Strategien zur Sicherung von Repräsentativität und Stichprobenvalidität bei kleinen Samples

von
Gerald Prein, Susann Kluge und Udo Kelle

Arbeitspapier Nr. 18

Herausgeber: Der Vorstand des Sfb 186
2. Auflage, Bremen 1994

Gerald Prein
Susann Kluge
Udo Kelle
Universität Bremen
Sonderforschungsbereich 186
Bereich Methoden und EDV
Wiener Str. – FVG-West
28359 Bremen
Tel.: x49 421 218-4168/-4169
FAX: x49 421 218-4153
Email: gprein@sfb186.uni-bremen.de
ukelle@sfb186.uni-bremen.de

Inhaltsverzeichnis

Vorwort

In diesem Arbeitspapier werden Teilergebnisse der Arbeit des Bereichs "Methoden und EDV" des Sonderforschungsbereichs 186 vorgestellt. Ausgehend von der Konzeption der *forschungsbegleitenden Methodenentwicklung*, bei der forschungspraktische Probleme der Teilprojekte den Ausgangspunkt methodischer bzw. methodologischer Reflexion bilden, stand die Frage der Validitätssicherung im Kontext der unterschiedlichen Methodologien im Zentrum der Arbeit. Im vorliegenden Papier wird hierbei vor allem auf Problembereiche statistisch-quantitativer Analysen eingegangen.

Angemessene inferenzstatistische Methoden für kleine Stichproben und Verfahren zur Absicherung von deren Validität werden in der sozialwissenschaftlichen Methodenliteratur noch immer stiefmütterlich behandelt, obwohl bei der statistischen Modellierung sozialwissenschaftlicher Daten aus zwei Gründen hier ein großer Bedarf besteht: (1.) Wenn die untersuchten gesellschaftlichen Phänomene auf komplexe, multikausale Bedingungskonstellationen verweisen, wird das Problem kleiner Fallzahlen auch dann virulent, wenn anfänglich umfangreiche Samples gezogen wurden. (2.) Bei der Befragung von Bevölkerungsgruppen, deren Erreichbarkeit oder Antwortbereitschaft – wie etwa im Falle delinquenter Jugendlicher oder Nichtseßhafter – als gering anzusehen ist, kommt es häufig zu hohen Ausfällen bzw. bei Wiederholungsbefragungen zu einer hohen Panelmortalität.

Kleine Fallzahlen bringen grundsätzlich zwei verschiedene methodische Fragestellungen mit sich: zum einen nach der *Repräsentativität der Stichprobe*, zum anderen nach adäquaten *Inferenzstrategien*. Der Bereich "Methoden und EDV" hat in enger Kooperation mit verschiedenen Teilprojekten des Sonderforschungsbereichs 186 zu diesen Fragenkomplexen Lösungsansätze entwickelt, die in diesem Arbeitspapier vorgestellt werden.

Ansgar Weymann
Stellvertretender Sprecher des Sfb 186

Many small-scale experiments with local control and choice of measures is in many ways preferable to giant national experiments with a promised standardisation that is neither feasible nor even desirable from the standpoint of making irrelevancies heterogeneous. (COOK, CAMPBELL 1979, S. 80)

1. Einleitung

1.1 Zum Problem der Repräsentativität kleiner Stichproben

Der Begriff der Repräsentativität wird – selbst in großen Teilen der gängigen Methodenliteratur – wenig problematisiert. Insbesondere suggerieren zahlreiche Umfrageforschungen mit der Zusatzbemerkung *"auf der Basis einer Repräsentativbefragung von x Bundesbürgern"*, daß die Sozialforschung ein passendes "Rezept" gefunden hätte, das eine problemlose Generalisierung der Ergebnisse erlaubt.

Dabei ist die Verwendung des Repräsentativitätsbegriffs in der methodologischen Diskussion in hohem Maße unscharf und kann zu tautologischen Aussagen führen wie das folgende Beispiel zeigt: *"Eine Stichprobe ist repräsentativ, wenn aus ihr der Schluß auf die zugrunde gelegte Grundgesamtheit erlaubt ist."* (KREIENBROCK 1989, S. 9) Eine solche Definition führt nicht weiter, da es gerade das Ziel einer repräsentativen Stichprobenauswahl ist, von einer Stichprobe auf die zugrunde gelegte Grundgesamtheit zu schließen.

Diese unbefriedigende Situation hat dazu geführt, daß der Repräsentativitätsbegriff inzwischen kritisch diskutiert wird. Hierbei kann grob zwischen einer prinzipienorientierten und einer relativistischen Position unterschieden werden:

- erstere bezeichnet ihn als *"ungenau und unnötig"*, da als entscheidendes Kriterium für die Beurteilung der Güte einer Stichprobe in der Sozialforschung nur gelten könne, ob die Auswahl eine *Zufallsauswahl* darstellt oder nicht (vgl. SCHNELL ET AL. 1989, S. 281);

- letztere bezieht den Begriff der Repräsentativität auf die Relevanz für die Forschungsfragestellung, die durch die Untersuchung der jeweiligen Stichprobe beantwortet werden soll: Es gäbe keine Repräsentativität als solche, da eine Stichprobe immer nur im Hinblick auf bestimmte Merkmale oder Merkmalskombinationen repräsentativ sein könne. Der entscheidende Punkt sei die Relevanz dieser Merkmale für die Forschungsfrage (vgl. KRIZ, LISCH 1988, S. 220).

Hierbei wird deutlich, daß sich auch in dieser Diskussion zwei Ebenen vermischen:

- zum einen geht es um eine *methodische Strategie*, mit deren Hilfe Repräsentativität sichergestellt werden soll: in Stichproben, die über Verfahren der Zufallsziehung gewonnen wurden, ist die Wahrscheinlichkeit der Unverzerrtheit bezüglich aller möglichen Merkmale hoch *und* Konfidenzintervalle statistischer Parameter sind berechenbar;

- zum anderen wird Repräsentativität als *empirisches Konzept* benutzt: Repräsentativität einer Stichprobe bezeichnet in diesem Sinne die Tatsache, daß sie ein verkleinertes, unverzerrtes Abbild einer Grundgesamtheit in bezug auf die theoretisch relevant erachteten Merkmale darstellt.

Um die Sprachverwirrung zumindest in diesem Papier zu begrenzen, bezeichnet der Begriff der **statistischen Repräsentativität einer Stichprobe** hier die Tatsache, daß sich *alle* Merkmale von Elementen sowie deren Kombinationen in der Stichprobe mit einer bestimmbaren Wahrscheinlichkeit wie in der Grundgesamtheit verteilen, da *alle* Elemente der Grundgesamtheit aufgrund des Ziehungsverfahrens die *gleiche Chance* haben, in die Stichprobe aufgenommen zu werden. Der Begriff der statistischen Repräsentativität ist damit ein probabilistisches Konzept, das methodische Implikationen der Stichprobenauswahl beinhaltet. Demgegenüber soll der Begriff der

theorieorientierten Repräsentativität ausdrücken, daß eine Stichprobe bezüglich aller theoretisch relevanten Merkmale unverzerrt ist.

In der Diskussion um den Repräsentativitätsbegriff wird neben der Frage des Auswahlverfahrens die notwendige *Stichprobengröße* diskutiert. Allgemein bekannt ist, daß sich im Falle einer Zufallsstichprobe mit wachsendem Umfang die Konfidenzintervalle geschätzter statistischer Parameter verkleinern. Schnell, Hill und Esser folgern daraus, daß es für die Einschätzung des notwendigen Stichprobenumfangs wesentlich ist, wie stark die Variabilität der beobachteten Merkmale ist und welche Konfidenzintervalle als tolerabel angesehen werden können. Stichprobengrößen zwischen 1.000 und 3.000 werden von ihnen – in Anlehnung an Scheuch (1974, S. 51f) – i.d.R. für Aussagen über die Gesamtbevölkerung als ausreichend erachtet (SCHNELL ET AL. 1989, S. 259f). Diese Richtzahlen können jedoch – wie Schnell, Hill und Esser bemerken – im Falle von stark streuenden Merkmalen dennoch unzureichend sein.

Eine generelle Gleichsetzung "repräsentative Stichprobe = große Stichprobe" wird auch von Kriz und Lisch kritisiert: *"Der Begriff der Repräsentativität wird häufig falsch gebraucht. So wird er immer wieder im Sinne großer Stichproben verwendet, obwohl große Fallzahlen nicht automatisch Repräsentativität bedeuten."* (KRIZ, LISCH 1988, S. 220) Auch große Stichproben können systematische Verzerrungen beinhalten und deshalb nicht valide sein. Generell sagt die Größe *per se* nichts über die Stichprobenvalidität aus. Umgekehrt sind jedoch Schlüsse von kleinen Zufallsstichproben auf Grundgesamtheiten immer mit dem Problem der unbeobachteten Heterogenität behaftet: in kleinen Stichproben ist die Wahrscheinlichkeit, alle relevanten Gruppenunterschiede zu beobachten, aus rein praktischen Gründen geringer als in großen Stichproben, da die Wahrscheinlichkeit aus einer Population zufällig eine bestimmte Merkmalskombination zu ziehen, bei einer kleinen Stichprobe geringer ist als bei einer großen.

Aus diesem Grunde ist es notwendig, in Fällen, in denen nur kleine Stichproben gezogen werden können, andere Verfahren zur Sicherung der Stichprobenvalidität anzuwenden als bei großen Stichproben:

- zum einen Samplingverfahren, die nicht dem Prinzip einer globalen Zufallsauswahl folgen, sondern die notwendige Heterogenität durch eine Schichtung in bezug auf die theoretisch relevanten Merkmale sicherstellen;
- zum anderen Verfahren der Abschätzung der Stichprobenqualität wie etwa Vergleiche mit anderen Datensätzen oder Anpassungstests.

1.2 Zum Problem der Inferenzstrategien bei kleinen Stichproben

Neben der Frage der Repräsentativität in bezug auf die Stichprobenauswahl stellt sich bei kleinen Stichproben ein weiteres Problem: das der adäquaten Inferenz- oder – allgemeiner – Modellierungsstrategien: Mit abnehmender Stichprobengröße nimmt bei statistischen Tests die Wahrscheinlichkeit, einen Fehler zweiter Ordnung (β-Fehler) zu begehen, zu. Dies hat zur Folge, daß auch stärkere Zusammenhänge im Datenmaterial bei kleinen Stichproben nicht zur Zurückweisung der Nullhypothese führen, da die Teststärke der verwendeten Verfahren nicht ausreicht. Im Falle kategorialer Daten wird dieses Problem noch verschärft, da die für größere Stichproben zulässigen asymptotischen Approximationen von Teststatistiken (etwa χ^2) bei vielen schwach besetzten Zellen unzulässig werden und somit traditionell erfolgreiche Testverfahren bei kleinen (und/oder schief verteilten[1]) Stichproben unzulässig werden (vgl. AGRESTI 1990; MEHTA, PATEL 1983).

Die Entwicklung exakter Tests oder die Anwendung von Monte-Carlo-Simulationen zur Schätzung von Prüfverteilungen für kategoriale Daten (vgl. BOYETT 1979; PATEFIELD 1981; BÜSSING, JANSEN 1988; MEHTA, PATEL 1983) hat entscheidend dazu beigetragen, daß Inferenzstrategien auch für kleine und/oder schief verteilte Stichproben entwickelt werden konnten. Wenngleich aber etwa Mehta immer wieder darauf hinweist, daß die Teststärke bei der Anwendung solcher Verfahren im Falle kleiner Stichproben deutlich über der asymptotischer Approximationen liegt, bleibt das grundlegende Problem, daß bei kleinen Stichproben die Teststärke per

[1] Auf das spezifische Problem schiefer Verteilungen wird in Teil 3 expliziter eingegangen.

se geringer ist als bei großen und damit bei traditionellen Inferenzstrategien nicht identifizierbare Bereiche der "Unentscheidbarkeit" (vgl. WITTE 1980, S. 86ff) entstehen können, auch mit diesen Verfahren ungelöst.

1.3 Übersicht

Im vorliegenden Papier soll aufgezeigt werden,

- daß unter bestimmten Voraussetzungen kleine Fallzahlen nicht automatisch fehlende Stichprobenvalidität nach sich ziehen, wenn das Samplingverfahren bestimmten Kriterien genügt, und
- daß alternative Inferenzstrategien statistische Schlüsse auf der Grundlage solcher Stichproben ermöglichen.

Hierbei soll neben der einschlägigen Literatur auf Beispiele aus den Arbeiten von Teilprojekten des Sfb 186 zurückgegriffen werden, mit denen der Bereich "Methoden und EDV" im Rahmen seiner Konzeption einer *"forschungsbegleitenden Methodenentwicklung"*[2] eng kooperiert hat: In diesem Kontext entstand während der zweiten Förderphase des Sfb 186 eine Arbeitsgruppe *("Small Sample AG")*, deren Aufgabe in der Entwicklung rationaler Auswertungsstrategien auf der Basis kleiner Stichproben bestand.

Die Arbeit mit kleinen Stichproben muß keinesfalls grundsätzlich gegen Standards empirischer Sozialforschung verstoßen: Cook und Campbell etwa beschreiben die Möglichkeiten, über *quasi-experimentelle Forschungsdesigns* mit bewußt heterogen, theoretisch geschichteten Stichproben oder Auswahlen zu arbeiten (vgl. COOK, CAMPBELL 1979, S. 76ff). Sie orientieren sich dabei am pragmatischen Kriterium der Durchführbarkeit empirischer Studien: *"Deliberate purposive sampling for heterogeneity is usually more feasable than random sampling for representativeness."* (COOK, CAMPBELL 1979, S. 77)

---

[2] D.h. das Aufgreifen praktischer, methodischer und methodologischer Fragestellungen in Kooperationsvorhaben zwischen einzelnen Teilprojekten und dem Methodenbereich.

- Auf die Frage, wie solche Auswahlverfahren in Anlehnung an die Arbeiten von Cook und Campbell begründet werden können und wie sie in verschiedenen Forschungsbereichen des Sonderforschungsbereichs 186 realisiert worden sind, wird im folgenden Teil 2.1 ausführlicher eingegangen.

- In dem Teil 2.2 werden Verfahren diskutiert, mit denen die Qualität einer gezogenen kleinen Stichprobe abgeschätzt werden kann.

- Schließlich wird im Teil 3 am Beispiel einfacher log-linearer Modelle eine Inferenzstrategie vorgestellt, die das Problem geringer Teststärke bei der Modellbildung mit kleinen Stichproben berücksichtigt. Ein hierzu entwikkelter Monte-Carlo-Algorithmus ist im Anhang abgedruckt.

2. Strategien zur Sicherung der Stichprobenvalidität

2.1 Auswahlstrategien

2.1.1 Zufallsstichproben als generelle Lösung?

In der quantitativ-orientierten Methodenliteratur der Sozialwissenschaften existiert weitgehend die Konvention, nur dann von repräsentativen Samples zu sprechen, wenn es sich um Zufallsstichproben (random samples) handelt.[3]

> "Die Bezeichnung einer Stichprobe als 'repräsentativ' ist somit nur im Sinne des Prinzips der Zufallsauswahl zu verstehen: beide Begriffe sind im obigen Sinne synonym." (SCHNELL ET AL. 1989, S. 280)

Nur Zufallsstichproben sollen gewährleisten können, daß es sich bei einer gezogenen Stichprobe um ein repräsentatives Sample handelt, so daß also von der Verteilung aller Merkmale in der Stichprobe auf die Verteilung dieser Merkmale in der Grundgesamtheit geschlossen werden kann. Nur bei einer Auswahl nach dem Zufallsprinzip soll dieser "Repräsentativitätsschluß" (SCHNELL ET AL. 1989, S. 280) gezogen werden können. Kromrey (1987, S. 479) faßt die üblichen Definitionen zur Repräsentativität in statistisch orientierten Argumentationszusammenhängen folgendermaßen zusammen[4]:

> "Wenn mit dem Blick auf eine präzise abgrenzbare Grundgesamtheit mit endlicher Zahl von empirisch definierbaren Elementen ein Plan zur Durchführung einer *kontrollierten Zufallsauswahl* erstellbar ist, und wenn bei der Durchführung der Stichprobenziehung sowie bei der Datenerhebung *verzerrende Einflüsse* ausgeschaltet werden können, dann ist das Ergebnis mit angebbarer Wahrscheinlichkeit ein *verkleinertes Abbild* dieser Grundgesamtheit. Das Sample ist dann mit angebbarer Wahrscheinlichkeit repräsentativ hinsichtlich sämtlicher Merkmale und Merkmalskombinationen – mit der (gewünschten) Konsequenz, daß aus den Stichprobenstatistiken auch auf *unbekannte* Parameter der Grundgesamtheit geschlossen werden kann."

Diese Definition verdeutlicht die zahlreichen Voraussetzungen (exakt definierte Grundgesamtheit mit endlicher Zahl von empirisch definierbaren Elementen,

---

3 Siehe außerdem MAYNTZ ET AL. 1969, S. 70; SAHNER 1971, S. 14; KERLINGER 1978, S. 109.

4 Vgl. auch KERLINGER 1978, S. 108; SCHNELL ET AL. 1989, S. 249 f.

detaillierter Plan für die Stichprobenziehung, kontrollierte Zufallsauswahl, Ausschalten verzerrender Einflüsse), um von einer Zufallsstichprobe sprechen zu können.

Dagegen weisen etliche Methodologen darauf hin, daß das grundsätzliche Problem, daß Zufallsstichproben zufälligerweise auch nicht repräsentativ sein können (vgl. KERLINGER 1978, S. 110), nur mit größter Sorgfalt zu lösen ist, da häufig praktische Probleme bei Stichprobenziehung und -erhebung auftreten, die trotz eines ausgefeilten Stichprobenplans für eine Zufallsauswahl Verzerrungen erzeugen:

a) *Stichprobenziehungsverfahren*

Selbst wenn ein ausgefeilter Stichprobenziehungsplan besteht, treten zumeist Probleme bei der Ziehung der Stichproben auf, die mit dem in wissenschaftlichen Untersuchungen häufig benutzten ADM-Design[5] zusammenhängen. Zu dessen Nachteilen zählt insbesondere, daß für bestimmte Merkmalsausprägungen ausreichend große[6] Fallzahlen in den Samples nicht zu realisieren sind und bestimmte Bevölkerungsgruppen nicht berücksichtigt werden[7]. Dies führt laut Schnell dazu, daß trotz ausge-

---

[5] Das ADM-Master-Sample ist ein Musterstichprobenplan des Arbeitskreises Deutscher Marktforschungsinstitute, eine erstmals in den siebziger Jahren erstellte Datenbank auf der Basis von 49.380 Stimmbezirken (auch synthetische und zusammengefaßte). Dies sind Primäreinheiten, die 22.147.308 Haushalte umfassen, aus denen (Zufalls-) Auswahlen getroffen worden sind.

[6] Genau diese Problematik wird an den verschiedenen Untersuchungsgruppen der Teilprojekte des Sfb 186 (SonderschülerInnen, SozialhilfeempfängerInnen etc.) deutlich: zum einen stellen sie nur einen kleinen Anteil an der Gesamtbevölkerung dar, so daß die Untersuchungsgruppe in der Gesamterhebung entsprechend klein ausfallen würde; zum anderen reduzieren sich die Untersuchungsgruppen bei einer differenzierten Datenanalyse selbst nach "einfachsten" Merkmalen wie Geschlecht oder Alter nochmals, so daß aufgrund einer solchen Datenlage nur eingeschränkt aussagekräftige Ergebnisse erwartet werden können.

[7] Hierzu zählen lt. SCHNELL (1991) insbesondere:
- Deutsche im Ausland (mindestens 11.000),
- Ausländer und Ausländerinnen (ca. 4,85 Mio.),
- Anstaltsbevölkerung (ca. 800.000: StudentInnen, PolizistInnen, Soldaten, Zivildienstleistende, Ordensmitglieder, Strafgefangene, PatientInnen psychiatrischer Einrichtungen, Personen in Einrichtungen der Altenhilfe),
- Nichtbefragbare (ältere und pflegebedürftige Personen: ca. 260.000; Psychisch Kranke: ca. 1,2 Mio; Speziell Behinderte wie Sehbehinderte, Gehörlose; AnalphabetInnen: 500.000 bis 3 Mio.),
- Personen in ungewöhnlichen Arbeits-, Lebens- und Wohnsituationen: bestimmte Berufsgruppen: BinnenschifferInnen, deutsche Seeleute, SchaustellerInnengewerbe, BauarbeiterInnen auf Montage; mobile Gruppen: Roma und Sinti; UntermieterInnen, BewohnerInnen von Hinterhäusern, Obdachlose und Nichtseßhafte; "Eliten".

bauter "Sozialindikator-Infrastruktur" eine starke Unkenntnis über statistische Größen vieler "Sonderpopulationen" besteht (SCHNELL ET AL. 1991, S. 135).

Außerdem ist es eine ADM-Design begründete Problematik, daß Personensamples per se disproportional sind, da Personen aus kleineren Haushalten eine größere Chance besitzen, in die Auswahl zu gelangen. Auch dies kann unter bestimmten Umständen zu problematischen Verzerrungen führen[8].

b) *Ausfälle: Verweigerung und Nicht-Erreichbarkeit*

Neben den Problemen, die mit dem Stichprobenziehungsplan verbunden sind, scheitert die reine Zufallsstichprobe zudem an der Datenerhebung, da es so gut wie nie gelingt, **alle** in die Stichprobe gezogenen Elemente zu erheben. Neben der Verringerung des realisierten Stichprobenumfanges kann es durch diese Ausfälle zu Verzerrungen der Stichprobenergebnisse kommen, was jedoch als unproblematisch gilt, wenn diese Ausfälle (völlig) zufällig erfolgen (missing at random bzw. missing completely at random). Das Hauptproblem liegt hierbei darin, daß die Ausfälle mit Merkmalen korrellieren können, die mit dem Untersuchungsziel zusammenhängen und die Ausfälle zum Teil in hohem Maße mit einer Vielzahl von Faktoren der persönlichen Lebensumstände, des Sozialverhaltens, von Erfahrungen und Einstellungen der für die Befragung ausgewählten Personen korrespondieren: "*Selbst bei Personen also, die für freiwillige Datenerhebungen nicht grundsätzlich unzugänglich sind, sind selektive Ausfallprozesse festzustellen, die mit jeder Welle zu zunehmend ernsthafteren Verzerrungen führen.*" (ESSER ET AL. 1989, S. 144)

Ausfallprobleme und die damit verbundene Gefahr systematischer Verzerrungen treten laut Esser et al. bei allen Datenerhebungen auf: "*Demnach bestimmen Ausmaß und Selektivität der Ausfälle, ob die Daten-*

[8] Zum ADM-Design siehe auch: KIRSCHNER 1984a, 1984b, 1985; RÖSCH 1985.

> *qualität freiwilliger Erhebungen ausreichend ist oder nicht.*" (ESSER ET AL. 1989, S. X) Wichtig ist es, über das Ausmaß der Verzerrungen Aussagen zu treffen[9].

Hiermit wird deutlich, daß in der praktischen Umsetzung auch Verfahren der Zufallsauswahl nicht *automatisch* die Ziehung unverzerrter, valider Stichproben *garantieren* können und diese vor allem dann problematisch werden können, wenn bei der Erhebung mit hohen *Ausfällen* zu rechnen ist oder wenn die o.g. *"Sonderpopulationen"* untersucht werden sollen, da die Gefahr systematischer Verzerrungen in diesen Fällen relativ hoch ist. Dieses wird bei der Ziehung kleiner Stichproben noch dadurch verstärkt, daß hier per se der zufallsbedingte Standardstichprobenfehler größer ist und damit der Gesamtstichprobenfehler einen nicht mehr tolerierbaren Umfang annehmen kann. In solchen Fällen ist es für die Sicherung der Stichprobenvalidität weniger entscheidend, daß die Verteilung *aller möglichen* Merkmale in der Stichprobe deren Verteilung in der Grundgesamtheit entspricht, sondern daß zumindest die *theoretisch bedeutsam erscheinenden* Merkmale in der Stichprobe auf ihre Verteilung und somit auf systematische Verzerrungen hin kontrolliert werden können (vgl. PREIN, KELLE, KLUGE 1993, S. 40 ff).

Während es bei großen Bevölkerungsstichproben also anzustreben ist, die heutigen Stichprobenpläne und Erhebungsverfahren dahingehend zu verbessern, daß die o.g. Verzerrungen in geringerem Maße auftreten, bedeutet dies für kleine Stichproben, daß eine geschichtete, evtl. partielle (Zufalls-) Auswahl oder Verfahren des theorie-geleiteten Quotensamplings einem einfachen Zufallsverfahren überlegen sind, da hierbei zumindest die Verteilung der Schichtungs- bzw. Quotierungsmerkmale konstant gehalten werden kann. Dies ist natürlich nur die zweitbeste aller theore-tisch denkbaren Lösungen, da etwa – im Gegensatz zur gelungenen Zufallsauswahl mit einer ausreichend großen Stichprobe – die Heterogenität der Population in bezug auf andere als die kontrollierten Merkmale nicht abgeschätzt werden kann. Dennoch bewirkt eine geschichtete Auswahl zumindest, daß die Validität der Stichprobe in bezug – allerdings auch nur in bezug – auf die vorher definierten, theoretisch relevant

---

[9] "Der Vergleich von Volkszählungsbegleit-Panel und Sozio-ökonomischem Panel belegt also, daß Nichtteilnahmen in verschiedenen Bevölkerungsgruppen je nach Untersuchungsgegenstand variieren und daß nicht vom Verhalten in einer Befragung auf das Verhalten in anderen Befragungen generalisiert werden kann." (ESSER ET AL. 1989, S. 143)

erscheinenden Merkmale gesichert ist, selbst wenn die Vertei-lung anderer, theoretisch marginal erscheinender Merkmale in der Stichprobe mög-licherweise verzerrt ist.

2.1.2 Die Problematik kleiner Stichproben

Wenn es das Ziel eines empirischen Forschungsprojektes ist, Daten und Aussagen über größere Bevölkerungsgruppen oder die Gesamtbevölkerung zu ermitteln, sollte grundsätzlich auf der Grundlage von ausreichend großen Zufallsstichproben (wie sie der ALLBUS, das SOEP oder der Mikrozensus darstellen) gearbeitet werden, so daß bewährte Verfahren der Datenauswertung angewendet werden können. Ein solches Vorgehen stößt jedoch automatisch dort auf seine Grenzen, wo sich eine Forschungsfrage auf **ausgewählte Bevölkerungsgruppen** wie SozialhilfeempfängerInnen, Haupt- und SonderschülerInnen, Auszubildende in bestimmten Lehrberufen, Rehabilitanden, Ausbildungsgruppen bestimmter Jahrgänge usw. bezieht, wie dies im Rahmen des Sonderforschungsbereichs 186 der Fall ist. Für solche Untersuchungen ist ein Rückgriff auf die o.g. repräsentativen Datensätze problematisch, da für die interessierenden Untergruppen nur noch extrem kleine Fallzahlen zur Verfügung stehen. Damit sind zur Bearbeitung von Fragen, die mit diesen Bevölkerungsgruppen zusammenhängen, solche repräsentativen Bevölkerungsumfragen weitgehend ungeeignet, da Subsamples, die aus diesen Stichproben gebildet werden können, in der Regel zu klein sind, um methodisch verantwortbare Auswertungen durchzuführen.

Daß Stichproben grundsätzlich klein und/oder schief verteilt sein können, hat in den verschiedenen Feldern sozialwissenschaftlicher Forschung sehr unterschiedliche Gründe, ohne daß dies sogleich als methodischer “Kunstfehler” zu betrachten wäre:

- *Referenzpopulationen* selbst können sehr *klein* sein: So findet sich etwa für eine Untersuchung von GERHARDT (1986) nur eine begrenzte Anzahl von chronisch Nierenkranken.

- *Die Erreichbarkeit von Personen* aus bestimmten Zielpopulationen kann schwer abschätzbar sei: ein unerwartet hoher Mobilitätsgrad

(berufsbedingt oder auf Grund eigener Familiengründung nach der Ausbildung; Teilpro-jekte A1, A3) oder ungewöhnliche Lebensumstände (unstetes Leben aufgrund "abweichenden" Verhaltens, Teilprojekt A3) können aufgrund der schlechten Erreichbarkeit zu einem reduzierten Sample führen. – Ebenso sind Stichproben von Folgeerhebungen, deren Samples mit Datensätzen früherer Erhebungen verknüpft sind bzw. deren Auswahlkriterien auf diesen basieren, in der Regel deutlich kleiner als das Ursprungssample (Teilprojekt B1).

- Die Stichprobengröße reduziert sich durch Ausfälle aufgrund von *Verweigerungen und Nicht-Erreichbarkeit.* Hier spielen auch die Befragtengruppen eine wichtige Rolle: sowohl SozialhilfeempfängerInnen, Haupt- und SonderschülerInnen sowie Personen mit delinquentem Verhalten sind in der Regel Personengruppen, die schwieriger zu erreichen und zu einer Teil-nahme zu motivieren sind. In diesem Zusammenhang ist auch das Problem der Panel-Attrition zu sehen.

- Annahmen über *komplexe Zusammenhänge*, die ihren Ausdruck in *multivariaten Modellen* mit zahlreichen erklärenden Variablen finden, führen oftmals dazu, daß im untersuchten Sample – insbesondere extreme – Untergruppen sehr klein werden. Es kann also auch in Fällen, in denen die Samplegröße zunächst ausreichend erscheint, zu einer Situation kommen, in der große Stichproben für die Schätzung bestimmter Modelle "klein werden".

- Zuletzt ergeben sich kleine Stichproben fast automatisch in bestimmten methodologischen Traditionen: *Qualitativ-fallrekonstruktive Analysen* mit repräsentativen Zufallsstichproben durchzuführen, ist allein aufgrund des höheren Auswertungsaufwandes der meisten Verfahren[10] in der qualitativen

[10] Eine Ausnahme stellt hier die sog. "hermeneutisch-klassifikatorische Inhaltsanalyse" (vgl. ROLLER, MATHES 1993) dar, die in der Lage sein soll, mit qualitativ-hermeneutischen Verfahren repräsentative Stichproben (mit n=500) zu bearbeiten.

> Sozialforschung illusorisch. Dennoch muß auch in diesem Rahmen aufzeig-bar sein, daß die Ziehung der Untersuchungsstichprobe nicht willkürlich stattfindet.

Die Tatsache, daß für bestimmte Forschungsfelder nur die Ziehung kleiner Stichproben möglich ist oder solche Stichproben aufgrund der Spezifika des Forschungsfeldes leicht zustandekommen können, sollte einerseits nicht dazu führen, daß empirische Sozialforschung sich nur deshalb nicht mit diesen Realitätsbereichen beschäftigt, weil dort keine großen Zufallsstichproben erhoben werden können. Andererseits sollten die Vorteile, die große Stichproben mit sich bringen, nicht dazu führen, daß nur solche Untersuchungen durchgeführt und akzeptiert werden, die auf der Grundlage von *repräsentativen Bevölkerungsstichproben* durchgeführt werden, da in solchen Stichproben oftmals relevante Untergruppen extrem klein sind: Bei bestimmten Fragestellungen kann es demgegenüber viel wichtiger sein, *systematisch* theoretisch relevant erscheinende Subpopulationen zu untersuchen, die etwa regional geschichtet sind, um z.B. bestimmte "Störvariablen" konstant zu halten.

Hierbei muß allerdings geprüft werden, durch welche Samplingstrategien das Zustandekommen von Forschungsartefakten verhindert werden kann. Dazu gehören einerseits in bezug auf die Stichprobenziehung ein theorie- und empiriegeleiteter Stichprobenplan, über den Einflüsse intervenierender Variablen systematisch überprüft werden können, sowie die Auswahl von in bezug auf theoretisch formulierbare Kriterien kontrastierende Gruppen. Ein solches Verfahren der Stichproben-ziehung kann mit der von Cook und Campbell (1979) vorgeschlagenen *heteroge-nen Auswahl* verglichen werden, wobei sich die ausgewählten Subpopulationen einerseits nach theoretischen Kriterien unterscheiden sollten, und andererseits durch empirische Voruntersuchungen sichergestellt werden muß, daß es sich bei den ausgewählten Gruppen nicht nur um "Ausreißer" handelt.

2.1.3 Die “bewußt heterogene Auswahl” und die “Auswahl nach Modalkategorien” als Alternativen bei kleinen Stichproben?

Die Konstruktion heterogener, theoretisch geschichteter Stichproben läßt sich methodologisch auf die Arbeiten von Cook und Campbell über quasi-experimentelle Forschungsdesigns (1979) zurückführen. Diese beiden Autoren unterscheiden drei Samplingstrategien, mit denen in unterschiedlichem Umfang die externe Validität von Forschungsergebnissen gesichert werden kann:

1. *Repräsentative Zufallsauswahl:* Cook und Campbell würdigen das Modell der Zufallsauswahl prinzipiell als das beste. Sie stellen jedoch gleichzeitig heraus, daß es erstens aus forschungsökonomischen Gründen oftmals nicht angewandt werden kann, zweitens auch Zufallsstichproben nur Verallgemeinerungen in bezug auf eingeschränkte Kontexte erlauben, deren Interesse oftmals begrenzt erscheint (vgl. COOK, CAMPBELL 1979, S. 75 f).

2. *Eine bewußt heterogene Stichprobenauswahl* stellt demgegenüber eine nicht in jedem Fall zufallsgenerierte Auswahl dar, die in bezug auf theoretisch bedeutsam erscheinende Merkmale möglichst stark divergierende Gruppen umfaßt: *“When one samples nonrandomly, it is usually advantageous to obtain opportunistic samples that differ as widely as possible from each other.”* (COOK, CAMPBELL 1979, S. 76)

3. Die *impressionistische Auswahl nach Modalkategorien* basiert darauf, min-destens einen Fall aus jeder untersuchten Gruppe zu erfassen, der eher durchschnittliche Ausprägungen enthält. Dieser kann – soweit hierzu Daten verfügbar sind – auf der Grundlage übergreifender Statistiken oder – falls diese nicht existieren – auf der Basis von Expertengesprächen ausgewählt werden (vgl. COOK, CAMPBELL 1979, S. 77).

Die beiden Autoren stellen ausdrücklich fest, daß Zufallsstichproben eine wichtige Funktion in der empirischen Forschung haben, weil nur auf ihrer Basis Effekte in bezug auf eine große Anzahl von Subpopulationen abschätzbar sind. Da sie allerdings eine negative Beziehung zwischen *“'inferential' power”* und *“feasibility”* – also Durchführbarkeit – von Samplingstrategien sehen, konstatieren sie, daß *“the model of heterogeneous instances would be most useful, particularly if great care is made to include impressionistically modal instances among the heterogeneous*

ones" (COOK, CAMPBELL 1979, S. 78). Heterogene Auswahlen können also laut Cook und Campbell durchaus nutzbringend sein, insbesondere wenn darauf geach-tet wird, häufig auftretende Kategorien zu berücksichtigen. Statt dabei stehen zu bleiben, die Zufallauswahl als einzig methodologisch gerechtfertigten Weg der Stichprobenkonstruktion anzusehen, geht es Cook und Campbell um eine Be-schreibung und Rekonstruktion praktisch durchführbarer Verfahrensweisen sowie deren pragmatischen Nutzen. Daher postulieren sie: *"Many small-scale experi-ments with local control and choice of measures is in many ways pre-ferable to giant national experiments with a promised standardisation that is neither feasible nor even desirable from the standpoint of making irrelevancies heterogeneous."* (COOK, CAMPBELL 1979, S. 80)

Eine Samplingstrategie, die auf der Auswahl heterogener Untersuchungseinheiten basiert, erlaubt laut Cook und Campbell die Generalisierung von Forschungser-gebnissen auf Grundgesamtheiten nur unter bestimmten Bedingungen: auf ihrer Grundlage ist es im Gegensatz zur rein impressionistischen Auswahl von Modal-kategorien allerdings möglich, Differenzen zwischen verschiedenen Subpopulatio-nen zu identifizieren, sowie zu prüfen, ob sich Untergruppen in bezug auf die untersuchten Merkmale überhaupt unterscheiden: *"Hence one cannot – technically speaking – generalize from the archieved sample to any formally meaningful populations. All one has are purposive quotas of persons with specified attributes. These quotas permit one to conclude that an effect has or has not been obtained across the particular variety of samples of persons, settings, and times that were under study."* (COOK, CAMPBELL 1979, S. 76)

Cook und Campbell stellen damit eine praktikable Vorgehensweise der Stichpro-benkonstruktion vor, die auf der Auswahl von Fällen aus möglichst heterogenen Subpopulationen basiert. Was bei ihrem Verfahren allerdings nicht deutlich wird, sind die Kriterien zur Bestimmung solcher Subpopulationen. Sie deuten dies zwar mit ihrer Bemerkung an, daß Modalkategorien einbezogen werden sollten, stellen jedoch nur undeutlich dar, nach welchen Verfahrensweisen aus diesen Gruppen die heterogenen Untersuchungseinheiten definiert werden können.

2.1.4 Auswahlstrategien und kleine Stichproben am Sfb 186

In vielen Teilprojekten des Sonderforschungsbereichs 186 werden bei quantitativen und qualitativen Erhebungen Forschungsdesigns eingesetzt, die mit den von Cook und Campbell vorgeschlagenen vergleichbar sind. Das Konzept der bewußt heterogenen Auswahl ist dabei allerdings um zwei Elemente erweitert worden, um eine stärker *theorieorientierte Repräsentativität*[11] der Stichproben, d.h. deren Verallgemeinerbarkeit in bezug auf theoretisch relevant erscheinende Merkmale, sicherzustellen:

- *Empirisch und theoretisch begründete Stichprobenpläne:* Um sicherzustellen, daß empirisch *und* theoretisch relevante, heterogene Untergruppen und nicht etwa "Ausreißer" in die Stichprobe aufgenommen werden, ist die Strategie des *Matrjoschka-Sampling* insbesondere zur Ziehung kleiner und/oder qualitativer Stichproben[12] entwickelt worden. Dabei werden auf der Grundlage theoretischer Annahmen Schichtungsmerkmale definiert und deren Verteilung anhand empirischer Daten des Forschungsbereichs überprüft. Letzteres kann in einer Auswertung amtlicher Statistiken bestehen oder besser – da diese häufig nicht existieren bzw. unzureichend sind – in der Durchführung einer auf den Variablensatz zugeschnittenen Piloterhebung. Auf dieser Grundlage läßt sich überprüfen, ob eine bestimmte theoretische Kategorie empirisch bedeutsam ist oder etwa nur marginale Bedeutung in der untersuchten Population hat, ob in bezug auf die Herstellung von Heterogenität wesentliche Kategorien übersehen worden sind etc. Erst in einem Folgeschritt und nach einer evtl. Überarbeitung des theoretisch formulierten Stichprobenplans wird aus dem Sample der Piloterhebung bzw. aus der Population, auf die sich die amtliche Statistik bezieht, eine Stichprobe gezogen.[13]

- *Zufallsauswahl innerhalb der heterogenen Untergruppen:* Um die Variation von Merkmalen innerhalb der Subpopulationen möglichst wenig zu

---

[11] Das eingeschränkte Konzept der theoriebezogenen Repräsentativität darf hier natürlich nicht mit dem der statistischen Repräsentativität (s.o.) verwechselt werden!

[12] Auf die Problematik von Samplingstrategien für qualitative Stichproben wird ausführlicher im Arbeitspapier Nr. 19 (PREIN, KELLE, KLUGE 1993, S. 40ff) eingegangen.

[13] Daher auch das Bild der Matrjoschka - der russischen Holzpuppe - bei der sich jeweils aus einer großen Puppe eine kleine ziehen läßt.

verzerren und damit zumindest Repräsentativitätsschlüsse auf die jeweilige Subpopulation zu erlauben, werden bei quantitativen Untersuchungen innerhalb der einzelnen Gruppen in der Regel Zufallsstichproben gezogen oder – wenn möglich – Vollerhebungen durchgeführt.

Exemplarisch für den Kontext des Sonderforschungsbereichs soll hier die Vorgehensweise des Projekts B1 beschrieben werden:

In der ersten Phase des Projektes B1 "Statussequenzen von Frauen zwischen Erwerbsarbeit und Familie" wurden in der quantitativen Untersuchung Sequenzmuster von Erwerbs- und Familienarbeit in den Lebensläufen von Frauen untersucht. Hierzu wurde folgender Stichprobenplan erarbeitet:

1. Da Unterschiede in bezug auf normative Orientierungen zwischen ländlichen und städtischen Gebieten sowie von Gebieten mit protestantischer und römisch-katholischer Prägung vermutet wurden, wählte das Projekt zwei Arbeitsmarktregionen – Bremen als evangelisch-städtische und Koblenz als katholisch-ländliche Region – aus. In diesen Arbeitsmarktbezirken wurde über die Industrie- und Handwerkskammern eine *Vollerhebung* aller Frauen durchgeführt, die in den Jahren 1948 und 1949 eine Lehre beendet hatten. Durch diesen Untersuchungsschritt erlangte das Projekt Informationen über den Umfang der Untersuchungseinheit sowie die Verteilung der Ausbildungsberufe.

2. Da das Projekt weiterhin Unterschiede zwischen verschiedenen Ausbildungsberufen annahm, zugleich aber auch Modalkategorien berücksichtigt werden sollten, wählte das Projekt auf der Grundlage weiterer amtlicher Statistiken für die Bundesrepublik die fünf seinerzeit am stärksten von Frauen absolvierten Ausbildungen aus (Sampling nach Modalkategorien). Diese Berufe sollten weiterhin möglichst stark in bezug auf Arbeitsmarktchancen, Berufsrollendefinitionen und Vereinbarkeitsmöglichkeiten von Kinderbetreuung und Berufsarbeit variieren (Sampling nach weitestgehender Heterogenität). Während bspw. der Beruf der Schneiderin bedeutungslos wurde, expandierten die Arbeitsmarktchancen in kaufmännischen Berufen. Durch die Einbeziehung der Modalkategorien Friseurinnen, kaufmännische Angestellte, Kinderpflegerinnen, Verkäuferinnen und Schneiderinnen war auch das Kriterium weitgehender Heterogenität erfüllt.

3. Auf dieser Grundlage war geplant, in einem als Ex-post-facto-Experiment angelegten Design je Berufsgruppe und Region eine Zufallsstichprobe von 100 Frauen zu ziehen. Da z.T. nicht aktualisierbare Adressen vorlagen, konnte die Sollzahl nicht für alle fünf Gruppen realisiert werden. Dennoch gelang es, in etwa gleichgewichtige Samples zu ziehen und 800 Fragebögen zu verschicken. Von diesen kamen 220 auswertbar zurück. Aus Datenschutzgründen waren Nachfaß- oder Nachziehungsaktionen nicht möglich.

Es handelte sich bei dieser Stichprobe also um eine *dysproportional geschichtete Zufallsstichprobe*: *innerhalb* der Gruppen wurde nach dem Prinzip der *Zufallsauswahl* erhoben; bei der *Auswahl der Gruppen* wurde nach einem *Stichprobenplan* vorgegangen, der nach den Prinzipien weitgehender *Heterogenität* in bezug auf theoretisch relevant erscheinende Variablen (Arbeitsmarktchancen und Prestige des Berufs; regionale Kontexte) sowie weitgehende Einbeziehung von *Modalkategorien* aufgebaut war. Auf der Grundlage dieser Stichprobe wurden systematische Vergleiche zwischen den Berufsgruppen durchgeführt; dadurch war deren quantitativ unterschiedliche Verteilung innerhalb der Grundgesamtheit irrelevant.

Wir sehen an diesem Beispiel – das stellvertretend für das Vorgehen bei der Ziehung kleiner, quasi-experimenteller Samples am Sonderforschungsbereich 186 stehen kann -, daß solche Stichproben, sobald ihre Ziehung methodisch kontrolliert erfolgt, valide sein können[14]. Entscheidend hierfür ist allerdings die Aufstellung eines Stichprobenplans,

1. dessen Schichtungskriterien *theoretisch bedeutsamen Kategorien* entsprechen (hier etwa: normative Orientierungen, Arbeitsmarktchancen ...),

2. der durch die Variation dieser Kategorien die *Heterogenität* des Samples zumindest in bezug auf diese Merkmale sicherstellt (hier: Kontrastierung von Gruppen nach o.g. Kategorien),

---

[14] Ähnliche Verfahren sind in der Literatur zum Teil ansatzweise beschrieben worden (vgl. etwa HEINZ ET al. 1985). Im Gegensatz zu der Vorgehensweise, die Heinz und andere damals entwickelt haben, werden beim hier dargestellten Verfahren allerdings bewußt keine Klumpenstichproben gezogen.

3. der zugleich aber die empirische Bedeutung dieser theoretisch bedeutsamen Kategorien *quantitativ* überprüft und *Modalkategorien* berücksichtigt (hier: vorangehende Vollerhebung, Bestimmung der häufigsten Ausbildungsberufe für Frauen),

4. der *innerhalb der Schichten* methodologisch und forschungsökonomisch vertretbare *Auswahlkriterien* definiert (hier: *Zufallsauswahl*, in anderen Projekten, z.B. A1 und A3, *z.T. Vollerhebung*) und

5. der eine Stichprobe definiert, die innerhalb einer *empirisch beschreibbaren und bestimmbaren Grundgesamtheit bzw. eines repräsentativen Sample*s verortet werden kann ("Matrjoschka-Sampling", hier: Grundgesamtheit aller Frauen der gewählten Jahrgänge 1. in der BRD, 2. in den beiden Arbeitsamtsbezirken).

Bei der Anwendung einer solchen Samplingstrategie muß allerdings – ebenso wie bereits für Zufallsstichproben beschrieben – überprüft werden, inwieweit systematische Verzerrungen bei der Datenerhebung entstanden sind. Dies ist insbesondere dann der Fall, wenn etwa ein hoher Grad von Ausfällen zu erwarten ist. Zur Abschätzung von Verzerrungen, die bei der Datenerhebung entstanden sind, werden in der Literatur Anpassungstests von empirischen Verteilungen an repräsentative Datensätze vorgeschlagen. Auf die Möglichkeiten, die solche Verfahren in bezug auf kleine Stichproben bieten, sowie die Probleme, die in diesem Kontext entstehen können, soll im folgenden Teil eingegangen werden.

2.2 Datenvergleich

Aus dem Vorherigen wurde deutlich, daß bei der praktischen Umsetzung von Prinzipien der Zufallsauswahl eine Verallgemeinerbarkeit der Ergebnisse nicht *automatisch* gegeben ist. Etwa durch Ausfälle bei der Datenerhebung kann es auch bei Erhebungen, die auf Zufallsauswahlen beruhen, zu Verzerrungen kommen, so daß ein maßstabsgetreues Abbild der Grundgesamtheit durch die Stichprobe nicht zu erreichen ist. Dennoch lassen sich auf Grund der Stichprobe Aussagen über die Grundgesamtheit treffen, wenn man etwas über das **Ausmaß** der Verzerrungen weiß.

> "Die Nicht-Repräsentativität von Stichproben (Nicht-Repräsentativität im zuvor definierten Sinn) ist für sich genommen noch KEIN hinreichendes Argument gegen die Möglichkeit, die Befunde zu verallgemeinern. Ausgeschlossen ist zunächst lediglich die sozusagen 'automatische' Verallgemeinerung mit Hilfe wahrscheinlichkeitsstatistischer Verfahren." (KROMREY 1987, S. 485)

Um das Ausmaß eventueller Verzerrungen zu bestimmen, werden im Sfb selbst erhobene Stichproben mit Daten aus der amtlichen Statistik oder bundesweiten Repräsentativerhebungen abgeglichen, d.h. einzelne Stichprobenparameter werden mit bekannten Parametern der Grundgesamtheit verglichen.[15] Im Mittelpunkt des Vergleichs stehen dabei neben Indikatoren für theoretisch relevante Sachverhalte zumeist soziodemographische Variablen wie Geschlecht, Alter, Bildungsstand.

Im folgenden soll das Vorgehen detailliert beschrieben werden, da es einige Voraussetzungen zu beachten gilt, wenn der Vergleich aussagekräftig sein soll. Wir beziehen uns dabei auf Arbeiten der Sfb-Projekte, einen detaillierten Forschungsbericht des Sonderforschungsbereichs 3 (BLOSSFELD 1985) sowie die Arbeiten von Arminger (1989, 1990).

---

[15] Dies ist ein in der Sozialforschung übliches Verfahren; vgl. etwa BLOSSFELD 1985 (Sfb-3-Daten und Mikrozensus); ERBSLÖH, KOCH 1988, S. 36 ff (Vergleich von Teilnehmern und Nicht-Teilnehmern, ALLBUS und Mikrozensus); KIRSCHNER 1980, S. 83; PORST 1984 (eigene Daten und ALLBUS; ALLBUS und amtliche Statistik); SCHIMPL-NEIMANNS 1991 (ALLBUS und Mikrozensus); ESSER ET AL. 1989, S. 133 (ALLBUS, Mikrozensus und amtliche Statistiken); ARMINGER 1989 (Mikrozensus); HARTMANN 1990 (ALLBUS, Mikrozensus); HARTMANN, SCHIMPL-NEIMANNS 1992.

2.2.1 Statistische Verfahren des Datenvergleichs

Beim Vergleich der Daten einer Stichprobe mit Datensätzen der amtlichen Statistik oder anderer, repräsentativer Bevölkerungsumfragen werden unterschiedliche statistische Prüfverfahren genutzt: Zumeist sind dies der *Prozentsatzvergleich*, χ^2-basierte Techniken wie der χ^2-*Anpassungstest* oder aus dem Bereich der *Likelihood-Statistik* G^2 (vgl. BLOSSFELD 1985, S. 2 ff; HARTMANN, SCHIMPL-NEIMANNS 1992, S. 328) sowie *loglineare Modelle, Logit-Modell und logistische Regression* (vgl. ARMINGER 1989; HAAGENARS 1990).

In der entsprechenden Literatur finden sich schematisch folgende statistische Prüfschritte, die hier nur kusorisch dargestellt werden sollen:

1. Eindimensionale Vergleiche: Beurteilung des χ^2-Wertes aus dem Anpassungstest (vgl. etwa BORTZ, LIENERT, BOEHNKE 1990, S. 95ff für den Vergleich einer Stichprobe mit einer Grundgesamtheit; BLOSSFELD 1985, S. 15 für den Vergleich von zwei Stichproben), der Devianz-Teststatistik (ARMINGER 1989, S. 65), der relativen Differenzen und standardisierten Residuen von beobachteten und unter dem Modell der Unabhängigkeit erwarteten Werten (ebenda, S. 68); Vergleich der absoluten und relativen Differenzen zwischen den relativen Häufigkeiten in den beiden Stichproben (ebenda, S. 69);

2. Aufzeigen auftretender Abweichungsmuster auf der Basis eindimensionaler Vergleichsstatistiken; evtl. multivariate Modellierung der Abweichungen (vgl. ARMINGER 1989, S. 76 ff);

3. Formulierung möglicher Erklärungen für die Abweichungsmuster.

2.2.2 Probleme beim Datenvergleich

Häufig finden sich in empirischen Forschungberichten sogenannte "Repräsentativitätsnachweise" von Stichproben, die angeblich über Datenvergleiche mit amtlichen Datensätzen erbracht worden seien. Mit Esser et al. ist jedoch grundsätzlich davon auszugehen, daß

> "zur Beurteilung der Güte einer Stichprobe (...) daher die gelegentlich von Marktforschungsunternehmen sogenannten 'Repräsentanznachweise' (KAPLITZA 1982, S. 169; ein Beispiel geben NOELLE-NEUMANN und PIEL 1984, S. 223-231) nicht aus(reichen). (...) Der Nachweis, daß **bestimmte** Merkmale in der Grundgesamtheit mit derselben Häufigkeit vorkommen wie in der Stichprobe, **beweist** keinesfalls, daß die Stichprobe alle interessierenden Merkmale in der korrekten Häufigkeit wiedergibt. Dies wäre nur dann korrekt, wenn **alle** anderen Merkmale innerhalb der durch die überprüften Merkmale gebildeten Schichten vollständig homogen verteilt wären." (ESSER ET AL. 1989, S. 123 f)

Ein Problem des Datenvergleichs zwischen zwei Stichproben oder einer Stichprobe und einer Grundgesamtheit besteht immer darin, daß er sich konkret nur auf die in beiden in ähnlicher Weise **untersuchten** Variablen bezieht. Dies wird allerdings z.T. auch von Forschern, die solche Datenvergleiche durchführen, problematisiert:

> "Diese Befunde, und darauf ist zum Abschluß noch hinzuweisen, beziehen sich im strengen Sinne natürlich nur auf die in dieser Arbeit tatsächlich untersuchten Variablen und sagen nur bedingt etwas über mögliche Verzerrungen bei anderen Merkmalen der Sfb 3-Lebensverlaufsstudie aus." (BLOSSFELD 1985, S. 31)

Es besteht also weitgehende Einigkeit darüber, daß sich ein Vergleich von Datensätzen notwendigerweise auf die verfügbaren Variablen beschränken muß, und eine Unverzerrtheit in bezug auf diese Merkmale nicht automatisch garantiert, daß auch die anderen, in der Grundgesamtheit unbekannten Variablen ein akzeptables Abbild für diese liefern. Demgegenüber kann allerdings eine geringe Verzerrung *in bezug auf theoretisch relevante Variablen oder Indikatoren für theoretisch relevante Sachverhalte* als *ein Argument* oder *Indikator* – nicht Beweis! – zur Stützung der Annahme *theorieorientierter Repräsentativität* genutzt werden, indem sie zeigen, daß in bezug auf diese Variablen kein starker Bias besteht. Einer solchen Kontrolle auf grobe Verzerrungen stimmen auch Schnell, Hill und Esser (1989, S. 281) zu: "*Diese 'Repräsentanznachweise' (*KAPLITZKA *1982, S. 169)*

können jedoch als grobe Kontrollen des Ziehungsprozesses Verwendung finden, wenn Zufallsstichproben auf gröbste Verletzungen der Auswahlregeln überprüft werden sollen oder sehr simple Untersuchungen zu den Gründen von Ausfällen erfolgen." Damit liefert ein Datenvergleich zwar keine Garantien für die Güte der erhobenen Daten, er gibt aber zumindest *Anhaltspunkte* über das Ausmaß der Verzerrung bei den untersuchten Variablen.

Wenn in neuerer Zeit bspw. von Rendtel und Pötter (1993) gegen ein solches Vorgehen – wie in der Repräsentativitätsstudie von Hartmann und Schimpl-Neimanns (1992) – vehemente Kritik geübt wird, erscheint es im Rahmen konkreter Forschungspraxis sinnvoller, sich trotz aller bestehender Probleme dieser Techniken zu bedienen, da diese Kritiken bislang keine alternativen Vorgehensvorschläge beinhalten, wie auch Hartmann und Schimpl-Neimanns in ihrer Replik bemerken: *"Schlagen Rendtel und Pötter tatsächlich vor, externe Validierungen für unzulässig zu erklären? Wenn ja, welche Möglichkeiten gibt es für sie, die Güte einer realisierten Stichprobe festzustellen?"* (HARTMANN, SCHIMPL-NEIMANNS 1993, S. 364) Angesichts des Fehlens methodologischer Alternativen ist also beim Einsatz von Datenvergleichen und Anpassungstests zwar extreme Vorsicht geboten, wenn hieraus die generelle Unverzerrtheit von Datensätzen abgeleitet werden soll; es erscheint jedoch uneinsichtig, warum ein Verfahren, das in bestimmten Situationen in der Lage ist, Verzerrungen aufzudecken, in der Forschungspraxis keine Anwendung finden sollte.

2.2.3 Kriterien bei der Durchführung des Datenvergleichs

Damit Vergleiche zwischen Datensätzen solche Hinweise erbringen können, sind allerdings bestimmte Kriterien zu beachten:

a) *Die Grundgesamtheiten der verglichenen Stichproben müssen gleich bzw. zumindest vergleichbar sein:* Dies macht zunächst eine klare Definition der Zielpopulation (target population) der erhobenen Stichprobe erforderlich; wird diese nicht definiert, besteht die Gefahr, daß am Ende "*das Sample 'auf der Suche' nach der zugehörigen Grundgesamtheit ist*" (KROMREY 1987, S. 499). Sinnvoll ist ein Datenvergleich nur, wenn die Grundgesamtheiten übereinstimmen. Es muß also geprüft werden, inwieweit die target

population, auf die sich die Stichprobe beziehen soll, und die Grundgesamtheit der Vergleichsdaten übereinstimmen und ob eine eventuelle Abweichung akzeptabel ist (vgl. ESSER ET AL. 1989, S. 100 f; BLOSSFELD 1985, S. 13 f).

b) *Die Vergleichsdaten selbst müssen hohen Gütekriterien entsprechen:* In einem nächsten Schritt sollte geprüft werden, inwieweit die Vergleichsdaten selbst zuverlässig sind, da man davon nicht automatisch ausgehen kann: z.B. ist bekannt, daß in vielen Bevölkerungsumfragen ein deutlicher Mittelschichtsbias besteht[16]. Die Relevanz eines solchen Bias auf die untersuchte Fragestellung sowie die Frage, ob und inwieweit der gleiche Bias in der eigenen Erhebung durchgeschlagen haben könnte, muß auf jedem Fall vor der Durchführung eines Datenvergleichs geklärt werden.

c) *Die ausgewählten Vergleichsvariablen müssen für die Forschungsfragen relevant und zumindest ähnlich gemessen sein:* Es können für den Datenvergleich nur Variablen herangezogen werden, die auch in den Vergleichsdaten bekannt sind (BLOSSFELD 1985, S. 11) und deren Ausprägungen – evtl . durch sinnvolle Zusammenfassung – vergleichbar sind. Hierbei ist sicherzustellen, daß diese Variablen für die letztlich beabsichtigten Schlußfolgerungen von den Stichprobendaten auf die target population relevant sind bzw. ob Indikatoren für bestimmte theoretisch relevante Variablen existieren. Auch ein nur auf den Vergleich der “Standarddemographie” gerichtetes Vorgehen muß in bezug auf die untersuchte Forschungsfrage begründet werden.

2.2.4 Datenvergleiche bei kleinen Stichproben

Ein weiteres Problem, das insbesondere χ^2-Anpassungstests bei kleinen Stichproben oder stark streuenden Merkmalsausprägungen betrifft, ist das der Zulässigkeit asymptotischer Approximationen von Teststatistiken: Wenn eine große Anzahl erwarteter Häufigkeiten klein ist, ist es nicht mehr unproblematisch davon auszugehen, daß der aus beobachteten und erwarteten Zellenhäufigkeiten berechnete "χ^2-Wert" (etwa nach Pearson) auch asymptotisch χ^2-verteilt ist und somit Signifikanzprüfungen auf der Grundlage der tabulierten χ^2-Verteilung erfolgen können[17].

Für den univariaten Vergleich zweier Stichproben läßt sich dieses Problem durch den Einsatz von Spezialsoftware[18], die einen komfortablen Einsatz von exakten Tests oder Monte-Carlo-Simulationen zur Generierung von Prüfverteilung erlaubt, lösen. Sofern sich Fragen der Stichprobenanpassung mit dem Verfahren der logistischen Regression modellieren lassen, kann ebenfalls auf bestehende Software[19] zurückgegriffen werden.

Da dies jedoch nicht immer der Fall ist, wurden im Bereich Methoden und EDV für die Arbeit von Sfb-Teilprojekten Algorithmen entwickelt, die Anpassungstests auch für zwei weitere Problemkonstellationen erlauben, deren Bearbeitung mit bisheriger Software nicht möglich war:

1. den univariaten Vergleich zwischen Häufigkeitsverteilungen in einer Stichprobe und einer Grundgesamtheit (diese Konstellation kann dann auftreten, wenn amtliche Statistiken als Vollerhebungen gewertet werden können – wie etwa Daten von Schulbehörden);

2. den multivariaten Vergleich zwischen zwei Stichproben auf der Grundlage einfacher log-linearer Modelle.

Beide Algorithmen basieren auf der Anwendung von Monte-Carlo-Simulationen und sind im Anhang abgedruckt. Dort findet sich ebenso ein Beispiel aus dem Sfb,

---

16 Vgl. WIEDENBECK 1982; HARTMANN, SCHIMPL-NEIMANNS 1992, S. 324 ff; zum *"sogenannten"* Mittelschichtbias ESSER ET AL. 1989; zu amtlichen Statistiken: MAYER 1980; WILLMS 1984; HARTMANN, SCHIMPL-NEIMANNS 1992.

17 Auf dieses Problem wird in Teil 3 (Inferenzstrategien) ausführlicher eingegangen.

18 wie etwa StatXact

19 wie etwa LogXact

das einen univariaten Anpassungstest von Stichprobe und Grundgesamtheit (inclusive der entsprechenden asymptotischen Werte) zeigt. Hierbei wird deutlich, daß der entwickelte Prüfalgorithmus zuverlässige Resultate liefert. Multivariate Anpassungstests auf der Grundlage log-linearer Modelle werden ebenfalls an einem Beispiel im Anhang erläutert.

Da das praktische Vorgehen hier weitgehende Überschneidungen mit den im folgenden Kapitel diskutierten Inferenzstrategien hat, soll hier auf diese Aspekte nicht weiter eingegangen werden.

3. Inferenzstrategien bei kleinen Stichproben

Bei der Anwendung inferenzstatistischer Verfahren wird vorausgesetzt, daß es sich bei den erhobenen Merkmalsausprägungen in der Stichprobe um Zufallsvariablen handelt (vgl. etwa KREIENBROCK 1989, S. 124). Die Nutzung dieser Verfahren bei Quotenstichproben oder anderen Formen nichtzufälliger Auswahl wäre somit illegitim.

Dies bedeutet für kleine, nach den oben dargestellten Kriterien gezogene heterogene Stichproben, daß selbst dann, wenn durch die o.g. Anpassungstests eine grobe Verzerrung der Stichprobe auszuschließen ist, nicht automatisch die Voraussetzungen für einen Schluß von dieser Stichprobe auf die zugrundeliegende Grundgesamtheit gegeben sind. Ein solcher Schluß ist jedoch dann möglich, wenn Verfahren der Zufallsauswahl *innerhalb der erhobenen Schichten bzw. Gruppen* benutzt werden und sich Schlüsse auf den Vergleich der erhobenen Gruppen beschränken, nicht aber etwa Aussagen über die Varianz in der Grundgesamtheit beinhalten.

Ein solches Vorgehen, das auf der Auswertung partieller, geschichteter Zufallsstichproben basiert, ist vergleichbar mit experimentellen bzw. quasi-experimentellen Forschungsdesigns, in denen Gruppen, die einem bestimmten “treatment” ausgesetzt sind mit solchen verglichen werden, die diesem “treatment” nicht unterliegen.

Kleine Stichproben können jedoch auch in diesen Fällen zu Konstellationen führen, in denen traditionelle Vorgehensweisen der Inferenzstatistik als rational begründete Entscheidungsstrategien problematisch oder gar ungeeignet sind. Dies trifft insbesondere auf Verfahren kategorialer Datenanalyse zu.

3.1 Probleme kategorialer Datenanalyse bei kleinen Stichproben

Bei kategorialen Daten reduzieren sich Verfahren der Modellbildung oftmals auf mehr oder weniger ausgefeilte Techniken der Kontingenztafelanalyse (χ^2-Techniken, KFA, log-lineare Modelle etc.). Hierbei treten zwei unterschiedliche Problemstellungen auf:

1. Das Problem der *Anwendbarkeit asymptotischer Prüfverteilungen*: Bei der Analyse kategorialer Daten entstehen bei kleinen oder schief verteilten Stichproben oft Situationen, in denen einzelne Ausprägungen von Kategorien oft extrem gering besetzt sind und damit auch die erwarteten Häufigkeiten sehr klein werden. In solchen Fällen gilt die Anwendung asymptotischer Approximationen von Teststatistiken wie etwa der χ^2-Verteilung als problematisch. Wann dies der Fall ist, kann bislang nur auf der Basis von eingebürgerten "Faustregeln" entschieden werden; die am weitesten geteilte ("Cochran's Kriterium") für den χ^2-Wert scheint derzeit zu sein, daß keine erwartete Häufigkeit unter eins und höchstens ein Fünftel der erwarteten Häufigkeiten unter fünf liegen sollten, um asymptotische Approximationen zu erlauben.

2. Das Problem der *Teststärke*: Verglichen mit parametrischen Tests ist die Teststärke der hierbei angewandten Inferenzstrategien relativ gering – und damit die Wahrscheinlichkeit eines Fehlers zweiter Ordnung relativ hoch. Dies führt z.T. zu Situationen, in denen trotz bestehender mittlerer oder starker Zusammenhänge die Nullhypothese auf der Grundlage des vorhandenen empirischen Materials nicht zurückgewiesen werden kann.

Das erste Problem kann in vielen Fällen durch die Anwendung von Testverfahren, die exakte Prüfverteilungen benutzen oder diese mittels Monte-Carlo-Simulationen approximieren, gelöst werden. Dadurch werden kleine Stichproben in bezug auf die *mathematisch-statistischen* Grundannahmen von Inferenzstrategien unproblematisch. Wenngleich jedoch für viele bivariate Verfahren inzwischen Standardanwendungen zur Verfügung stehen, sind für multivariate Analysen Anwendungen nur in geringem Maße und mit z.T. recht weitgehenden Restriktionen vorhanden. In diesem Bereich besteht also ein Entwicklungsbedarf, insbesondere aufgrund der Tatsache, daß das Problem

"kleiner Stichproben" auch in solchen Konstellationen virulent wird, in denen kleine Zellenbesetzungen bei erwarteten Werten nicht durch eine zu geringe Fallzahl der Stichprobe, sondern durch ein komplexes Erklärungsmodell mit einer Vielzahl von Kovariaten entstehen.

Bei der Anwendung der o.g. exakten oder Monte-Carlo-basierten Testverfahren zeigt sich häufig, daß deren *Teststärke* bei kleinen Stichproben deutlich über der asymptotischer Approximationen liegt (vgl. hierzu MEHTA 1993). Was jedoch bislang keines dieser Verfahren ermöglicht, ist die konkrete Abschätzung der jeweiligen Teststärke bzw. des Fehlers zweiter Ordnung.

Zu dessen Bestimmung bzw. Schätzung ist es notwendig a) den Stichprobenumfang, b) das kritische α-Niveau und c) die Verteilung der relevanten statistischen Parameter unter der Altenativhypothese zu kennen. Letztere Bedingung führt häufig dazu, daß eine Bestimmung der Teststärke $(1-\beta)$ bzw. des Fehlers zweiter Ordnung (β) unterbleibt, denn gewöhnlich wird bei Signifikanztests nicht eine spezifizierbare Alternativhypothese überprüft, sondern die Nullhypothese ("Es besteht kein Zusammenhang zwischen ...) gegen eine unendliche Menge möglicher Alternativhypothesen ("Irgendein Zusammenhang in irgendeine Richtung von irgendeiner Stärke besteht.") getestet[20]. Hierdurch wird jedoch die Bedeutung des Fehlers erster Ordnung – der fälschlichen Zurückweisung der Nullhypothese – überbetont.

Die Spezifizierung einer Alternativhypothese erscheint allerdings dann unproblematisch, wenn nicht ein genereller bivariater Zusammenhang, sondern etwa ein log-lineares Modell auf Signifikanz der eingeschlossenen Effekte überprüft werden soll. In diesem Fall ist es problemlos möglich, auf der Grundlage der angenommenen Modellparameter ein Datenmodell unter der Alternativhypothese zu spezifizieren. Wenn dieses auf der Grundlage der vorhandenen Daten zurückgewiesen oder grundsätzlich modifiziert werden muß, auf der Basis zugrundeliegender Theorien allerdings dennoch weiterhin als plausibel erscheint, ist es sinnvoll, neben der Wahrscheinlichkeit eines α-Fehlers auch die Teststärke des benutzten Verfahrens anzugeben. Dies erlaubt eine Abwägung der Plausibilität des präsentierten (und "falsifizierten") Modells: möglicherweise entscheidet es sich deutlich von dem der Nullhypothese, allerdings nicht deutlich genug, um auf

---

[20] Alternativen zu diesem Testmodell wurden bereits in der Vergangenheit diskutiert (vgl. etwa WITTE 1980).

der Grundlage der vorliegenden Daten deren Zurückweisung begründen zu können. Eine solche Abwägung der Plausibilität ersetzt keineswegs Signifikanztests, noch setzt sie deren Ergebnisse außer Kraft. Die Konsequenzen können in solchen Fällen nur darin liegen, die vorgenommene Studie mit einer größeren Stichprobe oder einem anderen Design zu replizieren bzw. andere Strategien anzuwenden, um nach Evidenz bzw. Gegenevidenz für das getestete Modell zu suchen.

3.2 Asymptotische Schätzung der Teststärke bei χ^2-Tests

Bei der Modellbildung mit kategorialen Daten werden i.d.R. χ^2 oder G^2 als Teststatistiken genutzt. Für die asymptotische Schätzung der Teststärke existieren von verschiedenen Autoren Modelle (vgl. COHEN 1977; WITTE 1980; AGRESTI 1990), wobei das Grundprinzip jeweils gleich bleibt und sich die berechneten statistischen Parameter (w^2 oder λ) durch lineare Transformation ineinander überführen lassen. Aus diesem Grund soll hier nur das Verfahren vorgestellt werden, das Agresti beschreibt, da er explizit auf die Approximation der Teststärke bei multidimensionalen Kontingenztafeln und log-linearen Modellen eingeht.

Agrestis Vorgehen basiert auf einer asymptotischen Approximation der nichtzentralen χ^2-Verteilung (AGRESTI 1990, S. 98): Bei entsprechend großen Stichprobenumfängen sind Teststatistiken wie χ^2 und G^2 asymptotisch χ^2-verteilt mit einem Nichtzentralitätsparameter λ. Trifft eine getestete Nullhypothese zu, ist dies der Sonderfall der normalen, zentralen χ^2-Verteilung mit $\lambda=0$; je weiter die ”Wahrheit“ von der Nullhypothese entfernt ist, desto größer wird λ. Dies bedeutet für die Schätzung der Teststärke, daß dieser Parameter genutzt werden kann, um die Wahrscheinlichkeit zu berechnen, daß ein berechneter χ^2-Wert (als Teststatistik) in den Annahmebereich der Nullhypothese fällt.

Hieraus entwickelt er ein vierstufiges Schätzmodell für die Bestimmung der Teststärke bei χ^2-Tests für ein hypothetisches Modell M, das mit “wahren” Verteilungsparametern verglichen werden soll (vgl. AGRESTI 1990, S. 241ff):

1. Bestimmung der angenommenen wahren Zellenwahrscheinlichkeiten $\{\pi_i\}$;

2. Berechnung der Zellwahrscheinlichkeiten gemäß Modell M $\{\pi_i(M)\}$;

3. Berechnung des Nichtzentralitätsparameters λ als

$$\lambda = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left[\pi_{ij} - \pi_{ij}(M)\right]^2}{\pi_{ij}(M)}$$ für die Pearson χ^2-Statistik bzw.

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

4. Bestimmung der Wahrscheinlichkeit, einen χ^2-Wert zu beobachten, der über dem kritischen α-Niveau (d.h. im Ablehnungsbereich der Nullhypothese) liegt, d.h. $p[X^2_{\nu,\lambda} > \chi^2_\nu(\alpha)]$. Zur asymptotischen Bestimmung dieser Wahrscheinlichkeit liegen entsprechende Tabellen der nichtzentralen χ^2-Verteilung von Haynam et al. (1970) vor.

Bei der Modellbildung mit multidimensionalen Kontingenztafeln stellt sich das Problem der Teststärkebestimmung häufig nicht bezogen auf das Gesamtmodell, das etwa gegen die Nullhypothese völliger Unabhängigkeit getestet werden soll, sondern in bezug auf einzelne Parameter eines Modells, die gegen Null abgesichert werden sollen. So ist die Situation denkbar, in der ein Modell M_1, das einen bestimmten Modellparameter enthält, mit einem restringierten Modell M_2, das diesen Parameter nicht enthält, konkurriert. Durch die Berechnung von Partialchiquadraten für die einzelnen Modellparameter kann nun abgeschätzt werden, ob Unterschiede, die auf diesem Parameter beruhen, als signifikant angesehen werden können oder nicht.

Auch hier kann sich analog zur oben beschriebenen Situation das Problem der Teststärke stellen: Wie wahrscheinlich ist es, daß das restringierte Modell M_2 nicht “zurückgewiesen” werden kann, obwohl das “komplexere Modell” M_1 realitätsentsprechender ist? Wir können dies auf die o.g. Teststrategie übertragen, indem die Zellenwahrscheinlichkeiten gemäß dem komplexeren Modell M_1 $\{\pi_i(M_1)\}$ den Platz der “wahren” Zellenwahrscheinlichkeiten $\{\pi_i\}$ einnehmen und die Wahrscheinlichkeiten gemäß dem restringierten Modell M_2 $\{\pi_i(M_2)\}$ den der Wahrscheinlichkeiten gemäß Modell M $\{\pi_i(M)\}$.

3.3 Schätzung der Teststärke über Monte-Carlo-Simulationen

Das Vorgehen, das Agresti vorschlägt, ist nur in den Fällen angemessen, in denen die Stichprobengröße sowie die Verteilung der erwarteten Häufigkeiten asymptotische χ^2-Approximationen legitimieren. Dies ist allerdings nicht der Fall, wenn wir es mit kleinen oder schief verteilten Stichproben zu tun haben, bei denen aber das Problem der geringen Teststärke besonders relevant ist.

Bezüglich "normaler" Signifikanztests erlauben neuere Entwicklungen im Bereich exakter bzw. Monte-Carlo-basierter Tests, auch mit kleinen Stichproben umzugehen. Wenn wir solche Testverfahren jedoch bei der Bestimmung des Risikos eines Fehlers erster Ordnung nutzen, erscheint es unlogisch, bei der Bestimmung der Teststärke auf asymptotische Verfahren zurückzugreifen.

Aus diesem Grunde wurde im Bereich "Methoden und EDV" ein Algorithmus entwickelt, der – dem konkreten Bedarf entsprechend – die Monte-Carlo-Simulation einfacher, hierarchischer log-linearer Modelle (mit direkten Schätzern) erlaubt und somit in Kombination mit der von Agresti beschriebenen Methode für komplexe Kontingenztafeln eine Bestimmung der Teststärke von Partialchiquadraten ermöglicht[21].

Hierzu wurde eine in C++ geschriebene dynamische Link-Library (DLL) in ein Standard-Spreadsheet (MS-Excel 4.0) eingebunden. Dieses Vorgehen wurde gewählt, um die Vorteile dieses Spreadsheets (komfortable, graphische Benutzeroberfläche; Möglichkeit der Makroprogrammierung) mit denen einer schnellen und flexiblen Programmiersprache zu verbinden[22]. Der Algorithmus, der z.T. für die Generierung von Zufallstafeln implementiert ist, besteht aus einer modifizierten C++-Version der FORTRAN-Subroutine RCONT (BOYETT 1979). Die Verwendung von dynamischer Speicherverwaltung in C++ ermöglicht es hierbei, bestimmte Begrenzungen von Boyetts Algorithmus (etwa in bezug die zu testende Stichprobengröße) zu überwinden. Im Algorithmus RCONT wird die

---

[21] Die Anwendung exakter Tests wurde ursprünglich ins Auge gefaßt, aufgrund des immer noch relativ großen Stichprobenumfangs und den daraus resultierend unakzeptablen Rechnerzeiten jedoch wieder verworfen.

[22] Die von uns in C++ gechriebenen Routinen wären auch als Excel-Makro zu realisieren. Da hierdurch die Ausführungsgeschwindigkeit extrem verlangsamt wird - die Generierung von 2000 Tafeln würde selbst bei einfachen Modellen mehrere Tage auf einem PC mit Intel 80486 Prozessor benötigen - nahmen wir von diesem Vorgehen Abstand.

Ziehung von Kugeln (Fällen) aus einer Urne (ohne Zurücklegen) simuliert. Da dieser Algorithmus für die Analyse bivariater Häufigkeitsverteilungen programmiert wurde, sind die Möglichkeiten, die er in Kombination mit Excel bietet, auf die Simulation von Modellen mit direkten Schätzern – d.h. von Modellen, bei denen sich die erwarteten Zellenwahrscheinlichkeiten multiplikativ aus bestimmten Randwahrscheinlichkeiten berechnen lassen – begrenzt[23]. Hinzukommen in der Link-Library Funktionen zur Berechnung von χ^2 und G^2 sowie zur Generierung von Tafeln aus einer Multinominalverteilung.

Die Nutzung des Spreadsheets ermöglicht es, bivariate Tabellen in vielfacher Weise miteinander zu einer mehrdimensionalen Kontingenztafel zu kombinieren. Bei trivariaten Modellen können etwa für den Fall eines Modells {AB} {AC} für jede Ausprägung der Variablen C die bivariaten Tafel A x B gemäß den jeweiligen spezifischen Randsummen und Zellenwahrscheinlichkeiten generiert werden. Diese C Tabellen werden dann im Spreadsheet zu einer Gesamttabelle zusammengefaßt. Für die Simulation des Falls {A} {B} {C}, d.h. für das Modell statistischer Unabhängigkeit, wird zunächst eine Tafel A x B generiert, deren Zellenhäufigkeiten dann als Randverteilung für die Generierung der Tafel A/B x C genutzt werden. Für jede so generierte Gesamttabelle wird der χ^2- oder G^2-Wert berechnet.

Durch die mehrfache[24] Wiederholung dieser Operation kann eine Prüfverteilung simuliert werden, die bei großen Stichproben einer (nichtzentralen) χ^2-Verteilung entspricht.

Hierdurch wird die Übertragung des von Agresti vorgeschlagenen Vorgehens auf kleinere Stichproben möglich. Dabei ergeben sich bei der Bestimmung des β-Fehlers für einen Partialchiquadratwert folgende fünf Arbeitsschritte:

---

[23] Dies war für uns akzeptabel, denn konkret stellte sich uns das Problem der Teststärke im Kontext eines Modells, das diese Bedingungen erfüllte. Mittelfristig ist allerdings geplant, diese Restriktion durch die Integration von Newton-Raphson-Algorithmen zu überwinden. Die derzeit benutzten Algorithmen und Excel-Makros sind im Anhang abgedruckt.

[24] In der Praxis liefert die Generierung von 2.000 Tafeln zufriedenstellende Resultate.

1. Einfügen der empirischen Daten in ein Spreadsheet; Berechnung der erwarteten Werte gemäß eines restringierten Modells M_2, das den getesteten Modellparameter nicht enthält.

2. Bestimmung des kritischen χ^2-Wertes $\chi^2(\alpha)$ für ν Freiheitsgrade (des komplexeren Modells M_1).

3. Generierung von Zufallstafeln entsprechend den suffizienten Statistiken für Modell M_1.

4. Berechnung des χ^2- bzw. G^2-Wertes für jede dieser Tafeln (deren Gesamtverteilung entspricht der nichtzentralen χ^2-Verteilung in Agrestis Modell).

5. Bestimmung des Anteils der berechneten χ^2-Werte, die kleiner oder gleich dem kritischen χ^2-Wert sind. Die relative Häufigkeit dieser Werte gibt die Wahrscheinlichkeit eines Fehlers zweiter Ordnung an. Die Teststärke kann daraus als $1-\beta$ berechnet werden[25].

[25] Ein Bespiel für ein einfaches log-lineares Modell mit Daten aus dem Sfb 186 findet sind im Anhang.

3.4 Schlußfolgerungen für Inferenzstrategien

Inferenzstrategien, die die Schätzung der Teststärke mit einbeziehen, können zu vier unterschiedlichen Resultaten führen:

1. Teststärke hoch, $p(\chi^2)$ > kritischem α-Niveau:
   $\Rightarrow$ Verwerfe H_0, akzeptiere H_1

2. Teststärke hoch, $p(\chi^2)$ ### kritischem α-Niveau:
   $\Rightarrow$ Verwerfe nicht H_0, verwerfe H_1.

3. Teststärke niedrig, $p(\chi^2)$ > kritischem α-Niveau:
   $\Rightarrow$ Verwerfe H_0, H_1 ist evtl. nicht angemessen.

4. Teststärke niedrig, $p(\chi^2) \leq$ kritischem α-Niveau:
   $\Rightarrow$ Verwerfe nicht H_0, H_1 könnte jedoch auch plausibel sein.

In den Fällen eins und zwei sind die Schlußfolgerungen evident: Wenn die Wahrscheinlichkeit, einen Fehler zweiter Ordnung zu begehen, klein ist, können wir den gewohnten Verfahren der Signifikanztestung folgen, um zu entscheiden, ob Modelle oder Modellparameter als überzufällig akzeptiert werden können. Die Fälle drei und vier beziehen sich demgegenüber auf die problematischen Situationen, die bei der Anwendung traditioneller Inferenzstrategien nicht entdeckt werden können:

- Fall drei könnte etwa das Resultat eines stark simplifizierten Modells für einen umfangreichen Datensatz sein: die Annahme der Unabhängigkeit kann auf dem gewählten Signifikanzniveau beruhen, die Teststärke bleibt jedoch klein. Deshalb könnte ein komplexeres Modell angemessener sein.

- Fall vier könnte das Ergebnis sein, wenn kleine Stichproben und/oder komplexe Modelle zu starken Merkmalsstreuungen und/oder kleinen Zellenbesetzungen führen: die Nullhypothese (der vollständigen Unabhängig-

keit oder der Zufälligkeit des Effekts eines bestimmten Modellparameters) kann auf der Grundlage der vorliegenden Daten nicht zurückgewiesen werden. Da die Wahrscheinlichkeit eines Fehlers zweiter Ordnung allerdings auch hoch ist, sollte die Alternativhypothese nicht sofort zurückgewiesen, sondern die Entscheidung aufgeschoben werden. In diesem Falle könnte etwa die Untersuchung mit einer größeren Stichprobe wiederholt werden. Wenn dies – etwa aus forschungsökonomischen Gründen – nicht möglich ist, das komplexere Modell der Alternativhypothese theoretisch jedoch weiterhin als plausibel erscheint, sollte es zumindest Erwähnung finden, selbst wenn es – gemessen an "harten" Testkriterien – zurückgewiesen werden müßte. Dabei sollten jedoch zugleich die Teststatistiken vorgelegt werden, die den Lesern eine Beurteilung der Anpassungsgüte erlauben.

Die Strategie, die hier vorgeschlagen wird, soll also dazu dienen, die Fälle zu identifizieren, in denen "konservative" Teststrategien bei der Modellbildung zu katastrophalen Fehleinschätzungen führen können: insbesondere bei kleinen Stichproben können sie zur Zurückweisung theoretisch plausibler Alternativhypothesen führen. Eine Inferenzstrategie, die die Abschätzung der Teststärke mit einbezieht, kann demgegenüber dem Forscher helfen, die Fälle aufzufinden, in denen das empirische Material als Basis für eine rationale Entscheidung für oder gegen das getestete Modell nicht angemessen ist.

4. Zusammenfassung

Fassen wir die vorangehenden Kapitel thesenartig zusammen, so ergibt sich ein Modell für ein methodisch kontrolliertes Vorgehen bei der Arbeit mit kleinen Stichproben:

1. Können nur kleine Stichproben gezogen werden, so sind die Standardverfahren zur Ziehung von repräsentativen Bevölkerungsstichproben unangemessen. Kleine Stichproben müssen immer nach theoretisch relevanten Kriterien geschichtet sein, um die unbeobachtete Heterogenität der Stichprobe – zumindest in bezug auf diese Schichtungsmerkmale – zu minimieren. Für die Ziehung der Stichprobe ist ein sowohl theoretisch wie empirisch begründeter Stichprobenplan notwendig; ebenso müssen methodologisch begründet Kriterien zur Ziehung innerhalb der einzelnen Schichten formuliert werden (etwa: Zufallsauswahl innerhalb der Vergleichsgruppen), die Klumpungseffekte verhindern oder zumindest reduzieren.

2. Insbesondere bei kleinen Stichproben, die durch Ausfälle zustande gekommen sind, muß ex post durch Anpassungstests die Stichprobengüte abgeschätzt werden. Anpassungstests bieten dabei keine nachträgliche Garantie für die Qualität einer Stichprobe, stellen aber Indikatoren dar, durch die grobe Verzerrungen ausgeschlossen werden können. Alternative Verfahren stehen allerdings nicht zur Verfügung. Bei Anpassungstests ist bei kleinen Stichproben generell zu beachten, daß häufig die Voraussetzungen für die Anwendung asymptotischer Verfahren der Signifikanzberechnung nicht erfüllt sind und damit auf alternative Testverfahren (exakte Tests, Monte-Carlo-Simulationen) zurückgegriffen werden muß.

3. Kleine Stichproben erfordern alternative Inferenzstrategien insbesondere bei der Modellierung. Hierbei müssen die o.g. alternativen Testverfahren Anwendung finden. In Fällen, in denen die Ablehnung theoretisch plausibler Modelle zweifelhaft erscheint, sollten diese außerdem die Berechnung der Teststärke bzw. des Fehlers zweiter Ordnung beinhalten. Im Rahmen der Inferenzstrategien sind immer die Art der Stichpro-

benziehung und die damit verbundenen Möglichkeiten der Verallgemeinerung zu beachten.

Die o.g. Auflistung kann als Abfolge von Arbeitsschritten angesehen werden, die bei der Arbeit mit kleinen Stichproben zu beachten sind:

1. Schritt: Ziehung des Samples nach einem elaborierten Stichprobenplan, z.B. eine nach theoretisch und empirisch relevanten Merkmalen dysproportional geschichtete Zufallsstichprobe.

2. Schritt: Nach der Datenerhebung muß die Qualität der gezogenen Stichprobe mit den zur Verfügung stehenden Mitteln (Anpassungstests) kontrolliert werden.

3. Schritt: Zeigt diese Kontrolle keine groben Verzerrungen, kann die Stichprobe bearbeitet und die Hypothesen mit den o.g. Inferenzstrategien getestet werden.

5. Anhang

5.1 Monte-Carlo-Link-Library zur Generierung von Zufallstafeln (Source-Code)

```
typedef struct _FP
{
       unsigned short int rows;
       unsigned short int cols;
       double array[1];
} FP;

#include <stdlib.h>
#include <windows.h>
#include <stdio.h>
#include <time.h>
#include <stdarg.h>
#include <alloc.h>
#include <limits.h>
#define Word int
#define Double double
#pragma argsused

Word FAR PASCAL LibMain( HANDLE hInst, WORD wData, WORD cbHeap, LPSTR cmd )
{   return 1; }

/*****************************************************************/
/*                                                               */
/* Version 1: fixed row and column margins                       */
/*                                                               */
/*****************************************************************/

void FAR PASCAL _export fest
     (FP far *matrix, FP far *nrowt, FP far *ncolt)

/* matrix              random matrix
   nrowt            ....array with row totals
   ncolt    array with column totals                              */
{
Word              ntemp;             /* temorary variable          */
Word              noct;              /* random number              */
Word far         *nnvect;            /* temporary random array*    */
Word far         *nsubt;             /* array with subtotals       */
Word far         *nvect;             /* random array               */
Word              i=0;               /* index                      */
Word              j=0;               /* index                      */
Word              k=0;               /* index                      */
Word              ii=0;              /* index                      */
Word              ntotal=0;          /* total of cells             */
Word              ncol=0;            /* number of columns          */
Word              nrow=0;            /* number of rows..... .      */
FILE             *errlog;

errlog=fopen("cerror.log","a");
if (errlog==NULL){
    fprintf(stderr,"\n error opening cerror.log");
    return;
    }

srand((unsigned)(matrix->array[0] * USHRT_MAX));
ncol=Word(matrix->cols);
nrow=Word(matrix->rows);

if( (nsubt=(Word far*)farmalloc((unsigned long)(ncol*sizeof(Word))))==NULL){
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough memory for nsubt\n");
    fclose(errlog);
    return;
    }
```

```
if (nrow <= 0) {
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough rows, nrow=%d", nrow);
    fclose(errlog);
    return;
    }

if (ncol <= 1) {
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough columns, ncol=%d", ncol);
    fclose(errlog);
    return;
    }

for (i=0; i< nrow; i++) {
    ntotal+=Word(nrowt->array[i]+0.01);
    }

nsubt[0] = (Word)(ncolt->array[0]+0.01) ;

for (i=1; i < ncol; i++) {
    nsubt[i] = nsubt[i-1] + (Word)(ncolt->array[i]+0.01)  ;
    }

if( (nvect=(Word far*)farmalloc((unsigned long)(ntotal*sizeof(Word))))==NULL){
    matrix->array[0]=-1.0;
    fprintf(errlog,"\n\nnot enough memory for nvect");
    fclose(errlog);
    return ;
    }

for (i=0; i < ntotal; i++)
    nvect[i] = i+1;

if( (nnvect=(Word far*)farmalloc((unsigned long)(ntotal*sizeof(Word))))==NULL){
    matrix->array[0]=-1.0;
    fprintf(errlog, "\n\nnot enough memory for nnvect");
    fclose(errlog);
    return;
    }

for (i=0; i < ntotal; i++)
    nnvect[i] = nvect[i];

ntemp = ntotal;

for (i=0; i<ntotal; i++) {
    noct = rand() % ntemp ;
    nvect[i] = nnvect[noct];
    nnvect[noct] = nnvect[--ntemp];
    }

for (i=0; i<nrow; i++)
    for (j=0; j< ncol; j++)
       matrix->array[i*ncol+j] = 0.0;

for (i=0; i<nrow; i++) {
    for (k=0; k<Word(nrowt->array[i]+0.01); k++) {
       for (j=0; j<ncol; j++)
           if (nvect[ii] <= nsubt[j])
              break;
       ii++;
       matrix->array[i*ncol+j]+=1.0;
       }
    }
farfree(nnvect);
farfree(nvect);
farfree(nsubt);
fclose(errlog);

return;
}
```

```
/*************************************************************************/
/*                                                                       */
/* Version 2: fixed row margins, column probabilities             */
/*                                                                       */
/*************************************************************************/

void FAR PASCAL _export Zeilen
     (FP far *matrix, FP far *nrowt, FP far *colprob)

/* matrix               random matrix
   nrowt                array with row totals
   ncolprob             array with column probabilities                    */
{
Double  far  *nprob;            /* array with cum. cell probabilities      */
Double        rando=0.0         /* random number from 0 to 1               */
Word          i=0;              /* index                                   */
Word          j=0;              /* index                                   */
Word          k=0;              /* index                                   */
Word          ntotal=0;         /* total of cells                          */
Word          ncol=0;           /* number of columns                       */
Word          nrow=0;           /* number of rows                          */
FILE         *errlog;

errlog=fopen("cerror.log","a");
if (errlog==NULL){
    matrix->array[0]=-1.0;
    fprintf(stderr,"\n error opening cerror.log");
    return;
    }

srand((unsigned)(matrix->array[0] * USHRT_MAX));

ncol=Word(matrix->cols);

nrow=Word(matrix->rows);

if (nrow <= 0) {
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough rows, nrow=%d", nrow);
    fclose(errlog);
    return;
    }
if (ncol <= 1) {
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough columns, ncol=%d", ncol);
    fclose(errlog);
    return;
    }
for (i=0; i< nrow; i++) {
    ntotal+=Word(nrowt->array[i]+0.01);
    }
*i=ncol*nrow;
if( (nprob=(Double  far*)farmalloc((unsigned long)(i*sizeof(Double ))))==NULL){
    matrix->array[0]=-1.0;
    fprintf(errlog, "\nnot enough memory for nprob\n");
    fclose(errlog);
    return;
    }

for (i=0;i<nrow;i++){
    nprob[i*ncol] = colprob->array[i*ncol];
    for (j=1; j < ncol; j++)
       nprob[i*ncol+j] = nprob[i*ncol+j-1] + colprob->array[i*ncol+j]  ;
    }

for (i=0; i<nrow; i++)
    for (j=0; j< ncol; j++)
       matrix->array[i*ncol+j] = 0.0;

for (i=0; i<nrow; i++)
    for (k=0; k<Word(nrowt->array[i]+0.01); k++) {
       rando=Double (rand());
       rando/=Double (RAND_MAX);
```

```
       for(j=0;j<ncol;j++)
           if(rando < nprob[i*ncol+j])
              break;
       matrix->array[i*ncol+j]+=1.0;
       }

farfree(nprob);
fclose(errlog);

return;
}

/*************************************************************************/
/*                                                                       */
/* Version 3: fixed margins, cell probabilities                          */
/*                                                                       */
/*************************************************************************/

void FAR PASCAL _export Zellen
     (FP far *matrix, FP far *nrowt, FP far *ncolt, FP far *cellprob)

/* matrix              random matrix
   nrowt               array containing the row totals
   nclot               array containing the column totals
   cellprob            array with cell probabilities                      */

{
Double  far  *nprob;            /* array with cum. cell probabilities      */
Double        rando=0.0;        /* random number from 0 to 1               */
Word    far  *coltemp;          /* temporary array for column totals       */
Word    far  *rowtemp;          /* temporary array for row totals          */
Word          i=0;              /* index                                   */
Word          j=0;              /* index                                   */
Word          k=0;              /* index                                   */
Word          l=0;              /* index                                   */
Word          ntotal=0;         /* total of cells                          */
Word          ncol=0;           /* number of columns                       */
Word          nrow=0;           /* number of rows                          */
FILE         *errlog;

errlog=fopen("cerror.log","a");
if (errlog==NULL){
    matrix->array[0]=-1.0;
    fprintf(stderr,"\n error opening cerror.log");
    fclose(errlog);
    return;
    }

srand((unsigned)(matrix->array[0] * USHRT_MAX));

ncol=Word(matrix->cols);

nrow=Word(matrix->rows);

if (nrow <= 0) {
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough rows, nrow=%d", nrow);
    fclose(errlog);
    return;
    }

if (ncol <= 1) {
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough columns, ncol=%d", ncol);
    fclose(errlog);
    return;
    }

if( (coltemp=(Word far*)farmalloc((unsigned long)(ncol*sizeof(Word))))==NULL){
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough memory for coltemp\n");
    fclose(errlog);
    return;
    }
```

```
for (i=0;i<ncol;i++)
    coltemp[i]=(Word)(ncolt->array[i]+0.01);

if( (rowtemp=(Word far*)farmalloc((unsigned long)(nrow*sizeof(Word))))==NULL){
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough memory for rowtemp\n");
    fclose(errlog);
    farfree(coltemp);
    return;
    }

for (i=0;i<nrow;i++)
    rowtemp[i]=(Word)(nrowt->array[i]+0.01);

for (i=0; i< nrow;i++)
    ntotal+=(Word)(nrowt->array[i]+0.01);

i=ncol*nrow;
if( (nprob=(Double  far*)farmalloc((unsigned long)(i*sizeof(Double ))))==NULL){
    matrix->array[0]=-1.0;
    fprintf(errlog,"\nnot enough memory for nprob\n");
    fclose(errlog);
    farfree(rowtemp);
    farfree(coltemp);
    return;
    }

nprob[0] = cellprob->array[0];

for (i=1;i<nrow*ncol;i++)
    nprob[i] = nprob[i-1] + cellprob->array[i]  ;

for (i=0; i<nrow; i++)
    for (j=0; j< ncol; j++)
       matrix->array[i*ncol+j] = 0.0;

for (i=0; i<ntotal; i++) {
    rando=Double (rand());
    rando/=Double (RAND_MAX);
    for(j=0;j<ncol*nrow;j++)
        if(rando < nprob[j])
           break;
    k=(int)(j/ncol);
    l=(int)(j%ncol);
    if (coltemp[l] > 0 && rowtemp[k] > 0) {
        coltemp[l]--;
        rowtemp[k]--;
       matrix->array[j]+=1.0;
       }
    else i--;
    }

fclose(errlog);
farfree(coltemp);
farfree(rowtemp);
farfree(nprob);
return;
}
```

Beispiel für ein Excel-Makro (s. Abschnitt 5.3)

POWERTEST

```
=ECHO(FALSCH)

repeat=2000

plr=0
```

```
=FÜR("i";1;repeat;1)
  test1= MONTECAR.XLM!Zellen(!Matrix1;!Zeilen1;!Spalten1;!prob)
  test2 = MONTECAR.XLM!Fest(!Matrix2;!Zeilen2;!Spalten2)
  test3 = MONTECAR.XLM!Fest(!Matrix3;!Zeilen3;!Spalten3)
= WENN(ODER(test0>0; test1 >0;test2>0))
=   MONTECAR.XLM!lrchi2(tester;!Matrix4;!m2_)
=   WENN(tester<=!ALPHA;NAMEN.ZUWEISEN("plr";plr+1); )
= SONST()
    i=i-1
=   ENDE.WENN()
= MELDUNG(WAHR;"Tafel "&i&" p="&plr/i)
= WEITER()
=ECHO(WAHR)
=FORMEL(plr/repeat;!LRp)
=RÜCKSPRUNG()
```

5.2 Beispiel für einen χ^2-Anpassungstest

Beim Vergleich von Daten aus dem Teilprojekt B1 mit denen der amtlichen Rentenstatistik handelt es sich nicht um den Vergleich zweier Stichproben, sondern einen χ^2-Anpassungstest, bei dem die von der amtlichen Statistik (als Vollerhebung) vorgegebenen Zellenwahrscheinlichkeiten zur Berechnung der erwarteten Häufigkeiten benutzt werden.

Daten des Projekts B1 und der amtlichen Rentenstatistik

Verrentungsalter	**unter 65**	**65**	**über 65**
Datensatz B1 (absolute Zahlen)	52	6	1
amtliche Rentenstatistik (Prozent)	81,5	16,4	1,9
erwartete Häufigkeiten	48,09	9,68	1,12

$\chi^2=1{,}73$ df=2 $p_{asympt.}=0{,}4214$ $p_{Monte\text{-}Carlo}=0{,}3705$

Ein Drittel der erwarteten Häufigkeiten liegt unter fünf. Die Voraussetzungen für eine asymptotische Approximation der χ^2-Verteilung liegen demnach nicht vor. Dies wird besonders deutlich, wenn man die Wahrscheinlichkeiten für einen größeren χ^2-Wert berechnet: mit dem Monte-Carlo-Algorithmus ergibt sich ein p von 0,3705, bei zwei Freiheitsgraden liegt das asymptotische p dagegen bei 0,4214. Beide Tests zeigen eine überzufällige Übereinstimmung, das asymptotische Verfahren überschätzt diese allerdings um mehr als fünf Prozent.

Entsprechen die erwarteten Häufigkeiten den genannten Kriterien für die Anwendung asymptotischer Tests, nähern sich die Ergebnisse von Monte-Carlo-Simulation und asymptotischer Approximation, wie folgendes Beispiel einer leicht modifizierten Tabelle mit fiktiven Daten deutlich macht, in dem die Differenz der beiden p unter einem halben Prozentpunkt liegt:

Daten des Projekts B1 und fiktive amtliche Daten			
Verrentungsalter	**unter 65**	**65**	**über 65**
Datensatz B1 (absolute Zahlen)	52	6	1
fiktive amtliche Daten (Prozent)	75,0	16,4	8,6
erwartete Häufigkeiten	44,25	9,68	5,07

χ^2=6,023 df=2 $p_{asympt.}$=0,0492 $p_{Monte\text{-}Carlo}$=0,045

5.3 Beispiel für die Bestimmung der Teststärke

Im Teilprojekt B1 wurde im Rahmen einer querschnittsorientierten Zwischenauswertung[1] ein loglineares Modell mit den drei Variablen Beruf des Mannes, erlernter Erstberuf der Frau und Berufsstatus der Frau (15 Jahre nach Ausbildungsabschluß im Erstberuf tätig oder nicht) geschätzt. Hieraus ergab sich folgende dreidimensionale Tabelle:

Beruf des Mannes (v1)	**Arbeiter**		**Angestellter Beamter**		**Selbständiger**	
Frau im Erstberuf (v2)	**ja**	**nein**	**ja**	**nein**	**ja**	**nein**
Erstberuf der Frau (v3)						
kaufm. Angestellte	2	1	3	13	0	3
andere Berufe	3	10	2	28	0	6

Geschätzt wurden zwei Modelle: Modell eins beinhaltet Interaktionen sowohl zwischen dem aktuellen Berufsstatus der Frau und dem Berufsstatus des Mannes wie zwischen dem aktuellen Berufsstatus der Frau und dem von ihr erlernten Beruf. Die erstgenannte Beziehung läßt sich aus der gängigen Literatur zu diesem Thema ableiten, die zweitgenannte war das Resultat der Analysen in der vorangegangenen Förderphase des Projekts. Dieses Modell hatte eine sehr gute Anpassung

[2]6 Das Projekt B1 hat im weiteren Verlauf seiner Arbeiten deutlich gezeigt, daß für eine adäquate Betrachtungsweise der Projektdaten eine Längsschnittsperspektive notwendig ist. Die vorgestellten Daten zeigen also nur Zwischenergebnisse, die normalerweise in Forschungszusammenhängen unpubliziert bleiben.

(s.u.), der relativ hohe Partial-χ^2 für v2*v3 mit einem p von 0,0687 läßt jedoch den Verdacht einer Überparametrisierung aufkommen.

Deshalb wurde Modell zwei geschätzt, das diesen Interaktionseffekt nicht beinhaltete. Die Anpassung sank, ohne auf ein inakzeptables Niveau zu fallen.

Aus diesem Grunde wurden die Teststärke bzw. das Risiko eines Fehlers zweiter Ordnung untersucht. Bei einem λ von 1,84 konnte die Teststärke asymtotisch auf 0,18 geschätzt werden, d.h. in etwa 80 % aller Fälle würde die Nullhypothese (“Der Effekt v2*v3 ist Null.”) *fälschlich* nicht abgelehnt werden, selbst wenn dieser Effekt bestünde.

	χ^2	**df**	**asymtotisch**	**Monte-Carlo**
Modell 1 {v1,v2}{v3,v2}	2,76	2	p=0,251	p=0,44
Modell 2 {v1,v2}{v3}	4,66	3	p=0,199	p=0,34
Partial-χ^2 {v2*v3} (Likelihood-ratio)	3,31	2	p=0,069	
λ=1,84			Power=0,18	Power=0,32

Aufgrund der kleinen Zellenbesetzungen wurden sowohl die Teststärke wie die Modellanpassung über Monte-Carlo-Simulationen geschätzt. In beiden Fällen zeigten sich deutlich erhöhte Werte: Die Modellanpassung für beide Modelle ist deutlich besser, als asymptotische Verfahren angeben (0,44 statt 0,25 bzw. 0,34 statt 0,20); ebenso schätzen sie eine deutlich höhere Testpower (0,32 statt 0,18).

Dennoch ist im Rahmen der oben beschriebenen Inferenzstrategie auch aufgrund des höheren Teststärkewertes das Risiko eines Fehlers zweiter Ordnung mit etwa 70 % deutlich zu hoch, um eine rational begründete Entscheidung über das angemessene Modell zu erlauben. Diese konnte erst im weiteren Verlauf der Arbeit auf der Grundlage von längsschnittorientierten Auswertungen getroffen werden.

6. Literatur

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

ARMINGER, GERHARD (1989): *Alternative Erhebungsmethoden bei Bevölkerungsstichproben.* Statistische Grundlagen, Analysen und Konsequenzen am Beispiel des Mikrozensus. Wuppertal, Forschungsbericht

ARMINGER, GERHARD (1990): Pflicht- versus Freiwilligenerhebung im Mikrozensus. In: *Allgemeines Statistisches Archiv*, 74, S. 161-187

BLOSSFELD, HANS-PETER (1985): *Zur Repräsentativität der Sfb-3-Lebensverlaufsstudie – Ein Vergleich mit Daten aus der amtlichen Statistik*, Sfb 3 Arbeitspapier Nr. 163

BORTZ, JÜRGEN; LIENERT, GUSTAV A.; BOEHNKE, KLAUS (1990): *Verteilungsfreie Methoden in der Biostatik*. Berlin; Heidelberg, New York; London; Paris; Tokyo; Hong Kong; Barcelona: Springer

BOYETT, JAMES M. (1979): Algorithm AS 144. Random R x C Tables with Given Row and Column Totals. In: *Applied Statistics*, Volume 28, 1979, S. 329-332

BÜSSING, RENÉ; JANSEN, BERTHOLD (1988): Exact Tests of Two-Dimensional Contingency Tables: Procedures and Problems. In: *Methodika*, Vol. II, Issue 1, S. 27-39

COHEN, JACOB (1977): *Statistical Power Analysis for the Behavioral Sciences.* New York; San Francisco; London: Academic Press

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

DENZIN, NORMAN K. (1978): *The Research Act*. 2. Aufl., New York: McGraw Hill

ERBSLÖH, BARBARA; KOCH, ACHIM (1988): Die Non-Response-Studie zum ALLBUS 1986: Problemstellung, Design, erste Ergebnisse. In: *ZUMA-Nachrichten*, Nr. 22, S. 31-44

ESSER, HARTMUT; GROHMANN, HEINZ; MÜLLER, WALTER; SCHÄFFER, KARL-AUGUST (1989): *Mikrozensus im Wandel*. Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung. Band 11 der Schriftenreihe Forum der Bundesstatistik, Herausgeber: Statistisches Bundesamt Wiesbaden, Stuttgart: Metzler-Poeschel

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

GLASER, BARNEY. G.; STRAUSS, ANSELM. L. (1967, zuerst): *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago 1974

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

HANEFELD, U. (1987): *Das Sozio-ökonomische Panel.* Grundlagen und Konzeption. Frankfurt/M.

HARTMANN, PETER H. (1990): Wie repräsentativ sind Bevölkerungsumfragen? Ein Vergleich des ALLBUS und des Mikrozensus. In: *ZUMA-Nachrichten,* Nr. 26, S. 7-30

HARTMANN, PETER H.; SCHIMPL-NEIMANNS, BERNHARD (1992): Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, Jg. 44, Heft 2, S. 315-340

HARTMANN, PETER H.; SCHIMPL-NEIMANNS, BERNHARD (1993): Affirmative Repräsentativitäts"beweise" oder Test konkreter Hypothesen zu Verteilungsabweichungen. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, Jg. 45, Heft 2, S. 359-365

HAYNAM, G. E.; GOVINDARAJULU, Z.; LEONE, F.C. (1970): Tables of the cumulative non-central chi-square distribution. In: HARTER, H. L.; OWEN, D. B. (Hg.): *Selected Tables in Mathematical Statistics.* Chicago: Markham

HEINZ, WALTER; KRÜGER, HELGA; RETTKE, URSULA; WACHTVEITL, ERICH; WITZEL, ANDREAS (1985): *Hauptsache eine Lehrstelle.* Jugendliche vor den Hürden des Arbeitsmarkts. Weinheim; Basel: Beltz

HELBERGER, CHRISTOF (1987): Zur Repräsentativität des Sozio-ökonomischen Panels am Beispiel der Ergebnisse zur Struktur der Erwerbstätigen. In: KRUPP, H.-J.; HANEFELD, U. (Hrsg.): *Lebenslagen im Wandel*. Analysen. Frankfurt/M., S. 273-294

HERBERGER, LOTHAR (1985): Aktualität und Genauigkeit der repräsentativen Statistik der Bevölkerung und des Erwerbslebens. In: *Allgemeines Statistisches Archiv 69,* S. 16-55

KAPLITZKA, G. (1982): Die Stichprobe. In: HOLM, Kurt (Hg.): *Die Befragung*. Bd. 1, München, 2. Aufl., S. 136-186

KELLE, UDO; KLUGE, SUSANN; PREIN, GERALD (1993): *Strategien der Geltungssicherung in der qualitativen Sozialforschung. Zur Validitätsproblematik im interpretativen Paradigma.* Bremen: Arbeitspapiere des Sfb 186 Nr. 24

KERLINGER, FRED N. (1978): *Grundlagen der Sozialwissenschaften,* Bd. 1, Weinheim; Basel, 2. Aufl.

KIRSCHNER, HANS-PETER (1980): Komplexe Bevölkerungsstichproben: Eine Fallstudie. In: STENGER, HORST (Hg.): *Praktische Anwendungen von Stichprobenverfahren*. Sonderhefte zum Allgemeinen Statistischen Archiv, Göttingen, S. 69-84

KIRSCHNER, HANS-PETER (1984a): ALLBUS 1980. Stichprobenplan und Gewichtung. In: MAYER, KARL ULRICH.; SCHMIDT, PETER (Hg.): *Allgemeine Bevölkerungsumfragen der Sozialwissenschaften*, Frankfurt; New York: Campus, S. 114-182

KIRSCHNER, HANS-PETER (1984b): Zu Stichprobenfehlerberechnungen im Rahmen des ADM-Stichprobenplans. In: *ZUMA-Nachrichten,* Nr. 15, Mannheim, S. 40-71

KIRSCHNER, HANS-PETER (1985): Stichprobenprobleme in der Bundesrepublik – Bestandsaufnahme, Perspektiven, Thesen. In: KAASE, MAX; KÜCHLER, MANFRED (Hg.): *Herausforderungen der Empirischen Sozialforschung.* Beiträge aus Anlaß des zehnjährigen Bestehens des Zentrums für Umfragen, Methoden und Analysen, ZUMA e.V. Mannheim, S. 146-157

KREIENBROCK, LOTHAR (1989): *Einführung in die Stichprobenverfahren.* Lehr- und Übungsbuch der angewandten Statistik. München; Wien: Oldenbourg

KREUTZ, HENRIK (1970/71): Die tatsächliche Repräsentativität soziologischer Befragungen. Eine begriffliche und empirische Analyse der Nichtbeteiligung an Befragungen. In: *Angewandte Sozialforschung.* Informationen der Arbeitsgemeinschaft für interdisziplinäre angewandte Sozialforschung (AIAS), Heft 3/4, S. 242-262

KRIZ, JÜRGEN; LISCH, RALF (1988): *Methodenlexikon für Mediziner, Psychologen, Soziologen.* München; Weinheim: Psychologie Verlags Union.

KROMREY, HELMUT (1987): Zur Verallgemeinerbarkeit empirischer Befunde bei nicht-repräsentativen Stichproben. Ein Problem sozialwissenschaftlicher Begleitung von Modellversuchen und Pilotprojekten, illustriert am Bildschirmtext-Feldversuch Düsseldorf/Neuss. In: *Rundfunk und Fernsehen,* Jg. 35, Heft 4, S. 478-499

MAYER, KARL ULRICH (1980): *Amtliche Statistik und Umfrageforschung als Datenquellen der Soziologie*. Mannheim, VASMA-Arbeitspapier Nr. 16

MAYNTZ, RENATE; HOLM, KURT; HÜBNER, PETER (1969): *Einführung in die Methoden der empirischen Soziologie.* Opladen: Westdeutscher Verlag

MEHTA, CYRUS R. (1993): *Exact Nonparametric Inference.* Skript zum Tutorium auf der SoftStat '93, Heidelberg, 18. März 1993, Ms.

MEHTA, CYRUS R.; PATEL, NITIN R. (1983): A Network Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. In: *JASA 78,* S. 427-434

MEINEFELD, WERNER (1985): Die Rezeption empirischer Forschungsergebnisse – eine Frage von Treu und Glaube? Resultate einer Analyse von Zeitschriftenartikeln. In: *Zeitschrift für Soziologie,* Jg. 14, Heft 4, S. 297-314

NOELLE-NEUMANN, ELISABETH; PIEL, E. (Hg.)(1984): *Allensbacher Jahrbuch der Demoskopie 1978-1983.* Band VIII, München

PATEFIELD, W. M. (1981): Algorithm AS 159. An Efficient Method of Generating Random R x C Tables with Given Row and Column Totals. In: *Applied Statistics,* Vol. 30, S. 91-97

PORST, ROLF (1984): Haushalte und Familien 1982. Zur Erfassung und Beschreibung von Haushalts- und Familienstrukturen mit Hilfe repräsentativer Bevölkerungsumfragen. In: *Zeitschrift für Soziologie*, Jg. 13, Heft 2, S. 165-175

PREIN, GERALD; KELLE, UDO (1993): *Estimation of Beta-error in Multivariate Modelling with Small Samples.* Vortrag auf der SOFTSTAT '93. Heidelberg. Erscheint in: FAULBAUM, FRANK (Hg.): *SOFTSTAT '93 – Fortschritte der Statistik-Software.* Stuttgart: Gustav Fischer Verlag

PREIN, GERALD; KELLE, UDO; KLUGE, SUSANN (1993): *Strategien zur Integration quantitativer und qualitativer Auswertungsverfahren.* Bremen: Arbeitspapiere des Sfb 186 Nr. 19

PROJEKTGRUPPE "DAS SOZIO-ÖKONOMISCHE PANEL" (1990): Das Sozio-ökonomische Panel für die Bundesrepublik Deutschland nach fünf Wellen. In: *Vierteljahreshefte zur Wirtschaftsforschung,* Deutsches Institut für Wirtschaftsforschung (DIW), Heft 2/3, S. 141-151

PROJEKTGRUPPE "DAS SOZIO-ÖKONOMISCHE PANEL" (1991): Das Sozio-ökonomische Panel (SOEP) im Jahre 1990/91. In: *Vierteljahreshefte zur Wirtschaftsforschung*, Deutsches Institut für Wirtschaftsforschung (DIW), Heft 3/4, S. 146-155

RENDTEL, ULRICH; PÖTTER, ULRICH (1993): "Empirie" ohne Daten. Kritische Anmerkungen zu einer Repräsentativitätsstudie über den Allbus. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*. 45. Jg., Heft 2, S. 350-358

RÖSCH, GÜNTHER (1985): ADM-Design und Einwohnermelderegister – Ein Kommentar aus der Praxis. In: KAASE, MAX; KÜCHLER, MANFRED (Hg.): *Herausforderungen der Empirischen Sozialforschung.* Beiträge aus Anlaß des zehnjährigen Bestehens des Zentrums für Umfragen, Methoden und Analysen, ZUMA e.V., Mannheim, S. 159-169

ROLLER, EDELTRAUT; MATHES, RAINER (1993): Hermeneutisch-klassifikatorische Inhaltsanalyse. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie.* 45 Jg., Heft 1, S. 56-75

ROTH, ERICH (1987): *Sozialwissenschaftliche Methoden.* Lehr- und Handbuch für Forschung und Praxis. 2. Auflage, München; Wien: Oldenbourg

ROTHE, GUNTER; WIEDENBECK, MICHAEL (1987): Stichprobengewichtung: Ist Repräsentativität machbar? In: *ZUMA-Nachrichten,* Nr. 21, S. 43-58

SAHNER, HEINZ (1971): *Schließende Statistik.* Stuttgart: Teubner

SCHATZMAN, LEONARD; STRAUSS, ANSELM L. (1973): *Field Research: Strategies for a Natural Sociology.* Englewood Cliffs, NJ: Prentice Hall Inc.

SCHEUCH, ERWIN K. (1974): Auswahlverfahren in der Sozialforschung. In: KÖNIG, RENÉ (HRSG.): *Handbuch der empirischen Sozialforschung,* Bd. 3a, 3. Aufl., Stuttgart: Enke, S. 1-96

SCHIMPL-NEIMANNS, BERNHARD (1991): *Zur Repräsentativität soziodemographischer Merkmale des ALLBUS im Vergleich mit dem Mikrozensus. Schaubilder und Tabellen zum Vortrag,* ZUMA e.V. Mannheim

SCHNELL, RAINER (1991): Wer ist das Volk? Zur faktischen Grundgesamtheit bei "allgemeinen Bevölkerungsumfragen": Undercoverage, Schwererreichbare, Nichtbefragbare. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, Jg. 43, Heft 1, S. 106-137

SCHNELL, RAINER; HILL, PETER B.; ESSER, ELKE (1989): *Methoden der empirischen Sozialforschung*, 2. überarb. und erweit. Aufl., München; Wien: Oldenbourg

STRAUSS, ANSELM L. (1991): *Grundlagen qualitativer Sozialforschung*. Datenanalyse und Theoriebildung in der empirischen soziologischen Forschung. München: Fink

WIEDENBECK, MICHAEL (1982): Zum Problem repräsentativer Querschnitte von kleinen Teilgruppen der Bevölkerung am Beispiel des Projektes "Lebensverläufe und Wohlfahrtsentwicklung". In: *ZUMA-Nachrichten*, Nr. 10, Mannheim, S. 21-34

WIEDENBECK, MICHAEL (1984): *Zur Repräsentativität bundesweiter Befragungen.* Ein systematischer Mittelstands-Bias? Mannheim

WILLMS, ANGELIKA (1984): *Die Erforschung sozialer Tatsachen mit amtlichen Statistiken*. Mannheim, VASMA-Arbeitspapier Nr. 39

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

ZETTERBERG, HANS L. (1965): *On Theory and Verification in Sociology*, New York