheim: Deutscher Studien Verlag.

KASPRZYK, DANIEL; DUNCAN, GREG; KALTON, GRAHAM; SINGH, M.P. (eds.) (1989): *Panel Surveys.* New York: John Wiley & Sons.

LAKATOS, IMRE (1982): *Die Methodologie der wissenschaftlichen Forschungsprogramme*. Philosophische Schriften, Bd.1 Wiesbaden: Vieweg.

MEHTA, C.R.; PATEL, N.R. (1983): A Network Algorithm for performing Fisher´s Exact Test in r * c Contingency Tables. *Journal of the American Statistical Association.* 78, S. 427-434.

NEYMAN, J; PEARSON, E.S. (1967): *Joint Statistical Papers.* Cambridge: University press.

PREIN, GERALD; KLUGE, SUSANN; KELLE, UDO (1993): *Strategien zur Sicherung von Repräsentativität und Stichprobenvalidität bei kleinen Samples.* Arbeitspapier Nr. 18 des Sonderforschungsbereichs 186. Bremen

PREIN, GERALD; KELLE, KLAUS-UDO (1994): Estimation of β-Error in Multivariate Modelling with Small Samples. In: FAULBAUM, FRANK (Hg.): *SOFTSTAT '93 - Fortschritte der Statistik-Software.* Stuttgart: Enke (im Erscheinen).

POPPER, KARL (1989): *Logik der Forschung.* (1934) Tübingen: J.C.B. Mohr.

RESCHER; NICHOLAS (1987): *Induktion. Zur Rechtfertigung induktiven Schließens.* München, Wien: Philosophia.

WITTE, ERICH H. (1980): *Signifikanztest und statistische Inferenz.* Analysen, Probleme, Alternativen. Stuttgart: Enke