MPIfG Working Paper 03/7, Juli 2003

Augäpfel, Murmeltiere und Bayes: Zur Auswertung stochastischer Daten aus Vollerhebungen

von Andreas Broscheid und Thomas Gschwend

Max-Planck-Institut für Gesellschaftsforschung
Paulstraße 3
50676 Köln
Germany

Telephone	0221/2767-0
Fax	0221/2767-555
E-Mail	info@mpi-fg-koeln.mpg.de
Website	www.mpi-fg-koeln.mpg.de

Zusammenfassung

In diesem Papier diskutieren wir theoretisch-methodologische Grundlagen zur Analyse so genannter Vollerhebungen, also Datensätze, die Beobachtungen aller Elemente einer Population enthalten. Solche Datensätze spielen vor allem in quantitativen Makro-Analysen politischer und sozialer Systeme eine Rolle, und ihre inhärenten Probleme führen oft zu methodischer Verwirrung, die wir mit dem vorliegenden Essay verringern wollen. Da Vollerhebungen nicht das Resultat einer Zufallsstichprobe sind, ist die Anwendung frequentistischer Wahrscheinlichkeitskonzeptionen zur Begründung inferentieller statistischer Methoden nicht gegeben; außerdem kann die statistische Unabhängigkeit der Beobachtungen voneinander nicht ohne weiteres angenommen werden. Dennoch werden Vollerhebungsdaten durch stochastische Komponenten oder „Fehler“ beeinflusst. Wir argumentieren, dass die Stochastizität der Daten in die Analyse einbezogen werden muss, etwa in Form von Parameter-Varianzen, Signifikanztests, oder Konfidenzintervallen. Wir diskutieren verschiedene theoretische Strategien, mit denen Analysen der Stochastizität begründet werden können, wobei wir vor allem für die Annahme von Superpopulationen oder die Anwendung bayesianischer Ansätze plädieren.

Abstract

This paper discusses the theoretical and methodological foundation for the analysis of apparent populations, i.e., data sets that include observations of all elements of a population. Such data sets are commonly used in quantitative macro studies of political or social systems. Our essay tries to reduce methodological problems resulting from the fact that apparent population data are not the result of random sampling designs. First, the lack of random sampling prevents the use of the frequentist interpretation of probability commonly employed to justify inferential statistical methods. Second, with apparent populations, we cannot assume that observations are statistically independent. We argue that apparent population data are subject to a variety of stochastic processes, or “errors,” that have to be part of the analysis, for example through the investigation of parameter variances, significance tests, or confidence intervals. We discuss several theoretical strategies to justify the analysis of stochastic components of apparent populations, emphasizing in particular the concept of superpopulations and the usefulness of Bayesian approaches.

Inhalt

1 Einleitung

Makrodaten werden immer leichter verfügbar. Das eröffnet neue Chancen und Risiken für die Sozialwissenschaften. Die Chancen sind evident: Neben qualitativen werden zunehmend auch quantitative Makroanalysen ermöglicht. Von vergleichenden Analysen zwischen Staaten, Organisationen, Verbänden oder sonstigen Akteuren können alle Teildisziplinen der Sozialwissenschaften erheblich profitieren. Das Risiko besteht jedoch darin, dass der kumulative Fortschritt in sozialwissenschaftlicher Methodenlehre nicht mit dem Tempo der leichteren Verfügbarkeit der Makrodatensätze Schritt halten kann. Ziel dieses Papiers ist es deshalb, einige Probleme aufzuzeigen, die sich bei der Analyse bestimmter Makrodaten ergeben, und auf Ansätze zur Lösung dieser Probleme hinzuweisen.

Eine Charakteristik von vielen Makroanalysen ist, dass die zugrunde liegende Analyseebene nur wenige Ausprägungen hat, so dass bei der Datenerhebung vernünftigerweise auf alle verfügbaren Daten zurückgegriffen wird. Man spricht dabei von einer *Vollerhebung*: Stichproben, bei denen die Stichprobengesamtheit der Grundgesamtheit entspricht; alle Elemente der Grundgesamtheit sind in der Vollerhebung enthalten. Während der Umgang mit Daten aus Zufallsstichproben in der sozialwissenschaftlichen Methodenlehre unstrittig ist, gibt es erstaunlich konträre Ansichten in der Forschung, die in verschiedensten sozialwissenschaftlichen Zeitschriftenbeiträgen, Kolloquien und Dossiers von Politikberatern zu finden sind, wie denn statistische Analysen basierend auf Vollerhebungen durchzuführen und zu interpretieren seien. Ein Beispiel für die Unklarheit darüber, wie statistische Schätzwerte, die auf einer Vollerhebung basieren, interpretiert werden sollen, ist die Kunz-Obinger-Debatte, die in der Kölner Zeitschrift für Soziologie und Sozialpsychologie „ausgetragen“ wurde (Kunz 2000, 2001; Obinger 2001). Teil der Auseinandersetzung zwischen Kunz und Obinger war die Frage, ob bestimmte Schätzparameter signifikant waren und ob Signifikanztests bei Vollerhebungen (Kunz analysierte OECD-Daten) überhaupt anwendbar seien.

Der Dissens über den Umgang mit Vollerhebungsdaten schlägt sich in der Existenz zwei konträrer, aber üblicher Analysestrategien nieder. Die erste Strategie könnte als die *SPSS-Strategie* bezeichnet werden.[1] Die Daten werden – etwas überspitzt ausgedrückt – durch eines der üblichen Softwarepakete geschickt, Regressionsmodelle (oder ähnliche) werden geschätzt und die Outputs wie üblich berichtet. In der Regel werden dabei Signifikanztests berechnet, die auf der Annahme einer Zufallsstichprobe beruhen; die möglichen Besonderheiten von Vollerhebungsdaten werden schlichtweg ignoriert.

Die zweite Strategie könnte als die *Determinismus-Strategie* bezeichnet werden. Wissenschaftler, die dieser Strategie folgen, argumentieren, dass bei Vollerhebungen kein Fehler aus der Stichprobenziehung herrührt (was korrekt ist) und man sich deshalb bei der Interpretation auf „substanzielle Effekte“ der errechneten Koeffizienten beschränken kann (was – wie wir im zweiten Abschnitt zeigen – *nicht* korrekt ist). Signifikanztests, Fehlervarianzen oder Konfidenzintervalle sind nach dieser Strategie nicht zu berücksichtigen, da ja keine Stichprobenfehler vorhanden sind. Dahinter steckt die Annahme, dass Beziehungen innerhalb der Daten in Vollerhebungen fehlerfrei und damit deterministisch beschrieben werden können (Berk, Western und Weiss 1995). Übersehen

Für Anregungen und Kommentare danken wir Axel Becker, Steffen Ganghof, Martin Heipertz, Bernhard Kittel und Lothar Krempel.

1 Die Begriffswahl soll kein bestimmtes Statistikprogramm kritisieren, zumal die Autoren es selbst gerne verwenden. Wir benutzen hier *SPSS* lediglich als Beispiel eines Programms, das aufgrund seiner *Clickability* dazu verleiten kann, subtilere methodische Fragestellungen zu vernachlässigen.

wird allerdings, dass es viele weitere Fehlerquellen sowohl bei der Operationalisierung als auch bei der Datenerhebung gibt, die nichts mit Stichprobenfehlern zu tun haben.

Häufig – so auch in der Kunz-Obinger Debatte – werden diese Strategien iterativ eingesetzt. Zuerst wird die *SPSS-Strategie* angewandt, und die resultierenden signifikanten Ergebnisse, so sie interessant sind, veröffentlicht. Wenn die Ergebnisse dann mit dem Argument kritisiert werden, die Signifikanztests, auf denen sie beruhen, seien ungültig, so wird die *Determinismus-Strategie* eingesetzt: Da die Daten fehlerfrei (gemeint ist hier, frei von Stichprobenfehlern) seien, wäre ein Signifikanztest auch nicht nötig gewesen, was die Gültigkeit der berichteten Ergebnisse nur noch unterstreiche.

Uns geht es hier nicht darum, den Zeigefinger auf andere zu richten, da gerade auch wir uns gemäß der *SPSS-Strategie* „schuldig" gemacht haben (Broscheid und Teske 2003; Gschwend 2001; Gschwend, Johnston und Pattie 2003; Gschwend und Norpoth 2000, 2001). Dennoch wollen wir die gängige Praxis kritisieren und Lösungen aufzeigen. Vor allem wollen wir in diesem Beitrag die Leser davon überzeugen, dass die *Determinismus-Strategie* zu fehlerhaften Schlussfolgerungen führt, da ein wichtiges Element der kritischen Analyse quantitativer Ergebnisse ausgelassen wird – die Erkenntnis, dass Daten durch Stochastizität (man könnte auch sagen „Unschärfen") geprägt sind. Wenn wir die Unschärfen in den Daten ignorieren, ignorieren wir einen wesentlichen Teil der Information, die in den Daten enthalten ist und unseren Schlussfolgerungen möglicherweise widersprechen oder diese zumindest abschwächen könnte.

Unsere Argumentation ist wie folgt strukturiert. Im nächsten Abschnitt zeigen wir, dass die Unschärfe quantitativer Daten verschiedene Quellen hat, von denen die Stichprobenziehung nur eine ist. Im dritten Teil wollen wir klar machen, dass selbst bei Vollerhebungen noch genügend Quellen der Stochastizität in den Daten vorhanden sind, die in jeder vernünftigen statistischen Analyse berücksichtigt werden müssen. Im vierten Teil zeigen wir verschiedene theoretische Ansätze, die bei Vollerhebungen statistische Hypothesentests und damit die Verwendung von Unsicherheitsmaßen möglich machen. Dieser Teil bildet gewissermaßen das methodologische Fundament, das bisher Nutzern der *SPSS-Strategie* fehlte. Abschließend soll die Bedeutung unserer Argumentation nicht nur für die methodisch interessierte Leserschaft, sondern gerade auch für substanziell orientierte Sozialwissenschaftler zusammengefasst werden.

Unsere Argumentation richtet sich vor allem an substanziell orientierte Wissenschaftler. Eines unserer Hauptargumente ist, dass formale methodologische Aussagen auch substanzielle Aussagen darstellen. Die Analyse der Stochastizität von Daten ist nicht nur notwendige technische Arbeit, sondern führt uns zu substanziellen Erkenntnissen der sozialen Wirklichkeit, die wir erforschen. Obwohl wir zur Erläuterung unserer Argumente einige formale Notation einführen und wir einige Grundkenntnisse in quantitativer Methodik (etwa linearer Regressionsanalyse) voraussetzen, halten wir die Diskussion in allgemeinverständlicher Sprache und illustrieren unser Argument mit verschiedenen Beispielen.

2 Fehlerquellen – Quellen der Stochastizität

2.1 Formale Notation

Sozialwissenschaftler sind an substanziellen Aussagen über soziale Systeme, zum Beispiel Individuen, Gruppen, Wahlsysteme, Regierungen, interessiert. Normalerweise sind viele Charakteristiken eines sozialen Systems bekannt, messbar oder zumindest (ab)schätzbar. Ziel von Sozialwissenschaftlern ist es, bestimmte Charakteristika sozialer Systeme durch andere Charakteristika zu erklären; deshalb werden häufig jene als abhängige (Vektor Y) und diese als unabhängige Variablen (Matrix X) bezeichnet. Die abhängige Variable kann etwa die Höhe der Unternehmensbesteuerung in OECD-Staaten sein, während die unabhängigen Variablen die Stärke linker Parteien, Integration in internationale Märkte u.ä. sein können (siehe beispielsweise Garrett und Mitchell 2001).

Unter einem *statistischen Modell* versteht man eine formale Repräsentation des Prozesses, mit dem ein soziales System die untersuchten Daten generiert. Dabei ist es hilfreich, zwei Komponenten zu unterscheiden, aus denen statistische Modelle bestehen: eine *systematische* sowie eine *stochastische* Komponente.[2] Die stochastische Komponente ist ein wichtiger Teil des statistischen Modells. Sie legt explizit Annahmen offen, wie die Verteilung der abhängigen Variablen *Y* theoretisch modelliert wird (King 1989).

Die Diskussion kann durch die Einführung formaler Notation klarer gestaltet werden. Deshalb verwenden wir

$$Y \sim f(y \mid \theta), \tag{1.1}$$

um zu sagen, dass *Y f*-verteilt ist mit Dichteverteilung *f(y)* bei gegebenem Parametervektor θ. Falls die vertraute lineare Regression benutzt werden soll, um eine abhängige Variable zu modellieren, ist *f* bekanntlich die Normalverteilung mit den typischen Parametern μ für den Erwartungswert der abhängigen Variablen und σ^2 für die zu schätzende Varianz (demnach wäre also $\theta = (\mu, \sigma)$).

Die *systematische Komponente* eines statistischen Modells beschreibt den funktionalen Zusammenhang *g*, mit dem bestimmte Elemente von θ (das heißt, die zu schätzenden Parameter der abhängigen Verteilung von *Y*) als Funktion von den gemessenen Charakteristiken eines sozialen Systems, das heißt den unabhängigen Variablen *X* und zu schätzenden Effektparametern β abhängen. Formal,

$$\theta = g(X, \beta). \tag{1.2}$$

Alle so genannten verallgemeinerten linearen Modelle können beispielsweise durch Gleichungen (1.1) und (1.2) formal beschrieben werden. Für den parametrischen Zusammenhang, der durch die systematische Komponente eines statistischen Modells beschrieben ist, wird bei der Wahl einer linearen Regression als statistischen Modells für gewöhnlich der Erwartungswert μ der abhängigen Variablen als lineare Funktion der unabhängigen Variablen modelliert, das heißt also

2 Wir beschränken unser Argument hier auf Modelle, die zwischen abhängigen und unabhängigen Variablen unterscheiden, um die formale Sprache einfach zu halten. Unsere Argumentation ist aber insofern allgemein, als die überwiegende Mehrheit sozialwissenschaftlich genutzter quantitativer Modelle als Regressionsmodelle ausgedrückt werden können: Beispielsweise können ANOVA-Modelle und T-Tests als Regressionen mit binären unabhängigen Variablen modelliert werden (Hays 1988: 675 ff.), Logit-, Probit- und andere nichtlineare Modelle sind Regressionen latenter Variablen oder transformierter abhängiger Variablen („Generalized Linear Models“, Gill 2001a), und Faktoranalysen können als Regressionen verschiedener Indikatorvariablen auf gleiche latente Faktoren verstanden werden (Bollen 1989: 226 ff.).

$$\mu = X\beta = \beta_0 + \sum_{j=1}^{J} \beta_j x_j .$$

(wobei β_0 den konstanten Parameter der linearen Regression bezeichnet, und $j = 1,...,J$ die unabhängigen Variablen nummeriert). Bei einer linearen Regression wird der Parameter σ als Konstante modelliert, so dass $g = (X\beta, \sigma) = (\mu_i, \sigma)$.[3]

Die *stochastische Komponente* folgt aus der Tatsache, dass *Y* nicht eine deterministische Funktion von θ ist, sondern eine Wahrscheinlichkeitsverteilung mit θ als Parameter darstellt. Am einfachsten lässt sich der Zusammenhang zwischen systematischer und stochastischer Komponente im linearen Regressionsmodell darstellen. In der Gleichung

$$Y_i = \mu_i + \varepsilon_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_j + \varepsilon_i$$

stellt dann μ_i den systematischen Teil dar, der auch als „Vorhersage" für *Y* verstanden werden kann. Element ε_i wird dagegen häufig als „Fehler" bezeichnet[4] und ist rein stochastisch, also zufällig; wir nehmen an, dass es normalverteilt ist mit Durchschnitt (das heißt Erwartungswert) *0* und Varianz σ^2. Im Folgenden werden wir dieses lineare Regressionsmodell als Beispiel verwenden, da es allgemein bekannt ist und uns die Unterscheidung zwischen systematischer und stochastischer Komponente besonders deutlich darstellen lässt. Unser Argument besitzt allerdings allgemeine Gültigkeit auch in Bezug auf andere statistische Modelle (siehe Fußnote 2).

Allzu oft wird die stochastische Komponente, zurückgehend auf deterministische Sichtweisen des 18. Jahrhunderts in der Tradition von LaPlace, ignoriert oder als technische Spitzfindigkeit abgetan. Allerdings, wie Pollock schon 1979 trefflich formulierte: „[The] fundamental intellectual breakthrough that has accompanied the development of the modern science of statistical inference is the recognition that the random component has its own tenuous regularities that may be regarded as part of the underlying structure of the phenomenon" (Pollock 1979: 1). In der Tat hat die stochastische Komponente Auswirkungen auf den Grad der Unsicherheit einer jeden Schätzung und damit auf die inhaltliche Interpretation der daraus resultierenden Schlussfolgerungen.

In den folgenden Abschnitten dieses Kapitels wollen wir verschiedene „Fehlerquellen" vorstellen, mit denen typischerweise Daten als operationalisierte theoretische Konstrukte behaftet sind. Stichprobenfehler bilden aber nur einen bestimmten Fehlertyp, mit dem Daten unscharf werden. Vielmehr argumentieren wir, dass Stochastizität notwendigerweise Teil sozialwissenschaftlicher Daten ist. Sie stellt nicht nur eine Verfälschung ansonsten wahrer Ergebnisse dar, sondern ist ein wichtiger Teil unseres Untersuchungsgegenstands. Die verschiedenen Quellen von Stochastizität, die wir im Folgenden diskutieren, beeinflussen prinzipiell (wenn auch in individuell unterschiedlichem Ausmaß) alle Arten sozialwissenschaftlicher Daten – einzige Ausnahme bilden dabei die Stichprobenfehler, die nur in Daten anzufinden sind, die durch Stichprobenziehung ermittelt werden.

3 Dies bedeutet nicht, dass in erweiterten Modellen σ nicht auch als Funktion unabhängiger Variablen modelliert werden kann, siehe Alvarez und Brehm (1995), Franklin (1991), Gschwend (2001) oder Harvey (1976).

4 Wie wir sehen werden, ist eine der möglichen Ursachen für diese Art von „Fehler" auch tatsächlich in fehlerhaften Messungen zu suchen; bei Datensätzen, die auf Stichproben basieren, besteht eine der Fehlerquellen auch in der Tatsache, dass Stichprobe und Universum der Untersuchungseinheiten nicht übereinstimmen.

2.2 Stichprobenfehler

Das Ziehen einer Zufallsstichprobe aus einer Grundgesamtheit ist mit Unsicherheiten behaftet, da *a priori* nicht festgelegt ist, welche Elemente der Grundgesamtheit ausgewählt werden. Die Folge ist, dass diese zufälligen Fehler oder Unsicherheiten sowohl Reichweite wie auch Präzision von sozialwissenschaftlichen Ergebnissen beeinträchtigen, sofern von Ergebnissen der Stichproben auf die Grundgesamtheit geschlossen wird. Nehmen wir einmal an, das bekannte Meinungsforschungsinstitut ALLESWIS zieht eine einfache Zufallsauswahl der bundesdeutschen Wahlbevölkerung, in der alle Wahlberechtigten die selbe Chance haben befragt zu werden, und erreicht eine perfekte Ausschöpfung, so dass die Stichprobe tatsächlich repräsentativ ist. Wir lernen, dass von 1.000 Befragten der Stichprobe 570 angeben, mit der Arbeit der amtierenden Bundesregierung zufrieden zu sein. Unter solchen Umständen verallgemeinert das Institut das Ergebnis auf die Grundgesamtheit und verkündet stolz, dass 57% der Deutschen (Wahlberechtigten) mit der Arbeit der Regierung zufrieden sind. Nun führt das Institut NIXWIS am selben Tage eine ebensolche Umfrage durch, errechnet aber nur 55%. Nehmen wir an, dass alle anderen Institute ebenso sauber arbeiten und auch Werte für diese konkrete Frage erhalten. Die verschiedenen Ergebnisse über die Zufriedenheit mit der Arbeit der amtierenden Regierung beschreiben eine Verteilung. Der Mittelwert dieser Verteilung, sofern nur genügend (das heißt unendlich) viele einfache Zufallsstichproben unabhängig voneinander gezogen werden, beschreibt den tatsächlichen Zufriedenheitswert mit der amtierenden Bundesregierung in der bundesdeutschen Wahlbevölkerung. Die Streuung dieser Verteilung um den Mittelwert wird durch die Varianz beschrieben. Je kleiner die Streuung um den Mittelwert, desto sicherer können wir sein, dass der Mittelwert auch dem wahren Wert in der Bevölkerung entspricht.[5]

Wenn man aber, wie so oft, höchstens *eine* repräsentative Umfrage der gewünschten Grundgesamtheit hat, nimmt man eben den Mittelwert der Stichprobe als beste Schätzung für den Mittelwert der Grundgesamtheit. Diese Schätzung ist aber mindestens mit zufälligem Stichprobenfehler behaftet, weil wir selbst bei perfekt durchgeführter Stichprobe nicht sicher sein können, dass der Mittelwert der Stichprobe auch dem der Grundgesamtheit entspricht.[6] Der Grad dieser Unsicherheit wird durch die Länge des Konfidenzintervalls um den Mittelwert der Stichprobe, also letztlich durch die Varianz der Stichprobe, ausgedrückt.

Bei Vollerhebungen wird oft angenommen, dass Stichprobenfehler keine Rolle spielen, da keine *Stich*probe gezogen wird, sondern die Grundgesamtheit vollständig erfasst wird. Bei Makrodaten – etwa gesamtwirtschaftlichen Daten wie dem Bruttosozialprodukt, aggregierten Länderdaten aus den Eurobarometern, dem World Value-Survey oder dem Comparative Studies of Electoral Systems (CSES)-Projekt – sind Stichprobenfehler dennoch von Bedeutung, da die einzelstaatlichen Werte, die in den Datensatz eingehen, häufig selbst auf Stichproben beruhen. Während die Verteilung von Stichprobenfehlern im Kontext eines auf einer Zufallsstichprobe basierenden Datensatzes relativ verlässlich geschätzt werden kann (mehr davon unten), ist dies problematisch, wenn einzelne Datenpunkte des Datensatzes auf unterschiedlichen Stichproben beruhen.

---

5 Wir beziehen uns hier nicht auf Abweichungen, die durch das *Nonresponse*-Problem in der Umfragepraxis auftreten; siehe hierzu etwa Schnell (1997). Nonresponse-Fehler führen häufig zu Verfälschungen der Ergebnisse; uns geht es im Gegensatz dazu um solche Fehler, die als nicht verfälschende Ungenauigkeiten der Messung dargestellt werden können.

6 In der Tat, wir können sogar sicher sein, dass das nicht so ist, falls es sich um eine kontinuierliche Variable handelt. Die Wahrscheinlichkeit, dass beide Werte übereinstimmen, ist null.

2.3 Messfehler

Messfehler sind eines der wichtigsten methodischen Probleme sozialwissenschaftlicher Studien. Die Frage ist nicht, *ob* ein Messfehler auftritt, wenn ein theoretisches Konzept gemessen wird: Kein Maß wird jemals absolut fehlerfrei das dahinter stehende theoretische Konzept messen können. Die Frage ist demnach nur, *wie groß* der Fehler ist und welche inhaltlichen Konsequenzen er mit sich bringt. Die Konsequenzen von Messfehlern sowohl für die Eindeutigkeit als auch die Tragweite der Ergebnisse sozialwissenschaftlicher Studien sind von entscheidender Bedeutung (Berry und Feldman 1985).

Im Allgemeinen unterscheidet man zwischen zufälligen und nichtzufälligen Messfehlern. Nichtzufällige Messfehler treten auf, wenn ein Indikator eines theoretischen Konzepts *systematisch* etwas anderes misst. Dabei handelt es sich also um einen Fehler des erhobenen Indikators. Die Konsequenz ist jedoch, dass Schlussfolgerungen, die auf solchen Indikatoren beruhen, verzerrt und daher substanziell schwer interpretierbar sind. Grundsätzlich ist das eine Frage der *Validität* des besagten Indikators, was wir hier aber nicht weiter verfolgen wollen. Für unsere Fragestellung direkt relevant sind aber die nicht vorhersehbaren und daher zufälligen Messfehler, die die *Reliabilität*[7] unserer Messungen, das heißt ihre Genauigkeit beeinflussen.

Zufällige Messfehler sind demnach Ungenauigkeiten, die sich beim Messen eines Konzeptes in *unsystematischer* Weise einschleichen und sich somit in der stochastischen Komponente widerspiegeln. Also angenommen wir haben einen validen Indikator eines Konzepts. Werte (Y) für dieses Konzept werden manchmal zu groß, manchmal zu klein gemessen, aber im Durchschnitt sind sie korrekt, das heißt, der Erwartungswert von Y ist μ, da zufällige Messfehler sich asymptotisch im Mittel gegeneinander aufheben (das heißt, der Erwartungswert von ε ist *0*).

Es gibt verschiedene Gründe, warum Variablen als Zufallsvariable definiert werden, das heißt, weshalb sie selbst bei identischem wahrem Wert von Messung zu Messung zufällig variieren können. In Umfragen können Befragte beispielsweise Antworten raten, statt sie wirklich zu wissen. Vor allem faktische Fragen, die sich auf das Erinnerungsvermögen der Befragten verlassen, sind notorisch fehlerbehaftet (Burton und Blair 1991). Zudem können Antwortkategorien für die Befragten schwer verständlich oder uneindeutig sein (Achen 1975; Converse und Presser 1986). Daneben gibt es fast unzählige Fehlerquellen beim Aufnehmen (zum Beispiel Interviewereffekte), Eingeben und Übertragen von Daten. Nicht erst seit den amerikanischen Präsidentschaftswahlen im Jahre 2000 gibt es genügend Gründe zu vermuten, dass selbst amtliche Datenhandbücher nicht davor gefeit sind. Im Grunde schleichen sich zwangsläufig Fehler ein, wenn wir etwas messen wollen, vor allem wenn ein Konzept nicht direkt beobachtbar ist. Die schon erwähnte Tatsache, dass viele Makrodaten auf der Aggregation von Stichprobendaten beruhen, kann auch als eine Form von Messfehler verstanden werden.

Es ist schlichtweg eine Illusion, dass wir ein Konzept perfekt messen können. Sozialwissenschaftliche Schlussfolgerungen sind daher inhärent unsicher. Aufgabe wissenschaftlicher Untersuchungen ist es aber gerade, diese Unsicherheiten abzuschätzen (King, Keohane und Verba 1994: 7–9; Schnell, Hill und Esser 1999: 6). Deshalb ist es unbedingt notwendig, dass Analysen quantitativer Daten nicht nur „substanzielle“ Effekte und Zusammenhänge verschiedener Variablen benen-

7 Als Reliabilität eines Indikators wird der Anteil der Varianz des Indikators bezeichnet, der der „wahren“ Varianz entspricht.

nen, sondern auch die Unsicherheit, das heißt die Varianz dieser Zusammenhänge. Dies gilt natürlich auch für Daten, die auf Vollerhebungen basieren.

2.4 Stochastische Wirklichkeit

*Mess*fehler beeinflussen lediglich unsere Wahrnehmung der Wirklichkeit, und damit die Schlussfolgerungen, die wir daraus ziehen; sie können deshalb als Folge der Unvollkommenheit menschlicher Wahrnehmung gesehen werden. Wenn dies die einzige Fehlerquelle wäre, müssten wir die Wirklichkeit als eine fixe Größe ansehen, zu der der Akt des Messens Varianz hinzufügt. Unseres Erachtens ist dies allerdings eine allzu simple Sichtweise sozialer Phänomene. In der Tat erscheint es uns überzeugender anzunehmen, dass die soziale Wirklichkeit selbst stochastisch ist, die Varianz in unseren Daten also zum Teil in der Natur unseres Untersuchungsgegenstands begründet ist und deshalb Teil unseres substanziellen Untersuchungsinteresses sein sollte.[8]

Was bedeutet es, wenn wir behaupten, dass die soziale Wirklichkeit stochastisch ist? Beginnen wir mit zwei Beispielen:

Beispiel 1: Nehmen wir einmal an, wir hören ein Musikstück, einen klassischen Sonatensatz etwa, den ersten Satz von Beethovens Klaviersonate in D-Dur, op. 28. Wir hören, dass Takte 2 bis 162 Ton für Ton wiederholt werden.[9] Bei genauem Hinhören stellen wir allerdings fest, dass die Wiederholung nicht genau der ersten Vorführung gleicht. Wenn wir die musikalischen Parameter genau messen, werden wir unzählige Unterschiede feststellen. Dennoch würden wir sagen, dass die Musik dieser 160 Takte zweimal gespielt worden ist. Irgendwie ist es uns also möglich, zwischen dem tatsächlichen genauen Ereignis und einer zugrunde liegenden Struktur zu unterscheiden.

Beispiel 2: Wir befragen eine Person nach ihrer Einschätzung verschiedener Persönlichkeiten des politischen Lebens. Wir verwenden eine Skala von 0 bis 10, wobei „0“ für „sehr schlecht“ und „10“ für „sehr gut“ steht. Die befragte Person gibt dem Bundeskanzler die Wertung „7“. Nehmen wir einmal an, wir wiederholten die Befragung einen Monat später, und die befragte Person gäbe nun dem Kanzler die Wertung „6“. Hat nun die befragte Person ihre Meinung geändert? Es ist durchaus möglich, dass der oder die Befragte neue Informationen über den Kanzler erhalten hat, die die Einschätzung änderte. Es ist aber auch denkbar, dass die Antwort der Person zufälligen Fluktuationen unterworfen ist, zumal die substanzielle Bedeutung der Wertung „7“ im Vergleich zur Wertung „6“ nicht unbedingt klar ist.

Fluktuationen des Antwortverhaltens können durchaus als Messfehler eingeordnet werden (Achen 1975), zumal bekannt ist, dass sie durch Kontextfaktoren, etwa die Art der Fragestellung oder die Person des Interviewers, beeinflusst werden können. Die Frage wäre dann aber, was denn die „wahre“ Meinung der oder des Befragten sei, die durch den Messfehler verfälscht wird. Diese Frage ist nicht zu beantworten: Wenn wir annehmen, dass die befragte Person in beiden Instanzen die gleiche zugrunde liegende Meinung hat, dann können wir nicht entscheiden, welche Antwort denn nun wahr ist und welche verfälscht.[10]

---

8 Darüber lässt sich natürlich trefflich philosophisch streiten. Jedoch bleibt die Konsequenz beider Sichtweisen dieselbe: Daten sind inhärent unscharf. Damit beschäftigt sich auch Abschnitt 2.6.

9 Wenn der Pianist nicht die Wiederholung der Exposition weg lässt. Über die Frage, ob solche Wiederholungen im Zeitalter der Reproduktion von Musik auf elektronischen Tonträgern noch zeitgemäß sind, tobt ein heftiger ideologischer Kampf in der Musikwelt.

10 Nur in Panel-Befragungen mit vielen Wellen ist es annähernd möglich, solche Fragen zu beantworten.

Im Gegensatz zu dieser Sichtweise argumentieren wir, dass Meinungen nie außerhalb kontextueller Faktoren existieren. Meinungen sind situationsabhängig. Daher gibt es keinen wahren Meinungs-Wert zu entdecken; wir müssen uns mit einem stochastisch fluktuierenden Wert begnügen. Ähnlich ist es mit der Interpretation eines Musikstücks: Es mag bis zu einem bestimmten Grade möglich sein, eine Vorführung als verfälschend zu bezeichnen (etwa wenn explizit notierte Elemente außer Acht gelassen werden). Aber es ist nicht möglich, eine wahre Interpretation des Musikstücks zu benennen.

Wir behaupten, dass nicht nur Meinungen, sondern soziale Phänomene im Allgemeinen stochastisch sind, da menschliches Handeln nicht deterministisch ist. Die Nichtdeterminiertheit menschlichen Handelns kann auf zweierlei Weise verstanden werden. Zum einen können wir menschliches Handeln als letztlich nicht vollständig erklärbar begreifen. Zum anderen können wir argumentieren, dass unser Handeln durch solch eine Vielzahl von Faktoren beeinflusst wird, dass eine komplette Erklärung nicht möglich ist. Mit der zweiten Sichtweise befasst sich Abschnitt 2.5.

Wenn menschliches Handeln also stochastisch ist, dann ist es unmöglich, selbst mit perfekten Messinstrumenten die stochastische Fluktuation in unseren Daten zu eliminieren. Deshalb können wir diese Art der Stochastizität nicht fehlerhafter Datenermittlung unterschieben; die Wirklichkeit selbst ist stochastisch. Bedeutet dies, dass die Untersuchung sozialer Phänomene unmöglich ist? Wir behaupten, dass dies die falsche Schlussfolgerung ist, solange stochastische Phänomene identifizierbare Tendenzen, also systematische Komponenten aufweisen. Vergleiche zwischen sozialen Phänomenen sind immer noch möglich, die Frage ist nur, ob beobachtbare Unterschiede dem systematischen Teil unseres Gegenstands ($X\beta$) zuzuschreiben sind, oder ob sie durch den stochastischen Teil (ε) erklärt werden können. Wir können uns mit dieser Frage, die die zentrale substanzielle Frage unserer Datenanalyse ist, nur beschäftigen, wenn wir die Varianz unserer Daten mit in die Analyse einbeziehen. Daher gilt das auch für Daten aus Vollerhebungen.

2.5 Deterministische Fehler: Viele kleine ignorierte Faktoren

Die bisher diskutierten Quellen von Stochastizität basierten entweder auf der Unfähigkeit, alle Einheiten einer Grundgesamtheit zu beobachten, auf der Fehlerhaftigkeit von Messungen oder auf der Stochastizität des Beobachtungsgegenstandes sozialwissenschaftlicher Forschung. Eine weitere Quelle der Stochastizität liegt in der Natur empirisch testbarer Erklärungen: Deren Komplexität muss geringer sein als die Komplexität des Untersuchungsgegenstandes. Aus erkenntnistheoretischer Perspektive ist dies notwendig, da sonst Erklärung und Beschreibung identisch wären, womit der sozialwissenschaftliche Anspruch auf nomothetische Aussagen nicht erfüllt werden könnte. Einige interpretative Ansätze in den Sozial- und Geschichtswissenschaften besitzen dagegen nicht unbedingt Anspruch auf Allgemeingültigkeit und können bisweilen die Komplexitätsbeschränkung der Theorie umgehen. Allerdings entziehen sich solche Erklärungen der quantitativen Analyse, da die Zahl der Erklärungsfaktoren die Zahl der Beobachtungen übersteigt; damit sind mehrere Theorien mit den Daten vereinbar.

Was bedeutet es konkret, wenn wir sagen, dass die Komplexität unserer Erklärungen geringer ist als die Komplexität unseres Untersuchungsgegenstands? Ein Beispiel: Nehmen wir einmal an, dass wir eine Theorie des Wahlverhaltens testen wollen, die die Wahrscheinlichkeit der Wahlbeteiligung von folgenden Faktoren abhängig macht: sozialer Status, Häufigkeit des Kirchgangs,

Schulabschluss und Einkommen. Nun gibt es natürlich weitere Faktoren, die im Einzelfall das Wahlverhalten beeinflussen, etwa Krankheit: Eine plötzliche Erkrankung kann jemanden, der ansonsten vorhatte zu wählen, davon abhalten. Im Prinzip können wir zumindest einzelne solcher Faktoren in unserer Theorie mit berücksichtigen. In der Tat fordert Mayntz (2002: 35 ff.), solche *Interferenzen koinzidentieller Effekte* in die Analyse einzelner sozialer Makrophänomene mit einzubeziehen. Dem ist auch grundsätzlich nicht zu widersprechen. Dennoch wird es immer koinzidentielle Effekte geben, die nicht explizit in die Analyse einbezogen werden können, um eben die Komplexität der Erklärung unter der der Wirklichkeit zu halten; dies gilt vor allem für quantitative Studien. So entschließen wir uns wahrscheinlich im genannten Beispiel, Krankheit und ähnliche Faktoren auszuschließen, obwohl diese real und im Einzelfall das Wahlverhalten durchaus beeinflussen mögen.[11]

In der Notation der linearen Regression können wir sagen, dass unsere Theorie durch den systematischen Teil $X\beta$ repräsentiert wird, während die ausgeschlossenen Faktoren sich im ε-Term sammeln. Auch wenn ε durchaus individuell bedeutsame Faktoren beinhaltet, stellt er aus unserer theoretischen Perspektive einen Fehlerterm dar: Selbst wenn unsere Theorie wahr ist, werden die realen Daten von den Werten, die unsere Theorie vorhersagt, abweichen, das heißt „fehlerhaft" sein. Wir empfehlen, das Paradoxon dieses Vorgehens, das wir nicht vermeiden können, auszukosten: Der „stochastische" Term ist aus dieser Sichtweise nicht wirklich stochastisch, sondern besteht aus einer Vielzahl von Faktoren, die durchaus erklärungskräftig sind, uns aber – zumindest im Moment – nicht interessieren.

Die Tatsache, dass die Faktoren, die den ε-Term ausmachen, nicht Teil der systematischen Komponente sind, entlässt uns nicht aus der Notwendigkeit, den ε-Term selbst in unsere Theorie einzubeziehen! Am deutlichsten wird dies in Zeitreihenanalysen, in denen ungemessene Faktoren, die aufeinander folgenden Beobachtungen gemein sind, zu serieller Korrelation der Beobachtungen führen, die dann in statistischen Modellen berücksichtigt werden muss. Aber auch in Abwesenheit serieller Korrelation ist eine Theorie des Fehlerterms notwendig. In einfachen linearen Regressionsmodellen ist etwa die Annahme möglich, dass der Fehlerterm aus der Summe einer Vielzahl von Faktoren besteht, deren Erwartungswert jeweils Null ist (dies wird durch die Auswahl der unabhängigen Variablen erreicht) und die nicht mit den unabhängigen Variablen korrelieren. Ein Fehler, der solchermaßen aus vielen „kleinen" unabhängigen Abweichungen vom Erwartungswert besteht, ist die lineare Funktion eines Durchschnitts. Wenn wir etwa *k* unterschiedliche Fehlerfaktoren haben, die wir mit ε_j bezeichnen, dann ist $\varepsilon_i = \sum_{j=1}^{k} \varepsilon_{ij} = k\bar{\varepsilon}_i$, wobei $\bar{\varepsilon}_i$ der Durchschnitt aller *k*-Fehlerfaktoren für Beobachtung *i* ist. Wenn *k* groß genug ist, und wenn die *k*-Fehlerfaktoren unabhängig voneinander sind, ist $\bar{\varepsilon}_i$ annäherungsweise normal verteilt.

Die Tatsache, dass unsere Theorien gegenüber der Wirklichkeit unzureichend sein müssen, zwingt uns dazu, unsere Schätzungen als unvollkommen anzusehen. Diese Unvollkommenheit schlägt sich in stochastischen Ergebnissen nieder, die Varianzen und Standardabweichungen unserer Schätzwerte messen die Unvollkommenheit unserer Theorie. Deshalb müssen sie integraler Bestandteil empirischer Tests unserer Theorien sein, selbst wenn wir Daten aus Vollerhebungen benutzen.

11 Natürlich steht es uns frei, Krankheit in unsere Theorie mit einzubeziehen; geringes Einkommen und geringe Bildung kann etwa das Aufkommen von Krankheit erhöhen und dadurch zusätzlich das Wahlverhalten beeinflussen.

2.6 Stochastische Wirklichkeit oder nicht beachtete Faktoren?

Auf den ersten Blick erscheint es notwendig, sich zu entscheiden, ob sozialwissenschaftliche Daten stochastische Wirklichkeit reflektieren, oder ob sie aufgrund nicht gemessener Faktoren nur stochastisch erscheinen, oder ob beides der Fall ist. Philosophisch Interessierte mögen dieser spannenden Debatte gerne nachgehen (siehe dazu etwa die Debatte über objektive und subjektive Wahrscheinlichkeit; Mises 1951, de Finetti 1981, Hennig 2001).

Hier wäre aber darauf hinzuweisen, dass unabhängig davon, ob wir der Annahme stochastischer Wirklichkeit oder stochastischer Wahrnehmung folgen, wir diese wie auch immer konzeptionalisierte Stochastizität in unserer Analyse berücksichtigen müssen. Beide Perspektiven entsprechen Kings Unterscheidung zwischen der Modellierung stochastischer abhängiger Variablen einerseits und der Aufteilung abhängiger Variablen in systematische und stochastische Komponenten („Fehler“) andererseits (King 1989: 8). Beide Modellversionen sind methodologisch äquivalent!

3 Was ist das Problem bei Vollerhebungen?

Vollerhebungen sind Stichproben, bei denen die Stichprobengesamtheit der Grundgesamtheit entspricht. Alle Elemente der Grundgesamtheit sind in der Vollerhebung enthalten. Bei (einfachen) Zufallsstichproben werden die Elemente der Grundgesamtheit hingegen zufällig ausgewählt; deshalb sind nicht alle Elemente der Grundgesamtheit in der Stichprobe enthalten. Das Prinzip der einfachen Zufallsauswahl ist daher besonders effektiv, um Schätzwerte für große Populationen zu erhalten. Präzise Schätzungen hängen von der Größe der Stichprobenauswahl ab, nicht von der Größe der Grundgesamtheit. Eine repräsentative Zufallsstichprobe aller Einwohner Kölns wird daher im Wesentlichen so viele Befragte umfassen wie eine repräsentative Zufallsstichprobe aller Einwohner Deutschlands. Das Prinzip der einfachen Zufallsauswahl ist weniger effektiv in kleinen Grundgesamtheiten (Kalton 1983: 14). In kleinen Populationen ist der Mehraufwand, eine Vollerhebung durchzuführen statt eine Stichprobe zu ziehen, nur noch marginal.

Werden nun Datensätze analysiert, die auf Vollerhebungen statt auf Zufallsstichproben basieren, ändern sich mindestens zwei entscheidende Dinge: die Struktur des Fehlerterms und die Interpretation des Wahrscheinlichkeitskonzepts, das beispielsweise Signifikanztests zugrunde liegt.

Die erste entscheidende Änderung betrifft den Fehlerterm, das heißt die stochastische Komponente. Bei (einfachen) Zufallsstichproben werden die Beobachtungen der Stichprobe aus der Grundgesamtheit zufällig ausgewählt, das heißt, die Auswahl eines Elements der Grundgesamtheit ist unabhängig von der Auswahl eines anderen Elements. Daher sind nicht nur die Beobachtungen sondern auch die Fehler der verschiedenen Beobachtungen unabhängig voneinander, was wiederum bedeutet, dass der zentrale Grenzwertsatz angewandt werden kann. Kleinste-Quadrate-Schätzwerte, wie im gängigen linearen Regressionsmodell, lassen sich daher problemlos berechnen, und ihre Signifikanz kann ermittelt werden.

Bei Vollerhebungen jedoch sind die Fehler der Beobachtungen demgegenüber *nicht* notwendigerweise unabhängig voneinander; serielle oder räumliche Korrelationen der Fehlerterme sind daher nicht auszuschließen. Die Theorie der Zeitreihenanalyse versucht statistisch, sich des Problems seriell korrelierter Fehlerterme anzunehmen, das immer dann auftritt, wenn der Fehlerterm späterer Beobachtungen eine Funktion der Fehlerterme früherer Beobachtungen ist. Räumliche

Korrelationen von Fehlern treten zwangsläufig auf, wenn (räumlich) nah gelegene Beobachtungen sich ähnlicher sind als weiter entfernte Beobachtungen. So wird beispielsweise in einer Vollerhebung aller OECD-Länder das Wirtschaftswachstum Deutschlands im Jahre 2002 besser erklärt durch das deutsche Wirtschaftswachstum des Jahres 2001 als durch den entsprechenden Wert Portugals. Räumliche Korrelationen sind in Vollerhebungen beispielsweise aller deutschen Bundestagswahlkreise nahe liegend. Benachbarte Wahlkreise sind sich wahrscheinlich ähnlicher als zwei beliebige Wahlkreise im Westen und im Osten Deutschlands.

Wenn Fehler nicht voneinander unabhängig sondern seriell oder räumlich korreliert sind, dann hat das für Modellschätzungen bedeutende *inhaltliche* Konsequenzen. Die geschätzten kausalen Effekte erscheinen sicherer, als sie eigentlich sind (Berry und Feldman 1985). Intuitiv bedeuten korrelierte Fehlerterme, dass die Komplexität der Wirklichkeit nicht so einfach auf stochastische Prozesse reduziert werden kann. Diese Komplexität kann aber in statistische Modelle integriert werden. Unter dem Stichwort „Maximum Likelihood" (ML) firmiert ein Lösungsansatz, der die explizite Modellierung korrelierter Fehlerterme ermöglicht. Allerdings stehen der erweiterten Flexibilität erhöhte Kosten gegenüber. ML-Schätzungen sind zwar konsistent, aber in kleinen Datensätzen führen sie zu Verzerrungen der Parameter gegenüber ihren Erwartungswerten;[12] Vollerhebungen werden hingegen häufig nur gemacht, wenn die Grundgesamtheit wenige Fälle umfasst.

Die zweite Änderung, die sich bei Vollerhebungen ergibt, folgt aus theoretischen und konzeptionellen Schwierigkeiten der Interpretation von Wahrscheinlichkeiten. Bei (einfachen) Zufallsstichproben wird häufig auf das *frequentistische* Wahrscheinlichkeitsmodell Bezug genommen: Demnach sind Wahrscheinlichkeiten relative Häufigkeiten von Ereignissen bei (annäherungsweise) unendlich häufiger Wiederholung der Messung. Da bei Zufallsstichproben annäherungsweise unendlich viele unabhängige Wiederholungen der Stichprobenziehung möglich – wenn auch unpraktisch – sind, erscheint dieses Wahrscheinlichkeitsmodell als plausibel.

Bei Vollerhebungen hingegen gibt es nur die eine Stichprobe – Wiederholung ausgeschlossen. Das entspricht dem sattsam bekannten Münzwurfexperiment als Daten generierendem Prozess, nur dass die Münze nach einmaligem Werfen zwar noch auf „Kopf" (oder „Zahl") liegen bleibt – was wir pflichtbewusst notieren – sie dann aber aus unserer Hand rutscht und auf nimmer Wiedersehen verloren ist. Der Daten generierende Prozess ist somit nicht replizierbar. Damit kann aber streng genommen das frequentistische Wahrscheinlichkeitsmodell nicht als theoretischer Rahmen zur Konzeptualisierung der Ergebnisse herangezogen werden (Berk, Western und Weiss 1995). Aussagen über die Signifikanz von Schätzergebnissen, die auf Aussagen über ihre Wahrscheinlichkeit unter verschiedenen hypothetischen Zuständen beruhen, erscheinen somit als bedeutungslos.

12 ML-Schätzungen sind nur asymptotisch ohne Verzerrung. Als Faustregel gilt, dass man mit über hundert Fällen konsistente Ergebnisse erhält.

4 Was sind die Lösungen?

4.1 Eyeballing: Substanzielle Effekte und subjektive Wahrheit

Eine Reaktion auf die Probleme, die sich bei der Interpretation von Vollerhebungsdaten auftun, sind natürlich die oben erwähnten *Determinismus-* und die *SPSS-Strategien*. Hier befassen wir uns kurz mit Varianten der *Determinismus-Strategie*, die unseres Erachtens keine überzeugende Lösung anbietet.

Die reine *Determinismus-Strategie*, bei der die Größe und Vorzeichen sozusagen als wahre Werte genommen werden, ist vergleichsweise selten; dennoch lohnt es sich, die Probleme einer solchen Strategie zu analysieren unter der Erkenntnis, dass zusätzlich zu Stichprobenfehlern weitere Quellen der Stochastizität existieren. Als Beispiel nehmen wir die Analyse von Parameterwerten einer linearen Regression, die nach dem OLS-Verfahren geschätzt worden sind, gemäß dem Modell:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \quad .^{13}$$

Die einfache Version der *Determinismus-Strategie* nimmt an, dass die Werte der β-Parameter die wirklichen Zusammenhänge zwischen Variablen wiedergeben; für Hypothesentests wäre es demnach ausreichend, die Direktionalität dieser Parameter zu betrachten. Wenn eine Hypothese beispielsweise einen positiven Zusammenhang zwischen abhängiger Variable und unabhängiger Variable x_3 behauptet, dann wird diese Hypothese widerlegt, wenn β_3 negativ ist.

Diese Strategie des Hypothesentests ist problematisch. Erstens, Hypothesentests, die lediglich auf der Direktionalität von β_3 beruhen, sind gegenüber Hypothesentests, die Varianzen in Betracht ziehen (selbst wenn diese falsch geschätzt sind!), häufig zu unkritisch. Wir können dies am Fehler ersten Typs festmachen. Üblicherweise wird bei einem Signifikanztest eine so genannte Null-Hypothese angenommen. Wenn unter Annahme der Null-Hypothese die Wahrscheinlichkeit gering ist, den beobachteten Schätzparameter zu erhalten, dann wird die Null-Hypothese verworfen. Ein Fehler ersten Typs besteht nun in der Wahrscheinlichkeit, die Null-Hypothese zu verwerfen, obwohl sie richtig ist. Wenn unsere substanzielle Hypothese nun beispielsweise ist, dass β_3 positiv ist, implizieren wir die Null-Hypothese $\beta^0{}_3 \leq 0$. Wenn die real existierende stochastische Komponente symmetrisch um den Schätzparameter verteilt ist, und wir nehmen an, dass die Null-Hypothese durch eine β_3-Schätzung, die größer als Null ist, widerlegt wird, dann ist die Wahrscheinlichkeit eines Fehlers ersten Typs bis zu 0,5. In anderen Worten, wir haben eine bis zu fünfzigprozentige Wahrscheinlichkeit, auf einen positiven Zusammenhang zwischen Y und x_3 zu schließen, wenn in Wirklichkeit der Zusammenhang negativ ist. Selbst ein Signifikanztest, der auf verfälschten Fehlerschätzungen beruht, wäre konservativer.

Das zweite Problem der einfachen *Determinismus-Strategie* besteht darin, dass sie in der Praxis einen Test der Hypothese ausschließt, dass kein Zusammenhang zwischen zwei Variablen besteht. Lediglich in extremen und in der Praxis sehr seltenen Fällen ist die Wahrscheinlichkeitsdichte des Parameterwertes $\beta_i=0$ positiv. Da die β-Parameter in der Regel kontinuierlich sind,

13 In der Praxis wird diese Gleichung nie perfekt aufgehen, sofern die Zahl der unabhängigen Variablen geringer ist als die Zahl der Beobachtungen, oder $k\text{-}1<n$, was natürlich die einfache Version der *Determinismus-Strategie* naiv erscheinen lässt.

strebt die Wahrscheinlichkeit eines einzelnen Wertes gegen Null. Dadurch wird die Brauchbarkeit der empirischen Analysen in hohem Maße eingeschränkt.

Eine subtilere Variante der *Determinismus-Strategie* bezieht sich nicht nur auf die Direktionalität, sondern auch auf die Größe des geschätzten Parameterwerts, das heißt des geschätzten marginalen Einflusses der unabhängigen auf die abhängige Variable. Ist der marginale Effekt „substanziell", dann wird von einem Zusammenhang ausgegangen. Dieses Vorgehen ist natürlich insoweit sinnvoll, als die substanzielle Bedeutung der Schätzergebnisse immer Teil der Analyse sein sollte, auch wenn Signifikanztests verwendet werden (King 1986). Für den Test eines positiven Zusammenhangs zwischen zwei Variablen impliziert solch ein „Substanztest" einen kritischen Wert, der größer als Null ist und den der geschätzte Parameterwert überschreiten muss, um als substanziell anerkannt zu werden. Solch ein Substanztest ist im Grunde ein Signifikanztest ohne wahrscheinlichkeitstheoretische Grundlage. Er führt zu einem kleineren Fehler ersten Typs als der reine Direktionalitätstest, aber er macht keine Aussage über die Größe dieses Fehlers.

Der Substanztest hat nur eine eingeschränkte intersubjektive Gültigkeit, da die Festlegung, welcher Parameterwert ein substanziell bedeutendes Ergebnis impliziert, *ad hoc* vom Wissenschaftler festgelegt wird. Ein Versuch, dieses Problem zu umgehen, besteht in der Verwendung standardisierter Regressionskoeffizienten. Kunz (2000) interpretiert beispielsweise alle standardisierten Regressionskoeffizienten, die größer als 0,3 oder kleiner als -0,3 sind, als substanzielle Effekte. Dies führt allerdings zu weiteren Problemen, da der standardisierte Regressionskoeffizient nicht nur den Zusammenhang zwischen abhängigen und unabhängigen Variablen misst, sondern auch die Varianzen dieser Variablen. King (1986: 672) präsentiert ein Beispiel, in dem bei gleich bleibendem unstandardisiertem Regressionskoeffizient der standardisierte Regressionskoeffizient größer ist, wenn die Zahl der Datenpunkte und die Varianz der unabhängigen Variablen ansteigt. Das heißt, dass ein standardisierter Koeffizient, der größer als 0,3 ist, nicht unbedingt auf einen stärkeren Zusammenhang zwischen abhängiger und unabhängiger Variable verweist als ein Koeffizient, der kleiner als 0,3 ist; es ist möglich, dass die Differenz lediglich auf die Varianzen der unterschiedlichen Variablen zurückzuführen ist. Der Substanztest verliert so seine Substanz.

Im Gegensatz zum Substanztest, der allein vom Wert der Parameterschätzung ausgeht, beruht ein Signifikanztest auf einem Vergleich des Parameterschätzwerts und der Varianz der Schätzfehler. Ist die letztere, verglichen mit dem Parameterwert, groß, dann ist es relativ wahrscheinlich, dass die Parameterschätzung auch ohne tatsächlichen Zusammenhang in der Wirklichkeit zustande kommt. Auch wenn „relativ wahrscheinlich" wieder eine arbiträre Setzung notwendig macht, so gibt es mittlerweile sozialwissenschaftliche Konventionen, die eine intersubjektive (das heißt, replizierbare) Festlegung von Signifikanzschwellen ermöglichen. Der nächste Abschnitt schlägt Annahmen vor, die die Verwendung von Signifikanztests auch bei Vollerhebungen ermöglicht.

4.2 Und täglich grüßt das Murmeltier: Superpopulationen und die Rettung der SPSS-Strategie

Eine häufig praktizierte Strategie bei Vollerhebungen ist, die resultierenden Daten mit klassischen statistischen Methoden zu analysieren, und so zu tun, als ob sie das Ergebnis einer Zufallsstichprobe seien; wir haben dieses Vorgehen die *SPSS-Strategie* getauft. Diese ist theoretisch möglich, wenn man glaubwürdig argumentieren kann, dass

a) die Daten durch stochastische Prozesse gekennzeichnet sind,
b) die Beobachtungen unabhängig voneinander sind (bzw. identisch und unabhängig voneinander verteilt sind), oder
c) wenn Bedingung b) nicht zutrifft, aber die Abhängigkeiten im statistischen Modell berücksichtigt sind.

Das Problem mit Kriterium c) ist, dass Vollerhebungen häufig eine geringe Zahl von Beobachtungen aufweisen. Da klassische statistische Methoden auf asymptotischen Eigenschaften von Schätzparametern beruhen, stellt dies ein Problem der Anwendung klassischer Methoden dar. Bei komplexeren stochastischen Prozessen und Schätzmodellen sind klassische Methoden deshalb oft nicht anwendbar, als Alternative bietet sich ein bayesianisches Vorgehen an, von dem weiter unten noch die Rede sein soll. Wenn serielle oder räumliche Fehlerkorrelationen aber keine Rolle spielen oder einfach kontrolliert werden können (zum Beispiel mit *Fixed-Effects*-Modellen, Greene 1990: 466), bietet die folgende Argumentation eine Möglichkeit, die oft einfacheren Methoden der klassischen Statistik anzuwenden.

Wie können wir die Gültigkeit von Annahmen a) und b) und/oder c), das heißt, wie können wir die SPSS-Strategie glaubhaft begründen? Ein Beispiel aus der Populärkultur kann uns helfen: In der amerikanischen Filmkomödie „Groundhog Day“ (deutscher Titel: „Und täglich grüßt das Murmeltier“) durchlebt ein misanthropischer Reporter einen Albtraum: Er wacht immer wieder am selben Tag auf und muss diesen Tag durchleben. Obwohl die Hauptereignisse des Tages immer die gleichen sind, gibt es kleine Veränderungen (die am Ende des Films den Charakter der Hauptfigur – ganz in Hollywoodmanier zum Positiven – verändern).

Dem phantastischen Sujet des Films liegt ein Gedankenspiel zu Grunde, das auch für uns brauchbar sein kann: Stellen wir uns vor, Geschichte würde sich wiederholen, und wir könnten eine Gesamtpopulation mehrfach messen, wobei die Messungen unabhängig voneinander sind (das heißt, abhängig nur von gewissen messbaren Faktoren). Nähmen wir dann eine dieser Messungen und die resultierenden Daten, so hätten wir eine Stichprobe aus einer größeren Grundgesamtheit – einer so genannten *Superpopulation* (Berk, Western und Weiss 1995). Die Stichprobe wäre gekennzeichnet durch systematische Charakteristika des sozialen Phänomens, das wir messen, und durch stochastische Fehler: Messfehler und Zufallsfluktuationen des Phänomens selbst und durch die Tatsache, dass unsere Messungen lediglich Zufallsstichproben aus der Superpopulation darstellten. Da unsere Messungen unabhängig von einander sind, sind unsere Beobachtungen identisch und unabhängig verteilt und wir können frequentistische Annahmen machen: Wenn sich die Geschichte unendlich häufig wiederholte, und wir jede Wiederholung messen würden, würde der Fehler einen Durchschnitt von Null haben, und seine Varianz würde gegen Null streben. Wenn die Beobachtungen alle unabhängig sind, kann der zentrale Grenzwertsatz angewandt werden, nach dem Mittelwerte annähernd normal verteilt sind.

Kritiker dieses Ansatzes argumentieren, dass sich Geschichte nicht wiederholt, und dass deshalb das Superpopulations-Argument nicht überzeugt.[14] Verfechter des Ansatzes geben hingegen zu bedenken, dass reale Populationen nicht viel weniger spekulativ sind als Superpopulationen. Bei der Verwendung von Zufallsstichproben realer Populationen ist es zwar theoretisch möglich (wenn auch praktisch oft sehr schwer), weitere unabhängige Zufallsstichproben zu messen, dies geschieht aber so gut wie nie. Stattdessen reicht das Argument, dass eine alternative Stichprobe zur Verfügung hätte stehen können, über die wir theoretische Aussagen treffen, die dann wiederum die Basis statistischer Inferenz bilden. Dies ähnelt dem Vorgehen bei Superpopulationen: Wir nehmen an, dass unsere Vollerhebung auch zufällig andere Daten hätte produzieren können, und stützen unsere inferenziellen Schätzungen auf diese Annahmen.

Ein zusätzliches Argument für die theoretische Gültigkeit von Superpopulationen weist darauf hin, dass inferenzielle Methoden grundsätzlich auf imaginären Annahmen, nämlich Hypothesen, basieren. So basiert etwa ein Signifikanztest auf einem Gedankenexperiment, bei dem angenommen wird, dass die Null-Hypothese wahr ist; der Test errechnet dann die Wahrscheinlichkeit, dass die beobachteten Daten rein zufällig gemessen werden.

Auch wenn sie nicht unbedingt als solche bezeichnet werden, so sind Superpopulationen gang und gäbe in einem Bereich der statistischen Schätzung, in dem es nur (wiederholte) Vollerhebungen gibt: der Zeitreihenanalyse. Jeder Datenpunkt einer Zeitreihe besteht aus einem Datenpunkt, der die Gesamtpopulation an einem bestimmten Zeitpunkt misst. Im Gegensatz zu anderen Vollerhebungen wird die Population in einer Zeitreihe wiederholt gemessen. Durch ihren Ursprung in einer identischen (oder als identisch angenommenen) Population sind die aufeinander folgenden Werte nicht notwendigerweise unabhängig voneinander, aber sie sind durch stochastische Faktoren beeinflusst. Nicht von ungefähr schreiben McCleary und Hay,

> An observed time series is a realization of some underlying stochastic process. In this sense, the relationship between realization and process in time series analysis is analogous to the relationship between *sample* and *population* in cross-sectional analysis. (1980: 30)

Wir halten die Verwendung von Signifikanztests für die überzeugendere Strategie als die *ad hoc* (oder durch standardisierte Koeffizienten) bestimmte substanzielle Bedeutung von Parameterwerten. Dennoch sind solche Tests nicht ohne Problematik: Sie beruhen auf hypothetischen Annahmen so genannter Null-Hypothesen; ihre Interpretation ist nicht intuitiv und wird häufig verfälscht; Signifikanz impliziert nicht notwendigerweise Substanz – um nur ein paar Probleme zu nennen. Signifikanztests haben wiederholt im Zentrum heftiger methodologischer Auseinandersetzungen gestanden (vgl. beispielsweise die Aufsätze in Morrison und Henkel 1970).

Es gibt alternative Strategien, die einige der Probleme vermeiden, die mit Signifikanztests verbunden sind. Es ginge zu weit, hier im Detail auf diese Strategien einzugehen; dennoch soll auf diese knapp hingewiesen werden. Im Zentrum der alternativen Herangehensweisen an Stochastizität steht das Konzept subjektiver Wahrscheinlichkeiten. Klassische statistische Analysemethoden beruhen auf hypothetischen Parameter-Eigenschaften, die gelten, wenn eine Analyse unendlich oft wiederholt wird. Die subjektive Wahrscheinlichkeitskonzeption dagegen beruht auf der intuitiven Erfassung von Risiken: Wir alle haben ein Gefühl dafür, wie wahrscheinlich bestimmte Ereignisse sind, selbst wenn sie einmalig sind. Wir wissen etwa, dass es sehr unwahrscheinlich ist, sechs

14 Siehe zu diesem Thema die Debatte um Berk et al. 1995 in *Sociological Methodology*.

Richtige im Lotto zu haben, dass es wahrscheinlicher ist, mit dem Auto zu verunglücken als mit dem Flugzeug, und so weiter. Solche subjektiven Wahrscheinlichkeiten können zur Grundlage statistischer Analysen gemacht werden. Dabei ist es dann im Gegensatz zu frequentistischen Wahrscheinlichkeitskonzepten irrelevant, ob Ereignisse oder Stichproben wiederholt werden können; was zählt, ist die Adäquatheit unserer subjektiven Wahrscheinlichkeitsannahmen.

Selbst bei klassischen Methoden können subjektive Wahrscheinlichkeiten durchaus eine Rolle spielen. Maximum-Likelihood-Ansätze etwa gehen von einer hypothetischen Wahrscheinlichkeitsverteilung der Daten aus, als Funktion bestimmter zu schätzender Parameter; die angenommene Verteilung kann subjektiv begründet werden. Maximum-Likelihood-Methoden sind vor allem dann brauchbar, wenn die Daten nicht voneinander unabhängig sind, und wenn sie nichtnormalen Wahrscheinlichkeitsverteilungen folgen. Die Flexibilität der ML-Modelle erlaubt es, solcherlei Annahmen zu modellieren, selbst wenn Kleinste-Quadrate-Methoden versagen.

Die Anwendung von ML-Methoden bei Vollerhebungen ist allerdings häufig problematisch, da die Eigenschaften der resultierenden Schätzparameter – ihre Verteilung und ihre Varianz etwa – nur asymptotisch definiert sind. Das heißt, dass wir nur dann Aussagen über die Signifikanz von Schätzwerten machen können, wenn die Zahl der Beobachtungen gegen unendlich strebt und die so genannten Normalitätsbedingungen erfüllt sind. Da Vollerhebungen häufig relativ kleine Datensätze produzieren, sind asymptotische Aussagen oft nicht nützlich. Außerdem beruht die ML-Methode darauf, die Parameter zu ermitteln, die die Likelihood-Funktion (das heißt die Wahrscheinlichkeit der Daten) maximieren. Bei komplexen Wahrscheinlichkeitsfunktionen führt dies häufig zu praktischen Schwierigkeiten: Oft sind die Wahrscheinlichkeitsfunktionen relativ „flach“, was dazu führt, dass die üblichen numerischen Maximierungs-Algorithmen nicht auf einen bestimmten Parametervektor konvergieren; oft hat eine Wahrscheinlichkeitsfunktion mehrere Maxima, und die Ermittlung eines (oder auch des globalen) Maximums ist nicht sonderlich informativ.

Bayesianische Methoden bieten alternative Strategien der Datenanalyse, die noch deutlicher auf subjektiven Wahrscheinlichkeiten beruhen und nicht die Schwächen des ML-Ansatzes besitzen (Berk, Western und Weiss 1995; Gill 2002; Western 2001; Western und Jackman 1994). Die Grundidee bayesianischer Schätzmethoden ist es, *a priori* bestehenden Annahmen über die zu schätzenden Parameter (etwa Regressions-Koeffizienten) formal in das Schätzmodell mit einzubeziehen. Der Schätzvorgang ermittelt dann, ob und in welcher Weise die existierenden Daten die A-priori-Annahmen revidieren. Da Annahmen immer in der Form von Wahrscheinlichkeitsverteilungen formuliert werden, besteht das Ergebnis bayesianischer Schätzungen in A-posteriori-Verteilungen von Parameterwerten. Die Resultate werden deshalb in der Regel nicht als Punktschätzungen ermittelt, sondern in Form von Konfidenzintervallen (bzw. Intervallen höchster Wahrscheinlichkeit). Da diese Intervalle immer mit Hilfe von Simulationsmethoden ermittelt werden können, sind die Ergebnisse bayesianischer Schätzungen auch bei geringer Datenzahl unverfälscht. Das bedeutet, dass, im Gegensatz zu ML-Methoden, bayesianische Schätzungen auch komplizierter nichtlinearer Modelle möglich sind, die etwa räumliche Korrelationen zwischen Beobachtungen in Betracht ziehen.

Eine Schwäche – so sehen es zumindest viele Kritiker – bayesianischer Ansätze ist die Notwendigkeit, Annahmen über die A-priori-Verteilung der Parameter eines Modells machen zu müssen.[15] Beruht eine Analyse auf einem umfangreichen Datensatz, ist dies nicht allzu bedeutsam, da

15 Wir danken Bernhard Kittel für den Hinweis, dass die Problematik der angenommenen A-priori-Verteilungen besonders deutlich ist wenn A-posteriori-Verteilungen existierender Studien als A-priori-Verteilungen neuer

die Daten die A-priori-Annahme dominieren: Unterschiedliche A-priori-Verteilungen beeinflussen die A-posteriori-Ergebnisse kaum. Bei kleinen Datensätzen können A-priori-Annahmen aber das Ergebnis entscheidend beeinflussen. In einer Sensitivitätsanalyse werden in solchen Fällen die Ergebnisse unter Annahme verschiedener A-priori-Verteilungen verglichen, um den Einfluss unserer Annahmen einzuschätzen. Werden (annähernd) uniforme A-priori-Verteilungen angenommen, sind die Ergebnisse bayesianischer und Maximum-Likelihood-Methoden äquivalent.

Wir sehen in der deutlichen Rolle von A-priori-Annahmen auch eine Stärke bayesianischer Analysen, denn sie zwingt uns dazu, den Einfluss unserer Vorannahmen auf unsere Schlussfolgerungen offenzulegen. Damit können wir abschätzen, welchen Erkenntnisgewinn unsere Daten bieten und wo ihre Grenzen liegen. Wenn wir etwa *a priori* glauben, dass ein bestimmter Parameter positiv ist, können wir dies mit einer entsprechenden A-priori-Verteilung in die Analyse einbringen. Zum Vergleich können wir auch die (skeptische) Vorstellung formalisieren, dass der Parameter negativ ist. Anhand der bayesianischen Datenanalyse können wir dann entscheiden, ob unsere Daten die Annahme eines positiven Werts bestätigen oder widerlegen und ob sie die Meinung des Skeptikers ändern können. Wir haben also die Möglichkeit, formal zu bestimmen, was wir von unseren Daten lernen und was wir nicht lernen. Im Vergleich dazu sind die Vorannahmen von ML-Modellen selten explizit Teil der Analyse; die Gültigkeit der Normalitätsbedingungen etwa werden selten thematisiert, sie sind technische Annahmen über die Daten und von geringer substanzieller Bedeutung.

Es ist nicht Ziel dieses Artikels, eine detaillierte Einführung in die bayesianische Datenanalyse zu bieten (siehe dazu etwa Gill 2002). Wir möchten allerdings darauf hinweisen, dass bayesianische Methoden häufig die theoretisch und methodisch einzig „korrekten“ Lösungen der Analyse von Vollerhebungen darstellen. Da Beobachtungen in Vollerhebungen häufig nicht unabhängig voneinander verteilt sind, sind die Gauss-Markov-Bedingungen, die der Kleinsten-Quadrate-Methode zugrunde liegen, häufig nicht erfüllt. Da Vollerhebungen oft geringe Fallzahlen haben, sind die Varianzen der Schätzparameter in Maximum-Likelihood-Modellen nicht korrekt schätzbar. Bayesianische Methoden basieren zwar auf relativ „starken“ Annahmen, aber diese Annahmen sind substanziell nachvollziehbar und dadurch theoretisierbar; außerdem bieten sie die einzige Möglichkeit, verlässliche Schlussfolgerungen aus unseren Daten zu ziehen.

5 Was tun? Einige Schlussfolgerungen

Ausgangspunkt dieser Erörterung war ein Unbehagen an der Art und Weise, wie üblicherweise mit der Analyse von Vollerhebungen umgegangen wird. Zum einen werden häufig Kleinste-Quadrate-Methoden, die auf der Annahme unabhängiger Datenpunkte beruhen, unreflektiert angewandt; zum anderen wird mit Hinweis auf die Tatsache, dass ein Datensatz Beobachtungen aller Einheiten einer Population einschließt, die Relevanz inferenzieller Analysemethoden negiert und durch Ad-hoc-Diskussionen von Parameterwerten ersetzt. Unser Ziel war es deshalb, Strategien aufzuzeigen, die es uns ermöglichen, verlässliche Schlüsse aus Vollerhebungen zu ziehen. Dabei haben wir drei Argumente vorgetragen:

Studien verwandt werden und es somit einen „lock-in“ geben kann, bei dem Ergebnisse früherer Studien die Ergebnisse folgender Studien determinieren. Wir glauben aber, dass dieses Problem durch den Vergleich verschiedener A-priori-Verteilungen umgangen werden kann.

1. Auch wenn Vollerhebungen keine Stichprobenfehler aufweisen, so werden die resultierenden Daten durch eine Reihe stochastischer Prozesse beeinflusst, die ihre Ursache etwa in Messfehlern haben oder der inhärenten Stochastizität menschlichen Handelns.
2. Dadurch ist die Annahme, dass Daten aus Vollerhebungen keinen stochastischen „Fehler" aufweisen, nicht haltbar. Analysen von Vollerhebungen müssen deshalb stochastische Elemente, das heißt Varianzen und Standardfehler, in Betracht ziehen.
3. Es gibt verschiedene Möglichkeiten, inferenzielle Statistik in Vollerhebungen einzusetzen; es ist dabei aber notwendig, die zugrunde liegenden Annahmen genau zu spezifizieren.

Was ist also zu tun? Der Geist des Determinismus verfolgt uns leider noch immer. Nach den Naturwissenschaften sollten auch die Sozialwissenschaften anerkennen, dass wir in einer „Welt der Propensitäten" (Popper 1990) agieren. Oder glauben wir tatsächlich noch immer, dass wenn wir die Geschichte wiederholen und dabei die erklärenden Variablen unserer Theorien konstant halten könnten, wir absolut sicher sind, identische Phänomene zu beobachten? Wenn sozialwissenschaftliche Daten grundsätzlich stochastisch sind, müssen wir dies in unsere Analysen einbeziehen. Wenn wir nämlich stochastische Elemente – das heißt: Varianzen und Standardfehler – ignorieren, vergeben wir ein wichtiges Element der Kritikfähigkeit unseren Daten gegenüber.

Wenn wir etwa die Hypothese, dass ein Parameterwert positiv ist, immer dann bestätigt sehen, wenn der entsprechende Schätzwert größer als Null ist, akzeptieren wir immer eine höhere Wahrscheinlichkeit eines Fehlers erster Art (also die Bestätigung unserer Hypothese wenn sie tatsächlich falsch ist) als im Falle eines wie auch immer gearteten Signifikanztests. Trotz aller berechtigten Kritik an Fisherschen und Neyman-Pearsonschen Signifikanztests (Gill 2001b, Morrison und Henkel 1970) sind solche immer noch besser als die unkritische Annahme positiver oder negativer Parameterwerte. Auch die Verwendung substanzieller Schwellenwerte für die Parametergrößen hilft nicht wesentlich weiter, weil solche Verfahren zu Ad-hoc-Entscheidungen führen, die nicht auf die Stochastizität der Daten eingehen (unabhängig von den Problem der mit diesem Vorgehen häufig verbundenen Anwendung standardisierter Parameterwerte).

Da bei der Verwendung von Daten, die nicht auf einer Zufallsstichprobe beruhen, Annahmen ausfallen, die häufig zur Bestimmung stochastischer Charakteristika der Daten verwandt werden, ist es notwendig, explizit Annahmen über die Verteilung der Daten in die Analyse einzubeziehen. In diesem Zusammenhang schlagen wir den Rekurs auf subjektive Wahrscheinlichkeiten vor. Dies zwingt uns, unsere theoretischen Vorstellungen über die Form stochastischer Datenanteile offenzulegen und explizit in unsere Analyse einzubeziehen, und damit der Kritik zu unterwerfen; Stochastizität ist dann nicht einfach ein technisches Detail, es ist ein substanzieller Aspekt unserer theoretischen und empirischen Untersuchung.[16]

Von besonderem Nutzen sind dabei bayesianische Ansätze. Aus technischer Perspektive ermöglichen diese uns den Umgang mit relativ kleinen Datensätzen, da sie asymptotische Annahmen durch explizite Annahmen über Verteilungen und die empirische Beschreibung der geschätzten Parameterverteilungen ersetzen. Zudem beziehen bayesianische Methoden die Annahmen über die Form stochastischer Elemente explizit in die Analyse ein. Mittlerweile ist auch die Implementation einfacher bayesianischer Schätzungen (etwa von Modellen mit binären abhängigen Variablen)

16 Wenn wir etwa binäre Daten analysieren, ist es nicht möglich, eine Normalverteilung anzunehmen, und wir sind gezwungen, uns für ein anderes stochastisches Modell zu entscheiden. Dies kann etwa auf Basis eines entscheidungs- oder spieltheoretischen Modells geschehen, das beispielsweise zur Definition eines Probit-Modells oder eines Quantal-Response-Modells (Signorino 1999) führen kann.

durch allgemeinverständliche Lehrbücher (Gill 2002) und relativ einfach einzusetzende Programme wie WinBugs (Spiegelhalter et al. 2003) möglich.

Aber selbst die *SPSS-Strategie* ist häufig begründbar, wenn die Annahmen annäherungsweise unabhängiger Beobachtungen gerechtfertigt werden kann. Dabei ist es nicht wesentlich, ob die stochastischen Prozesse als subjektive Wahrscheinlichkeiten verstanden werden, oder ob auf die Theorie der „apparent Populations" Rekurs genommen wird.

Wir sollten zum Schluss darauf hinweisen, dass es uns nicht darum geht, die blinde Anwendung komplizierter Analysemethoden zu propagieren. Im Gegenteil: Achens Plädoyer für gut durchdachte, theoretisch wohlbegründete und wohldefinierte Hypothesentests mit simplen statistischen Methoden halten wir für überaus sinnvoll (Achen 2000). Solch ein Vorgehen erleichtert es eben auch, die stochastische Natur sozialwissenschaftlicher Daten in die Analyse einzubeziehen, da die zu testenden Zusammenhänge nicht durch zahlreiche intervenierende Variablen und Faktoren verworren werden. Ob solch „simple" Methoden möglich sind (in Wirklichkeit sind sie auf theoretischer Ebene ausgesprochen subtil), hängt allerdings von der Problemstellung, der Theorie und den Daten ab.

Unser Appell ist nicht reduziert auf die Analyse von Vollerhebungen, sondern gilt generell. Er richtet sich an alle Sozialwissenschaftler(innen) als Produzenten von Ergebnissen, Gutachter von Manuskripten und nicht zuletzt Mitarbeiter in den Redaktionen der Fachzeitschriften. Die substanzielle Interpretation statistischer Modelle ist nicht haltbar, wenn nicht auf Fehlervarianzen eingegangen wird. Um es klar zu sagen: Bitte keine Interpretationen von Regressionsmodellen mehr ohne explizite Bezugnahme auf Standardfehler!

6 Literatur

Achen, Christopher H. 1975. Mass Political Attitudes and the Survey Response. *American Political Science Review* 69:1218–1231.

Achen, Christopher H. 2000. Warren Miller and the Future of Political Data Analysis. *Political Analysis* 8 (2):142–146.

Alvarez, R. Michael, und John Brehm. 1995. American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values. *American Journal of Political Science* 39:1055–1082.

Berk, Richard A., Bruce Western, und Robert E. Weiss. 1995. Statistical Inference for Apparent Populations. *Sociological Methodology* 25:421–458.

Berry, William D., und Stanley Feldman. 1985. *Multiple Regression in Practice*. Newbury Park, CA: Sage.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Broscheid, Andreas, und Paul Teske. 2003. Public members on medical licensing boards and the choice of entry barriers. *Public Choice* 114 (3–4):445–459.

Burton, Scot, und Edward Blair. 1991. Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly* 55:50–79.

Converse, Jean M., und Stanley Presser. 1986. *Survey Questions. Handcrafting the Standardized Questionnaire*. Beverly Hills: Sage.

de Finetti, Bruno. 1981. *Wahrscheinlichkeitstheorie*. Wien: Oldenbourg.

Franklin, Charles H. 1991. Eschewing Obfuscation? Campaigns and the Perception of U.S. Senate Incumbents. *American Political Science Review* 85 (4):1193–1214.

Garrett, Geoffrey, und Deborah Mitchell. 2001. Globalization, Government Spending and Taxation in the OECD. *European Journal of Political Research* 39:145–177.

Gill, Jeff. 2001a. *Generalized Linear Models. A Unified Approach*. Edited by M. Lewis-Beck, *Quantitative Applications in the Social Sciences*. Thousand Oaks: Sage.

Gill, Jeff. 2001b. Whose Variance is it Anyway? Interpreting Empirical Models with State-Level Data. *State Politics and Policy Quarterly* 1 (3):318–338.

Gill, Jeff. 2002. *Bayesian Methods for the Social and Behavioral Sciences*. Boca Raton: Chapman & Hall/CRC.

Greene, William H. 1990. *Econometric Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Gschwend, Thomas. 2001. *Strategic Voting in Mixed-Electoral Systems*. Ph.D. Thesis, State University of New York, Stony Brook, NY.

Gschwend, Thomas, Ron Johnston, und Charles Pattie. 2003. Split-Ticket Patterns in Mixed-Member Proportional Election Systems: Estimates and Analyses of Their Spatial Variation at the German Federal Election, 1998. *British Journal of Political Science* 33:109–127.

Gschwend, Thomas, und Helmut Norpoth. 2000. Soll und Haben: Die deutsche Wählerschaft rechnet mit den Parteien ab. In *50 Jahre Empirische Wahlforschung in Deutschland. Entwicklung, Befunde, Perspektiven, Daten*, Hrsg. von M. Klein, W. Jagodzinski, E. Mochmann and D. Ohr. Wiesbaden: Westdeutscher Verlag, 389–409.

Gschwend, Thomas, und Helmut Norpoth. 2001. „Wenn am nächsten Sonntag ...“: Ein Prognosemodell für Bundestagswahlen. In *Wahlen und Wähler: Analysen aus Anlass der Bundestagswahl 1998*, Hrsg. von M. Kaase and H.-D. Klingemann. Wiesbaden: Westdeutscher Verlag, 473–499.

Harvey, Andrew. 1976. Estimating Regression Models with Multiplicative Heteroskedasticity. *Econometrica* 44:461–465.

Hays, William L. 1988. *Statistics*. 4. Auflage. Fort Worth: Harcourt Brace.

Hennig, Christian. 2001. *Was ist Wahrscheinlichkeit? Antworten, konstruktivistisch betrachtet.* Manuskript.

Kalton, Graham. 1983. *Introduction to Survey Sampling*. Beverly Hills: Sage.

King, Gary. 1986. How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*:666–687.

King, Gary. 1989. *Unifying Political Methodology. The Likelihood Theory of Statistical Inference*. Cambridge: Cambridge University Press.

King, Gary, Robert O. Keohane, und Sidney Verba. 1994. *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Kunz, Volker. 2000. Kulturelle Variablen, organisatorische Netzwerke und demokratische Staatsstrukturen als Determinanten der wirtschaftlichen Entwicklung im internationalen Vergleich. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 52 (2):195–225.

Kunz, Volker. 2001. „Do Institutions Matter?“ Politische Bestimmungsfaktoren des Wirtschaftswachstums in demokratischen Industriegesellschaften. Antwort auf Herbert Obinger. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 53 (1):149–1665.

Mayntz, Renate. 2002. Zur Theoriefähigkeit makro-sozialer Analysen. In *Akteure – Mechanismen – Modelle. Zur Theoriefähigkeit makro-sozialer Analysen*, Hrsg. von R. Mayntz. Frankfurt: Campus, 7–43.

McCleary, Richard, und Richard Hay. 1980. *Applied Time Series Analysis for the Social Sciences*. Beverly Hills: Sage.

Mises, Richard von. 1951. *Wahrscheinlichkeit, Statistik und Wahrheit*. 3. Auflage. Vienna: Springer.

Morrison, Denton E., und Hamon E. Henkel, Hrsg. 1970. *The Significance Test Controversy. A Reader*. Chicago: Aldine.

Obinger, Herbert. 2001. Verteilungskoalitionen und demokratische Staatsstrukturen als Determinanten der wirtschaftlichen Entwicklung. Eine Replik auf Volker Kunz. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 53 (1):136–148.

Pollock, David S. G. 1979. *The Algebra of Econometrics*. Chichester: Wiley.

Popper, Karl. 1990. *A World of Propensities*. Bristol: Thoemes.

Schnell, Rainer. 1997. *Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen*. Opladen: Leske + Budrich.

Schnell, Rainer, Paul B. Hill, und Elke Esser. 1999. *Methoden der empirischen Sozialforschung*. München und Wien: R. Oldenbourg.

Signorino, Curtis S. 1999. Strategic Interaction and the Statistical Analysis of International Conflict. *American Political Science Review* 93 (2):279–297.

Spiegelhalter, David J., Andrew Thomas, Nicky Best, und Wally R. Gilks. 2003. *WinBUGS 1.4*. Cambridge: MRC Biostatistics Unit.

Western, Bruce. 2001. Bayesian Thinking about Macrosociology. *American Journal of Sociology* 107 (2):353–378.

Western, Bruce, and Simon Jackman. 1994. Bayesian Inference for Comparative Research. *American Political Science Review* 88 (2):412–423.