Kontingente Lebensverläufe

Soziologische und statistische Aspekte ihrer Beschreibung und Erklärung

Götz Rohwer

Bremen. Oktober 1994

Vorwort

Die vorliegende Arbeit beschäftigt sich mit statistischen Methoden zur Beschreibung und Modellierung von Lebensverlaufsdaten. Das Ziel liegt darin, diese Methoden im Hinblick auf soziologische Fragestellungen darzustellen und zu diskutieren. Ausgangspunkt bildet die Überzeugung, daß die Statistik nicht als ein anwendungsneutrales Methodenarsenal verstanden werden kann, sondern daß die von ihr entwickelten Methoden im Hinblick auf den jeweils intendierten Anwendungsbereich interpretiert werden müssen. Für den vorliegenden Text bedeutet dies, daß die Darstellung statistischer Methoden auf die Frage bezogen wird, wie durch sie soziologische Einsichten in Lebensverläufe und ihre sozialen Bedingungen gewonnen werden können.

Als Bezugspunkt dient die soziologische Lebensverlaufsforschung. Innerhalb dieses Forschungsprogramms gibt es unterschiedliche Perspektiven und Schwerpunktsetzungen. In der vorliegenden Arbeit orientiere ich mich hauptsächlich an denjenigen Teilen der Lebensverlaufsforschung, die sich im Kontext einer „Sozialstrukturanalyse“ bewegen. Die Diskussion statistischer Verfahren und Modelle wird insofern auf die Frage bezogen, wie ausgehend von empirischen Daten über jeweils individuelle, kontingente Lebensverläufe Einsichten in die „Sozialstruktur“ gesellschaftlicher Verhältnisse gebildet werden können.

Der Text besteht aus fünf Kapiteln, die in Abschnitte und ggf. Unterabschnitte gegliedert sind, und aus einer Zusammenfassung.[1] Das erste Kapitel ist eine Einleitung, die die Intentionen und den wissenschaftstheoretischen Hintergrund der vorliegenden Arbeit darstellt. Lebensverläufe werden als kontingente Prozesse charakterisiert; ein wesentliches soziologisches Interesse wird darin gesehen, daß ein Verständnis gesellschaftlicher Verhältnisse als Bedingungen individueller Lebensverläufe gewonnen werden soll. Im zweiten Kapitel werden die Grundbegriffe einer statistischen Beschreibung von Lebensverläufen diskutiert. Im dritten Kapitel werden einfache statistische Modelle behandelt. Es wird die Auffassung vertreten, daß die mithilfe statistischer Modelle im Kontext der soziologischen Lebensverlaufsforschung gewinnbaren Aussagen nicht als Hypothesen über

[1] Verweise erfolgen durch Angabe von Kapitelnummern oder durch Angabe von Kapitel- und Abschnittsnummern. Zum Beispiel ist Abschnitt 2.3 der dritte Abschnitt in Kapitel 2; Abschnitt 2.5.3 ist der dritte Unterabschnitt in Abschnitt 2.5. Tabellen werden durch Angabe einer Kapitel-, Abschnitts- und Tabellennummer markiert; dasselbe gilt für Abbildungen. Formeln werden nur durch die Angabe einer Abschnitts- und einer Formelnummer gekennzeichnet. Dies ist sinnvoll, da die meisten Verweise auf Formeln sich auf den Kontext des jeweils gegebenen Kapitels beschränken. Zum Beispiel ist Formel (5.2) die zweite Formel im Abschnitt 5 desjenigen Kapitels, in dem diese Formel auftritt. Wenn auf Formeln in anderen Kapiteln bezug genommen wird, wird das entsprechende Kapitel explizit angegeben oder gelegentlich auch die Kapitelnummer in die Formelreferenz aufgenommen; zum Beispiel (4-5.2) um auf die Formel (5.2) in Kapitel 4 zu verweisen.

„allgemeine Gesetzmäßigkeiten“, sondern als *Beschreibungen* gesellschaftlicher Verhältnisse angesehen werden sollten. Ausgehend von dieser Auffassung werden dann die Konstruktions- und Schätzverfahren für einfache Übergangsratenmodelle behandelt. Im Mittelpunkt des vierten Kapitels steht die Frage, in welcher Weise von *Bedingungen* individueller Lebensverläufe gesprochen werden kann. Als formales Hilfsmittel werden mehrdimensionale (parallele und interdependente) Prozesse diskutiert. Bei allen diesen Erörterungen wird zunächst davon abgesehen, daß in praktischen Anwendungen in der Regel nur Daten aus Stichproben verfügbar sind; die Modellbildung wird unmittelbar auf eine endliche Grundgesamtheit von Individuen bezogen. Einige Aspekte des statistischen Inferenzproblems werden in Kapitel 5 behandelt.

Der Text enthält eine Reihe von Beispielen, deren Zweck jedoch ausschließlich darin liegt, die theoretischen Überlegungen zu illustrieren. Alle für die Beispiele verwendeten Lebensverlaufsdaten stammen aus dem Datenbestand des Sozio-ökonomischen Panels (SOEP). Sowohl für das Datenmanagement als auch für alle statistischen Berechnungen wurde das vom Autor entwickelte Programmpaket TDA (Transition Data Analysis) verwendet.

In soziologischen Texten wird häufig nur verkürzt oder andeutungsweise auf die mathematische Sprache bezug genommen, in der die verwendeten Methoden und Modelle entwickelt worden sind und diskutiert werden. Ich glaube jedoch, daß dies gerade für Leser, die diese Methoden und Modelle nicht bereits gut kennen, das Verständnis erschwert. Ich habe mich deshalb bemüht, den unvermeidlichen Rückgriff auf eine mathematische Sprache möglichst ausführlich vorzunehmen.

Erste Überlegungen zur vorliegenden Arbeit entstanden im Rahmen meiner Tätigkeit am Hamburger Institut für Sozialforschung, dem ich an dieser Stelle danken möchte. Von wesentlicher Bedeutung war dann meine Arbeit im Rahmen des Projekts „Household Dynamics and Social Inequality“, zunächst am European University Institute in Florenz, dann an der Universität Bremen. Ich möchte vor allem Hans-Peter Blossfeld, dem Leiter dieses Projekts, danken, nicht nur dafür, daß er mir im Rahmen dieses Projekts die Gelegenheit gegeben hat, die vielfältigen Probleme kennenzulernen, die sich bei der Verwendung statistischer Methoden in der soziologischen Forschungspraxis stellen, sondern auch für wesentliche Anregungen und Kommentare zur vorliegenden Arbeit. Für nützliche Diskussionen möchte ich schließlich auch den Mitarbeitern des genannten Projekts, Sonja Drobnic, Andreas Timm und Immo Wittig, danken, außerdem Marco Becht (Madrid), Christian Dustmann (London), Gøsta Esping-Andersen (Trento), Søren Leth-Sørensen (Kopenhagen) und Wolfgang Voges (Bremen).

Bremen, Oktober 1994 Götz Rohwer

Inhalt

Kapitel 1

Einleitung

In der Soziologie dominierte lange Zeit eine statische Betrachtungsweise, d.h. gesellschaftliche Verhältnisse wurden primär als festgefügte Strukturen von relativ zeitlos und unabhängig von ihren subjektiven Trägern definierbaren Positionen beschrieben. Zwar gab es immer wieder theoretische Kritik an diesem Gesellschaftsbild, die empirische Erforschung dynamischer Aspekte gesellschaftlicher Verhältnisse ist jedoch vergleichsweise jung. Einen wesentlichen Aufschwung hat es erst in den letzten etwa 25 Jahren gegeben.

Eine besonders wichtige Bedeutung hat in diesem Zusammenhang die soziologische Lebensverlaufsforschung. Ihr Ausgangspunkt ist die Überlegung, daß die empirische Erforschung gesellschaftlicher Verhältnisse bei den individuellen Lebensverläufen der Gesellschaftsmitglieder anzusetzen habe, und daß die Dynamik individueller Lebensverläufe gewissermaßen als Kern der Dynamik gesellschaftlicher Verhältnisse angesehen werden könne. Ausgehend von diesen beiden Basisüberlegungen ist in den vergangenen Jahren ein intensiv bearbeitetes Forschungsprogramm entstanden.[1]

Wie alle sich rasch entwickelnden Wissenschaftsgebiete ist auch die Lebensverlaufsforschung kein theoretisch homogenes Forschungsprogramm. Es gibt nicht nur zahlreiche unterschiedliche Formen der Bezugnahme auf Problemstellungen der soziologischen Theoriebildung, sondern auch durchaus unterschiedliche wissenschaftstheoretische Positionen. Etwas überspitzt könnte man sagen, daß die zahlreichen Beiträge zur Lebensverlaufsforschung bisher hauptsächlich nur zwei Gemeinsamkeiten aufweisen: erstens in der Erhebung und Präsentation von Lebensverlaufsdaten, zweitens in der Verwendung eines gewissen Spektrums statistischer Methoden zur Analyse und Modellierung dieser Daten.[2] Selbst diese relative Homogenität verschwindet, wenn man zusätzlich noch die „Biographieforschung“

[1] Als Einführungen in die soziologische Lebensverlaufsforschung vgl. u.a. Elder [1985], Mayer [1987, 1990], Becker [1990], Hagestad [1991], Dex [1991]; zur Geschichte dieses Forschungsprogramms s. Elder und Caspi [1990].

[2] Grundlegend für die methodische Vereinheitlichung der Lebensverlaufsforschung waren insbesondere die Arbeiten von Coleman [1981], Tuma, Hannan und Groeneveld [1979] und Tuma und Hannan [1984]. Inzwischen gibt es zahlreiche Einführungen in die in der Lebensverlaufsforschung vornehmlich verwendeten statistischen Methoden, u.a. Allison [1984], Diekmann und Mitter [1984], Blossfeld et al. [1986, 1989], Andreß [1985, 1992], Yamaguchi [1991]. Eine Methodeneinführung im Hinblick auf demographische Anwendungen haben Courgeau und Lelievre [1992], eine Darstellung aus ökonometrischer Sicht hat Lancaster [1990] gegeben.

mit ihrer Betonung „interpretativer Verfahren“ betrachtet.[3]

Bei der Verknüpfung der empirischen Lebensverlaufsforschung mit soziologischer Theoriebildung können drei Problembereiche unterschieden werden. Zunächst kann ein wesentlicher Beitrag der Lebensverlaufsforschung darin gesehen werden, daß sie einen empirischen Zugang zu einer dynamischen Beschreibung gesellschaftlicher Verhältnisse liefert. Schon hierin liegt, wie ich glaube, eine erhebliche *theoretische* Bedeutung. Mit den Ergebnissen der Lebensverlaufsforschung können nicht nur überlieferte statische Vorstellungen über die Verfassung gesellschaftlicher Verhältnisse infrage gestellt werden, sondern sie erzeugen darüberhinaus eine zentrale theoretische Aufgabenstellung: einen Begriff gesellschaftlicher Verhältnisse zu gewinnen, der die in ihnen lebenden Individuen als Subjekte von Lebensverläufen reflektierbar macht.

Ein zweiter Problembereich liegt darin, gesellschaftliche Verhältnisse *als Bedingungen* individueller Lebensverläufe sichtbar zu machen. Dies ist natürlich keine neue Fragestellung; die Vorstellung, daß Menschen durch ihre gesellschaftlichen Verhältnisse geprägt und – mehr oder weniger weitgehend – „bestimmt“ werden, ist eine Basisüberzeugung fast aller soziologischen Theoriebildungen. Eine wesentliche Bedeutung der Lebensverlaufsforschung kann jedoch darin gesehen werden, daß sie einen neuen empirischen Zugang zu dieser klassischen Fragestellung ermöglicht. Der neue Zugang liegt darin, daß die Individuen nicht als Träger sozial fixierter Positionen, sondern als Subjekte von Lebensverläufen wahrgenommen werden. Die klassische Fragestellung gewinnt dann eine neue Form. Die Gesellschaft erscheint nicht mehr als eine Maschine (oder Organismus), die sich ihre individuellen Mitglieder funktional unterwirft, sondern als ein sozialer Rahmen für die Entwicklung individueller Lebensverläufe. Es ist klar, daß ein solcher Perspektivenwechsel auch wesentliche Reformulierungen traditioneller Fragestellungen in bezug auf soziale Ungleichheit zur Folge hat. Die Betrachtung ungleicher Positionen muß auf die Vielfalt der sich in einer Gesellschaft entwickelnden Lebensverläufe bezogen werden.[4]

Der dritte Problembereich kreist um die Frage, in welcher Weise Individuen nicht nur als abhängig von ihren gesellschaftlichen Verhältnissen, sondern auch als Erzeuger ihrer gesellschaftlichen Verhältnisse angesehen werden können. Die traditionelle soziologische Theoriebildung neigte bekanntlich dazu, entweder (meistens) die erste oder (gelegentlich) die zweite dieser beiden Betrachtungsweisen einseitig zu betonen. Zwar kann auch die Lebensverlaufsforschung keine Patentlösung für eine theoretisch befriedigende Synthese beider Betrachtungsweisen liefern. Sie liefert jedoch in mehrfacher Hinsicht einen neuen Zugang zur Reflexion des zugrundeliegenden Problems. Zunächst vor allem dadurch, daß sie die Individuen als

[3] Exemplarisch sei auf Kohli [1986] und Brose [1990] verwiesen.

[4] Diese Fragestellung steht zwar nicht im Mittelpunkt der vorliegenden Arbeit. Einige kurze Anmerkungen werden jedoch in Abschnitt 2.6 gegeben.

Subjekte ihrer Lebensverläufe wahrnimmt. Diese sind dadurch bereits in ihrer soziologischen Konzeption nicht in erster Linie Geschöpfe ihrer gesellschaftlichen Verhältnisse, sondern Akteure. Darüberhinaus erleichtert die Lebensverlaufsforschung eine Reflexion des Verhältnisses von Individuen und Gesellschaft dadurch, daß sie sowohl die Individuen als auch ihre gesellschaftlichen Verhältnisse als temporale Verläufe sichtbar macht. Aus der Perspektive der Lebensverlaufsforschung müssen Einsichten in gesellschaftliche Verhältnisse aus Informationen über individuelle – und mithin in jeweils spezifischen historischen Situationen sich entwickelnde – Lebensverläufe konstruiert werden, denn nur auf diese Weise ist ein empirischer Zugang möglich. Der resultierende Begriff gesellschaftlicher Verhältnisse ist infolgedessen bereits in seiner theoretischen Konzeption dynamisch und historisch. Eine Beschreibung sozialen Wandels wird dann auf relativ einfache Weise möglich, indem zwei sich überschneidende Zeitdimensionen unterschieden und in ihrem Zusammenhang reflektierbar gemacht werden: einerseits die Zeitdimension der jeweils individuellen Lebensverläufe, die mit der Geburt ihrer Subjekte beginnen, und andererseits eine historische Zeitdimension, in der unterschiedliche Generationen von Individuen aufeinander folgen und sich ablösen. Ein solcher kohortenanalytischer Forschungsansatz liefert zwar zunächst nur einen deskriptiven Zugang zur Beschreibung sozialen Wandels. Aber die durch ihn ermöglichten Einsichten in den Ablauf sozialen Wandels bilden sicherlich einen sinnvollen Ausgangspunkt, um unterschiedliche theoretische Deutungen zu reflektieren und eine empirische Annäherung an die Frage zu finden, ob und ggf. wie die Subjekte individueller Lebensverläufe zugleich als Erzeuger ihrer gesellschaftlichen Verhältnisse verstanden werden können.[5]

Diese kurzen Andeutungen mögen hier ausreichen, um deutlich zu machen, daß die Lebensverlaufsforschung zahlreiche Bezüge zu Kernfragen der soziologischen Theoriebildung aufweist, denn im Mittelpunkt der vorliegenden Arbeit stehen nicht unmittelbar diese theoretischen Bezugsprobleme, vielmehr die statistischen Methoden und Modelle, die in der Lebensverlaufsforschung verwendet werden. Dabei geht es mir allerdings nicht in erster Linie um die technischen Aspekte dieser Methoden; dazu existiert bereits eine breite Literatur.[6] Es geht mir vielmehr um die Frage, in welcher Weise die in der Lebensverlaufsforschung verwendeten statistischen Begriffe, Methoden und Modelle als Instrumente *soziologischer* Wissensbildung verstanden werden können. Im Kontext dieser Fragestellung wird dann auch auf die oben angedeuteten theoretischen Bezüge der Lebensver-

[5] Eine gute Diskussion der Schwierigkeiten soziologischer Theorien sozialen Wandels findet sich in den Arbeiten von Boudon [1983, 1984].

[6] Vgl. die in Anmerkung 2 auf Seite 5 angegebenen Hinweise. Darüberhinaus gibt es eine umfangreiche statistische Literatur, die sich weitgehend unabhängig von speziellen Anwendungsgebieten mit Methoden zur Analyse von Verlaufsdaten beschäftigt. Als Einführungen können u.a. genannt werden: Kalbfleisch und Prentice [1980], Lawless [1982], Cox und Oakes [1984].

laufsforschung Bezug genommen.

Die Frage nach der soziologischen Bedeutung der in der Lebensverlaufsforschung verwendeten statistischen Methoden und Modelle steht in einem gewissen Spannungsverhältnis zu einer auch in den Sozialwissenschaften verbreiteten Ansicht, der zufolge die Statistik ein weitgehend anwendungsneutrales Methodenwissen liefert.[7] Die meisten Lehrbücher der Statistik unterstützen diese Ansicht und betonen die Allgemeinheit der statistischen Methoden und die Unabhängigkeit ihrer theoretischen Begründung von den jeweiligen Anwendungsfeldern.[8] Im Unterschied zu dieser Auffassung gehe ich in dieser Arbeit davon aus, daß ein jeweils spezifischer Gegenstands- bzw. Problembereich sowohl Sinnvoraussetzungen als auch Anwendungsgrenzen statistischer Methoden erzeugt. Als Leitfaden für die Darstellung und Erörterung statistischer Methoden zur Beschreibung und Modellierung von Lebensverlaufsdaten dient hier deshalb die Frage, welchen Sinn ihre Verwendung in der empirischen Sozialforschung haben soll. Darauf kann dann Bezug genommen werden, um die Methoden zur Konstruktion statistischer Modelle – d.h. die Begründungen für das durch sie intendierte Wissen – zu verstehen.

Ich betone diese Bezugnahme auf Erkenntnisinteressen, weil sich die statistische Theorie (und weitgehend auch die sie begleitende neuere wissenschaftstheoretische Diskussion) hauptsächlich auf naturwissenschaftliche Anwendungen bezieht. Die leitende Vorstellung besteht typischerweise darin, daß mithilfe reproduzierbarer Experimente oder experiment-ähnlicher Beobachtungen zeitlos gültige (nur an die reproduzierbaren oder wiederholt beobachtbaren Bedingungen des Experiments gebundene) Gesetzmäßigkeiten gefunden werden können, deren Sinn darin liegt, daß sie konditionale Prognosen ermöglichen.[9] Insofern mit soziologischer Theoriebildung ein vergleichbares Ziel verfolgt wird – Einsichten zu gewinnen, die

[7] Zahlreiche Belege und eine kritische Diskussion geben Gigerenzer et al. [1989, insb. S. 105ff, 286ff].

[8] Vgl. exemplarisch McPherson [1990, S. XVII]: „While the involvement of Statistics in scientific studies is varied both in respect of areas of application and forms of statistical analysis employed, there is a general structure which provides a unified view of Statistics as a means of comparing *data* with a *statistical model*.“ Sehr dezidiert äußert sich auch Schaich [1990, S. 2]: „Ihrer Eigenart nach sind die statistischen Methoden in ihrer theoretischen Grundlegung genau so wie in ihrer konkreten Anwendung vom *Erfahrungsbereich*, in welchem sie angewendet werden, *unabhängig*.“ Dagegen waren ältere Lehrbücher zur Statistik in den Sozialwissenschaften gelegentlich wesentlich differenzierter; vgl. zum Beispiel Anderson [1965].

[9] Manche Autoren sehen in der Bezugnahme auf wiederholbare Experimente eine notwendige Voraussetzung für die Anwendung statistischer Methoden; zum Beispiel sagt McPherson [1990, S. 3]: „Those scientific investigations in which Statistics might be employed are characterized by the fact that (i) they generate *data*, and (ii) there is an *experimental set-up* which is the source of the data. Scientists construct *models* of the experimental set-up and use the data to accept or modify those models.“ Auf die Frage, ob randomisierte Stichprobenziehungen als Experimente aufgefaßt werden können, wird in Kapitel 5 näher eingegangen.

sich für konditionale Prognosen über die gesellschaftliche Entwicklung verwenden lassen –, liegt es nahe, diese Leitvorstellung zu übernehmen, bloß mit der Einschränkung, daß die soziologische Forschung im wesentlichen nicht auf reproduzierbare Experimente gegründet werden kann, sondern sich mit nur begrenzt kontrollierbaren Beobachtungen behelfen muß.[10]

Ich will die Berechtigung dieses Interesses an konditionalen Prognosen über die gesellschaftliche Entwicklung nicht bestreiten. Fragwürdig erscheint mir jedoch, wenn darin das einzige Ziel einer empirischen Soziologie gesehen wird, sowie die von der „Analytischen Wissenschaftstheorie“ verbreitete Ansicht, daß sich die empirische Sozialforschung unmittelbar an den Naturwissenschaften orientieren könne, daß sie – vergleichbar mit den Naturwissenschaften – ihr Ziel darin sehen sollte, Gesetzmäßigkeiten zu ermitteln, die das soziale Leben der Menschen beherrschen. Demgegenüber gehe ich in dieser Arbeit davon aus, daß eine zentrale – und vorgängig konzipierbare – Aufgabe der empirischen Sozialforschung darin liegt, gesellschaftliche Verhältnisse *zu beschreiben* und sie als historisch sich wandelnde Bedingungen individueller Lebensverläufe sichtbar zu machen.[11]

Der Rest dieser Einleitung verfolgt hauptsächlich den Zweck, diese Aufgabenstellung zu präzisieren, ihre *theoretische* Bedeutung sichtbar zu machen und auf einige mit ihr verbundene Probleme hinzuweisen.

1.1 Lebensverlaufsdaten

Lebensverlaufsdaten beziehen sich nicht unmittelbar auf gesellschaftliche Verhältnisse, sondern auf Individuen. Sie erfassen – für jedes Individuum aus einer gegebenen Grundgesamtheit oder Stichprobe – eine spezifische Abfolge von Zuständen, die den Lebensverlauf des Individuums charakterisiert. Aus soziologischer Sicht geht es dabei um sozial bedeutsame Zustände: zum Beispiel Aspekte des sozialen Zusammenlebens, Beteiligung an Schul- und Berufsausbildung, unterschiedliche Formen der Erwerbstätigkeit. Eine genaue Abgrenzung sozial bedeutsamer Zustände ist vermutlich nicht möglich; grundsätzlich können Lebensverläufe von Individuen im Hinblick auf alle Aspekte sozialer Verhältnisse thematisiert werden. Die einzige aus soziologischer Perspektive wesentliche Bedingung liegt

[10] In einer durchaus typischen Formulierung sagt zum Beispiel Pfanzagl [1983, S. 91]: „Die Möglichkeit, unter ‘kontrollierten’ Bedingungen zu arbeiten, ist jedoch bei wirtschaftlichen und soziologischen Analysen beschränkter als in den meisten anderen Wissenschaften.“ Ob es sich hier nur um einen graduellen Unterschied handelt, ist allerdings umstritten, vgl. die Diskussion bei Lieberson [1985].

[11] Diese Aufgabenstellung korrespondiert mit folgender Charakterisierung der soziologischen Lebensverlaufsforschung durch Mayer [1987, S. 54]: „Die Lebensverlaufsforschung als ein soziologischer Arbeitsschwerpunkt befaßt sich mit der gesellschaftlichen Prägung von Lebensverläufen, der Verteilung und Ungleichheit von Lebensverläufen innerhalb einer Gesellschaft sowie deren Veränderungen im Kontext des gesellschaftlichen Wandels.“

darin, daß es sich um *sozial definierte* Zustände handeln muß, so daß die Individuen einer Gesellschaft (oder eines sinnvoll abgrenzbaren Teils) im Hinblick auf diese Zustände verglichen und klassifiziert werden können.[12]

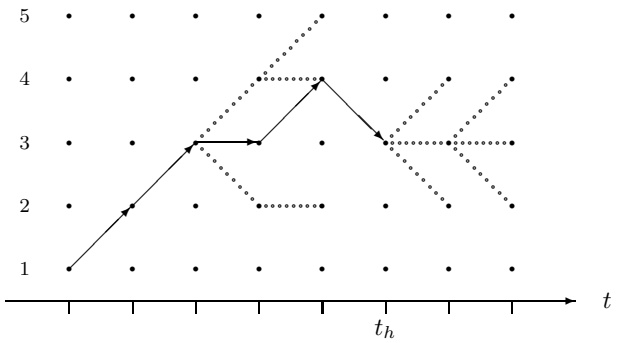
Um Lebensverlaufsdaten zu charakterisieren, ist es zweckmäßig, sie von Querschnittsdaten zu unterscheiden. Querschnittsdaten beziehen sich auf einen bestimmten Zeitpunkt, sie liefern infolgedessen keine Information über Veränderungen gesellschaftlicher Verhältnisse. Auch wenn Querschnittsdaten für eine Reihe von Zeitpunkten, etwa $t_1, t_2, \ldots, t_n$, verfügbar sind, bleibt die Information über Veränderungen in diesem Zeitraum wesentlich beschränkt. Handelt es sich zum Beispiel um eine Untersuchung der Arbeitslosigkeit, erfährt man, wie sich das Ausmaß der Arbeitslosigkeit in der Abfolge der Zeitpunkte $t_1, \ldots, t_n$ verändert hat; es ist jedoch in der Regel nicht möglich, daraus Einsichten zu gewinnen, wie die individuellen Gesellschaftsmitglieder *in ihren Lebensverläufen* von der Arbeitslosigkeit betroffen sind.

Die Erhebung von Lebensverlaufsdaten zielt demgegenüber ausdrücklich darauf, Informationen nicht nur über transitorische Zustände, sondern über Lebensverläufe zu gewinnen. Die praktischen Formen, in denen dies erreicht werden kann, sind dabei zunächst unwichtig. Entscheidend ist, daß man schließlich Informationen gewinnt, mit denen (wie auch immer begrenzte) Aussagen über Lebensverläufe – im Unterschied zu einmaligen Aussagen über transitorische Zustände – getroffen werden können. Einige Informationen dieser Art können bereits durch Querschnitts*erhebungen* gewonnen werden. Erfährt man dabei das Geburtsdatum der Personen, liefert jede Feststellung von Zuständen, die zum Erhebungszeitpunkt eingenommen werden, bereits eine gewisse Information über den bisherigen Lebensverlauf der befragten Personen. Wesentlich mehr Informationen erhält man natürlich, wenn die Querschnittserhebung mit ausführlichen Retrospektivbefragungen über die jeweils bis zum Befragungszeitpunkt realisierten Lebensverläufe verbunden wird. Zusätzliche Informationen können gewonnen werden, wenn man – im Rahmen einer Panelerhebung – die weitere Entwicklung der Lebensverläufe zu erfassen versucht.

[12] In der deutschsprachigen Literatur zur Lebensverlaufsforschung wird gelegentlich unterschieden zwischen *Lebensverläufen* im Sinne von Sequenzen objektiv feststellbarer Ereignisse bzw. Zustände und *Biographien* im Sinne von subjektiv gedeuteten Lebensgeschichten, vgl. zum Beispiel Kohli [1978, 1985, 1986]. Um dieser Unterscheidung Rechnung zu tragen, spreche ich in der vorliegenden Arbeit durchgängig von Lebensverläufen. Ich möchte jedoch offen lassen, ob dieser Unterscheidung eine systematische Bedeutung gegeben werden kann. Denn einerseits beruht natürlich auch die soziologische Theoriebildung, in die die empirische Lebensverlaufsforschung üblicherweise eingebettet wird, auf Deutungen und Interpretationen von Lebensverläufen, die häufig nicht weniger spekulativ, sondern nur schematischer sind als die Deutungen von Lebensverläufen durch ihre Subjekte. Und andererseits hat bereits weitgehend ein Prozeß der „Verwissenschaftlichung“ eingesetzt, d.h. eine Übernahme soziologischer Vorstellungen und Redewendungen in die alltagspraktische Deutung von Lebensverläufen. Beides macht die intendierte Unterscheidung fragwürdig.

Zur formalen Charakterisierung von Lebensverlaufsdaten kann (vorläufig) folgende Notation dienen. Es gibt eine endliche Anzahl von Individuen, identifizierbar durch Indizes $i = 1, \ldots, N$, und eine Zeitachse $\mathcal{T}$ mit einem beliebigen Nullpunkt, der sie mit der üblichen Kalenderzeit verbindet. Die Zeitachse wird zunächst als diskret angenommen, so daß von einer *Folge* von Zuständen gesprochen werden kann; man kann sich zum Beispiel vorstellen, daß die Zeit in Tagen gemessen wird. Schließlich gibt es noch eine endliche Menge von Zuständen, in denen sich die Individuen befinden können; sie wird mit $\mathcal{Y}$ bezeichnet. Es wird angenommen, daß dieser Zustandsraum so definiert ist, daß sich jedes Individuum zu jedem Zeitpunkt in genau einem der Zustände aus $\mathcal{Y}$ befindet. Um jedem Individuum zu jedem Zeitpunkt einen Zustand zuordnen zu können, soll der Zustandsraum insbesondere die Quasi-Zustände *noch nicht geboren* und *bereits gestorben* enthalten. Der Zustand, in dem sich das Individuum i zum Zeitpunkt $t \in \mathcal{T}$ befindet, wird mit y_{it} bezeichnet. Unter Verwendung dieser Notation kann jeder Lebensverlauf durch eine Folge von Zuständen (y_{it}) repräsentiert werden.

Diese formale Notation erleichtert das Reden über Lebensverläufe bzw. Lebensverlaufsdaten, verschleiert jedoch einige wesentliche Aspekte. Um dies zu erläutern, kann folgendes Bild dienen:[13]



In diesem Beispiel besteht der Zustandsraum aus fünf Zuständen: $\mathcal{Y} = \{1, 2, 3, 4, 5\}$. Zustand 1 ist der Quasi-Zustand *noch nicht geboren*, und 5 ist der Quasi-Zustand *bereits gestorben*. Die durchgezogenen Linien beschreiben den bisher, bis zum Zeitpunkt t_h („heute“) realisierten Lebenslauf eines Individuums. Er beginnt zu einem gewissen Zeitpunkt mit einem Übergang aus dem Quasi-Zustand 1 in den Zustand 2, dann erfolgt ein Übergang in den Zustand 3, usw. Allgemein: ein Lebensverlauf kann als ein zeitlich-sequentielles Durchlaufen eines Zustandsraums betrachtet

[13] Die Anregung zu diesem Bild sowie wesentliche Überlegungen zum Verständnis von Kausalität und Kontingenz verdanke ich von Wright [1974].

werden.

Eine wichtige Unterscheidung kann nun zwischen möglichen und realisierten Lebensverläufen getroffen werden. Zahlreiche unterschiedliche Lebensverläufe sind möglich, aber jedes Individuum realisiert schließlich nur einen bestimmten Lebensverlauf. Einige mögliche Lebensverläufe sind im obigen Bild durch gepunktete Linien angedeutet worden; nur einige wenige, denn ihre Anzahl ist bereits bei wenigen Zuständen *sehr* groß. Allerdings ist nicht jeder Verlauf, der in ein sequentielles Zustandsmodell eingezeichnet werden kann, möglich. Es gibt logische und reale Restriktionen. Zum Beispiel ist es logisch unmöglich, in den Quasi-Zustand 1 zurückzukehren, nachdem er einmal verlassen worden ist. Ebenso ist es – die üblichen semantischen Regeln unserer Sprache vorausgesetzt – logisch unmöglich, vom Zustand *unverheiratet* unmittelbar in den Zustand *geschieden* überzugehen, ohne zuvor den Zustand *verheiratet* durchlaufen zu haben.

Reale Restriktionen können unter verschiedenen Aspekten betrachtet werden. Ein wesentlicher Bereich von Restriktionen kann erfaßt werden, wenn man Menschen als Bestandteil ihrer „natürlichen“ Umwelt betrachtet. Ein anderer Bereich von Restriktionen ergibt sich „aus den gesellschaftlichen Verhältnissen“, in denen die Menschen leben, und auf diese Restriktionen konzentriert sich das soziologische Interesse. Es ist jedoch problematisch, diese beiden Bereiche abzugrenzen, denn auch die „natürlichen“ Restriktionen, denen menschliche Lebensverläufe unterliegen, sind sozial überformt und insoweit historisch wandelbar. Infolgedessen ist es kaum möglich, allgemeine Aussagen über Restriktionen zu treffen, sondern solche Aussagen sollten historisch relativiert und auf die jeweils betrachteten Aspekte (möglichen Zustände) der Lebensverläufe bezogen werden.[14]

Die Unterscheidung zwischen möglichen und realisierten Lebensverläufen hat offenbar eine zeitliche Dimension. Jeweils bis zur Gegenwart (im obigen Bild der Zeitpunkt t_h, „heute“) gibt es einen realisierten Lebensverlauf; für die „heute“ beginnende Zukunft gibt es jedoch zahlreiche unterschiedliche Möglichkeiten. Es gibt zwar logische und reale Restriktionen, nicht zuletzt infolge des jeweils bereits realisierten Lebensverlaufs, gleichwohl gibt es immer mehr als nur eine Möglickeit, wie der Lebensverlauf fortgesetzt werden kann,[15] d.h. die jeweils „heute“ beginnende Zukunft individueller Lebensverläufe ist nicht vollständig durch in der Vergangenheit entstandene Bedingungen determiniert. Ich nenne dies im folgenden die *Kontingenz von Lebensverläufen*.[16]

[14] Es bleibt dann immer noch die Frage, in welcher Weise davon gesprochen werden kann, daß gesellschaftliche Verhältnisse Restriktionen individueller Lebensverläufe sind. Diese für die soziologische Theoriebildung zentrale Frage wird in den späteren Abschnitten dieser Einleitung und dann vor allem in Kapitel 4 behandelt.

[15] Vielleicht gibt es Extremsituationen, in denen diese Aussage nicht gilt. Von dieser Möglichkeit wird hier und auch in allen weiteren Überlegungen abgesehen.

[16] Dies ist zu unterscheiden von einem Sachverhalt, den K. U. Mayer [1986, S. 168] „cumulative contingencies“ nennt, womit gemeint ist, daß der jeweils bisher realisierte Lebenslauf zugleich die Möglichkeiten einschränkt, in denen er fortgesetzt werden kann. Mit der Formulierung, daß Lebensverläufe kontingent sind, soll demgegenüber zunächst nur auf die Tatsache verwiesen werden, daß es – von Extremsituationen abgesehen – stets mehrere Möglichkeiten gibt, den bisher realisierten Lebensverlauf fortzusetzen.

Man kann diese Kontingenz unter zwei komplementären Gesichtspunkten betrachten. Einerseits kann man sagen, daß Menschen zumindest einige *Entscheidungen* treffen können, die ihren Lebensverlauf beeinflussen, und daß solche Entscheidungen nicht vollständig durch ihre Veranlagung, ihren bisherigen Lebensverlauf und Umweltbedingungen erklärt werden können (wobei hier eine kausale Erklärung im Unterschied zu psychologischen oder normativen Rationalisierungen gemeint ist).[17] Einen komplementären Gesichtspunkt liefert die Einsicht, daß es nur begrenzte Möglichkeiten gibt, um die Folgen von Entscheidungen vorauszusehen; die Kontingenz besteht dann darin, daß man nicht sicher wissen kann, welcher der in der Zukunft noch möglichen Lebensverläufe schließlich realisiert wird, und zwar selbst dann nicht, wenn es sich um Menschen handelt, die sich um eine Planung ihres Lebenslaufs bemühen. Man kann dies mit vielerlei Argumenten begründen. Am einfachsten und überzeugendsten ist vielleicht das Argument, das Popper [1965, XIf.] gegen den Historizismus angeführt hat: daß man den Fortschritt des (wissenschaftlichen) Wissens nicht antizipieren kann; oder in einer direkter auf Lebensverläufe beziehbaren Formulierung von Danto [1965, S. 465]: „Da wir nicht wissen, wie unsere Handlungen aus dem Gesichtswinkel der Historie gesehen werden, ermangeln wir dementsprechend auch der Kontrolle über die Gegenwart.“[18]

Man könnte einwenden, daß es sich bei Kontingenz individueller Lebensverläufe um einen trivialen Sachverhalt handelt, der für die soziologische Beobachterperspektive unerheblich ist. Ich stimme dem zu, wenn damit gemeint ist, daß soziologische Fragestellungen sich nicht auf individuelle Lebensverläufe und deren Kontingenz richten, sondern auf die Verfassung der gesellschaftlichen Verhältnisse, gewissermaßen auf den Rahmen, in dem sich die Kontingenz der individuellen Lebensverläufe abspielt. Gesellschaftliche Rahmenbedingungen – eine Metapher, die noch zu erörtern ist – können jedoch nicht vollständig losgelöst von individuellen Lebens-

[17] Es sei betont, daß diese Bezugnahme auf Entscheidungen nicht auf die Unterstellung eines “rationalen“ Akteurs angewiesen ist. Eine Entscheidung ist zunächst nur eine spontane Fixierung einer Alternative, und in diesem Sinne kann der Begriff auf alle aus soziologischer Sicht relevanten Phasen von Lebensverläufen angewendet werden, unabhängig davon, welche Rationalisierungsmöglichkeiten aus der Sicht des Subjekts oder eines Beobachters existieren.

[18] Dantos Überlegung geht offensichtlich über diejenige Poppers hinaus. In der Fortsetzung seines Gedankengangs heißt es: „Wenn es so etwas wie Unausweichlichkeit in der Geschichte gibt, so ist sie nicht so sehr sozialen Prozessen zuzuschreiben, die sich aus eigener Kraft in Gang halten und den Eigengesetzlichkeiten ihrer Natur gemäß ablaufen, sondern eher dem Umstand, daß es zu dem Zeitpunkt, zu dem klar wird, was wir getan haben, zu spät ist, noch etwas daran zu ändern.“

verläufen begrifflich gefaßt werden. Sie sind stets auch das transitorische Resultat der individuellen Lebensverläufe. Insofern ist es eine zunächst offene Frage, welche Konsequenzen die Kontingenz individueller Lebensverläufe für unseren Begriff gesellschaftlicher Verhältnisse – und für die Bilder, mit denen wir die Verstrickungen der Individuen in ihre gesellschaftlichen Verhältnisse reflektieren – haben sollte.[19]

Ich betone diesen Sachverhalt der Kontingenz von Lebensverläufen hauptsächlich aus zwei Gründen. Erstens, um ein Bild individueller Lebensverläufe zu gewinnen, das nicht von vornherein eine bestimmte soziologische Deutung impliziert. In der Lebensverlaufsforschung wird häufig von vornherein, d.h. bereits bei der Definition des Begriffs *Lebensverlauf*, davon ausgegangen, daß es sich um einen „institutionalisierten" Sachverhalt handelt.[20] Demgegenüber glaube ich, daß es sinnvoller ist, dies zunächst als eine offene Frage anzusehen.[21] Offen sowohl in empirischer Hinsicht, d.h. im Hinblick auf die Frage, in welchem Ausmaß tatsächlich Regelmäßigkeiten in den individuellen Lebensverläufen festgestellt werden können; aber auch in theoretischer Hinsicht, d.h. im Hinblick auf die Frage, welche theoretische Bedeutung der *individuellen* Differenzierung von Lebensverläufen beigemessen werden sollte und welche Aspekte individueller Lebensverläufe für die soziologische Reflexion sozialer Ungleichheit als relevant angesehen werden sollten.[22]

Zweitens betone ich die Kontingenz individueller Lebensverläufe, weil

[19] Vgl. dazu die Arbeit von Bohman [1991], in der die „Indeterminiertheit sozialen Handelns" als zentrales Problem der soziologischen Theoriebildung dargestellt wird.

[20] Vgl. Kohli [1985]. Gelegentlich wird in der Litertur auch „der Lebensverlauf" als eine „Institution" bezeichnet, etwa von Kohli, der zum Beispiel [1985, S. 2] sagt: „Die Bedeutung des Lebensverlaufs als soziale Institution hat stark zugenommen." Ich verstehe Formulierungen dieser Art so, daß durch sie darauf hingewiesen werden soll, daß es für die Lebensverläufe der Individuen zahlreiche soziale Regelungen gibt, daß damit jedoch die konzeptionelle Differenz zwischen Lebensverläufen, d.h. Beschreibungen individueller Entwicklungsprozesse, und sozialen Regelungen (Institutionen) nicht aufgehoben werden soll.

[21] Tatsächlich ist auch nicht immer klar, was mit der Formulierung gemeint sein soll, daß es sich bei individuellen Lebensverläufen um einen „institutionalisierten" Sachverhalt handele. Gelegentlich gewinnt man den Eindruck, daß damit in erster Linie ein normativer und ideologischer Sachverhalt gemeint ist, der mit der Realität und Vielfalt der individuellen Lebensverläufe nur locker verbunden ist. Zum Beispiel heißt es bei Meyer [1986, S. 203]: „It is important that the standardized life course as we discuss it is an institutional system, a set of rules and preferences of a formalized but highly abstract kind. There is a great deal of slippage when it becomes translated into the actual experience of individuals, whose real life courses have more anomalies and unpredictabilities than the official system." Wenn in dieser Arbeit von Lebensverläufen gesprochen wird, sind demgegenüber stets die „real life courses" gemeint.

[22] Ein wichtiger Beitrag der Lebensverlaufsforschung kann sicherlich bereits darin gesehen werden, daß sie den Blick für die Vielfalt individueller Lebensverläufe öffnet (dies wird zum Beispiel von Sherrod und Brim [1986] betont) und dadurch zahlreiche Abstraktionen und Schematisierungen der soziologischen Theoriebildung empirisch hinterfragbar macht.

dieser Sachverhalt eine wesentliche Grenze für die soziologische Wissenbildung darstellt. Dies gilt bereits für die empirische Wissensbildung. Was wir in Form von Daten über Lebensverläufe wissen, bezieht sich immer nur auf die Vergangenheit, also auf realisierte Lebensverläufe. Mithilfe solcher Daten können individuelle Lebensverläufe und, wie noch zu überlegen sein wird, gesellschaftliche Verhältnisse beschrieben werden. Aber diese Beschreibungen beziehen sich stets auf die Vergangenheit und erreichen grundsätzlich nicht das „heute" der individuellen Lebensverläufe, in dem ihre Kontingenz jeweils neu beginnt.[23] Noch wichtiger ist, daß die Kontingenz individueller Lebensverläufe auch der Theoriebildung Grenzen setzt. Die Vorstellung, man könne allmählich eine immer bessere, vollständigere Kausalerklärung für die Entwicklung individueller Lebensverläufe erreichen, ist sicherlich nicht haltbar. Bestenfalls kann man erwarten, partielle und selbst dem sozialen Wandel ausgesetzte Einsichten in einige der Bedingungen individueller Lebensverläufe zu gewinnen.

1.2 Lebensverläufe und gesellschaftliche Verhältnisse

Auf welche Weise können Lebensverlaufsdaten über gesellschaftliche Verhältnisse Aufschluß geben? Die Antwort auf diese Frage hängt davon ab, in welcher Weise ein theoretisch sinnvoller Begriff gesellschaftlicher Verhältnisse gebildet werden kann. Daß es sich um einen theoretischen Begriff handeln muß, folgt daraus, daß der intendierte Sachverhalt nicht unmittelbar empirisch zugänglich ist. Man lebt zwar tagtäglich in den „kollektiven Phänomenen", für die sich die Soziologie interessiert, die dabei möglichen Beobachtungen liefern aber bestenfalls bruchstückhafte Hinweise auf gesellschaftliche Verhältnisse, nicht bereits den gemeinten Sachverhalt. Dies wird erst recht deutlich, wenn man beachtet, daß gesellschaftliche Verhältnisse einen temporalen Sachverhalt bilden, der sich im Zeitablauf verändert.

Einen Zugang zum Begriff gesellschaftlicher Verhältnisse erhält man, wenn man sich überlegt, wie der gemeinte Sachverhalt beschrieben werden kann. Dabei unterstelle ich, daß mit gesellschaftlichen Verhältnissen nicht eine theoretische Fiktion, sondern ein beschreibbarer, empirisch zugänglicher Sachverhalt gemeint ist. Trotz der vielfältigen Möglichkeiten, die sich hier bieten, gibt es, wie ich glaube, nur zwei grundlegende Formen der Beschreibung.

(1) Zunächst ist es möglich, gesellschaftliche Verhältnisse durch *Zustandsverteilungen* zu beschreiben. Vorausgesetzt ist dabei zweierlei. Er-

[23] Damit soll nicht behauptet werden, daß Lebensverlaufsforschung nur im Kontext von „Historie" konzeptionalisiert werden kann. Insoweit die soziologische Lebensverlaufsforschung sich mit den jeweils lebenden Individuen und ihren Lebensverläufen beschäftigt, ist sie gegenwartsbezogen und kann – durch Befragung und Beobachtung dieser Individuen – empirisches Wissen in einer Form bilden, die dem Historiker in der Regel nicht zur Verfügung steht; vgl. Goldthorpe [1991].

stens eine Definition der „Einheiten“, die die gesellschaftlichen Verhältnisse bilden. In dieser Arbeit gehe ich davon aus, daß es sich um Individuen handelt; es ist jedoch möglich, auch komplexere soziale Gebilde zu betrachten, zum Beispiel Haushalte und Unternehmen.[24] Zweitens benötigt man eine Definition von Zuständen, in denen sich die Individuen befinden können. Eine elementare Form der Beschreibung gesellschaftlicher Verhältnisse erhält man dann einfach dadurch, daß man feststellt, wie sich die Individuen auf die Zustände verteilen. Zum Beispiel kann man auf diese Weise die Einkommensverteilung in einer Gesellschaft beschreiben, oder die jeweils existierende Berufsstruktur. Wie bei jeder Beschreibung erhält man auch bei dieser Vorgehensweise nur ein partielles Bild. Die resultierende Beschreibung gesellschaftlicher Verhältnisse hängt zunächst von den vorausgesetzten Zustandsunterscheidungen ab, außerdem von dem Zeitraum, für den die Zustandsverteilung ermittelt wird. Beziehen sich die verfügbaren Informationen nur auf einen Zeitpunkt, erhält man eine Beschreibung gesellschaftlicher Verhältnisse nur für diesen Zeitpunkt. Sind Daten für einen längeren Zeitraum verfügbar, kann man feststellen, wie sich die gesellschaftlichen Verhältnisse in diesem Zeitraum verändert haben, und ggf. von einem sozialen Wandel sprechen.

(2) Durch Zustandsverteilungen und ihre Veränderungen im Zeitablauf können gesellschaftliche Verhältnisse auf einfache Weise beschrieben werden, aber die Beschreibung bleibt in gewisser Weise oberflächlich. Zu einer anderen Art der Beschreibung gelangt man, wenn man sich die Individuen, die die gesellschaftlichen Verhältnisse bilden, als soziale Akteure vorstellt. Man kann dann versuchen, gesellschaftliche Verhältnisse durch Regeln zu beschreiben, denen die Akteure in ihrem sozialen Verhalten folgen. Es gibt zahlreiche unterschiedliche Formen, in denen Verhaltensweisen, die einer Regel folgen oder auf eine Regel Bezug nehmen, beschrieben werden können; zum Beispiel kann man von Gewohnheiten, Gebräuchen, privaten Vereinbarungen, staatlichen Gesetzen und moralischen und technischen Normen sprechen.[25] An dieser Stelle erscheint es zunächst ausreichend,

[24] Die Frage, von welchen Einheiten eine sinnvolle Beschreibung gesellschaftlicher Verhältnisse ausgehen sollte, ist kompliziert. Geht man von der Zielsetzung aus, gesellschaftliche Verhältnisse als Interaktionsverhältnisse darstellen zu wollen, ist es zweckmäßig, nur solche sozialen Gebilde zu verwenden, sie sich als soziale Akteure interpretieren lassen. Da ich mich in dieser Arbeit auf individuelle Personen beschränke, ist diese Bedingung automatisch erfüllt. – Bei Längsschnittbetrachtungen kommt das Problem hinzu, wie die Identität der Basiseinheiten gesellschaftlicher Verhältnisse im Zeitablauf begriffen werden kann. Im Hinblick auf individuelle Personen liefert die Umgangssprache eine (für die meisten Fragestellungen) hinreichende Grundlage. Dagegen ist es bei komplexeren sozialen Gebilden in der Regel äußert schwer, eine temporale Identität zu definieren. Das Problem beginnt bereits bei der Frage, ob sich Haushalte als Einheiten für Längsschnittanalysen eignen; vgl. zum Beispiel Duncan und Hill [1985], Hanefeld [1987, S. 95f], Keilman und Keyfitz [1988, S. 272ff].

[25] Eine instruktive Diskussion gibt von Wright [1963, Kap. I]. Aus soziologischer Sicht liegt es natürlich nahe, an dieser Stelle insbesondere auch auf Max Weber zu verweisen,

diese verschiedenen Formen in einem unspezifischen Begriff *sozialer Regeln* zusammenzufassen. Gemeint sind Regeln, an denen die Akteure in einer Gesellschaft ihr soziales Verhalten *orientieren*.[26] Es ist leider kompliziert, diesen Begriff präzise zu fassen (einige Bemerkungen folgen weiter unten). Ich glaube jedoch, daß die Idee, gesellschaftliche Verhältnisse durch soziale Regeln zu beschreiben, vor allem aus drei Gründen grundlegend ist. Zunächst deshalb, weil sie einem wesentlichen Aspekt der üblichen, umgangssprachlichen Wahrnehmung und Reflexion gesellschaftlicher Verhältnisse entspricht. Zweitens weil die Fixierung sozialer Regeln, insbesondere in der Form privater Vereinbarungen und rechtlicher und technischer Normen, die primäre Form ist, in der sich die Selbstorganisation von Gesellschaften entwickelt.[27] Schließlich aber auch deshalb, weil mit dem Begriff sozialer Regeln, an denen sich die Menschen in einer Gesellschaft orientieren, zunächst noch keinerlei Implikationen für unseren Umgang mit der Kontingenz individueller Lebensverläufe verbunden zu werden brauchen. Die Annahme der Existenz einer sozialen Regel impliziert weder, daß sie ausnahmslos befolgt wird, noch, daß sie befolgt werden muß.[28]

In beiden Fällen erhält man einen jeweils spezifischen, *abstrakten* Begriff gesellschaftlicher Verhältnisse.[29] In welchem Verhältnis sie zueinan-

in dessen Konzeption ein Begriff sozialer Regeln eine zentrale Rolle spielt. Allerdings hat Weber den Begriff „soziale Regel“ unmittelbar mit der Bedingung verknüpft, daß die einer Regel folgenden Verhaltensweisen „verstehbar“ sein müssen. So heißt es zum Beispiel [1976, S. 6]: „Nur solche statistische Regelmäßigkeiten, welche einem *verständlichen* gemeinten Sinn eines sozialen Handelns entsprechen, sind (im hier gebrauchten Wortsinn) verständliche Handlungstypen, also: ‘soziologische Regeln’.“ Dieser definitorischen Verknüpfung soll hier nicht gefolgt werden. Weber bemerkt übrigens gleich im Anschluß an das angeführte Zitat: „Vorgänge und Regelmäßigkeiten, welche, weil unverstehbar, im hier gebrauchten Sinn des Wortes nicht als ‘soziologische Tatbestände’ oder Regeln bezeichnet werden, sind natürlich um deswillen nicht etwa weniger *wichtig*. Auch nicht etwa für die Soziologie im hier betriebenen Sinne des Wortes (der ja eine Begrenzung auf ‘*verstehende* Soziologie’, welche niemandem aufgenötigt werden soll und kann). Sie rücken nur, und dies allerdings methodisch ganz unvermeidlich, in eine andere Stelle als das verstehbare Handeln: in die von ‘Bedingungen’, ‘Anlässen’, ‘Hemmungen’, ‘Förderungen’ desselben.“

[26] Unter *sozialem* Verhalten verstehe ich ein Verhalten, bei dessen Beschreibung exlizit oder implizit die Existenz anderer Menschen berücksichtigt wird. Dies entspricht im wesentlichen der von Weber [1976, S. 1] gegebenen Definition. Statt von Verhalten spreche ich auch von Handlungen; von der gelegentlich diskutierten Unterscheidung zwischen diesen beiden Begriffen wird hier abgesehen.

[27] Vgl. zur Bedeutung staatlicher Regulierungen für die Entwicklung von Lebensverläufen insbesondere die Arbeit von Mayer und Müller [1986], in der dieser Aspekt nachdrücklich betont wird.

[28] Ich glaube also, daß es sinnvoll möglich ist, gesellschaftliche Verhältnisse durch soziale Regeln zu definieren, ohne dies – gewissermaßen auf der begrifflichen Ebene – bereits mit einer Aussage darüber zu verbinden, warum und in welcher Weise soziale Regeln gelten. Insbesondere soll zunächst offengelassen werden, ob bzw. in welcher Weise soziales Handeln durch soziale Regeln erklärt werden kann.

[29] Wieweit sich die soziologische Betrachtung auf das implizite Abstraktionsniveau be-

der stehen, soll später erörtert werden. Zunächst ist wichtig, daß in beiden Fällen ein empirischer Zugang zur Beschreibung gesellschaftlicher Verhältnisse möglich ist.

Im Hinblick auf die Definition gesellschaftlicher Verhältnisse durch Zustandsverteilungen ist dies unmittelbar ersichtlich, denn Lebensverlaufsdaten sind typischerweise bereits in einer dafür geeigneten Form verfügbar. Schwieriger ist es, wenn man gesellschaftliche Verhältnisse durch soziale Regeln beschreiben möchte, denn diese Regeln lassen sich nicht unmittelbar beobachten. Zwar entspricht diese Betrachtungsweise dem üblichen, alltäglichen Verständnis gesellschaftlicher Verhältnisse, und jeder kennt – aus seiner Perspektive und für seinen sozialen Kontext – eine Menge sozialer Regeln; insofern ist der intendierte Begriff gut verständlich. Sieht man jedoch eine soziologische Aufgabe darin, die jeweils geltenden Regeln und ihre Veränderungen festzustellen, kann man sich nur sehr begrenzt auf das alltägliche Wissen über soziale Regeln beziehen. Um festzustellen, welche Regeln tatsächlich befolgt werden und wie sie sich verändern, kann man sich nicht auf Meinungen über Regeln berufen, sondern muß das tatsächliche soziale Verhalten untersuchen.

Dann entsteht allerdings ein Problem. Wie kann man aus der Beobachtung sozialen Verhaltens Einsichten in die Regeln gewinnen, denen dieses Verhalten folgt? Es scheint so, daß dies in gewisser Weise nicht objektiv feststellbar ist, sondern daß die Feststellung, daß eine Regel befolgt wird, nicht nur bereits eine Kenntnis der Regel, sondern auch eine Deutung des fraglichen Verhaltens im Hinblick auf die Regel erfordert. Dieses von Philosophen, aber auch – anknüpfend an die Arbeiten von Wittgenstein und Winch [1958] – von einigen Soziologen ausgiebig erörterte Problem ist kompliziert, braucht jedoch an dieser Stelle nicht verfolgt zu werden.[30] Die hier wichtige Frage ist vielmehr, ob mit den in der empirischen Sozialforschung verfügbaren Lebensverlaufsdaten sinnvoll interpretierbare Hinweise auf die Existenz sozialer Regeln gewonnen werden können.

Einen Zugang zu dieser Frage ermöglicht die statistische Analyse von Lebensverlaufsdaten. Die Überlegung geht davon aus, daß aus soziologischer Perspektive soziale Regeln nur insoweit von Interesse sind, als sie tatsächlich befolgt werden. Um soziologisch davon sprechen zu können, daß es eine soziale Regel gibt, muß sie empirisch als eine Regelmäßigkeit im sozialen Verhalten nachweisbar sein. An dieser Stelle wird dann die statistische Analyse von Lebensverlaufsdaten bedeutsam: sie liefert empirische Einsichten in Regelmäßigkeiten im Ablauf individueller Lebensverläufe, die dann als Hinweise auf soziale Regeln zur Beschreibung gesellschaftlicher Verhältnisse interpretiert werden können.

schränken kann, ist eine offene Frage. Man kann auf jeden Fall darauf hinweisen, daß es eine Vielzahl von Wechselwirkungen zwischen den sozialen Akteuren und ihren „natürlichen“ Lebensbedingungen gibt. Vgl. zum Beispiel die Beiträge von Martin, Finch und Vogel in Sørensen et al. [1986].

[30] Dieses Problem hat auch, insbesondere unter dem Einfluß der neukantianischen Unterscheidung von Natur- und Kulturwissenschaften, in der sozialwissenschaftlichen Statistik eine Rolle gespielt; vgl. zum Beispiel Hartwig [1956] sowie auch die Diskussion bei Menges [1959]. Eine Einführung in die neuere Diskussion findet sich bei Bohman [1991, insb. Kap. 2].

Zum Beispiel kann man die sozialen Regeln untersuchen, von denen sich die Menschen in einer Gesellschaft bei der Auswahl ihrer Lebenspartner leiten lassen. Eine statistische Analyse kann dann etwa zeigen, daß sich üblicherweise in zahlreichen Aspekten ähnliche Lebenspartner zusammenfinden. Ein anderes Beispiel ist der Zeitpunkt für Heiraten. Es gibt zahlreiche empirisch feststellbare Regelmäßigkeiten; zum Beispiel daß die an einer nichtehelichen Lebensgemeinschaft Beteiligten üblicherweise dann heiraten, wenn sich die Geburt eines Kindes ankündigt, und daß üblicherweise nicht geheiratet wird, solange man sich noch in der Schul- und Berufsausbildung aufhält.[31] Wie diese Beispiele zeigen, kann eine statistische Analyse von Lebensverlaufsdaten dazu beitragen, soziale Regeln zu finden, die sich zur Beschreibung gesellschaftlicher Verhältnisse eignen. Allerdings stellen sich bei näherem Hinsehen zahlreiche Fragen, die einer kurzen Erwähnung bedürfen.

1.3 Regeln, Regelmäßigkeiten und Gesetzmäßigkeiten

a) In welchem Verhältnis stehen statistisch nachweisbare Regelmäßigkeiten zum theoretischen Begriff sozialer Regeln zur Charakterisierung gesellschaftlicher Verhältnisse? Ein wichtiger Unterschied liegt offenbar darin, daß bei der empirischen Feststellung sozialer Regelmäßigkeiten von allen normativen Aspekten sozialer Regeln abstrahiert wird. Es geht nicht um die Frage, ob es gewisse moralische oder staatlich sanktionierte Vorschriften für Verhaltensweisen gibt, es soll vielmehr festgestellt werden, welchen Regeln die Mitglieder einer Gesellschaft tatsächlich folgen. Ein anderer wichtiger Unterschied liegt darin, daß es nicht zu jeder empirisch feststellbaren und soziologisch bedeutsamen sozialen Regelmäßigkeit eine im Bewußtsein der Gesellschaftsmitglieder existierende soziale Regel geben muß. Insofern hat auch die Formulierung *einer Regel folgen* zwei unterschiedliche Bedeutungen. Einerseits, im alltäglichen Verständnis, setzt die Formulierung voraus, daß die Regel vergegenwärtigt werden kann; sie muß zwar nicht „bewußt“ befolgt worden sein, aber das Verhalten, mit dem man einer Regel folgt, muß im Hinblick auf die Regel reflektierbar und kommunizierbar sein. Bei der soziologischen Feststellung sozialer Regelmäßigkeiten ist diese Bedingung nicht notwendig. Der Soziologe kann Regelmäßigkeiten im sozialen Verhalten entdecken, an die die Gesellschaftsmitglieder zuvor noch nie gedacht haben. Wenn dies der Fall ist, entsteht allerdings die Möglichkeit, daß die empirisch ermittelten Regelmäßigkeiten als soziale

[31] Vgl. Blossfeld und Huinink [1989], Blossfeld und Jaenichen [1990], Blossfeld et al. [1993], Brüderl und Diekmann [1994b].

Regeln (oder als Folgen sozialer Regeln) reflektierbar werden. Beide Unterschiede führen dazu, daß eine soziologische Beschreibung gesellschaftlicher Verhältnisse von ihrer alltäglichen Wahrnehmung abweicht und daß die Soziologie, durch die Darstellung dieser Differenz, zur Aufklärung der Gesellschaftsmitglieder über ihre gesellschaftlichen Verhältnisse beitragen kann.[32]

b) Es muß jedoch sofort hinzugefügt werden, daß sich nicht alle empirisch feststellbaren Regelmäßigkeiten, die als Beschreibungen gesellschaftlicher Verhältnisse angesehen werden können, auch als soziale Regeln deuten lassen. Dies betrifft einerseits Regelmäßigkeiten, in denen zum Ausdruck kommt, daß die Menschen nicht nur abstrakt konzipierbare soziale Akteure, sondern zugleich von „natürlichen" Lebensbedingungen abhängige Wesen sind; das in unserem Kontext wichtigste Beispiel liefern die empirisch feststellbaren Regelmäßigkeiten für die menschliche Lebensdauer. Andererseits handelt es sich um Regelmäßigkeiten, die deshalb nicht als soziale Regeln (denen man folgen kann oder nicht) gedeutet werden können, weil sie aus der Interaktion einer Mehrzahl individueller Akteure resultieren.

Statistisch ermittelbare empirische Regelmäßigkeiten können also aus zwei Gründen nicht mit sozialen Regeln gleichgesetzt werden. Erstens gibt es, wie unter (a) ausgeführt wurde, eine begriffliche Differenz, zweitens können empirische Regelmäßigkeiten auch in Bereichen festgestellt werden, die das potentiell intentionale Handeln individueller Akteure transzendieren. Es ist jedoch, wie ich glaube, keine wesentliche Frage, ob infolgedessen der oben eingeführte Begriff gesellschaftlicher Verhältnisse erweitert werden muß. Man kann soziale Regelmäßigkeiten, die sich nicht als soziale Regeln deuten lassen, als einen Aspekt beschreibbarer gesellschaftlicher Verhältnisse ansehen; man kann sie aber auch als eine Folge gesellschaftlicher Verhältnisse ansehen, die primär durch soziale Regeln zu charakterisieren sind.

[32] Diese potentielle Funktion der empirischen Sozialforschung kann natürlich auch selbstreflexiv auf die soziologische Theoriebildung angewendet werden. Insbesondere in der Lebensverlaufsforschung wird häufig davon ausgegangen, daß die Mitglieder einer Gesellschaft gewissen normativen Regeln unterliegen, die ihre Lebensverläufe „strukturieren", man denke zum Beispiel an das „concept of a normative timetable" (Elder [1975]). Damit sind jedoch nicht empirisch nachweisbare Regeln bzw. Regelmäßigkeiten gemeint, sondern den Gesellschaftsmitgliedern unterstellte Meinungen darüber, wie konforme Lebensverläufe verlaufen sollten. Die Differenz kommt zum Beispiel in folgender Bemerkung Hogans [1978, S. 1978] zum Ausdruck: „While the transition to adulthood is a relatively diffuse process in American society, it is not an unregulated process. Each cohort faces normative regulations regarding the appropriate timing for each of these events (Elder, 1974). Thus, members of the society ordinarily are in broad agreement regarding the approximate age at which it is appropriate for a man or a woman to start working, first marry, have a first child, and so on." Vermutlich gibt es derartige Auffassungen; aber es ist nicht nur unklar, inwieweit es typische und einheitliche Auffassungen dieser Art gibt, sondern auch, welche Bedeutung sie für den faktischen Ablauf der Lebensverläufe haben. Anstatt den Ablauf der Lebensverläufe durch Verweis auf derartige „normative" Meinungen zu erklären, erscheint es interessanter, sie mit den tatsächlich stattfindenden „diffusen" Abläufen zu konfrontieren.

c) Eine zunächst noch offene Frage besteht auch darin, *wie* soziale Regelmäßigkeiten empirisch festgestellt werden können. In der empirischen Sozialforschung wird dies in erster Linie durch die Erhebung und statistische Analyse von Daten – über Individuen und deren Lebensumstände – zu erreichen versucht. Einige dieser Methoden werden in späteren Kapiteln ausführlich behandelt. An dieser Stelle beschränke ich mich auf zwei allgemeine Bemerkungen.

Erstens erscheint es bemerkenswert, daß die Verwendung von Daten und statistischen Methoden in gewisser Weise dazu führt, daß die oben (S. 15) eingeführte Unterscheidung von zwei Methoden zur Beschreibung gesellschaftlicher Verhältnisse verschwindet. Es ist dies eine Konsequenz der bereits erwähnten Tatsache, daß mithilfe statistischer Methoden nicht unmittelbar soziale Regeln, sondern nur Regelmäßigkeiten festgestellt werden können. Man kann zunächst damit beginnen, gesellschaftliche Verhältnisse durch Zustandsverteilungen zu beschreiben; dann kann man versuchen, darüber hinausgehende Einsichten in soziale Regelmäßigkeiten zu gewinnen. Dabei bleibt man jedoch auf den Informationsgehalt der zunächst gegebenen Zustandsverteilungen beschränkt. Mit statistischen Methoden kann keine neue Information gewonnen werden, man kann nur versuchen, die in den Zustandsverteilungen enthaltene Information zu reformulieren, um auf diese Weise zu soziologisch interessanten und einsichtigen Beschreibungen gesellschaftlicher Verhältnisse zu gelangen. Damit wird nicht gesagt, daß statistische Methoden bedeutungslos sind; im Gegenteil, sie sind ein wesentliches Hilfsmittel, um aus Daten interessante Beschreibungen gesellschaftlicher Verhältnisse zu gewinnen. Aber auch das schließlich fertige statistische Modell liefert nur eine Beschreibung, nicht bereits eine Erklärung gesellschaftlicher Verhältnisse, obwohl es natürlich eine wichtige Funktion im Rahmen einer theoretischen Deutung im Hinblick auf soziologische Fragestellungen einnehmen kann.

Die zweite Bemerkung betrifft die Frage, ob es Kriterien für die Feststellung empirischer Regelmäßigkeiten gibt. Leider ist diese Frage kompliziert, und es gibt, soweit ich sehen kann, keine vollständig befriedigende Antwort. Zur Verdeutlichung des Problems kann das oben erwähnte Heiratsbeispiel dienen. Da man feststellen kann, daß fast alle Heiraten erst dann erfolgen, wenn die beteiligten Personen ihre Schul- und Berufsausbildung abgeschlossen haben, kann man in diesem Fall sicherlich sinnvoll von einer sozialen Regelmäßigkeit sprechen. Wenn man aber das Heiratsverhalten bei nichtehelichen Lebensgemeinschaften untersucht, stellt man vielleicht fest, daß – angesichts der Geburt eines Kindes – 70 Prozent der Paare heiraten und 30 Prozent nicht heiraten. Dies wäre offenbar kein überzeugender Nachweis für die Behauptung, daß in nichtehelichen Lebensgemeinschaften üblicherweise dann geheiratet wird, wenn ein Kind geboren wird.[33] Es ist zwar klar, daß man nicht verlangen kann, daß alle

[33] Allerdings liegt darin ebenfalls kein Nachweis für die gegenteilige Behauptung, daß

oder fast alle Personen einer bestimmten Regel folgen, um davon sprechen zu können, daß es die Regel gibt; denn es gehört gewissermaßen zum Wesen einer Regel, daß sie auch nicht befolgt werden kann. Aber andererseits muß eine Regel von hinreichend vielen Menschen befolgt werden, damit man von ihrer Existenz sprechen kann.

d) Diese Überlegung führt, wie ich glaube, zu zwei wichtigen Schlußfolgerungen. Erstens zeigt sie, daß das Problem nicht einfach darin besteht, wie soziale Regelmäßigkeiten *festgestellt* werden können, sondern bereits die Definition des Sachverhalts betrifft, den man feststellen möchte. Erforderlich ist stets eine theoretische Interpretation, um davon sprechen zu können, daß Lebensverläufe gewissen Regeln bzw. Regelmäßigkeiten folgen. Statistische Methoden und Modelle können eine solche Interpretation unterstützen, aber nicht ersetzen.[34] Eine zweite Schlußfolgerung besteht darin, daß man nicht davon ausgehen kann, daß sich alle Aspekte individueller Lebensverläufe durch soziale Regeln bzw. Regelmäßigkeiten beschreiben lassen. Inwieweit sich solche Regeln bzw. Regelmäßigkeiten finden lassen, und inwieweit infolgedessen Lebensverläufe probabilistisch voraussagbar sind, ist eine empirisch offene Frage. Aber auch dann, wenn sich solche Regeln bzw. Regelmäßigkeiten nicht nachweisen lassen, erhält man Einsichten in gesellschaftliche Verhältnisse. Für ein Erkenntnisinteresse, daß sich hauptsächlich auf die Voraussagbarkeit individueller Lebensverläufe richtet, erscheinen solche Einsichten zwar nicht unbedingt wissenswert, wenn es jedoch darum geht, gesellschaftliche Verhältnisse zu charakterisieren, kommt ihnen offensichtlich eine wichtige Bedeutung zu.

e) Schließlich muß noch ein Problem kurz erwähnt werden, das eigentlich nicht soziologischer Natur ist, jedoch die soziologische Diskussion stets begleitet hat. Es verdankt sich der Tatsache (und diese empirische

es keine Regel gibt. Es ist keineswegs ausgeschlossen, daß auch in diesem Fall eine differenziertere Beschreibung der involvierten Ereigniszusammenhänge Hinweise auf die Existenz dementsprechend differenzierterer Regeln liefern kann.

[34] Gelegentlich wird versucht, das Problem dadurch zu lösen, daß man sich auf statistische Signifikanzkriterien beruft. Diese Kriterien betreffen jedoch nur die Frage, mit welcher „Sicherheit“ ein statistischer Zusammenhang angenommen werden kann; über die Frage, ob und ggf. wie der statistisch ermittelte Zusammenhang als eine soziale Regelmäßigkeit *interpretiert* werden kann, geben sie keinen Aufschluß. Bestenfalls erhält man Hinweise auf Unterschiede in der Situation und in den Verhaltensweisen von Individuen, die nicht als „zufällig“ angesehen werden können, vgl. hierzu Carlsson [1951]. In dem oben angeführten Beispiel könnte etwa mit einer hinreichend großen Stichprobe gezeigt werden, daß bei einer signifikant größeren Anzahl nichtehelicher Lebensgemeinschaften eine Heirat stattfindet, wenn ein Kind geboren wird. Hätte man nicht nur eine Stichprobe, sondern vollständige Informationen über alle Lebensgemeinschaften in der Grundgesamtheit, könnte man schließlich mit Bestimmtheit sagen, daß es genau 70 Prozent sind, die in einer solchen Situation heiraten. Aber die Frage, mit welcher Genauigkeit bzw. Sicherheit man auf der Grundlage beschränkter Daten Aussagen über die intendierte Grundgesamtheit machen kann, ist offensichtlich zu unterscheiden von der Frage nach der theoretischen Bedeutung der (mehr oder weniger genau) ermittelten Sachverhalte.

Tatsache bildet natürlich ein zentrales *soziologisches* Problem), daß sich sowohl soziale Regeln als auch soziale Regelmäßigkeiten in der historischen Entwicklung gesellschaftlicher Verhältnisse verändern können. Weder Regeln noch Regelmäßigkeiten können infolgedessen als „Gesetzmäßigkeiten des sozialen Lebens“ angesehen werden. Diese Einsicht ist keineswegs neu. Obwohl deterministische Gesellschaftsauffassungen in den Anfängen der empirischen Sozialforschung – etwa bei Quetelet – zeitweilig dominierten, sind sie bereits sehr früh kritisiert und zurückgewiesen worden. worden.[35] Zum Beispiel betonte Lexis [1903, S. 217]: „Wie die zwingenden Gesetze, so bringen auch Sitte, Gewohnheit, Mode gleichmäßige Massenhandlungen hervor. Aber alle diese Bestimmungsgründe des Handelns sind selbst wieder veränderlich und die auf ihnen beruhenden Massenerscheinungen haben daher keineswegs die Stabilität der sich wiederholenden Komplikationen von Naturerscheinungen, sondern sie zeigen einen *historischen* Charakter, teil mit erkennbarer Entwicklung in bestimmter Richtung, teil mit unregelmäßig veränderlichen Phasen.“[36]

Insofern sind die in den Anfängen der empirischen Sozialforschung ausgetragenen Kontroversen überholt, und man könnte sich mit der Feststellung begnügen, daß weder bei sozialen Regeln noch bei sozialen Regelmäßigkeiten von Gesetzmäßigkeiten gesprochen werden kann, es sei denn metaphorisch, um ein besonderes Ausmaß temporaler Stabilität anzudeuten. Leider ist jedoch die alte Kontroverse durch Soziologen, die sich an der „Analytischen Wissenschaftstheorie“ orientieren, gewissermaßen neu auferstanden. Ihrer Ansicht nach erfordert eine *Erklärung* sozialer Sachverhalte die Verwendung „allgemeiner Gesetze“, d.h. ihrem Anspruch nach zeitlos und ausnahmslos gültige Behauptungen über kausale Zusammenhänge zwischen Phänomenen.[37] Die scheinbar überholte Frage, welchen Charakter die mithilfe statistischer Modelle nachweisbaren sozialen Regelmäßigkeiten haben, ist infolgedessen zu einer *methodologischen* ge-

[35] Vgl. Porter [1987].

[36] Selbst G. von Mayr, der sich noch weitgehend in der Tradition Quetelets befand, hatte zu dessen deterministischer Gesellschaftsauffassung bereits ein gebrochenes Verhältnis. So heißt es zum Beispiel [1877, S. 64]: „Es ist übrigens ein ziemlich müßiger Wortstreit, ob man von statistischen ‘Gesetzen’ oder nur von ‘Gesetzmäßigkeiten’ und ‘Regelmäßigkeiten’ sprechen dürfe. Ist eine Regelmäßigkeit oder Gesetzmäßigkeit constant so groß, daß in einem gegebenen Falle unter gleichen Verhältnissen die Wiederkehr einer gleichen Erscheinung mit höchster Wahrscheinlichkeit erwartet werden muß, so darf man – wenn auch nicht mathematische Gewißheit gegeben ist – von einem statistischen Gesetze reden, während allerdings bei geringeren Graden der Wahrscheinlichkeit ein bescheidenerer Ausdruck am Platze ist.“

[37] Die Standardreferenz ist Hempel [1965], insbesondere die darin enthaltenen, gemeinsam mit Oppenheim verfaßten *Studies in the Logic of Explanation*. Eine systematische Darstellung im Hinblick auf soziologische Erklärungen hat Nowak [1977, Kap. 6] gegeben. Als ein aktuelles Plädoyer für diese Wissenschaftsauffassung in der Soziologie vgl. man zum Beispiel die Darstellung der „Soziologie. Allgemeine Grundlagen“ durch H. Esser [1993, insb. Kap. 4].

worden: ob und ggf. in welcher Weise Aussagen über soziale Regelmäßigkeiten als Hypothesen über „allgemeine Gesetze“ (oder „Quasi-Gesetze“) angesehen werden können.

Ob das durch die „Analytische Wissenschaftstheorie“ empfohlene deduktiv-nomologische Erklärungsschema einen sinnvollen methodologischen Rahmen für die empirische Sozialforschung bildet, ist umstritten.[38] So-

[38] Eine aus meiner Sicht gute Zusammenfassung zahlreicher problematischer Aspekte dieser Auffassung findet sich bei Donagan [1966], eine die seither geführte Diskussion berücksichtigende Einführung hat Bohman [1991] gegeben. An dieser Stelle sei nur kurz darauf hingewiesen, daß bereits unklar und umstritten ist, ob und ggf. in welcher Weise soziologische Erklärungen diesem Erklärungsschema“ folgen *können*. In der wissenschaftstheoretischen Literatur ist häufig festgestellt worden, daß soziale Regelmäßigkeiten, die mithilfe statistischer Modelle festgestellt werden können, nicht ohne weiteres als „allgemeine Gesetze“ angesehen werden können. Dies entspricht unserem üblichen Verständnis: es handelt sich um Beschreibungen gesellschaftlicher Verhältnisse, die sich im historischen Ablauf verändern. Dann stellt sich jedoch die Frage, ob es überhaupt „allgemeine Gesetze“ gibt, die das soziale Leben der Menschen beherrschen und sich deshalb für nomologische Erklärungen eignen. In einem gegenwärtig verbreiteten Lehrbuch über „Methoden der empirischen Sozialforschung“ von Schnell, Hill und E. Esser [1992, S. 48f] heißt es zum Beispiel im Anschluß an eine Darstellung des deduktiv-nomologischen Erklärungsschemas: „Zunächst wird man feststellen, daß in den Sozialwissenschaften keine Gesetze im obigen Sinne bekannt sind und deshalb die ideale Form der Erklärung zur Zeit kaum möglich ist.“ Anstatt jedoch die Frage zu behandeln, warum es sich so verhält und welche Schlußfolgerungen daraus vielleicht gezogen werden sollten, wird folgender Ratschlag gegeben: „Gerade aus dem Mangel an sozialwissenschaftlichen Gesetzen ergibt sich die zentrale Aufgabe der empirischen Sozialforschung als die *Suche* nach Gesetzen zur Erklärung sozialen Handelns.“ Es stellt sich natürlich die Frage, wie diese „Suche nach Gesetzen zur Erklärung sozialen Handelns“ gestaltet werden soll. Die übliche Antwort besteht darin, daß man versuchen soll, *Hypothesen* über diese Gesetze zu bilden, die dann mithilfe empirischer Daten überprüft werden können. Man kann sich jedoch leicht klarmachen, daß dies für die empirische Sozialforschung kein besonders sinnvoller Vorschlag ist. Denn in diesem Fall sprechen alle bisher gesammelten empirischen Einsichten dafür, daß die sozialen Regelmäßigkeiten, die mithilfe statistischer Modelle konstruiert werden können, keine allgemeinen Gesetze sind, daß es sich vielmehr um Merkmale historisch wandelbarer gesellschaftlicher Verhältnisse handelt. Das deduktiv-nomologische Erklärungsschema degeneriert dann zu einer unsinnigen rhetorischen Form; ihm zu folgen, läuft dann darauf hinaus, die gesellschaftlichen Verhältnisse so darzustellen, *als ob* es sich um Gesetzmäßigkeiten handelt, die das Leben der Menschen beherrschen. Die von H. Esser [1993] angeführten Beispiele für Erklärungen, die dem deduktiv-nomologischen Erklärungsschema folgen, können als Illustration dienen: „Beispielsweise sei zu erklären, warum es nach Wegfall der Grenzsperren in der ehemaligen DDR eine sehr stabile Rate an Wanderungen aus der ehemaligen DDR, aber kaum von der CSSR in die Bundesrepublik gab. Ein Gesetz zur Erklärung von Wanderungen könnte etwa behaupten, daß Wanderungen eine Funktion von Unterschieden im Bruttosozialprodukt zwischen Regionen sei.“ (S. 44) In einem anderen Beispiel beschäftigt sich Esser mit dem Anstieg der Scheidungsraten. Sein Vorschlag ist, zunächst von einem „Mikro-Modell“ auszugehen, dessen „nomologischer Kern“ in der Aussage besteht: „Je höher das Konfliktpotential in einer Ehe und je eher und leichter Alternativen zu der ehelichen Beziehung verfügbar sind, umso höher ist die Wahrscheinlichkeit für eine Ehescheidung.“ Dann wird auf den Prozeß der Verstädterung verwiesen und hinzugefügt: „Unter städtischen Verhältnissen gibt es mehr Möglichkeiten, seine Interessen außerhalb einer ehelichen Beziehung zu verfolgen. Und dieses führt zum Anstieg der ehelichen Streitigkeiten. Gleichzeitig bieten städtische Verhältnisse einen leichteren Zugang zu Alternativen [...].“ Aus beidem folgt dann – wie es das HO-Schema verlangt – der zu erklärende Sachverhalt. Ich will natürlich nicht bestreiten, daß es in beiden Überlegungen Aspekte gibt, die bei der Erklärung von Wanderungen bzw. Ehescheidungen berücksichtigt werden sollten (wobei es vielleicht sinnvoll wäre, anstelle von „Bruttosozialprodukt“ von subjektiv wahrgenommen Verdienstmöglichkeiten zu sprechen). Aber in beiden Fällen handelt es sich sicherlich nicht um allgemeine Gesetze des sozialen Lebens.

weit diese Frage nur die Form bzw. Rhetorik soziologischer Erklärungen betrifft, kann davon hier abgesehen werden. Ich möchte es auch offen lassen, ob und ggf. in welcher Weise die empirische Sozialforschung an der Aufgabe orientiert werden kann, nach „Gesetzmäßigkeiten“ zu suchen, die das soziale Leben der Menschen beherrschen. Eine vorgängig konzipierbare Aufgabe kann jedenfalls darin gesehen werden, die sozialen Regeln bzw. Regelmäßigkeiten zu beschreiben, denen die Menschen in einer Gesellschaft folgen. Dabei gibt es zumindest zwei wesentliche Besonderheiten. Erstens ist davon auszugehen, daß soziale Regeln bzw. Regelmäßigkeiten einem historischen Wandel unterliegen; woraus als eine wichtige Aufgabe für die empirische Sozialforschung folgt, diesen sozialen Wandel zu beschreiben. Zweitens ist zu berücksichtigen, daß sich soziale Regeln bzw. Regelmäßigkeiten auf Subjekte von Lebensverläufen beziehen. Dadurch entsteht eine Fragestellung, die typischerweise ausgeblendet wird, wenn sich das Erkenntnisinteresse unmittelbar auf „Gesetzmäßigkeiten“ richtet: in welcher Weise das soziale Verhalten von Individuen durch Verweise auf soziale Regeln bzw. Regelmäßigkeiten erklärt werden kann.

1.4 Formen der Bezugnahme auf Individuen

Es ist eine gängige Formulierung, daß die empirische Sozialforschung danach strebt, Verhaltensweisen und Lebensverläufe „der Individuen“ in einer Gesellschaft zu erklären. Aber es ist nicht ohne weiteres klar, in welcher Bedeutung diese Zielsetzung verstanden werden kann. Ein Aspekt des Problems kann, wie ich glaube, durch die Frage erfaßt werden, in welcher Weise individuelles Verhalten durch soziale Regeln bzw. Regelmäßigkeiten erklärt werden kann.

a) In gewisser Weise erscheint es unproblematisch, zu sagen: daß eine Erklärung sozialen Verhaltens dadurch erreicht werden kann, daß man die sozialen Regeln bzw. Regelmäßigkeiten in Erfahrung bringt, denen die Menschen in einer Gesellschaft folgen. Man kann zum Beispiel daran denken, in ein fremdes Land zu geraten, über dessen gesellschaftlichen Verhältnisse man nur wenig weiß. Wenn man Glück hat, findet man eine Person, die einem diese Verhältnisse *erklären* kann, d.h. die Regeln expliziert, an denen die Menschen in diesem Land ihr soziales Verhalten orientieren. In dem Maße, wie die Erklärung gelingt, kann man sich dann als Akteur in den gesellschaftlichen Verhältnissen bewegen, d.h. soziale Interaktion auf Erwartungen gründen und den Regeln folgen oder von ihnen abweichen.

Man könnte sagen, daß durch soziale Regeln bzw. Regelmäßigkeiten erklärt werden kann, *wie* sich die Menschen in einer Gesellschaft verhalten, *wie* sie ihre Lebensverläufe entwickeln. Solange sich die Erklärungsintention auf solche Wie-Fragen bezieht, erscheint sie unproblematisch; schwierig wird es erst, wenn man Warum-Fragen stellt.

b) In welcher Weise beziehen sich Erklärungen durch Regeln bzw. Regelmäßigkeiten auf Individuen? Eine wesentliche Besonderheit wird sichtbar, wenn man an Gesetzmäßigkeiten denkt, zum Beispiel an die Aussage, daß alle Menschen sterblich sind. Charakteristisch für solche Gesetzmäßigkeiten ist, daß sie auf eine spezifische Weise *für alle* Individuen (individuellen Sachverhalte) gelten, auf die sie sich beziehen. Aus dem Satz, daß alle Menschen sterblich sind, kann eine sinnvolle Aussage für jeden einzelnen Menschen abgeleitet werden. Dies gilt jedoch nicht für soziale Regeln bzw. Regemäßigkeiten. Sie lassen *in jedem einzelnen Fall* Ausnahmen zu. Sie sollten auch deshalb von Gesetzmäßigkeiten unterschieden werden.

c) Da jedes einzelne Individuum von einer sozialen Regel bzw. Regelmäßigkeit abweichen kann, liefertn Hinweise auf Regeln im allgemeinen keine befriedigenden Antworten auf Warum-Fragen. Angenommen, man könnte beobachten, daß die an einer nichtehelichen Lebensgemeinschaft beteiligten Partner in der Regel dann heiraten, wenn sich die Geburt eines Kindes ankündigt. Könnte man dann das Ereignis Heirat mit dieser Regel und der Voraussetzung, daß sich die Geburt eines Kindes ankündigt, erklären? Sicherlich nicht in der Weise, daß man sagt, daß die Schwangerschaft bzw. Geburt eine Ursache für die Heirat ist. Denn die an einer Lebensgemeinschaft beteiligten Partner sind auch in einer solchen Situation nicht gezwungen zu heiraten. Die beobachtbare Tatsache, daß in vielen nichtehelichen Lebensgemeinschaften dann geheiratet wird, wenn die Frau schwanger bzw. ein Kind geboren wird, liefert in jedem einzelnen Fall *keine* Erklärung im Sinne einer Antwort auf die Warum-Frage. Oder anders gesagt: der Hinweis auf die soziale Regel ist nicht das, was wir wissen möchten, nämlich eine Einsicht in die Gründe, aufgrund derer bei der Geburt eines Kindes geheiratet – oder nicht geheiratet – wird.[39]

d) Insoweit individuelle Lebensverläufe auch ein Ergebnis individuell getroffener Entscheidungen sind, benötigt man zu ihrer Erklärung (im Hinblick auf die Warum-Frage) Einsichten in die Motive und Überlegungen, die den Entscheidungen zugrunde liegen. Aus dieser Feststellung können allerdings unterschiedliche Konsequenzen gezogen werden.

Erstens könnte man die Ansicht vertreten, daß Fragen nach Handlungsmotiven nur im Hinblick auf konkrete, jeweils bestimmte Individuen sinnvoll gestellt werden können. Dies schließt nicht aus, daß möglicherweise generalisierbare Einsichten gewonnen werden können, aber darüber kann

[39] Dies schließt natürlich nicht aus, daß es Gründe gegeben kann, in denen die Existenz einer Regel als Argument auftritt; zum Beispiel wenn die Befolgung einer Regel damit begründet wird, daß ihre Nichtbefolgung negative Folgen hätte.

vorab nichts ausgesagt werden, und insofern kann man sagen, daß diese Ansicht nicht darauf abzielt, generalisierbare Erklärungen individuellen Verhaltens zu gewinnen.

Zweitens kann man versuchen, hermeneutische Verfahren so zu konzipieren, daß von einer Bezugnahme auf jeweils konkrete und bestimmte Individuen abstrahiert werden kann. Exemplarisch kann man an Max Webers Konzeption einer „verstehenden Soziologie“ denken, aber auch an Rational-Choice-Theorien, wenn bzw. soweit mit ihnen versucht wird, Einsichten in Gründe oder Motive sozialen Verhaltens zu geben.[40]

[40] Bei soziologischen Theorieansätzen, die von der Annahme ausgehen, daß Individuen rational handelnde soziale Akteure sind, erscheint es sinnvoll, zwei Verwendungsweisen dieser Annahme zu unterscheiden. Einerseits die Möglichkeit, gestützt auf diese Annahme Rational-Choice-Modelle zu konstruieren, um Einsichten in die Struktur sozialer Interaktionsprozesse zu gewinnen. Modelle dieser Art liefern zwar keine nomologischen Erklärungen sozialer Sachverhalte, sie können aber in der Art von Denkmodellen in vielen Fällen unser Verständnis sozialer Interaktionsprozesse fördern. Davon zu unterscheiden sind jedoch Versuche, ausgehend von Rationalitätsannahmen das tatsächlich in einer Gesellschaft beobachtbare Verhalten der Individuen zu erklären. Man kann dies als eine Variante bzw. als einen Spezialfall hermeneutischer Handlungserklärungen ansehen: die jeweils unterstellten Rationalitätsannahmen dienen als Schema, um die beobachtbaren Verhaltensweisen zu deuten. (Ich spreche hier von Rationalitätsannahmen im Plural, weil die Übernahme ökonomischer Rationalitätskonzepte in die soziologische Theoriebildung bereits zu einer Vielzahl unterschiedlicher soziologischer Varianten geführt hat; vgl. zum Beispiel die Ausführungen bei Esser [1993, Kap. 14].) So betrachtet, bilden rationalistische Handlungserklärungen keine besonderen Probleme, es handelt sich, wie in allen anderen Fällen, um hermeneutische Deutungen, deren Unterstellungen mehr oder weniger angemessen sein können. Fragwürdig werden rationalistische Handlungserklärungen jedoch dann, wenn sie als *nomologische* Erklärungen dargestellt werden, d.h. wenn Rationalitätsannahmen zur Deutung sozialen Verhaltens als Hypothesen über Gesetzmäßigkeiten formuliert werden. Zum Beispiel vertritt Esser [1987, S. 93f] die Ansicht, „daß eine nomologische Erklärung von Handlungen und die rationale Deutung nicht nur keine Gegensätze sind, sondern wahrscheinlich sogar ein einheitliches logisches Grundschema beinhalten.“ (Die wesentlichen Grundgedanken dieser Position finden sich bereits bei Hempel [1965], vgl. insb. den Beitrag „The Function of General Laws in History“, S. 231ff.) An anderer Stelle betont Esser [1989, S. 68], „daß der erforderliche (universale) *nomologische Kern* zur Erklärung sozialer Phänomene aller Art in einer *Theorie des Handelns* menschlicher Akteure zu suchen ist. Diese geht davon aus, daß menschliche Akteure externe (‘objektive’) Situationsmerkmale wahrnehmen, nach Maßgabe ihrer Erfahrungen (‘subjektiv’) bewerten und aufgrund der Einschätzung bestimmter Konsequenzen die Handlung auswählen, deren erwarteter Wert (insgesamt) höher ist als der erwartete Wert jeder anderen Alternative.“

Zwar soll nicht bestritten werden, daß es sich hierbei um ein in vielen Fällen angemessenes Schema für die Deutung sozialen Verhaltens handelt. Sicher ist jedoch auch, daß es in vielen Fällen nicht angemessen ist; vgl. zum Beispiel die Diskussion bei Elster [1989]. Ganz im Gegensatz zu Esser versucht Elster die These zu begründen, daß das Befolgen sozialer Normen grundsätzlich nicht hinreichend durch Rationalitätsannahmen erklärt werden könne. Andere Autoren, zum Beispiel Rowe [1989], haben dagegen versucht, auch für das Befolgen sozialer Normen rationale Gründe zu finden. Die Diskussion zeigt m.E. vor allem eines: daß es sich bei Rationalitätsannahmen nicht um Gesetzmäßigkeiten sozialen Verhaltens handelt, sondern um normative oder hermeneutische Prinzipien. Ich möchte betonen, daß sich diese Kritik nicht gegen Rational-Choice-Modelle richtet,

Drittens kann man nach einer Konzeption empirischer Sozialforschung suchen, bei der von der Frage, *warum* – aus welchen Motiven und Gründen – die Individuen ein jeweils spezifisches Verhalten entwickeln, abstrahiert werden kann. Sie beschränkt sich dann auf Wie-Fragen und versucht nur, durch den empirischen Nachweis sozialer Regeln bzw. Regelmäßigkeiten zu erklären, *wie* die Menschen ihre Lebensverläufe entwickeln.

Diese drei unterschiedlichen Auffassungen schließen sich wechselseitig nicht aus, sondern können miteinander kombiniert werden. Viele Beiträge zur Lebensverlaufsforschung können so verstanden werden, daß durch sie versucht wird, die zweite und dritte Erklärungsintention miteinander zu verbinden. Die Unterscheidung erscheint mir gleichwohl sinnvoll, um die Bedeutung statistischer Verfahren und Modelle in der Lebensverlaufsforschung zu verstehen. Wie ich im weiteren Verlauf dieser Arbeit zu zeigen versuche, kann man sie so verstehen, daß mit ihrer Hilfe Antworten auf Wie-Fragen gegeben werden können.

e) Es sei betont, daß sich die hier verwendete Unterscheidung von Wie- und Warum-Fragen auf Lebensverläufe bezieht, wobei davon ausgegangen wird, daß Lebensverläufe durch Subjekte entwickelt werden, d.h. auch von ihren Motiven, Wünschen, Präferenzen und Entscheidungen abhängen. Die Warum-Frage im hier gemeinten Sinn bezieht sich auf diese Abhängigkeit der Lebensverläufe von ihren Subjekten. Damit ist nicht gemeint, daß nur im Rahmen von Warum-Fragen von Bedingungen von Lebensverläufen gesprochen werden kann. Man kann, wie in Kapitel 4 zu zeigen versucht wird, Lebensverläufe als Sequenzen von sich bedingenden Ereignissen beschreiben und dadurch Einsichten in ihren Ablauf gewinnen. Die durch diese Betrachtungsweise möglichen Aussagen sind jedoch wiederum Aussagen über soziale Regeln bzw. Regelmäßigkeiten und sollten infolgedessen im Kontext von Wie-Fragen interpretiert werden.[41]

sondern gegen die Umdeutung ihrer Rationalitätsannahmen in nomologische Aussagen. Einer der interessantesten Aspekte der Rational-Choice-Modelle liegt aus meiner Sicht gerade darin, daß sie wissenswerte Einsichten in Interaktionsprozesse und deren Folgen liefern können, *ohne* Behauptungen darüber formulieren zu müssen, warum Menschen so handeln (oder sogar: so handeln müssen), wie für die Modellbildung angenommen wird. Dies hat gelegentlich zu dem Vorwurf geführt, daß diese Art der Modellbildung empirisch gehaltlos sei. Aber bei diesem Vorwurf werden zwei Dinge verwechselt bzw. fälschlicherweise gleichgesetzt. Einerseits der empirische Gehalt eines Modells, andererseits der Anspruch, beobachtbares Verhalten (kausal) erklären zu können. Wenn es möglich ist, soziale Regeln für das Verhalten von Individuen in einer Gesellschaft empirisch zu ermitteln, können sie zum Ausgangspunkt einer empirisch gehaltvollen Modellbildung verwendet zu werden. Es ist *dafür* weder erforderlich, die Existenz der für die Modellbildung verwendeten Regeln zu erklären, noch muß erklärt werden, warum sich die Individuen in einer Gesellschaft diesen Regeln entsprechend verhalten. Diese Betrachtungsweise könnte darüberhinaus einen Ausgangspunkt bilden, um Rational-Choice-Modelle mit empirischer Sozialforschung zu verknüpfen; vgl. als einen interessanten Diskussionsbeitrag zu dieser Frage Boudon [1979].

[41] Diese Betrachtungsweise führt dazu, daß insoweit Beschreibungen und Erklärungen nicht systematisch unterschieden werden können. Man könnte vielleicht sagen, daß aus einer Beschreibung eine Erklärung wird, sobald sie Bedingungen für einen Sachverhalt sichtbar macht. Da soziale Regeln stets an spezifische Situationen gebunden sind, die man als Bedingungen des durch sie geregelten Verhaltens ansehen kann, liefert insofern jede Beschreibung einer sozialen Regel bereits eine Erklärung.

f) Ein wesentliches Merkmal von Erklärungen, die im Rahmen der Wie-Frage durch den Nachweis sozialer Regeln bzw. Regelmäßigkeiten erbracht werden können, liegt darin, daß durch sie von der Kontingenz individueller Lebensverläufe abstrahiert werden kann. Zu sagen, daß eine Regel gilt, impliziert nicht, daß alle Individuen in der Situation, auf die sich die Regel bezieht, ihr folgen. Die empirische Aussage besteht nur darin, daß sich das Verhalten einer „hinreichend großen“ Anzahl von Individuen durch die Regel beschreiben läßt. Durch soziale Regeln bzw. Regelmäßigkeiten kann insofern das Verhalten „der Individuen“ erklärt werden, nicht jedoch das Verhalten jedes einzelnen Individuums.

Um diesen eigentümlichen Charakter von Erklärungen mithilfe von Regeln bzw. Regelmäßigkeiten zum Ausdruck zu bringen, könnte man sagen: sie erklären nicht das Verhalten der Individuen (als jeweils konkrete, bestimmte Individuen), sondern sie beschreiben gesellschaftliche Verhältnisse. Diese Formulierung liefert jedoch nur eine negative Abgrenzung, indem sie zeigt, welche Art von Erklärungen nicht erreicht werden können. Es bleibt die Frage, in welcher Weise gleichwohl davon gesprochen werden kann, daß soziale Regeln bzw. Regelmäßigkeiten (gesellschaftliche Verhältnisse) das Verhalten der Menschen in einer Gesellschaft erklären.

Diese Frage ist kompliziert, und es gibt, soweit ich sehen kann, keine befriedigende allgemeine Antwort. Zunächst können zwei extreme Betrachtungsweisen unterschieden werden. Einerseits die Vorstellung, daß soziale Regeln ein ihnen konformes Verhalten der Individuen erzwingen. Diese Ansicht ist jedoch offensichtlich problematisch, denn nicht nur kann jedes einzelne Individuum von den Regeln abweichen, sondern dies geschieht auch fortwährend. Die Abweichung von sozialen Regeln ist gewissermaßen das Medium, in dem sie sich verändern.[42] Eine andere, gleichermaßen extreme Ansicht liefert die Vorstellung, daß soziale Regeln keinerlei Einfluß auf das Verhalten der Individuen haben, daß sie gewissermaßen nur einen aus der Aggregation entstehenden ideologischen Reflex des jeweils individuellen Vollzugs kontingenter Lebensverläufe darstellen. Gegen diese Ansicht kann eingewendet werden, daß soziale Regeln nicht nur in der Einbildung von Soziologen existieren, sondern daß sich – vielleicht nicht immer, aber in vielen Fällen – die sozialen Akteure selbst an solchen Regeln orientieren, so daß insoweit davon gesprochen werden kann, daß soziale Regeln real exi-

[42] Damit soll nicht behauptet werden, daß sich alle Formen sozialen Wandels durch eine allmähliche Transformation sozialer Regeln – in der Literatur gelegentlich, zum Beispiel von Carlsson und Karlsson [1970], „discretionary change“ genannt – angemessen beschreiben lassen. Immerhin erfaßt die Formulierung eine grundlegende Form sozialen Wandels, die bereits im Begriff sozialer Regeln (genauer: in der Tatsache, daß sie stets interpretiert werden müssen) impliziert ist.

stierende Bedingungen sozialen Verhaltens sind. Eine „mittlere Position“ zwischen diesen beiden extremen Ansichten kann vermutlich nur gefunden werden, wenn man die Fragestellung im Hinblick auf unterschiedliche Typen sozialer Regeln differenziert.[43] Denn es hängt von der Art der sozialen Regeln ab, ob und ggf. wie man davon sprechen kann, daß sich die individuellen Akteure an ihnen orientieren.

g) Das Problem wird noch etwas komplizierter dadurch, daß die in der empirischen Lebensverlaufsforschung verwendeten statistischen Verfahren nicht unmittelbar Einsichten in soziale Regeln vermitteln, sondern nur statistische Regelmäßigkeiten aufzeigen. Zwei komplementäre Strategien, um mit dieser Differenz umzugehen, können unterschieden werden.

Einerseits kann man versuchen, die empirisch ermittelbaren statistischen Regelmäßigkeiten im Hinblick auf soziale Regeln theoretisch zu deuten. Findet man zum Beispiel, daß in nichtehelichen Lebensgemeinschaften in der Regel dann geheiratet wird, wenn sich die Geburt eines Kindes ankündigt, kann man diese statistische Regelmäßigkeit als Hinweis auf die Existenz einer sozialen Regel deuten, die sich – in den jeweils betrachteten gesellschaftlichen Verhältnissen – darauf bezieht, in welcher Form von Lebensgemeinschaften Kinder aufgezogen werden sollten. Diese Strategie verknüpft die zweite und dritte der in (d) unterschiedenen Konzeptionen, indem sie die statistische Ermittlung von Regelmäßigkeiten mit einer hermeneutisch orientierten theoretischen Deutung zu verbinden versucht, d.h. mit Aussagen über (in diesem Fall generalisierbare) Motive der sozialen Akteure.

Andererseits kann man versuchen, statistisch ermittelbare Regelmäßigkeiten vollständig im Kontext von Wie-Fragen, d.h. ohne Rückgriff auf hermeneutische Deutungen zu interpretieren. Zwei zentrale Begriffe – „Erwartung“ und „Chance“ – können für diese Strategie verwendet werden. Beide Begriffe werden sowohl in der Umgangssprache als auch in wissenschaftlichen Texten vielfältig verwendet, um Aussagen über Sachverhalte zu formulieren, deren Beschaffenheit nicht vollständig bekannt ist, oder über Ereignisse, die sich nicht sicher voraussagen lassen. In dieser Arbeit gehe ich davon aus, daß sich der Begriff „Erwartung“ auf Subjekte bezieht, die sich Erwartungen bilden, und daß demgegenüber mit dem Begriff „Chance“ eine Beschreibung von Eigenschaften der Situation gemeint ist, in der die nicht sicher voraussagbaren Ereignisse stattfinden.[44]

Statistische Regelmäßigkeiten im Kontext der Lebensverlaufsforschung beziehen sich typischerweise auf Situationen, in denen gewisse Ereignisse

[43] Vgl. die Diskussion dieses Problems durch Granovetter [1985].

[44] Statt von „Chancen“ wird häufig auch von „Risiken“ gesprochen. Abgesehen von unterschiedlichen subjektiven Bewertungen der Ereignisse, auf die jeweils Bezug genommen wird, scheint es jedoch keinen systematischen Unterschied zu geben; vgl. zur soziologischen Diskussion dieser Begriffe Bonß [1991]. In der vorliegenden Arbeit werden deshalb beide Begriffe als austauschbar angesehen.

(die zur Beschreibung von Lebensverläufen bedeutsam erscheinen) stattfinden können. Wir sprechen von statistischen Regelmäßigkeiten, weil wir uns *im Einzelfall* nicht sicher sein können, welche der möglichen Ereignisse tatsächlich realisiert werden. Betrachten wir noch einmal das Beispiel nichtehelicher Lebensgemeinschaften, in denen die Frau schwanger wird, und nehmen wir an, daß als statistische Regelmäßigkeit ermittelt werden kann, daß 70 % der Paare in einer solchen Situation heiraten. Man kann dann im Hinblick auf jede bestimmte Lebensgemeinschaft, die sich in einer solchen Situation befindet, eine Erwartung darüber bilden, ob die beteiligten Personen vermutlich heiraten werden. Die Kenntnis statistischer Regelmäßigkeiten ermöglicht es darüber hinaus, solche Erwartungen zu quantifizieren; zum Beispiel: in der hier exemplarisch genannten Situation kann mit einer Wahrscheinlichkeit von 70 % erwartet werden, daß eine Heirat stattfindet.

Allerdings sind solche Quantifizierungen problematisch, da zweckmäßigerweise bei der Bildung von Erwartungen über singuläre Ereignisse (d.h. hier, wenn man sich auf jeweils konkrete, bestimmte Individuen bezieht) nicht nur auf statistische Regelmäßigkeiten Bezug genommen werden sollte, sondern auf jeweils alle Informationen, die über den Gegenstand der Erwartungsbildung zur Verfügung stehen. Insofern ist Erwartungsbildung ein „subjektiver“ Vorgang, da abhängig von dem jeweils individuell verfügbaren Wissen.[45] Demgegenüber kann mit dem Begriff der Chance eine „objektivierbare“ Aussage getroffen werden. In der genannten Situation besteht eine Chance von 70 %, daß eine Heirat stattfindet. Der „objektive“ Charakter einer solchen Aussage wird dadurch erreicht, daß auf eine statistische Regelmäßigkeit (oder eine Vielzahl solcher Regelmäßigkeiten) Bezug genommen wird *und zugleich* von allen im Einzelfall (zum Beispiel für die

[45] In der Literatur wird dies als Gegensatz von „subjektiven“ und „objektiven“ Wahrscheinlichkeitsaussagen diskutiert. Darauf wird in Abschnitt 2.3 näher eingangen. Bereits an dieser Stelle kann jedoch ein einfaches, von Benenson [1984, S. 10] angegebenes Beispiel dazu dienen, den Unterschied deutlich zu machen. Das Zufallsexperiment besteht darin, ein Kartenspiel (52 Karten) zu mischen und verdeckt hinzulegen. Die Frage, die durch eine Wahrscheinlichkeitsaussage beantwortet werden soll, betrifft die Farbe der obersten, zuerst aufgedeckten Karte. In dieser Situation kommt man zunächst bei beiden Interpretationen des Wahrscheinlichkeitsbegriffs zu dem Ergebnis, daß die Wahrscheinlichkeit, daß an der obersten Stelle eine rote Karte liegt, 0.5 beträgt. Wird jedoch als zusätzliche Information angegeben, welche Farbe die zu unterst liegende Karte aufweist, gelangt man zu unterschiedlichen Wahrscheinlichkeitsaussagen. Da sich an der relativen Häufigkeit, mit der sich an der obersten Stelle eine rote Karte befindet, durch diese zusätzliche Information nichts ändert, beträgt die „objektive“ Wahrscheinlichkeit dieses Ereignisses nach wie vor 0.5; dagegen ist es offensichtlich ratsam, die subjektive Erwartung, die sich auf eine individuelle Realisierung des Zufallsexperiments bezieht, von der zusätzlich gegebenen Information abhängig zu machen und, je nachdem welche Farbe die unterste Karte aufweist, von der subjektiven Wahrscheinlichkeit 25/51 oder 26/51 auszugehen. Akzeptiert man, daß in beiden Fällen eine unterschiedliche Bedeutung von Wahrscheinlichkeitsaussagen gemeint ist, liegt darin kein Widerspruch. Es bleibt natürlich die Frage, welcher Typ von Wahrscheinlichkeitsaussagen dem jeweils vorliegenden theoretischen oder praktischen Kontext angemessen ist.

Lebensgemeinschaft im Nachbarhaus) verfügbaren zusätzlichen Informationen abstrahiert wird. Der Unterschied kann, wie ich glaube, folgendermaßen zum Ausdruck gebracht werden: der Erwartungsbegriff dient dazu, um Aussagen über kontingente singuläre Ereignisse zu formulieren; der Begriff der Chance dient demgegenüber einer Beschreibung der Situation, in der eine Vielzahl singulärer kontingenter Ereignisse stattfinden kann. Es handelt sich gewissermaßen um komplementäre sprachliche Formen, um Aussagen über kontingente Ereignisse machen zu können.

h) Wichtig erscheint, daß mit dem Begriff der Chance eine Beschreibung gesellschaftlicher Verhältnisse gegeben werden kann, bei der hermeneutische Deutungen der sozialen Regeln bzw. Regelmäßigkeiten, auf die jeweils Bezug genommen wird, nicht unbedingt erforderlich sind. Die Verwendung dieses Begriffs liegt deshalb insbesondere dann nahe, wenn es sich um soziale Regelmäßigkeiten handelt, die nicht unmittelbar als Ausdruck oder Folge sozialer Regeln (die hermeneutisch deutbar sind) interpretiert werden können. Dies gilt zum Beispiel für einige elementare demographische Regelmäßigkeiten, etwa für die menschliche Lebensdauer. Darin kann vermutlich ein Grund dafür gesehen werden, daß der Begriff der Chance, der zunächst im „subjektiven" Kontext von Entscheidungen unter Unsicherheit gebildet worden ist, in den Anfängen der empirischen Sozialforschung vor allem zur theoretischen Reflexion demographischer Phänomene verwendet worden ist.[46] Exemplarisch sei auf Lexis hingewiesen, der bereits 1875 den Chancenbegriff explizit für eine soziologische (wenn auch zunächst noch weitgehend auf demographische Sachverhalte beschränkte) Theoriebildung verwendet hat:

> Die unbekannte Verbindung zwischen dem Anfange und dem Endzustande der Elementarmasse, d.h. in unserem Beispiele zwischen der Zahl der im Ganzen beobachteten und der gestorbenen Individuen einer Elementargruppe in einer bestimmten Altersclasse denken wir uns also in der Form eines Chancensystems. [...] Demnach fasst sich die Aufgabe der socialphysiologischen Statistik in Folgendem zusammen: sie hat möglichst individualisirte Elementarmassen zu bilden und durch Wahrscheinlichkeitsverhältnisse die Chancensysteme zu charakterisiren, welche die bedeutsamen Veränderungen derselben bedingen; sie hat ferner zu untersuchen, wiefern durch die Verschiedenheit der Unterscheidungsmerkmale der Elementarmassen Verschiedenheiten ihrer Chancensysteme entstehen, und endlich festzustellen, ob die einzelnen Chancensysteme im Laufe der Zeit annähernd constant bleiben oder sich in einer bestimmbaren Weise ändern.[47]

In moderner Terminologie ausgedrückt, bezieht sich Lexis in diesen

[46] Ausführliche Darstellungen des historischen Prozesses, in dem sich probabilistische Anschauungen verbreitet und weitgehend durchgesetzt haben, finden sich u.a. bei Porter [1986] und Hacking [1990].

[47] Lexis [1875, S. 121].

Ausführungen auf das statistische Konzept einer Survivorfunktion, die beschreibt, wie in einer Gesamtheit von Individuen der Übergang von einem Ausgangszustand in einen Folgezustand als ein zeitlicher Prozeß abläuft. Jedes der beteiligten Individuen vollzieht den Übergang „auf seine Weise" und unter spezifischen Bedingungen. Einige dieser Bedingungen können vielleicht erfaßt werden, aber die Kontingenz der jeweils individuellen Ereignisfolgen bildet eine praktisch nicht erreichbare Grenze für die Formulierbarkeit statistischer Gesetzmäßigkeiten.[48]

i) Der Begriff der Chance erscheint aus mehreren Gründen geeignet, um soziologische Lebensverlaufsforschung mit einer Konzeption von Sozialstrukturanalyse zu verknüpfen. Erstens, wie erwähnt, weil er es erlaubt, von hermeneutischen Deutungen abzusehen, ohne sie auszuschließen.[49] Zweitens, weil mit seiner Hilfe Beschreibungen gesellschaftlicher Verhältnisse gegeben werden können, die zur Rationalisierung probabilistischer Aussagen über *individuelle* Lebensverläufe verwendet werden können. Dadurch erhält die soziologische Wissensbildung einen gewissermaßen reflexiven Bezug zu den sozialen Akteuren, auf deren Verhaltensweisen sie sich bezieht. Die sozialen Akteure können das soziologisch gewonnene Wissen zur Reflexion ihrer sozial bedingten Handlungsmöglichkeiten – ihrer „Chancen" und „Risiken" – verwenden.[50] Schließlich erlaubt der Chancenbegriff eine Darstellung gesellschaftlicher Verhältnisse, durch die beide der in (f) dargestellten, jeweils einseitigen Betrachtungsweisen vermieden werden können. Obwohl statistische Regelmäßigkeiten zunächst aus einer Beobachtung individueller Lebensverläufe gewonnen werden und insofern nur eine Beschreibung ihrer kontingenten Entwicklung liefern, kann man sie als Beschreibungen der gesellschaftlichen Verhältnisse deuten, in denen sich die individuellen Lebensverläufe vollziehen. Bedient man sich für den damit

[48] Lexis hat bereits ausdrücklich darauf hingewiesen, daß es sich nicht in erster Linie um ein „Datenproblem" handelt, sondern um ein essentielles Merkmal der statistischen Methode: daß sie sich durch die Art ihrer Begriffsbildung auf Gesamtheiten von Individuen beziehen muß. „Die Gruppen, welche die Statistik schliesslich als homogene Collectivwesen behandelt, müssen jedoch noch immer den Charakter von Massen besitzen, wie es die statistische Untersuchung ihrem Wesen nach verlangt. Diese Elementarmassen sind nach den bestimmbaren allgemeinen Einflüssen unterschieden und isolirt, auf die concreten Ursachen aber, welche innerhalb einer jeden die Sterbefälle im Einzelnen bedingen, kann die Statistik nicht zurückgehen." (Lexis [1875, S. 120])

[49] Man denke an die prominente Rolle, die der Chancenbegriff in Max Webers „verstehender Soziologie" spielt.

[50] Erreicht wird dies durch die komplementäre Bedeutung der Begriffe „Chance" und „Erwartung". Die soziologische Wissensbildung abstrahiert zwar von jeweils bestimmten, konkreten Individuen; in einer Formulierung von Lindsey [1973, S. 1]: „The sociologist is not concerned with determining how an individual will react in a given set of circumstances. Instead, more general relationships are developed to show that, under given circumstances, a certain proportion of individuals will do this, another proportion something else and so on, until all of the common responses or attributes have been included." Gleichwohl kann dieser abstrakte Bezug auf Individuen zur Erwartungsbildung für jeweils individuell bestimmte Situationen genutzt werden.

vollzogenen Schritt zur Theoriebildung des Chancenbegriffs, kann reflektiert werden, daß die gesellschaftlichen Verhältnisse – als Chancen – weder die individuellen Lebensverläufe determinieren, noch unabhängig von ihnen existieren, sondern zum Ausdruck bringen, welche Handlungsmöglichkeiten sich die sozialen Akteure in der Entwicklung ihrer Lebensverläufe erschließen.

j) Die logische Struktur des Gedankengangs, der zu dieser Interpretation führt, ist formal ähnlich zur sog. Propensity-Interpretation des Wahrscheinlichkeitsbegriffs.[51] Der Grundgedanke besteht darin, Wahrscheinlichkeitsaussagen nicht auf individuelle Ereignisse zu beziehen, sondern sie als eine Charakterisierung der Bedingungen anzusehen, unter denen die zufälligen Ereignisse stattfinden. Die Aussage zum Beispiel, daß die Wahrscheinlichkeit, mit einem gewissen Würfel eine gerade Augenzahl zu erzielen, 0.5 beträgt, ist dann eine Aussage über die Eigenschaften eines bestimmten Zufallsexperiments; sie charakterisiert das Zufallsexperiment, nicht jedoch seine zufälligen Realisierungen.

Verwendet man diesen Gedankengang zur theoretischen Deutung probabilistischer Aussagen über das soziale Verhalten der Individuen, folgt daraus, daß es sich nicht um Aussagen über individuelles Verhalten, *sondern* um Aussagen über Eigenschaften ihrer gesellschaftlichen Verhältnisse handelt. Der Gegensatz ist allerdings, wie stets bei Aussagen über „die Individuen“, zweideutig. Gemeint ist, daß mit einer solchen Interpretation deutlich werden kann, in welcher Weise durch probabilistische Aussagen von den kontingenten Umständen der jeweils individuell realisierten Lebensverläufe abstrahiert wird. Probabilistische Aussagen beziehen sich bei dieser Interpretation nicht auf Eigenschaften von Individuen, die ihnen individuell zurechenbar sind, sondern auf Eigenschaften einer Gesamtheit von Individuen, die wir uns als „Träger“ der jeweils existierenden gesellschaftlichen Verhältnisse vorstellen können.

k) Der theoretische Gewinn liegt zunächst darin, daß es keine logisch notwendigen Beziehungen zwischen den Eigenschaften gesellschaftlicher Verhältnisse und dem sozialen Verhalten der Individuen gibt. Insbesondere implizieren Aussagen über Eigenschaften gesellschaftlicher Verhältnisse keinerlei Annahmen oder Aussagen über die Kontingenz der individuell realisierten Lebensverläufe; die Theoriebildung beruht vielmehr auf einer systematischen Abstraktion von dieser Kontingenz. Darüberhinaus gewinnt man einen Begriff gesellschaftlicher Verhältnisse, der es erlaubt, sich sozialen Wandel als „Folge“ von sich verändernden individuellen Lebensverläufen vorzustellen, ohne sogleich das Programm des „methodologischen Individualismus“ übernehmen zu müssen. Wesentlich dafür ist, wie ich glaube, die Unterscheidung von Gesetzmäßigkeiten (im Sinne der „Analytischen Wissenschaftstheorie“) und sozialen Regeln bzw. Regelmäßigkeiten.

[51] Vgl. u.a. Popper [1960], Hacking [1965], sowie die Diskussion bei Kyburg [1974].

Der „methodologische Individualismus“ kann zunächst als eine sozialphilosophische Gegenposition zu einer in der Soziologie lange Zeit dominierenden Auffassung des Verhältnisses von Individuen und ihren gesellschaftlichen Verhältnissen angesehen werden. Diese traditionelle Auffassung wurde sehr pointiert von Berger [1963, S. 106] folgendermaßen formuliert: „Die Gesellschaft schreibt uns nämlich nicht nur vor, was wir zu tun, sondern auch, wer wir zu sein haben. Mit anderen Worten: Unser gesellschaftlicher Ort bestimmt nicht nur unser Verhalten, sondern auch unser Sein.“ Dagegen erscheint es durchaus plausibel, wenn zum Beispiel J. W. N. Watkins die Gegenposition folgendermaßen formuliert: „Was ich als das eigentliche Problem ansehe, ist etwa folgendes: Sozialwissenschaftler lassen sich grob und mit einiger Gewaltsamkeit in zwei Hauptgruppen einteilen: jene, die soziale Prozesse sich als sozusagen aus eigener Kraft fortbewegend denken, entsprechend ihrer eigenen Natur und ihrer Gesetze, die Menschen, die in sie verstrickt sind, mit sich fortreißend; und jene, die soziale Prozesse als das komplizierte Ergebnis des Verhaltens von Menschen ansehen.“[52] Bei dieser Formulierung fällt es nicht schwer, sich der Position des „methodologischen Individualismus“ anzuschließen. Es ist jedoch eine bemerkenswerte Tatsache, daß der Versuch, diese Position in ein empirisches Forschungsprogramm umzusetzen, gewissermaßen zu einer Verkehrung der Fronten geführt hat. Zunächst als eine sozialphilosophische Position angetreten, um die Handlungsfreiheiten sozialer Akteure *gegenüber* ihren gesellschaftlichen Verhältnissen zu verteidigen, führte der Versuch, „soziale Prozesse als das komplizierte Ergebnis des Verhaltens von Menschen“ zu erklären, schließlich zu der Auffassung, daß „die zentrale Aufgabe der empirischen Sozialforschung [...] die *Suche* nach Gesetzen zur Erklärung sozialen Handelns“ sei.[53] Zwar kann man nach der Aufklärung durch die „Analytische Wissenschaftstheorie“ auf eine rhetorische Unterscheidung verweisen: es sind nicht mehr gesellschaftliche Verhältnisse, sondern Quasi-Gesetzmäßigkeiten, die die Handlungen und Lebensverläufe der sozialen Akteure determinieren; aber im Ergebnis entsteht ziemlich genau das durch den „methodologischen Individualismus“ zunächst kritisierte Bild: soziale Akteure, die in der Erzeugung gesellschaftlicher Verhältnisse den Gesetzmäßigkeiten folgen, denen ihr soziales Handeln unterworfen ist.[54]

[52] Zitiert bei Danto [1965, S. 495f].

[53] Schnell, Hill und Esser [1992, S. 49].

[54] Zum Beispiel in einer Formulierung von Huckfeld et al. [1982, S. 7 und 9], die den Fortschritt durch eine „Verzeitlichung der Variablensoziologie“ gewissermaßen schon hinter sich hat: „Social reality is more than a set of substantively related but randomly occurring events. In this monograph we think of social phenomena as processes – structured series of events, operations, and activities whose logic is orderly and predictable. The goal of dynamic modeling is to specify the structure of such processes and to deduce the manner in which they generate social change.“ „Throughout this monograph we utilize the concept of the structure of a dynamic process. By this we mean the laws of change that define the time-dependence of a process.“

Zunächst kann man infolgedessen die Überlegungen des „methodologischen Individualismus“ übernehmen, um gegen seine Verwendung als ein (auf der Grundlage des deduktiv-nomologischen Erklärungsschemas re-interpretiertes) Programm der empirischen Sozialforschung zu argumentieren. Das entscheidende sozialphilosophische Argument ist, daß jede akzeptable Konzeption empirischer Sozialforschung mit der Kontingenz individueller Lebensverläufe – und mithin mit den Kommunikationsformen, in denen die sozialen Akteure die Kontingenz ihrer Lebensverläufe reflektieren – *vereinbar* sein muß.

Abgesehen von diesem sozialphilosophischen Einwand kann man jedoch auch die Basisannahme des „methodologischen Individualismus“ infrage stellen: daß gesellschaftliche Verhältnisse „das komplizierte Ergebnis des Verhaltens von Menschen“ sind. Genauer gesagt: Man kann dieser Basisannahme eine triviale Interpretation geben – daß es ohne soziale Akteure keine gesellschaftlichen Verhältnisse gäbe –; aber daraus folgt nicht, daß gesellschaftliche Verhältnisse als „Ergebnis des Verhaltens von Menschen“ *erklärt* werden können. Eine sinnvolle Erklärung würde verlangen, daß man zeigen kann, in welcher Weise soziale Akteure die Erzeuger gesellschaftlicher Verhältnisse sind. Vielleicht sind einige Aspekte gesellschaftlicher Verhältnisse auf diese Weise erklärbar, zum Beispiel die Regeln, die sich eine Gruppe von Menschen selbst gibt. Im allgemeinen können jedoch die durch die empirische Sozialforschung ermittelbaren sozialen Regeln und Regelmäßigkeiten nicht als in diesem Sinne durch soziale Akteure verursacht erklärt werden.[55] Gleichwohl läßt sich beschreiben und dadurch erklären, *wie* sozialer Wandel aus dem sich verändernden sozialen Verhalten der Individuen resultiert. Die soziologische Lebensverlaufsforschung zeigt dies bereits durch die von ihr verwendeten statistischen Verfahren, die von Beobachtungen über individuelle Lebensverläufe ausgehen, um Einsichten in soziale Regeln bzw. Regelmäßigkeiten zu gewinnen.

Zusammenfassung

Thema der vorliegenden Arbeit ist eine Diskussion statistischer Methoden und Modelle im Hinblick auf ihre Verwendung in der soziologischen Lebensverlaufsforschung. Das Ziel besteht darin, herauszufinden, welche Implikationen sich aus diesem Verwendungszusammenhang für ein angemessenes Verständnis statistischer Begriffe und Modelle ergeben. Zugrunde

[55] Es ist aufschlußreich, auf welche Weise Watkins [1957, S. 168f] eine scheinbare Lösung für dieses Problem vorschlägt: „The central assumption of the individualistic position – an assumption which is admittedly counter-factual and metaphysical – is that no social tendency exists which could not be altered *if* the individuals concerned both wanted to alter it and possessed the appropriate information.“ Die Annahme, daß *individuelle* Willensbildung und Informationsgewinnung bereits eine hinreichende Voraussetzung zur Veränderung gesellschaftlicher Verhältnisse sei (so daß man insoweit davon sprechen könnte, daß diese Verhältnisse durch soziale Akteure verursacht seien), ist sicherlich nicht haltbar.

liegt die (leider nur begrenzt realisierbare) Vorstellung, daß es eine Aufgabe der Soziologie ist, die von ihr zu verwendenden statistischen Verfahren zu konzipieren und theoretisch zu deuten, und daß sie dabei von ihren jeweils spezifischen Erkenntnisinteressen auszugehen hat. Aus den in dieser Einleitung vorgetragenen Überlegungen ergeben sich die folgenden Leitgedanken für die weitere Diskussion.

1. Ohne die Vielfalt möglicher Konzeptionen von Lebensverlaufsforschung zu ignorieren, wird in dieser Arbeit von der Perspektive einer Sozialstrukturanalyse ausgegangen. Das heißt, das erkenntnisleitende Interesse richtet sich auf einen Begriff gesellschaftlicher Verhältnisse, der sie als Bedingungen individueller Lebensverläufe sichtbar machen kann.

2. Ein zentrales Problem der Sozialstrukturanalyse wird darin gesehen, daß individuelle Lebensverläufe kontingent sind und sich einfachen deterministischen oder quasi-deterministischen Erklärungen entziehen. Die Aufgabe besteht insofern darin, eine Konzeption der Sozialstrukturanalyse zu entwickeln, die mit diesem Sachverhalt vereinbar ist. Ein gewisses Kriterium für diese Vereinbarkeit kann darin gesehen werden, ob soziologische Vorstellungen über die soziale Bedingtheit individueller Lebensverläufe mit den Kommunikationsformen vereinbar sind, in denen soziale Akteure als Subjekte von Lebensverläufe ihre Handlungsmöglichkeiten reflektieren und kommunizieren.

3. Zwei heuristische Überlegungen sollen dazu dienen, um eine solche Konzeption von Sozialstrukturanalyse vorstellbar zu machen. Erstens die Annahme, daß das Ziel nicht darin liegen sollte, *individuelle* Lebensverläufe zu erklären, sondern daß der Versuch, soziale Bedingungen individueller Lebensverläufe sichtbar zu machen, verlangt, daß von ihrer Kontingenz *abstrahiert* wird. Zweitens die Annahme, daß eine mit der Kontingenz individueller Lebensverläufe vereinbare theoretische Vorstellung über ihre soziale Bedingtheit durch einen Begriff sozialer Regeln gebildet werden kann, an denen sich die sozialen Akteure *orientieren*.

4. Von diesen beiden Annahmen wird ausgegangen, um ein dem soziologischen Verwendungszweck angemessenes Verständnis statistischer Begriffe und Modelle zu erreichen. Ihr soziologischer Sinn wird darin gesehen, daß mit ihrer Hilfe von der Kontingenz der jeweils individuell realisierten Lebensverläufe abstrahiert werden kann und daß sie Einsichten in soziale Regelmäßigkeiten vermitteln können, die sich durch soziale Regeln deuten lassen.

5. Dieser heuristische Leitfaden impliziert, daß versucht wird, statistische Modelle nicht als Hypothesen über allgemeine Gesetzmäßigkeiten zu betrachten, sondern als Beschreibungen gesellschaftlicher Verhältnisse, die einem sozialen Wandel unterliegen. Eine wesentliche Aufgabe wird insbesondere darin gesehen, mit ihrer Hilfe diesen sozialen Wandel beschreibbar zu machen.

6. Der hier verfolgte heuristische Leitfaden legt schließlich nahe, daß der – in diesem Kontext erreichbare – Erklärungsanspruch sich auf Wie-

Fragen beschränken sollte. Denn im Rahmen dieser Konzeption sind gewissermaßen beide Zugangsmöglichkeiten zu Warum-Fragen versperrt. Weder können Warum-Fragen durch einen Verweis auf allgemeine Gesetzmäßigkeiten gegenstandslos gemacht werden, noch kann auf die individuellen Akteure zurückgegriffen werden, denn von ihren individuellen Antworten auf Warum-Fragen wird abstrahiert.

Kapitel 2

Statistische Beschreibung von Lebensverläufen

Statistische Methoden spielen in der Lebensverlaufsforschung eine zentrale Rolle. Sie sind gewissermaßen das Werkzeug, um die jeweils verfügbaren Lebensverlaufsdaten für die intendierte soziologische Wissensbildung nutzbar zu machen. Der elementare Zweck dieser Methoden kann, mit den Worten R. A. Fishers [1922, S. 311], folgendermaßen charakterisiert werden: „Briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usally by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.“ Um statistische Methoden als ein Werkzeug soziologischer Wissensbildung zu verstehen, ist diese Formulierung jedoch nicht ausreichend. Es muß überlegt werden, wie die verfügbaren Daten mithilfe statistischer Methoden auf die zugrundeliegende soziologische Fragestellung bezogen werden können. Das Ziel besteht schließlich nicht darin, Einsichten in Daten zu gewinnen, sondern mithilfe von Daten Einsichten in gesellschaftliche Verhältnisse zu gewinnen.

Dies ist die Leitidee für die folgenden Erörterungen. Es geht um den soziologischen Bedeutungsgehalt der in der Lebensverlaufsforschung verwendeten statistischen Begriffe, Methoden und Modelle. Entsprechend den in der Einleitung angestellten Überlegungen versuche ich, sie nicht auf Individuen, *sondern* auf ihre gesellschaftlichen Verhältnisse zu beziehen. Zur Vermittlung dient die für die Statistik grundlegende Vorstellung einer „Gesamtheit“ von Individuen.

2.1 Der formale Rahmen: Zustandsraum – Zeitachse

Um statistische Methoden anwenden zu können, ist stets ein gewisser formaler Rahmen erforderlich, der es erlaubt, die Sprache der Statistik mit dem jeweils interessierenden Gegenstandsbereich zu verknüpfen. Im Kontext der Lebensverlaufsforschung muß zunächst formal präzisiert werden, in welcher Weise von Lebensverläufen gesprochen werden soll.

Die in der Lebensverlaufsforschung üblich gewordene Grundidee zur formalen Charakterisierung von Lebensverläufen wurde bereits in der Einleitung skizziert: Ein Lebensverlauf ist eine zeitlich geordnete Folge von Zuständen. Dies kann auf einfache Weise formal präzisiert werden. Es gibt

einen Zustandsraum $\mathcal{Y}$ und eine Zeitachse $\mathcal{T}$. Der Zustandsraum definiert die Menge der möglichen Zustände, in denen sich die Individuen befinden können. Der Zustand eines Individuums i zum Zeitpunkt $t \in \mathcal{T}$ kann dann durch $y_i(t)$ repräsentiert werden, wobei $y_i(t) \in \mathcal{Y}$. Als Funktion der Zeit betrachtet, liefert dann $y_i(t)$ eine Beschreibung des Lebensverlaufs dieses Individuums.

Wie ein geeigneter Zustandsraum beschaffen sein sollte, ist in erster Linie eine Frage der intendierten Theoriebildung. Denn von seiner Definition hängt ab, welche Aspekte von Lebensverläufen sichtbar werden können. Umfaßt der Zustandsraum zum Beispiel nur die Zustände *erwerbstätig* und *nicht erwerbstätig*, können Lebensverläufe nur im Hinblick auf diese beiden Zustände beschrieben werden. Der formale Rahmen schließt es jedoch nicht aus, beliebig komplexe Zustandsräume anzunehmen. Um $y_i(t)$ als eine Funktion der Zeit interpretieren zu können, muß nur vorausgesetzt werden, daß zu jedem Zeitpunkt ein im Zustandsraum $\mathcal{Y}$ eindeutig definierter Zustand eingenommen wird.

Der formale Rahmen kann jedoch auf einfache Weise so erweitert werden, daß davon gesprochen werden kann, daß sich Individuen gleichzeitig in mehreren unterschiedlichen Zuständen befinden können. Angenommen, man möchte Lebensverläufe im Hinblick auf m unterschiedliche Aspekte beschreiben. Man kann dann für jeden dieser Aspekte einen separaten Zustandsraum definieren: $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$; zum Beispiel können im Zustandsraum $\mathcal{Y}_1$ Arten der Erwerbstätigkeit unterschieden werden, im Zustandsraum $\mathcal{Y}_2$ kann das erzielte Einkommen klassifiziert werden, und im Zustandsraum $\mathcal{Y}_3$ können Migrationen erfaßt werden. Zu jedem Zeitpunkt gibt es dann einen mehrdimensionalen Zustand $(y_{i,1}(t), \ldots, y_{i,m}(t))$, wobei $y_{i,j}(t)$ der Zustand ist, den das Individuum i zum Zeitpunkt t im Zustandsraum $\mathcal{Y}_j$ einnimmt.

Solche mehrdimensionalen Zustandsräume werden in Kapitel 4 näher betrachtet, da sie einen geeigneten formalen Rahmen darstellen, um das Problem der Interdependenz von Ereignissen (Zustandswechseln) in Lebensverläufen zu untersuchen. Sieht man von dieser Problemstellung ab, können mehrdimensionale Zustandsräume auch als einfache (eindimensionale) Zustandsräume dargestellt werden, indem jede mögliche Kombination von Zuständen in $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$ als *ein* Zustand in einem eindimensionalen Zustandsraum $\mathcal{Y}$ repräsentiert wird.

Der Zustandsraum ist zunächst eine beliebige Menge von Zuständen, um Lebensverläufe zu charakterisieren. Primär wichtig ist, daß es sich um soziologisch sinnvolle Klassifikationen handelt. Die mathematische Repräsentation des Zustandsraums ist dem untergeordnet, sie ist nur ein sprachliches Hilfsmittel für die formale Darstellung. Ich betone dies, weil die Sprache der Mathematik Präzisierungen und Differenzierungen ermöglicht (und in gewisser Weise erfordert), für die es in der Umgangssprache und in der soziologischen Sprache keine sinnvollen Entsprechungen

gibt.[1] Zum Beispiel können aus mathematischer Sicht diskrete und stetige Zustandsräume unterschieden werden. Da in der Lebensverlaufsforschung meistens nur wenige Zustände unterschieden werden, ist die Vorstellung eines diskreten Zustandsraums fast immer ausreichend. Gelegentlich kann es jedoch sinnvoll sein, stattdessen von der Vorstellung eines stetigen Zustandsraums auszugehen. Will man zum Beispiel Änderungen im erzielten Einkommen beschreiben, könnte es sinnvoll sein, das Einkommen durch eine Variable zu repräsentieren, die kontinuierlich variieren kann. Natürlich ist dies strenggenommen nicht der Fall; das Einkommen ist wie alle anderen sozial definierten Variablen diskret. Gleichwohl kann es in diesem Fall *einfacher* sein, Einkommensänderungen so zu beschreiben, als ob sie in kontinuierlicher Weise möglich wären. Wichtig ist nur, die Differenz zwischen dem zu beschreibenden Sachverhalt und der zu ihrer Beschreibung verwendeten Sprache im Auge zu behalten.[2]

Das gleiche Problem tritt auf, wenn man versucht, die Vorstellung einer Zeitachse mathematisch zu präzisieren. Wiederum gibt es zwei Möglichkeiten. Man kann sich eine Zeitachse diskret, als eine Folge von Zeitpunkten oder Zeitintervallen vorstellen, oder man kann von der Vorstellung einer kontinuierlichen Zeitachse ausgehen. Die Mathematik stellt für beide Vorstellungen einen jeweils geeigneten Begriffsrahmen zur Verfügung. Welche Form sollte gewählt werden, um Lebensverläufe zu beschreiben?

Einige Autoren haben dafür plädiert, grundsätzlich von einer kontinuierlichen Zeitachse auszugehen. Das dafür hauptsächlich angeführte Argument ist, daß Zustandswechsel (Ereignisse) *jederzeit* eintreten können.[3] Die Vorstellung, daß es für Zustandswechsel im Rahmen von Lebensverläufen exakte Ereigniszeitpunkte auf einer kontinuierlichen Zeitachse gibt, ist jedoch problematisch. Der Grund ist, daß alle sozial relevanten Ereignisse eine ihnen inhärente Zeitdauer aufweisen; zum Beispiel die Geburt eines Kindes, eine Heirat, der Verlust eines Arbeitsplatzes. Die Umgangssprache trägt dem Rechnung, indem sie von einer diskreten Zeitachse ausgeht, die dem jeweiligen Ereignistyp und seiner intendierten Beschreibung ange-

[1] Man kann hierzu zwei unterschiedliche Haltungen einnehmen. Einerseits kann man die Mathematik als ein Medium betrachten, in dem eine „präzise“ Theoriebildung im Unterschied zu umgangssprachlichen Formulierungen erreicht werden kann; vgl. zum Beispiel die Darlegung dieser Auffassung durch Neyman [1952, S. 23ff]. Andererseits kann man die Mathematik als eine Sprache ansehen, mit deren Hilfe die Komplexität umgangssprachlicher Formulierungen vereinfacht und formal handhabbar gemacht werden kann. Ich folge hier dieser zweiten Betrachtungsweise. Insbesondere gehe ich davon aus, daß mathematische Begriffe und Modelle, um einer soziologischen Wissensbildung dienen zu können, durch umgangssprachlich verfügbare Vorstellungen *interpretierbar* sein müssen.

[2] Diese Differenz beruht natürlich auf der Möglichkeit, einen Sachverhalt auf mehrere unterschiedliche Weisen beschreiben zu können.

[3] Zum Beispiel Coleman [1981, S. 6]: „Changes can occur at any point in time and are not constrained to predetermined time points.“ Dies ist auch das wesentliche Argument bei Tuma und Hannan [1984, S. 21, 82f] und Andreß [1992, S. 26ff].

paßt wird. Um eine diskrete Zeitachse zu definieren, kann man von beliebig kleinen Zeiteinheiten ausgehen (der Unterschied zu einer stetigen Zeitachse besteht nur darin, daß ein Grenzübergang zu „infinitesimal“ kleinen Zeitpunkten ausgeschlossen wird). Bei den meisten Ereignissen, die zur Beschreibung von Lebensverläufen aus soziologischer Sicht sinnvoll definierbar sind, kann jedoch eine angemessene Zeitbestimmung bereits dadurch vorgenommen werden, daß man von Tagen ausgeht. Nur ausnahmsweise erscheint es sinnvoll, kleinere Zeiteinheiten (etwa Stunden oder Minuten) zu verwenden.

Diese Überlegung legt es nahe, grundsätzlich von einer diskreten Zeitachse auszugehen, um der Tatsache Rechnung zu tragen, daß Ereignisse (Zustandswechsel) typischerweise selbst eine gewisse Zeit benötigen. Eine diskrete Zeitachse ermöglicht es darüberhinaus, Abläufe als eine *Folge* von Zuständen zu beschreiben. Die Verwendung einer stetigen Zeitachse kann jedoch als eine idealisierende Approximation angesehen werden, die gelegentlich den Vorteil bietet, daß sie zu einfacheren mathematischen Formulierungen führt.[4] Wichtig erscheint mir nur, daß im Auge behalten wird, daß es sich um eine Idealisierung handelt und daß gewisse Merkmale statistischer Modelle, die von einer stetigen Zeitachse ausgehen, keine sinnvollen Entsprechungen in realen Lebensverläufen haben. Dies betrifft zum Beispiel das Problem der Gleichzeitigkeit von Ereignissen (in mehrdimensionalen Zustandsräumen). Auf der Grundlage einer stetigen Zeitachse ist die Wahrscheinlichkeit, daß zwei Ereignisse gleichzeitig auftreten, Null; aber in der Realität ist diese Möglichkeit natürlich gegeben und sollte deshalb bei der Modellbildung berücksichtigt werden.

Ich möchte betonen, daß es hier zunächst nur um die Frage geht, wie eine für die theoretische Konzeption von Lebensverläufen angemessene Zeitachse beschaffen sein sollte. Ein vollständig anderes Problem liegt darin, daß häufig nur *ungenaue* Informationen über Ereigniszeitpunkte verfügbar sind.[5] Bei der Erhebung von Lebensverlaufsdaten werden zum Beispiel in der Regel nur die Monate erfaßt, in denen gewisse Ereignisse

[4] In gewisser Weise hat dies allerdings nur historische und konventionelle Gründe. Der größte Teil der anwendungsorientierten Mathematik ist im Kontext naturwissenschaftlicher Anwendungen entstanden, bei denen von einer stetigen Zeitachse ausgegangen wird. Diese Mathematik erscheint einfach, weil sie üblich und weitestgehend entwickelt worden ist. Es ist infolgedessen auch verständlich, daß Sozialwissenschaftler in erster Linie auf diese Mathematik zurückgreifen. Aber ich glaube nicht, daß sich mit einem Verweis auf diese historische Entwicklung – so wie es zum Beispiel von Coleman [1968, S. 428ff] versucht worden ist – eine überzeugende Begründung dafür geben läßt, daß die Soziologie unbedingt von der Vorstellung einer stetigen Zeitachse ausgehen müsse.

[5] Diese Unterscheidung ist bei der Diskussion der Frage, ob diskrete oder stetige Zeitachsen angemessen sind, nicht immer klar. Eine berechtigte Kritik richtet sich häufig gegen die Verwendung von mehr oder weniger willkürlich gewählten Beobachtungszeitpunkten als Zeitachse für die statistische Modellbildung, insbesondere bei Panel-Daten (z.B. Singer und Spilerman [1976, S. 451]); daraus folgt jedoch nicht, daß die Verwendung einer stetigen Zeitachse eine adäquate Alternative ist.

stattgefunden haben (Geburt, Heirat, Verlassen der Schule, Beginn eines neuen Jobs, usw.). Dann liegt eine ungenaue Information vor, gemessen daran, daß für eine genaue Angabe der jeweilige Tag des Ereignisses anzugeben wäre. Gelegentlich sind die verfügbaren Zeitangaben noch wesentlich ungenauer. Dies hängt vor allem davon ab, in welcher Form Lebensverlaufsdaten erhoben werden. Bei ereignisorientierten Erhebungsdesigns ist es in der Regel möglich, Monatsangaben zu erreichen. Bei Panelerhebungen, die typischerweise nur einmal jährlich den jeweils eingenommenen Zustand ermitteln, erreicht man bestenfalls Jahresintervalle.[6]

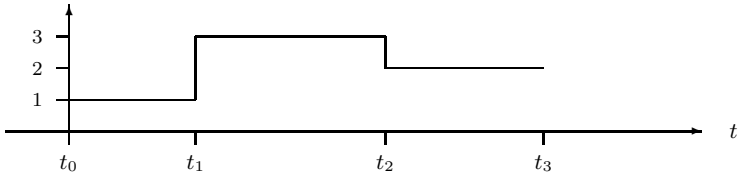
Tatsächlich sind die praktisch verfügbaren Lebensverlaufsdaten fast immer ungenau, gemessen an dem für die Theorie- bzw. Modellbildung wünschenswerten Genauigkeitsgrad. Es gibt dann zwei Möglichkeiten. Einerseits kann man versuchen, das theoretische Modell in den bei der Datenerhebung tatsächlich verwendeten Zeiteinheiten zu reformulieren. Andererseits kann man versuchen, das vorgegebene theoretische Modell mithilfe der nur ungenau verfügbaren Daten *näherungsweise* zu bestimmen. Dafür gibt es wiederum zwei Möglichkeiten. Man kann die Ungenauigkeiten in den Zeitangaben ignorieren; man verwendet dann die Daten so, *als ob* sie auf der durch das theoretische Modell vorgegebenen Zeitachse genau wären. Dies ist das bisher meistens übliche Vorgehen. Oder man kann versuchen, die Tatsache, daß nur ungenaue Zeitangaben vorliegen, bei der Modellschätzung explizit zu berücksichtigen. Beide Vorgehensweisen werden in Kapitel 3 näher diskutiert.

[6] Moderne Panelerhebungen enthalten allerdings häufig einige ereignisbezogene Abschnitte. Soweit dies der Fall ist, können natürlich genauere Angaben über Ereigniszeitpunkte ermittelt werden.

2.2 Zustands- und ereignisbezogene Darstellungen

Der in Abschnitt 2.1 eingeführte formale Rahmen erlaubt es, individuelle Lebensverläufe zu beschreiben. Dabei sind zwei unterschiedliche Formen der Darstellung möglich. Die erste Darstellungsform wurde bereits erwähnt: der Lebensverlauf eines Individuums i wird durch eine Funktion $y_i(t)$ beschrieben, die als eine Funktion der Zeit $t \in \mathcal{T}$ Werte in einem vorgegebenen Zustandsraum $\mathcal{Y}$ annehmen kann. Ich nenne dies eine *zustandsbezogene Darstellungsform*; jeder Lebensverlauf wird als eine zeitlich geordnete Menge von Zuständen beschrieben. Ist die Zeitachse diskret, kann anschaulich von einer *Folge* von Zuständen gesprochen werden.

Eine alternative Darstellungsform ergibt sich aus der Überlegung, daß – bei den meisten für soziologische Untersuchungen relevanten Zustandsräumen – Zustandswechsel nur selten auftreten. Dies legt es nahe, Lebensverläufe als Folgen von Episoden zu beschreiben. Die folgende Abbildung veranschaulicht diese Art der Beschreibung.



In diesem Beispiel beginnt der Lebensverlauf zum Zeitpunkt t_0 mit dem Zustand 1, dann erfolgt zum Zeitpunkt t_1 ein Übergang in den Zustand 3, und zum Zeitpunkt t_2 ein Übergang in den Zustand 2. Die Beobachtung endet zum Zeitpunkt t_3. Oder in einer etwas anderen Formulierung kann man davon sprechen, daß es drei Episoden gibt:

1. Episode: von t_0 bis t_1 im Zustand 1
2. Episode: von t_1 bis t_2 im Zustand 3
3. Episode: von t_2 bis t_3 im Zustand 2

Ich nenne dies eine *ereignisbezogene Darstellungsform*. Jeder individuelle Lebensverlauf wird als eine Folge von Episoden dargestellt, die durch Ereignisse eingeleitet bzw. abgeschlossen werden; in formaler Darstellung:

$$E_{il} \equiv (i, l, o_{il}, d_{il}, s_{il}, t_{il}) \qquad i = 1, \ldots, N \quad l = 1, \ldots, L_i$$

E_{il} ist die l.te Episode der Person i; L_i ist die Anzahl der Episoden, die für die i.te Person beobachtet werden können. o_{il} und d_{il} bezeichnen den Anfangs- bzw. Endzustand, s_{il} und t_{il} bezeichnen den Anfangs- bzw. Endzeitpunkt der Episode. Ersichtlich erlaubt eine ereignisbezogene Darstellungsform in der Regel wesentlich sparsamere Formulierungen und ist deshalb häufig zweckmäßiger als eine zustandsbezogene Darstellungsform. Beide Darstellungsformen liefern jedoch die gleiche Information.

Kalenderzeit und Prozeßzeit

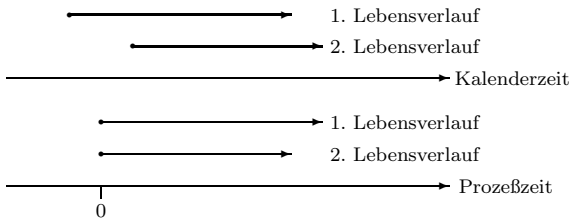
Ein Sachverhalt kann nur beschrieben werden, wenn es einen Beobachter gibt. Bei der Beschreibung zeitlicher Vorgänge kommt dem Beobachter eine besondere Bedeutung zu, denn es hängt dann insbesondere von seiner zeitlichen Position ab, wie der Sachverhalt beschrieben werden kann. Für die Beschreibung von Lebensverläufen resultieren daraus zwei Probleme.

Das erste Problem betrifft die zur Beschreibung zu verwendende Zeitachse. Zunächst hat jeder Beobachter seine eigene Zeitachse, definiert durch seine subjektive Unterscheidung zwischen seiner Vergangenheit, seiner Gegenwart und seiner Zukunft. Unter Verwendung dieser Zeitachse können der jeweils eigene Lebensverlauf und partiell die Lebensverläufe anderer Personen beschrieben werden. Die damit verbundene Art, Lebensverläufe zu beschreiben, ist umgangssprachlich verankert, aber für eine soziologische Beschreibung ungeeignet. Sie benötigt eine Zeitachse, die von der zeitlichen Position des Beobachters unabhängig ist. Ob eine solche Zeitachse existiert, ist in der Philosophie umstritten;[7] die soziologische Theoriebildung kann sich jedoch darauf berufen, daß in jeder Gesellschaft eine quasi-objektive, den Gesellschaftsmitgliedern gemeinsame Zeitachse faktisch existiert. In den modernen (gegenwärtigen) Gesellschaften ist dies die amtliche Kalenderzeit. Sie dient, als soziale Konvention, dazu, die unterschiedlichen Zeitverläufe (d.h. Lebensverläufe) der Individuen im Hinblick auf soziale Interaktion vergleichbar zu machen. Insofern die Soziologie auf eine Beschreibung gesellschaftlicher Verhältnisse zielt, ist es deshalb sinnvoll, daß sie sich ebenfalls dieser Konvention bedient. Die Berechtigung für die damit vollzogene Abstraktion von der subjektiven Zeit der individuellen Lebensverläufe kann darin gesehen werden, daß schließlich nicht eine Beschreibung individueller Lebensverläufe, sondern eine Beschreibung und Erklärung der gesellschaftlichen Verhältnisse angestrebt wird, in denen sich die individuellen Lebensverläufe entwickeln.

Das methodische Hilfsmittel ist eine *vergleichende* Beobachtung und Beschreibung individueller Lebensverläufe. Der Vergleich beruht auf zwei wesentlichen Abstraktionen. Erstens wird ein Zustandsraum konstruiert, und ein Vergleich wird nur im Hinblick auf die darin unterschiedenen Zustände vorgenommen. Zweitens muß, zumindest bis zu einem gewissen Grad, von den zeitlichen Unterschieden zwischen den individuellen Lebensverläufen abstrahiert werden. Der Vergleich erfolgt nicht unmittelbar in der Kalenderzeit, sondern wird auf einer Prozeßzeitachse vorgenommen.

Folgende Abbildung veranschaulicht, examplarisch für zwei Lebensverläufe, den Übergang von der Kalenderzeit zu einer Prozeßzeitachse:

[7] Grundlegende Überlegungen zu dieser Frage stammen von McTaggart [1905], vgl. die interessante Rekonstruktion seiner Überlegungen durch Dummett [1960].



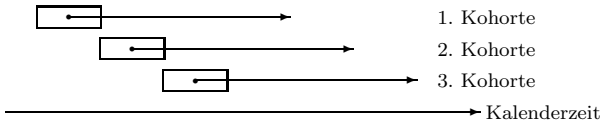
Auf der Kalenderzeitachse beginnen die Lebensverläufe zu jeweils unterschiedlichen Zeitpunkten, auf der Prozeßzeitachse beginnen sie zu einem gemeinsamen Zeitpunkt, üblicherweise mit ihrem Nullpunkt. Wenn, wie in diesem Beispiel, die Lebensverläufe der Individuen seit ihrer Geburt erfaßt werden, entspricht die Prozeßzeitachse dem Lebensalter, die Prozeßzeit ist das Lebensalter.

Die Konstruktion einer Prozeßzeitachse ist zunächst ein methodisches Hilfsmittel, um Lebensverläufe zeitlich vergleichbar zu machen; sie hängt deshalb davon ab, welcher Aspekt der individuellen Lebensverläufe verglichen werden soll. Möchte man zum Beispiel die Lebensverläufe daraufhin vergleichen, wie sie sich seit dem Abschluß der Schulbildung entwickelt haben, kann eine Prozeßzeitachse verwendet werden, die mit diesem Ereignis beginnt. Im allgemeinen beginnt also die Prozeßzeit mit einem Ereignis. Sobald dieses Ereignis eintritt, beginnt für alle in den Vergleich einbezogenen Individuen eine ihnen gemeinsame Prozeßzeit. Die Konstruktion einer Prozeßzeitachse dient darüberhinaus dazu, die vergleichende Beschreibung von Ereignissequenzen mit einer Theoriebildung zu verknüpfen. Einen möglichen Zugang erhält man durch die Überlegung, daß eine Prozeßzeitachse dann beginnt, sobald die Chance bzw. das Risiko (zunächst nur: die Möglichkeit) besteht, daß ein Ereignis eintreten kann.[8] Einen sehr allgemeinen begrifflichen Rahmen liefert die Vorstellung, daß Lebensverläufe

[8] Exemplarisch kann man daran denken, daß die Aufgabe darin besteht, soziale Regeln bzw. Regelmäßigkeiten für das Heiratsverhalten zu ermitteln. Eine lebensverlaufbezogene Betrachtung könnte zu dem Gedanken führen, daß die Risikoperiode (für das Ereignis *Heirat*) bei jedem Individuum mit einem unterschiedlichen Alter beginnt. Diese Vorstellung erscheint sinnvoll, wenn die Beschreibung auf jeweils bestimmte Individuen zielt, zum Beispiel im Kontext einzelfallbezogener Biographieforschung. Wenn dagegen eine Einsicht in soziale Regeln bzw. Regelmäßigkeiten intendiert ist, geht es nicht darum, für jedes Individuum eine besondere Regel zu finden, sondern anhand der individuell unterschiedlichen Lebensverläufe verallgemeinerbare, d.h. *soziale* Regeln zu finden. Dies setzt voraus, daß zur Definition des Ereignisses, das eine Risikoperiode (formal: eine Episode) einleitet, auf vergleichbare Eigenschaften der Individuen bzw. ihrer Situation zurückgegriffen werden muß. Dies muß nicht unbedingt das Alter sein, jedes beliebige Ereignis kann verwendet werden, wenn es sich eignet, um eine soziale Situation zu charakterisieren, im Hinblick auf die sinnvoll von möglichen sozialen Regeln bzw. Regelmäßigkeiten gesprochen werden kann. Im Hinblick auf das Heiratsverhalten könnte man zum Beispiel eine Ausgangssituation durch die Beendigung des Schulbesuchs definieren. Tatsächlich gibt es meistens zahlreiche unterschiedliche Möglichkeiten, d.h. eine Vielzahl möglicher Regeln ist vorstellbar und ggf. empirisch ermittelbar, um Einsichten in die situationsbezogene Entwicklung von Lebensverläufen zu gewinnen.

Sequenzen von sich in zeitlicher Folge bedingenden Ereignissen sind.[9]

Die Konstruktion einer Prozeßzeitachse dient dazu, die sich in der Kalenderzeit asynchron entwickelnden Lebensverläufe vergleichbar zu machen. Um dabei die Tatsache berücksichtigen zu können, daß es wesentliche Unterschiede zwischen Lebensverläufen geben kann, je nachdem in welcher historischen Situation sie realisiert werden, ist es in der Lebensverlaufsforschung üblich, *Kohorten* zu bilden. In einer Definition von Glenn [1977, S. 8]: „a cohort is defined as those people within a geographically or otherwise delineated population who experienced the same significant life event within a given period of time.“ Es werden also jeweils diejenigen Personen zu einer Kohorte zusammengefaßt, bei denen das die Prozeßzeitachse konstituierende Ereignis in der gleichen historischen Situation (formal definiert durch ein Kalenderzeitintervall) stattgefunden hat. Dadurch wird es möglich, die Prozeßzeit mit der Kalenderzeit zu verknüpfen, etwa so, wie es in der folgenden Abbildung für drei Kohorten illustriert wird.



Die Unterscheidung von Kohorten ist in der Lebensverlaufsforschung ein wichtiges Hilfsmittel, um Einsichten in sozialen Wandel zu gewinnen.[10] Darauf wird in Abschnitt 2.5 näher eingegangen.

Unvollständige Beobachtungen

Ein zweites wesentliches Problem bei der Beschreibung von Lebensverläufen besteht darin, daß die verfügbaren Beobachtungen in der Regel unvollständig und ungenau sind. Drei Aspekte können unterschieden werden. Erstens ist es in der Regel unmöglich, die Gesamtheit der in einer Gesellschaft realisierten Lebensverläufe zu beobachten. Insbesondere dann, wenn Lebensverlaufsdaten nur durch Interviews gewonnen werden können, kann man nur einen Teil, eine Stichprobe der Personen aus der Grundgesamtheit befragen. Zweitens können die interessierenden Lebensverläufe in der Regel nicht vollständig von ihrem Anfang bis zu ihrem Ende beobachtet werden. Dies ist eine unmittelbare Folge dessen, daß ein zeitlicher Prozeß nur beobachtet werden kann, solange sich der Beobachter synchron zu diesem Prozeß bewegt. Da die für die soziologische Forschung interessierenden Lebensverlaufsdaten meistens nur durch Interviews gewonnen wer-

[9] Diese „ereignisanalytische“ Betrachtungsweise wird in Abschnitt 4.1 ausführlich diskutiert.

[10] Dazu grundlegend waren insbesondere die Beiträge von Ryder [1964, 1965]. Als empirische Anwendung vgl. exemplarisch Blossfeld [1989].

den können, bildet der Zeitraum der Befragung eine zeitliche Begrenzung möglicher Beobachtungen. Drittens sind die über Lebensverläufe verfügbaren Daten fast immer mehr oder weniger ungenau.

Alle drei Aspekte sind offenbar wichtig, wenn man aus Lebensverlaufsdaten einigermaßen verläßliche Informationen gewinnen möchte. Im weiteren Verlauf dieser Arbeit werden deshalb alle drei Aspekte noch mehrfach thematisiert.

Beispiel: SOEP-Daten über Lebensgemeinschaften

Um die Erhebung und Darstellung von Lebensverlaufsdaten zu illustrieren, beziehe ich mich auf das Sozio-ökonomische Panel (SOEP). Es handelt sich um eine Kombination aus einer Retrospektiv- und einer Panelerhebung.[11] Die Datenerhebung wurde 1984 mit einer „repräsentativen“ Stichprobe aus der Bevölkerung der damaligen Bundesrepublik begonnen, um Auskünfte über die gegenwärtigen Lebensbedingungen und einige Aspekte der bisher realisierten Lebensverläufe zu gewinnen. Soweit möglich, wurden dann die in die Stichprobe einbezogenen Personen in den folgenden Jahren erneut befragt, um zu erfahren, wie sich ihre Lebensverläufe seither entwickelt haben. Gegenwärtig (1994) sind die Ergebnisse aus 9 Befragungswellen (1984 – 1992) verfügbar.

Als Beispiel sei der Frage nachgegangen, welche Informationen das SOEP über die Bildung und Auflösung von Lebensgemeinschaften liefert. Es gibt im wesentlichen drei Arten von Informationen. (a) In jeder Welle wurde der Familienstand zum Befragungszeitpunkt erfaßt. Dies umfaßt sowohl eheliche als auch nichteheliche Lebensgemeinschaften. (b) Ab der 2. Welle (1985) wurden, jeweils seit dem Beginn des Vorjahres, die Zeitpunkte familiärer Veränderungen auf einer Monatsbasis erfragt. Dies umfaßt: Heiraten und Bildung von Lebensgemeinschaften, Scheidungen, Trennungen und Tod des Lebenspartners, Geburten von Kindern. (c) In der 2. Welle wurde einmalig retrospektiv eine Familienbiographie erfragt. Sie umfaßt die Geburtsjahre für alle bis zum Befragungszeitpunkt geborenen Kinder, sowie Beginn, Ende und Status von bis zu drei Eheschließungen.

Um eine an Lebensverläufen orientierte Darstellung zu erreichen, müssen zunächst diese Informationsquellen zur Bildung von Familienbiographien bis zum Ende der jeweiligen Teilnahme am SOEP zusammengefaßt werden. Dies ist schwierig, weil die Zeitangaben teilweise auf Jahresbasis und teilweise auf Monatsbasis vorliegen. Außerdem sind Angaben über nichteheliche Lebensgemeinschaften nur für den Befragungszeitraum verfügbar. Ich beschränke mich hier deshalb auf eheliche Lebensgemeinschaften.[12]

[11] Vgl. als Einführung in die Konzeption dieser Datenerhebung Hanefeld [1984, 1986, 1987].

[12] Ein Beispiel, das sich auf nichteheliche Lebensgemeinschaften bezieht, wird in Abschnitt 4.2.3 gegeben.

Für die Darstellung wird eine Prozeßzeitachse auf Monatsbasis verwendet. Monat 0 ist der Juni desjenigen Jahres, in dem eine Person geboren wurde, Monat 1 ist der Juli des gleichen Jahres, usw. Ereignisse, bei denen nur das Jahr bekannt ist (insbesondere die Geburtszeitpunkte), werden auf den Juni des betreffenden Jahres datiert. Die ehelichen Lebensgemeinschaften beginnen mit dem Ereignis *Heirat*, sie enden mit einem der Ereignisse *Scheidung*, *Tod des Partners* oder *Tod der Bezugsperson*.[13] Um die Darstellung zu vereinfachen, werden zunächst alle Beobachtungen beim jeweiligen Ende der Teilnahme am SOEP als rechts zensiert angesehen; das Ereignis *Tod der Bezugsperson* – in statistischer Sprache der Übergang in einen absorbierenden Endzustand – wird also nicht gesondert erfaßt. Der Zustandsraum, wobei bis zu vier Ehen unterschieden werden,[14] sieht dann folgendermaßen aus:

0	Vor der ersten Heirat
1	1. Ehe
2	1. Scheidung
3	1. Tod des Lebenspartners
4	2. Ehe
5	2. Scheidung
6	2. Tod des Lebenspartners
7	3. Ehe
8	3. Scheidung
9	3. Tod des Lebenspartners
10	4. Ehe
11	Rechts zensierte Beobachtung oder Tod der Bezugsperson

Es ist zweckmäßig, die Festlegung eines Zustandsraums durch die explizite Angabe der möglichen Zustandswechsel zu ergänzen. Ich nenne dies im folgenden ein *Biographieschema*. Es zeigt, auf der Grundlage eines gegebenen Zustandsraums, die Gesamtheit der möglichen Lebensverläufe. Für unser Beispiel zeigt Abbildung 2.2.1 ein solches Biographieschema. Jeder mögliche Lebensverlauf beginnt im Zustand 0, *vor der ersten Heirat*; als Ereignis, das zu diesem Zustand führt, kann zum Beispiel die Geburt oder das Erreichen des 15. Lebensjahr angesehen werden. Und jeder mögliche Lebensverlauf endet schließlich mit dem Erreichen eines absorbierenden

[13] Da in der retrospektiv erfragten Familienbiographie des SOEP nur Scheidungen und Tod des Lebenspartners erfaßt werden, wird auf eine Unterscheidung von *Scheidungen* und *Trennungen* verzichtet. Tatsächlich wird jedoch ggf. auf Angaben über Trennungen zurückgegriffen, wenn – im Rahmen der laufenden Panelerhebungen – vor dem Beginn einer neuen Ehe nur Angaben über eine Trennung, nicht über eine zuvor erfolgte Scheidung gemacht worden sind.

[14] Dies ist die maximale Anzahl von Heiraten, die in der SOEP-Stichprobe beobachtet werden können.

Endzustands, in unserem Beispiel der Tod *oder* das Ende der Beobachtung des Lebensverlaufs. Wichtig ist, daß bei der Definition eines Biographieschemas berücksichtigt wird, daß jederzeit ein Übergang in einen absorbierenden Endzustand möglich ist.

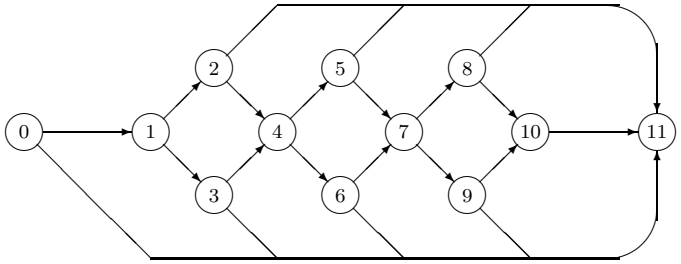


Abbildung 2.2.1 Biographieschema für die im SOEP erfaßten Heiraten.

Nach diesen begrifflichen Vorbereitungen kann schließlich beschrieben werden, welche Lebensverläufe von den durch das SOEP beobachteten Personen realisiert worden sind. Allerdings ist zuvor noch festzulegen, auf welche Längsschnittstichprobe sich die Darstellung beziehen soll. Für dieses und die folgenden Beispiele gehe ich von folgender Abgrenzung aus: Es werden nur Personen betrachtet, die (a) in den Jahren 1918 – 1968 geboren worden sind und (b) an mindestens den ersten drei Wellen des SOEP teilgenommen haben.[15] Unsere Stichprobe besteht dann aus insgesamt 8663 Personen. Da es hier nur um eine Illustration von Daten geht, werden zunächst keine weiteren Differenzierungen vorgenommen.[16]

[15] Das zweite Abgrenzungskriterium ist problematisch, weil die Teilnahme am SOEP einem nur begrenzt kontrollierbaren Selektionsprozeß unterliegt. Es ist gleichwohl zweckmäßig, da erst in der zweiten Welle retrospektive Familienbiographien erhoben worden sind, und da erst in der dritten Welle der Zeitpunkt für den Beginn von Erwerbstätigkeiten ermittelt worden ist.

[16] Bei analytisch orientierten Verwendungen dieser Daten muß berücksichtigt werden, daß das SOEP eine mehrfach geschichtete Stichprobe ist. Es gibt zunächst eine Schichtung in zwei Teilstichproben (s. SOEP-Benutzerhandbuch, B.1-1), nämlich Stichprobe A: „Personen in Privathaushalten, deren Haushaltsvorstand nicht die türkische, griechische, jugoslawische, spanische oder italienische Staatsangehörigkeit besitzt.“ Und Stichprobe B: „Personen in Privathaushalten, deren Haushaltsvorstand die türkische, griechische, jugoslawische, spanische oder italienische Staatsangehörigkeit besitzt, sowie die Anstaltsbevölkerung dieser fünf Nationalitäten.“ Die Stichprobe B ist darüberhinaus in 5 weitere Teilstichproben geschichtet, entsprechend der 5 in ihr repräsentierten Nationalitäten. Die für unsere Beispiele verwendete Stichprobe besteht aus 6416 Personen aus der Teilstichprobe A und aus 2247 Personen aus der Teilstichprobe B.

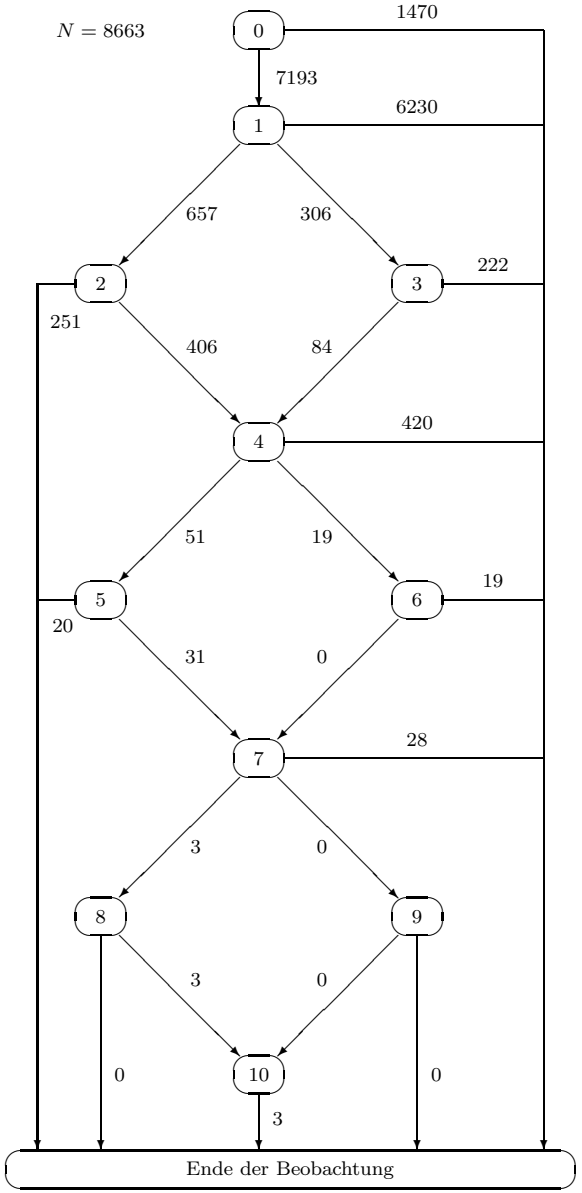


Abbildung 2.2.2 Biographieschema fur die im SOEP erfaßten Heiraten. Stichprobenabgrenzung: alle Personen, die im Zeitraum 1918 – 1968 geboren worden sind und an mindestens den ersten drei Befragungswellen des SOEP teilgenommen haben.

Abbildung 2.2.2 zeigt für diese 8663 Personen, wie sich bei ihnen die durch das SOEP *bisher* beobachteten Lebensverläufe im Hinblick auf Bildung und Auflösung von Ehen entwickelt haben. Die meisten dieser Personen (7193 Personen, 83 %) haben mindestens einmal geheiratet. Bei 963 (657 plus 306) Personen kann eine Auflösung dieser ersten Ehen beobachtet werden, und es beginnt dann eine sich mehr und mehr verzweigende Biographie.

Abbildung 2.2.2 liefert eine erste Einsicht in die Bildung und Auflösung von Lebensgemeinschaften bei den Personen unserer SOEP-Stichprobe. Tatsächlich fehlen jedoch noch zwei wesentliche Aspekte. Erstens zeigt das Biographieschema in dieser Abbildung zwar die beobachtbaren Zustandswechsel und ihre zeitliche Reihenfolge, es zeigt jedoch nicht, *wie lange* sich die Personen in den jeweiligen Zuständen aufhalten. Zweitens verschleiert das Schema die Tatsache, daß die verfügbaren Beobachtungen wesentlich unvollständig sind. Auf beide Aspekte wird in den folgenden Abschnitten genauer eingegangen. An dieser Stelle soll zunächst nur kurz gezeigt werden, inwiefern die verfügbaren Informationen unvollständig sind. Entsprechend den oben vorgenommenen Unterscheidungen gibt es drei Aspekte.

a) Zunächst sind die verfügbaren Informationen unvollständig, weil es sich nur um eine Stichprobe handelt, die verfügbaren Daten sind infolgedessen selektiv. Nehmen wir an, daß die uns interessierende Grundgesamtheit aus der Gesamtheit der Personen besteht, die im Zeitraum 1918 – 1968 geboren worden sind. Dann gibt es zwei wesentliche Selektionsprozesse. Erstens bereits auf der Ebene der Stichprobenkonzeption: offenbar sind nur diejenigen Personen potentielle Mitglieder unserer Stichprobe, die zum Zeitpunkt der ersten Erhebungswelle des SOEP (1984) im Gebiet der damaligen Bundesrepublik tatsächlich gelebt haben. Über Personen, die bereits vorher das Gebiet der BRD verlassen haben oder gestorben sind, erhalten wir – aus dieser Stichprobe – keine Information. Zweitens gibt es ein Selektionsproblem auf der Ebene der Bereitschaft zur Teilnahme am Befragungsprogramm des SOEP. Dies betrifft bereits die Teilnahme an der ersten Erhebungswelle des SOEP. Ursprünglich wurde mit 10638 Haushaltsadressen begonnen, schließlich wurden nur 5969 Interviews realisiert (vgl. SOEP-Benutzerhandbuch, B.2-2), teilweise weil die ausgewählten Adressen nicht auffindbar waren oder nicht dem Stichprobenplan entsprachen, zum größeren Teil jedoch, weil die Interviews nicht erfolgreich durchgeführt werden konnten. Dieses aus Querschnittserhebungen bekannte Selektionsproblem verschärft sich jedoch bei Panelerhebungen. Denn von Welle zu Welle gibt es einen durchaus erheblichen Anteil von Personen, mit denen keine weiteren Interviews mehr durchgeführt werden können, sei es weil sie nicht mehr erreicht werden können oder weil sie nicht länger zur Teilnahme an der Befragung bereit sind. Tabelle 2.2.1 illustriert diesen Sachverhalt am Beispiel der ersten 9 Wellen des SOEP.[17]

[17]Dieser Ausfallprozeß wurde, im Hinblick auf das SOEP, in zahlreichen Arbeiten von Rendtel untersucht; vgl. u.a. Rendtel [1990, 1993, 1994]. An dieser Stelle beschränke ich mich auf einige allgemeine Bemerkungen. (a) Zunächst zeigt Tabelle 2.2.1, daß das Ausmaß der Stichprobenausfälle im Verlauf der Datenerhebung geringer geworden ist. Außerdem ist sichtbar, daß das Problem der Ausfälle in der Teilstichprobe B ein größeres Ausmaß hat als in der Teilstichprobe A, hauptsächlich wegen der in der Teilstichprobe B wesentlich größeren Anzahl von Umzügen ins Ausland. (b) Insbesondere jüngere Personen verlassen die SOEP-Stichprobe vorzeitig. Als eine Konsequenz daraus resultiert, daß sich die Besetzung der Altersgruppen im Verlaufe des Panels nicht nur dadurch verändert, daß die Stammpersonen älter werden und in einigen Fällen sterben; sondern es kommt als ein signifikanter Faktor hinzu, daß insbesondere jüngere Personen überdurchschnittlich schnell ausscheiden. (c) Ein weiterer wichtiger Faktor ist die Haushaltsgröße. Vor allem bei 1-Personen-Haushalten gibt es eine überdurchschnittlich hohe Wahrscheinlichkeit für ein vorzeitiges Verlassen der SOEP-Stichprobe. Zahlreiche unterschiedliche Ursachen können dafür vermutet werden, zum Beispiel eine überdurchschnittlich hohe räumliche Mobilität der alleinlebenden Personen. Sicherlich spielt auch die Tatsache eine Rolle, daß die Entscheidung zum Verlassen des Panels im Haushaltskontext getroffen wird; in den meisten Fällen scheiden ganze Haushalte aus der SOEP-Stichprobe aus. (d) Man könnte vermuten, daß auch das Bildungsniveau eine gewisse Rolle spielt. Zumindest in der Teilstichprobe A bestätigt sich jedoch diese Vermutung nicht. Dagegen weisen die Personen aus der Teilstichprobe B, die nicht über eine abgeschlossene Schulausbildung verfügen, eine deutlich höhere Wahrscheinlichkeit für ein vorzeitiges Verlassen des Panels auf. (e) Bei den Personen aus der Teilstichprobe B könnte man vermuten, daß auch ihre Aufenthaltsdauer in der BRD eine gewisse Rolle spielt. Dies ist tatsächlich der Fall; und zwar ist die Wahrscheinlichkeit für ein vorzeitiges Verlassen des Panels bei denjenigen Personen besonders groß, die erst relativ spät (etwa in dem Jahrzehnt, das der SOEP-Stichprobe unmittelbar vorausgegangen ist) in die BRD gekommen sind. Dem entspricht, daß diejenigen Personen das Panel besonders schnell verlassen, deren geplante Aufenthaltsdauer in der BRD besonders kurz ist (bis 12 Monate).

Tabelle 2.2.1 Dauer der Teilnahme (bis zum ersten Ausscheiden) von Stammpersonen am SOEP. Nur Befragungspersonen, differenziert nach den Teilstichproben A und B und nach einigen Gründen für das Ausscheiden. Ausgeschieden in Welle t bedeutet, daß in Welle $t+1$ keine Teilnahme mehr stattgefunden hat.

		Welle								
Stichprobe A		1	2	3	4	5	6	7	8	9
Personen	N	9076	7997	7231	6781	6281	5830	5477	5211	4945
ausgeschieden	N	1079	766	450	500	451	353	266	266	—
	%	11.9	9.6	6.2	7.4	7.2	6.1	4.9	5.1	—
davon										—
- Verstorben		83	92	67	87	71	64	69	61	—
- Umzug Ausland		24	10	6	8	9	5	12	5	—
- Umzug Inland		96	74	63	79	53	58	30	33	—
- Sonstiges		876	590	314	326	318	226	155	167	—
Stichprobe B		1	2	3	4	5	6	7	8	9
Personen	N	3169	2566	2254	2083	1891	1733	1614	1514	1423
ausgeschieden	N	603	312	171	192	158	119	100	91	—
	%	19.0	12.2	7.6	9.2	8.4	6.9	6.2	6.0	—
davon										—
- Verstorben		8	4	3	2	3	5	3	4	—
- Umzug Ausland		192	84	53	65	44	38	25	21	—
- Umzug Inland		36	19	22	12	21	14	13	15	—
- Sonstiges		367	205	93	113	90	62	59	51	—

b) Zweitens sind die durch das SOEP verfügbaren Daten unvollständig, weil nur ein zeitlich beschränkter Ausschnitt aus den Lebensverläufen der Stichprobenmitglieder beobachtet werden kann. Da das SOEP eine Retrospektivbefragung mit einer Panelerhebung kombiniert, gibt es dafür zwei Gründe. Einerseits ist die Beobachtungsdauer davon abhängig, wann eine Person geboren worden ist; andererseits ist sie auch von der Anzahl der Wellen abhängig, in denen eine Teilnahme am Panel stattgefunden hat. Abbildung 2.2.3 zeigt die Verteilung der Beobachtungsdauern für die 8663 Personen aus der hier gewählten Teilstichprobe. Die Darstellung erfolgt in der Form einer Survivorfunktion. Alle Personen können mindestens bis zu ihrem 19. Lebensjahr beobachtet werden (da unsere Stichprobenabgrenzung eine Teilnahme an mindestens den ersten drei Wellen verlangt), mit länger werdender Beobachtungsdauer wird dann die Anzahl der Stichprobenmitglieder immer kleiner.

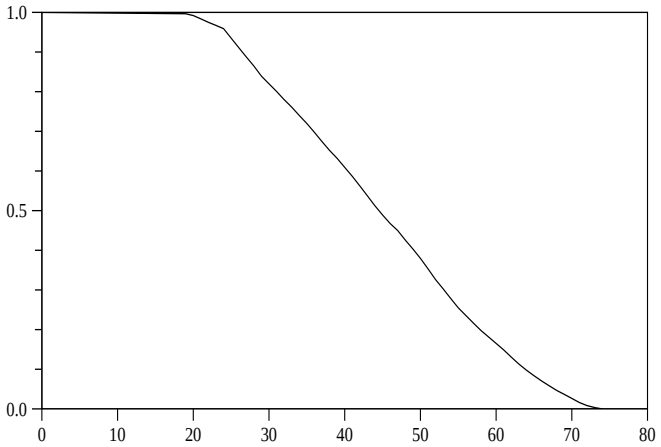


Abbildung 2.2.3 Beobachtungsdauer der SOEP-Stammpersonen, dargestellt durch eine Survivorfunktion. Abgrenzung: alle Personen, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen des SOEP teilgenommen haben. Abszisse: Beobachtungsdauer (Alter) in Jahren.

Die Beobachtungsdauer hängt in erster Linie vom Geburtsjahr ab. Um dies sichtbar zu machen, zeigt Abbildung 2.2.4 die Beobachtungsdauer differenziert nach Geburtskohorten. Jede Geburtskohorte wird durch eine Linie dargestellt. Der durchgezogene Teil der Linie repräsentiert die minimale, der gestrichelte Teil die maximale Beobachtungsdauer für jede Geburtskohorte.

c) Ein dritter Aspekte betrifft schließlich die Frage der Genauigkeit bzw. Zuverlässigkeit der verfügbaren Daten. Wie bereits erwähnt worden

ist, verwendet das SOEP zwei unterschiedliche Einheiten, um Ereigniszeitpunkte zu bestimmen. Für Ereignisse, die während der Laufzeit des Panels stattfinden, werden Monatsangaben erhoben, für weiter zurückliegende Ereignisse, insbesondere für die retrospektive Erhebung von Familienbiographien, werden Jahresangaben verwendet. Außerdem ist zu vermuten, daß Zeitangaben für in der Vergangenheit liegende Ereignisse mit Erinnerungsfehlern behaftet sind.[18]

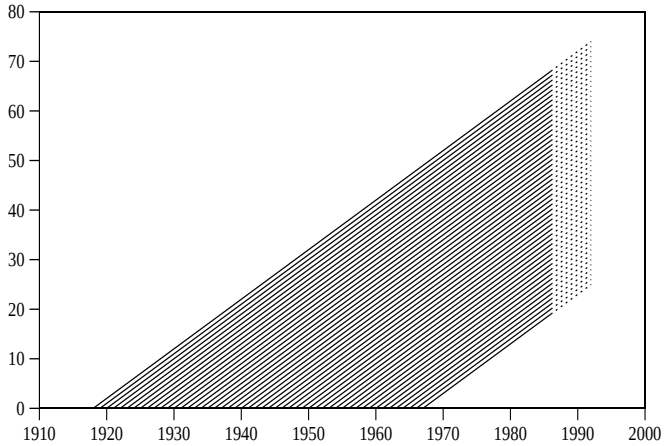


Abbildung 2.2.4 Beobachtungsdauer der SOEP-Stammpersonen, differenziert nach Geburtsjahrgängen. Abgrenzung: alle Personen, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen des SOEP teilgenommen haben. Abszisse: Kalenderzeit, Ordinate: Beobachtungsdauer (Alter) in Jahren.

[18] Zu dieser Frage existiert eine inzwischen breite Literatur, vgl. u.a. Featherman [1980], Bernard et al. [1984], Brückner [1990], Janson [1990]. In dieser dieser Diskussion ist u.a. festgestellt worden, daß die Genauigkeit von Erinnerungen in erster Linie von der (subjektiven) Bedeutung der Ereignisse, weniger vom Ausmaß ihres zeitlichen Zurückliegens abhängig ist.

2.3 Zufallsvariablen und Wahrscheinlichkeitsaussagen

Fast alle Begriffsbildungen der Statistik beziehen sich, direkt oder indirekt, auf die Vorstellung von Zufallsvariablen. Dementsprechend wird bei der statistischen Modellierung individueller Lebensverläufe häufig unmittelbar davon ausgegangen, daß es sich bei den Zuständen, in denen sich die Individuen befinden, um Realisationen von Zufallsvariablen handelt. Zum Beispiel heißt es bei Bartholomew [1977, S. 145]: „For example, individuals in human populations may change in regard to their income, residence, habits and opinions to mention only a few possibilities. The variables thus have to be thought of as depending on time in a probabilistic way. In more technical terms the population now has to be viewed as a collection of individual random, or stochastic, processes." Es ist jedoch nicht unmittelbar klar, wie Aussagen dieser Art verstanden werden können und welche Auffassung der Individuen und ihrer gesellschaftlichen Verhältnisse durch sie zum Ausdruck gebracht werden soll. Das übliche Vorverständnis – „intuitively, a random variable is any uncertain quantity to which one is willing to attach probability statements"[19] – ist offenbar an dieser Stelle nicht sehr hilfreich. Insbesondere sollte überlegt werden, in welcher Weise mit dem Begriff „Zufallsvariable" auf Gesamtheiten von Ereignissen oder Sachverhalten Bezug genommen wird.

Zuständig für die Definition dieses Begriffs ist zunächst die Wahrscheinlichkeitstheorie. Allerdings beschränkt sie sich auf eine rein formale Definition ihrer Grundbegriffe. Seit ihre durch Kolmogorow [1933] vorgeschlagene Axiomatisierung allgemein anerkannt worden ist, ist sie zu einem Teilgebiet der Mathematik geworden.[20] Ihre weitere Entwicklung konnte dadurch von den bis heute anhaltenden Kontroversen über die Interpretation ihrer Grundbegriffe abgekoppelt werden.[21] Insofern bietet die Wahrscheinlichkeitstheorie einen festen, aber nur begrenzt hilfreichen Ausgangspunkt. Um die von ihr geschaffenen Begriffe und Überlegungen für empirische Anwendungen nutzbar zu machen, bedarf es einer jeweils spezifischen Interpretation. Die Wahrscheinlichkeitstheorie und die darauf aufbauende statistische Theorie liefern einen formalen Rahmen; der *empirische* Sinn der in diesem Rahmen formulierbaren Aussagen kann nur durch eine Bezugnahme auf ein jeweils spezifisches Anwendungsfeld expliziert werden.

Um dies etwas deutlicher zu machen, beginne ich mit einer kurzen Erinnerung an die axiomatischen Grundlagen der Wahrscheinlichkeitstheorie.[22] Ausgangspunkt ist der Begriff eines Wahrscheinlichkeitsraums $(\Omega, \mathcal{A}, \mathrm{P})$, der aus drei Komponenten besteht:

1. Es gibt eine Basismenge Ω, die aus einer beliebigen Anzahl von zunächst nicht näher spezifizierten Elementen besteht.
2. Weiterhin gibt es ein Mengensystem $\mathcal{A}$, das aus Teilmengen von Ω besteht und folgenden Bedingungen genügt: $\Omega \in \mathcal{A}$; wenn $A \in \mathcal{A}$, dann auch das Komplement von A, also $\Omega - A \in \mathcal{A}$; für jede beliebige Folge (A_i) aus $\mathcal{A}$ gilt, daß auch ihre Vereinigungsmenge in $\mathcal{A}$ enthalten ist, also $A_1 \cup A_2 \cup \ldots \in \mathcal{A}$. Das Mengensystem $\mathcal{A}$ wird dann als eine σ-Algebra bezeichnet.
3. Schließlich gibt es ein Wahrscheinlichkeitsmaß P, das jedem Element $A \in \mathcal{A}$ eine reelle Zahl $\mathrm{P}(A)$ zuordnet und folgenden Bedingungen genügt:

$$\mathrm{P}(A) \geq 0 \qquad \text{für jedes } A \in \mathcal{A}$$
$$\mathrm{P}(\Omega) = 1$$
$$\mathrm{P}(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

für jede Folge paarweise disjunkter Mengen $A_i \in \mathcal{A}$.

Ist ein solcher Wahrscheinlichkeitsraum gegeben, wird $\mathrm{P}(A)$ als Wahrscheinlichkeit der Menge $A \in \mathcal{A}$ bezeichnet. Offensichtlich handelt es sich um eine rein formale Definition. Um die Wahrscheinlichkeitstheorie für empirische Forschungen nutzbar zu machen, muß man sich über den empirischen Bedeutungsgehalt der zunächst bloß formal fixierten Grundbegriffe verständigen.

Ob und wie empirisch gehaltvolle Definitionen des Wahrscheinlichkeitsbegriffs gegeben werden können, ist umstritten.[23] In dieser Arbeit gehe ich davon aus, daß Wahrscheinlichkeitsaussagen in unterschiedlichen Kontexten formuliert werden können und daß ihre Bedeutung vom jeweiligen Kontext abhängig ist.[24] Ich unterscheide drei Arten von Wahrscheinlichkeitsaussagen: deskriptive Wahrscheinlichkeitsaussagen, einzelfallbezogene Wahrscheinlichkeitsaussagen und Aussagen über Zufallsgeneratoren.

[19] Pratt und Gibbons [1981, S. 3].

[20] „Der leitende Gedanke des Verfassers – schreibt Kolmogoroff [1933, S. III] – war dabei, die Grundbegriffe der Wahrscheinlichkeitsrechnung, welche noch unlängst für ganz eigenartig galten, natürlicherweise in die Reihe der allgemeinen Begriffsbildungen der modernen Mathematik einzuordnen."

[21] Eine informative Darstellung einiger historischer Umstände der Axiomatisierung der Wahrscheinlichkeitstheorie findet sich bei Maistrov [1974, Kap. V].

[22] Ausführliche Darstellungen finden sich in zahlreichen Büchern zur Wahrscheinlichkeitstheorie und theoretischen Statistik. Eine mathematisch orientierte Einführung gibt z.B. Bauer [1978], eine Einführung anhand zahlreicher Beispiele findet sich bei Pfeiffer [1978].

[23] Einen breit angelegten Überblick gibt Fine [1973]. Einführungen in die unterschiedlichen Auffassungen geben u.a. Salmon [1966], Gottinger [1980].

[24] Die Auffassung, daß es nicht nur einen, sondern mehrere unterschiedliche Wahrscheinlichkeitsbegriffe gibt, wurde zuerst von Carnap [1945] betont; dann hat Mackie [1973]

2.3.1 Deskriptive Wahrscheinlichkeitsaussagen

Eine sehr einfache Interpretation von Wahrscheinlichkeitsaussagen ist dann möglich, wenn die Basismenge Ω endlich ist. Wahrscheinlichkeitsaussagen können dann unmittelbar als Aussagen über relative Häufigkeiten verstanden werden. Diese Interpretation bietet sich insbesondere an, um die mithilfe statistischer Verfahren und Modelle in der Lebensverlaufsforschung gewinnbaren Aussagen zu interpretieren. Denn in diesem Kontext beziehen wir uns stets auf eine endliche Gesamtheit von Individuen, die wir uns als „Träger“ gesellschaftlicher Verhältnisse vorstellen. Als Interpretation des die weitere Modellbildung konstituierenden Wahrscheinlichkeitsraums ergibt sich dann:

1. Die Basismenge Ω ist eine gegebene, endliche Grundgesamtheit von Individuen, wodurch (implizit oder explizit) zugleich ein bestimmter Raum-Zeit-Bezug hergestellt wird.
2. Die σ-Algebra $\mathcal{A}$ ist die Menge aller Teilmengen von Ω.
3. Als Wahrscheinlichkeitsmaß wird eine Gleichverteilung angenommen, also

$$\mathrm{P}(A) = \frac{|A|}{|\Omega|} \quad \text{für} \quad A \in \mathcal{A} \tag{2.1}$$

Ich nenne dies im folgenden einen *deskriptiven Wahrscheinlichkeitsraum für die Grundgesamtheit* Ω.[25] Das Wahrscheinlichkeitsmaß läßt sich in diesem Fall auf einfache Weise interpretieren: $\mathrm{P}(A)$ ist der Anteil der Mitglieder der Menge A an der Gesamtheit der Individuen in der Menge Ω. Diese einfache deskriptive Interpretation von Wahrscheinlichkeitsaussagen ist möglich, weil es sich um eine endliche Grundgesamtheit von Individuen handelt. Dies ermöglicht es, die Wahrscheinlichkeit eines Merkmals mit seiner relativen Häufigkeit in der endlichen Grundgesamtheit von Individuen zu identifizieren. Wie man sich leicht überzeugt, erfüllt die in (2.1) angegebene Definition die oben angegebenen Bedingungen für Wahrscheinlichkeitsmaße.

Zufallsvariablen

Im Rahmen der Wahrscheinlichkeitstheorie dienen Zufallsvariablen zur Repräsentation von Merkmalen (empirisch feststellbaren Eigenschaften) der

vorgeschlagen, daß mindestens fünf verschiedene Kontexte für Wahrscheinlichkeitsaussagen unterschieden werden sollten. Im Hinblick auf *statistische* Aussagen hat vor allem Kyburg [1980] sinnvolle Unterscheidungen vorgeschlagen, die den Charakter der Grundgesamtheit betreffen, auf die eine Bezugnahme erfolgt.

[25] In der Literatur wird auch die Bezeichnung *Laplacescher Wahrscheinlichkeitsraum* verwendet, um den Zusammenhang zur klassischen Wahrscheinlichkeitsdefinition zu betonen; vgl. etwa Bauer [1978, S. 131].

Elemente der Basismenge Ω. Eine Zufallsvariable X wird formal als eine meßbare Abbildung von Ω in einen geeigneten Meßraum definiert. In den meisten praktischen Anwendungen genügt es, als Meßraum die Menge der reellen Zahlen $\mathbf{R}$ mit einer Borelschen σ-Algebra anzunehmen.[26] Eine Zufallsvariable X ist dann eine Abbildung

$$X : \Omega \longrightarrow \mathbf{R} \tag{2.2}$$

so daß für alle Borelschen Teilmengen $M \subseteq \mathbf{R}$ gilt:

$$\{\omega \mid X(\omega) \in M\} \in \mathcal{A} \tag{2.3}$$

$X(\omega)$ ist eine reelle Zahl, die eine bestimmte, am Element $\omega \in \Omega$ festgestellte Eigenschaft repräsentiert. Die in (2.3) angegebene Meßbarkeitsbedingung besagt, daß die Menge der Individuen $\omega \in \Omega$, für die das durch X gemessene Merkmal in M liegt, im Rahmen des zugrundeliegenden Wahrscheinlichkeitsraums existiert, so daß dort von der Wahrscheinlichkeit dieser Menge (von Individuen) gesprochen werden kann.

Ersichtlich ist dieses Konzept sehr allgemein und kann zum Beispiel verwendet werden, um die Ergebnisse von Zufallsexperimenten zu beschreiben. Diese Standardinterpretation motiviert die Wortwahl: *Zufallsvariable*. Der Begriff kann jedoch gleichermaßen verwendet werden, um im Rahmen eines deskriptiven Wahrscheinlichkeitsraums die Eigenschaften der Mitglieder einer Grundgesamtheit von Individuen zu beschreiben. Zum Beispiel kann eine Zufallsvariable X_1 gebildet werden, um das Geburtsjahr der Individuen aus Ω zu erfassen. Ganz analog können auch nicht-numerische Eigenschaften erfaßt werden, indem man ihnen einen numerischen Kode gibt. Um das Geschlecht der Individuen aus Ω zu erfassen, kann z.B. eine Zufallsvariable X_2 definiert werden, die bei Männern den Wert 1 und bei Frauen den Wert 2 annimmt.

Die wahrscheinlichkeitstheoretische Definition verlangt, daß Zufallsvariablen *meßbar* sein müssen. Dadurch wird erreicht, daß sich Wahrscheinlichkeitsaussagen auch für die Ergebnisse von Zufallsvariablen formulieren lassen. Diese Wahrscheinlichkeitsaussagen haben die Form

$$\mathrm{P}\left(X \in M\right) \quad \text{für} \quad M \subseteq \mathrm{R}$$

womit gemeint ist: die Wahrscheinlichkeit, daß die Zufallsvariable X einen Wert in der Menge M annimmt.[27] Wenn X meßbar ist, kann diese Wahr-

[26] Eine genaue Definition dieser σ-Algebra ist hier nicht erforderlich. Es genügt die Vorstellung, daß es sich um ein System von Teilmengen aus der Menge der reellen Zahlen handelt, das alle praktisch relevanten Teilmengen enthält, insbesondere die einelementigen Teilmengen und offene, geschlossene und halboffene Intervalle.

[27] Ich verwende in dieser Arbeit die Bezeichnung $\mathrm{P}(\cdot)$ für Wahrscheinlichkeitsaussagen über Zufallsvariablen, die auf der Grundlage eines deskriptiven Wahrscheinlichkeitsraums definiert sind. Später, bei der Beschreibung von Stichprobendesigns, wird eine andere Notation verwendet, da es sich dort um Wahrscheinlichkeitsaussagen eines anderen Typs handelt.

scheinlichkeit durch das im zugrundeliegenden Wahrscheinlichkeitsraum verfügbare Wahrscheinlichkeitsmaß definiert werden:[28]

$$\mathrm{P}\,(X \in M) = \mathrm{P}(\,\{\omega \mid X(\omega) \in M\}\,)$$

Die Wahrscheinlichkeit, daß die Zufallsvariable X einen Wert in M annimmt, ist dadurch zurückgeführt auf die Wahrscheinlichkeit derjenigen Teilmenge von Ω, bei deren Elementen die Variable X einen Wert in M annimmt.

Geht man von Zufallsvariablen aus, die in einem deskriptiven Wahrscheinlichkeitsraum für eine endliche Grundgesamtheit von Individuen definiert sind, erhält man wiederum eine sehr einfache empirische Interpretation für Wahrscheinlichkeitsaussagen über Zufallsvariablen, nämlich:

$$\mathrm{P}\,(X \in M) = \frac{|\{\omega \mid X(\omega) \in M\}|}{|\Omega|}$$

Die Wahrscheinlichkeit, daß die Zufallsvariable X einen Wert in M annimmt, ist der Anteil der Personen aus der Grundgesamtheit Ω, bei denen die Messung von X einen Wert in M ergibt. Wenn zum Beispiel X das Geburtsjahr bedeutet, ist $\mathrm{P}(X < 1960)$ der Anteil derjenigen Personen aus der Grundgesamtheit Ω, die vor 1960 geboren sind.

Eine wichtige Konsequenz aus der wahrscheinlichkeitstheoretischen Definition des Begriffs *Zufallsvariable* liegt darin, daß sich dieser Begriff stets auf eine Gesamtheit von vergleichbaren Ereignissen oder Sachverhalten bezieht. Zufallsvariablen sind Funktionen, Abbildungen einer Grundgesamtheit Ω in einen Meßraum. Sie dürfen also nicht mit gewöhnlichen Variablen verwechselt werden, wie sie in der Mathematik verwendet werden. Terminologisch unterscheidet man deshalb eine Zufallsvariable X (als Funktion) von den Werten (Realisierungen) $X(\omega)$, die sie bei bestimmten Elementen ω aus der Grundgesamtheit annimmt.[29] Geht man von einem deskriptiven Wahrscheinlichkeitsraum aus, ist der Sinn dieser Begriffsbildung unmittelbar einsichtig. Die Definition von Zufallsvariablen dient dann dazu, um von Aussagen über Individuen (individuelle Ereignisse oder Sachverhalte) zu Aussagen über eine Gesamtheit zu gelangen, der diese Individuen angehören.

Diese Methode kann als *deskriptive Abstraktion* bezeichnet werden. Aussagen über eine Zufallsvariable sind Aussagen über ihre *Wahrscheinlichkeitsverteilung* (kurz: *Verteilung*). Man kennt eine Zufallsvariable dann vollständig, wenn man die Wahrscheinlichkeiten kennt, mit denen sie ihre

[28] Da wir von einem deskriptiven Wahrscheinlichkeitsraum mit einer endlichen Basismenge Ω ausgehen, ist die Meßbarkeitsvoraussetzung stets erfüllt.

[29] Um diese Unterscheidung zu erleichtern, werden Zufallsvariablen meistens mit Großbuchstaben bezeichnet: X, Y, Z usw. (ggf. mit Indizes versehen), ihre Realisierungen mit den korrespondierenden Kleinbuchstaben: x, y, z usw.

möglichen Werte annimmt. Offensichtlich wird dabei von den Individuen, die der Grundgesamtheit angehören, auf eine spezifische Weise abstrahiert. Wird das durch eine Zufallsvariable zu erfassende Merkmal bei zwei Individuen ω_i und ω_j vertauscht, ändert sich ihre Verteilung nicht. Man könnte sagen, daß bei der Beschreibung von Individuen durch Zufallsvariablen gewissermaßen von ihrer Identität abstrahiert wird. In einer Formulierung von Lexis [1875, S. 1]: „Bei der Bildung von Massen für die statistische Beobachtung verschwindet das Individuum als solches, und es erscheint nur noch als eine Einheit in einer Zahl von gleichartigen Gliedern, die gewisse Merkmale gemein haben und von deren sonstigen individuellen Unterschieden abstrahirt wird.“ Man kann dies so zusammenfassen: Insofern Wahrscheinlichkeitsaussagen über Zufallsvariablen Aussagen über deren Verteilung sind, beziehen sie sich nicht auf einzelne Individuen, sondern auf Gesamtheiten von Individuen.[30]

Wegen der grundlegenden Bedeutung von Zufallsvariablen für den gesamten theoretischen Aufbau der Wahrscheinlichkeitstheorie und Statistik gibt es zahlreiche ergänzende Begriffsbildungen. Einige, auf die im folgenden immer wieder zurückgegriffen wird, seien kurz genannt.

a) Bei der formalen Definition einer Zufallsvariable kann man sich darauf beschränken, sie als eine Abbildung in die Menge der reellen Zahlen zu charakterisieren, vgl. (2.2), und offenlassen, welche Werte sie tatsächlich annehmen kann. Um die Charakterisierung von Zufallsvariablen zu vereinfachen, ist es häufig nützlich, sich nur auf diejenigen Werte zu beziehen, die eine Zufallsvariable tatsächlich annehmen kann. Ich nenne dies im folgenden ihren *Wertebereich* und verwende zur Notation in der Regel $\mathcal{X}$ für den Wertebereich der Zufallsvariable X, $\mathcal{Y}$ für den Wertebereich der Zufallsvariable Y, usw.

b) Innerhalb eines Wahrscheinlichkeitsraums können mehrere, beliebig viele Zufallsvariablen definiert werden. Zum Beispiel kann bei einer Menge von Individuen mithilfe von Zufallsvariablen das Geburtsjahr, das Geschlecht, der Erwerbsstatus zu einem bestimmten Zeitpunkt t_1, der Erwerbsstatus zu einem anderen Zeitpunkt t_2, und der Wohnort erfaßt werden; in formaler Schreibweise:

$$X_j : \Omega \longrightarrow \mathcal{X}_j \qquad \mathcal{X}_| \subseteq \mathbf{R} \qquad j = 1, \ldots, m$$

zur Definition von m Zufallsvariablen. Um zum Ausdruck zu bringen, daß sich diese Zufallsvariablen auf denselben Wahrscheinlichkeitsraum beziehen, wird dieses System von m Zufallsvariablen häufig eine m-dimensionale Zufallsvariable genannt.

c) Die Verteilung einer Zufallsvariable kann vollständig durch ihre *Verteilungsfunktion* beschrieben werden. Für eine Zufallsvariable X mit dem Wertebereich $\mathcal{X}$ ist die Verteilungsfunktion folgendermaßen definiert (ich verwende, wie allgemein üblich, die Schreibweise $[a, b]$ als Bezeichnung für das abgeschlossene Intervall aller reellen Zahlen von a bis b):

$$F_X : \mathcal{X} \longrightarrow [0, 1] \quad \text{wobei} \quad F_X(x) = \mathrm{P}\,(X \leq x) \tag{2.4}$$

[30] Die Abstraktion von den Individuen beginnt bereits bei der Datenerhebung. Üblicherweise findet eine Anonymisierung statt. Das *International Statistical Institute* [1986, S. 238] hat dies in einer „Declaration of Professional Ethics“ folgendermaßen formuliert: „Statistical data are unconcerned with individual identities. They are collected to answer questions such as 'how many?' or 'what proportions?', not 'who?'. The identities and records of co-operating (or non-cooperating) subjects should therefore be kept confidential, whether or not confidentiality has been explicitly pledged.“

Analog wird die Verteilungsfunktion bei mehrdimensionalen Zufallsvariablen durch

$$F_{X_1,\ldots,X_m} : \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \longrightarrow [0,1]$$
$$\text{wobei} \quad F_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = \mathrm{P}\,(X_1 \leq x_1, \ldots, X_m \leq x_m)$$

definiert (der Ausdruck $\mathcal{X}_1 \times \ldots \times \mathcal{X}_m$ bezeichnet das kartesische Produkt der Wertebereiche $\mathcal{X}_j$, d.h. die Menge aller geordneten Kombinationen von Werten aus $\mathcal{X}_1, \ldots, \mathcal{X}_m$); sie gibt die Wahrscheinlichkeit an, daß gleichzeitig (bei einem Element $\omega \in \Omega$) die Zufallsvariable X_1 einen Wert $\leq x_1$, die Zufallsvariable X_2 einen Wert $\leq x_2$, usw. annimmt. Man kennt dann die *gemeinsame Verteilung* dieser Zufallsvariablen in der Grundgesamtheit.

d) Es ist üblich, diskrete und stetige Zufallsvariablen zu unterscheiden. Von einer diskreten Zufallsvariable spricht man, wenn ihr Wertebereich endlich (oder abzählbar unendlich) ist; von einer (absolut) stetigen Zufallsvariable spricht man, wenn ihre Verteilungsfunktion durch ein Integral einer stetigen Dichtefunktion ausgedrückt werden kann. Bei einer stetigen Zufallsvariable X, deren Wertebereich aus der Gesamtheit der reellen Zahlen besteht, könnte also die in (2.4) definierte Verteilungsfunktion folgendermaßen geschrieben werden:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$$

f_X wird dann als ihre Dichtefunktion bezeichnet. Strenggenommen sind alle Zufallsvariablen, die innerhalb eines deskriptiven Wahrscheinlichkeitsraums definiert werden können, diskret, da bei einer endlichen Menge von Individuen nur eine endliche Anzahl unterschiedlicher Merkmale auftreten kann. Gelegentlich, wenn eine Zufallsvariable sehr viele unterschiedliche Werte annehmen kann, ist es jedoch sinnvoll, stetige Zufallsvariablen als vereinfachende Approximationen für diskrete Zufallsvariablen zu verwenden. In der vorliegenden Arbeit werden stetige Zufallsvariable grundsätzlich nur unter dem Aspekt einer Approximation der Verteilung diskreter Zufallsvariablen betrachtet; dazu bemerkt Kolmogoroff [1933, S. 14]: „Bei einer Beschreibung irgendwelcher wirklich beobachtbarer zufälliger Prozesse kann man nur endliche Wahrscheinlichkeitsfelder erhalten. Unendliche Wahrscheinlichkeitsfelder erscheinen nur als idealisierte Schemata reeller zufälliger Prozesse. Wir beschränken uns dabei willkürlich auf solche Schemata, welche dem Stetigkeitsaxiom [der σ-Additivität] genügen."

Als ein Beispiel für deskriptive Wahrscheinlichkeitsaussagen über Zufallsvariable beziehe ich mich auf einen Aspekt der in Abschnitt 2.2 beschriebenen Familienbiographien: das Alter bei der ersten Heirat. Formal kann dies als eine Zufallsvariable betrachtet werden, und ihre Verteilung kann zum Beispiel durch eine Survivorfunktion beschrieben werden. Um dies zu präzisieren, muß zunächst ein deskriptiver Wahrscheinlichkeitsraum angegeben werden, insbesondere die Basismenge Ω, auf die wir uns beziehen wollen. Wie die Definition vorgenommen werden sollte, hängt sowohl von den intendierten Aussagen als auch von den jeweils verfügbaren Daten ab; darauf wird in Abschnitt 2.4.1 näher eingegangen. An dieser Stelle nehmen wir an, daß zunächst nur eine Aussage über die verfügbaren Daten intendiert ist, also für die in der SOEP-Stichprobe tatsächlich erfaßten Individuen. Da es sich um eine geschichtete Stichprobe handelt, beziehen wir uns nur auf die Teilstichprobe A. Es kann dann folgende Zufallsvariable

definiert werden:

$$T : \Omega_A \longrightarrow \mathcal{T}$$

T ist das Alter bei der ersten Heirat, gemessen auf einer Zeitachse $\mathcal{T}$ (in Monaten bzw. Jahren). Ω_A ist die Gesamtheit der Individuen aus der Teilstichprobe A des SOEP, wobei von den in Abschnitt 2.2 (S. 50) beschriebenen Abgrenzungen ausgegangen wird; insgesamt $N = 6416$ Personen. Dies vorausgesetzt, kann die Verteilung dieser Zufallsvariablen beschrieben werden. Dafür gibt es verschiedene, formal äquivalente Möglichkeiten. Zur Beschreibung von Zufallsvariablen, die die Zeitdauer bis zum Eintreten eines Ereignisses repräsentieren, ist es häufig am anschaulichsten, ihre Verteilung durch eine Survivorfunktion zu beschreiben, die folgendermaßen definiert ist:

$$G(t) = \mathrm{P}\,(T > t) \qquad t \in \mathcal{T}$$

d.h. durch eine Funktion, die für jeden möglichen Zeitpunkt t die Wahrscheinlichkeit dafür angibt, daß die Zufallsvariable T einen Wert größer als t annimmt. In unserem Beispiel ist T die von der Geburt an gemessene Zeitdauer bis zur ersten Heirat, $G(t)$ ist also die Wahrscheinlichkeit, daß bis zu einem Alter von t (Jahren, Monaten) noch keine Heirat stattgefunden hat; deskriptiv interpretiert: der Anteil der Personen aus Ω_A, die bis zum Alter t noch nicht geheiratet haben.

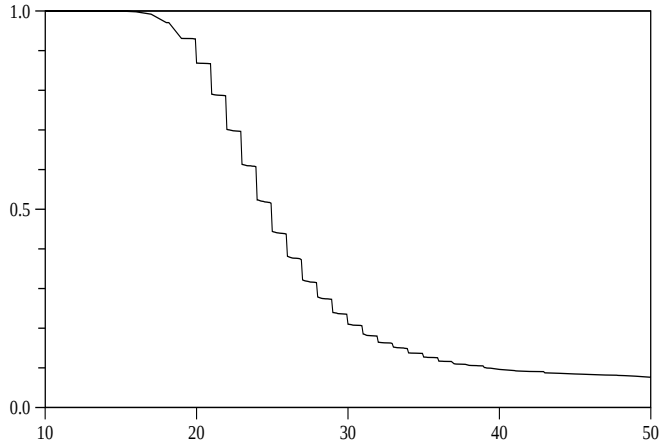


Abbildung 2.3.1 Survivorfunktion für das Alter bei der ersten Heirat. Kaplan-Meier-Schätzung für die Stammpersonen des SOEP, die der Teilstichprobe A angehören, im Zeitraum von 1918 – 1968 geboren wurden und an mindestens den ersten drei Wellen des SOEP teilgenommen haben ($N = 6416$). Abszisse: Alter in Jahren.

Abbildung 2.3.1 zeigt eine Schätzung dieser Survivorfunktion.[31] Die Abszisse zeigt das Alter, die Ordinate den Anteil der noch unverheirateten Personen.

Was für eine Information erhält man durch die Kenntnis einer solchen Survivorfunktion? Zunächst liefert sie eine deskriptive Information über das Heiratsverhalten der in Ω_A erfaßten Personen. Betrachtet man Ω_A als eine Stichprobe aus einer größeren Grundgesamtheit, kann man einen Schritt weiter gehen und die anhand von Ω_A gewonnene Survivorfunktion als *Schätzung* einer für diese Grundgesamtheit definierten Survivorfunktion betrachten. Da es sich um eine endliche Grundgesamtheit handelt, kann der für sie geschätzten Survivorfunktion immer noch eine deskriptive Interpretation gegeben werden: sie beschreibt das Heiratsverhalten der Individuen in dieser Grundgesamtheit.

Die Konstruktion der in Abbildung 2.3.1 gezeigten Survivorfunktion liefert ein einfaches Beispiel dafür, wie mithilfe statistischer Methoden aus einer komplexen Menge an Daten eine überschaubare Information gewonnen werden kann. Wichtig ist, daß damit die Information gewissermaßen ihren Charakter geändert hat. Die zunächst verfügbaren Daten liefern eine Information über individuelle Lebensverläufe. Nachdem diese komplexen und im einzelnen unüberschaubaren Informationen zu einer Survivorfunktion verdichtet worden sind, handelt es sich nicht länger um eine Information über konkrete, jeweils bestimmte Individuen, sondern um eine Aussage über „die Individuen“ in einer Gesamtheit von Individuen. Insofern hat sich der Gegenstand der Beschreibung verändert. An die Stelle einer Menge konkreter, jeweils bestimmter Individuen ist eine Gesamtheit von Individuen getreten. Die in einem deskriptiven Wahrscheinlichkeitsraum formulierbaren Wahrscheinlichkeitsaussagen beziehen sich in ihrer Bedeutung auf Eigenschaften einer Gesamtheit von Individuen, nicht auf Eigenschaften der je individuellen Mitglieder der Gesamtheit.[32]

Dem üblichen Sprachgebrauch folgend kann man diese Art von Beschreibungen von Gesamtheiten von Individuen auch als Beschreibungen ihrer gesellschaftlichen Verhältnisse ansehen. Im Hinblick auf eine soziologische Theoriebildung erscheint es jedoch sinnvoll, Gesamtheiten von Individuen und ihre gesellschaftlichen Verhältnisse begrifflich zu unterscheiden. Gesamtheiten von Individuen entstehen durch eine begriffliche Zusammenfassung ihrer Elemente. Ihre Existenz ist abhängig von der Existenz ihrer Mitglieder. Sie verändern sich, wenn sich ihre Mitglieder verändern. Aus der Perspektive der Lebensverlaufsforschung ist bereits die Formulierung

[31] Die Schätzung wurde mit dem Kaplan-Meier-Verfahren durchgeführt, um rechts zensierte Beobachtungen berücksichtigen zu können. Dies wird in Abschnitt 3.4.1 näher beschrieben.

[32] Für die ältere Sozialstatistik war dies, worauf bereits in der Einleitung hingewiesen wurde, selbstverständlich; aber auch noch, zum Beispiel, für R. A. Fisher, der etwa [1970, S. 1f] feststellt: „Nevertheless, in a real sense, statistics is the study of populations, or aggregates of individuals, rather than of individuals.“

Gesamtheit von Individuen problematisch; besser wäre es, von einer Gesamtheit von Lebensverläufen zu sprechen.

Demgegenüber ist der Begriff *gesellschaftliche Verhältnisse* eine theoretische Konstruktion. Er dient einer theoretischen Deutung der in einer Gesamtheit von Individuen beobachtbaren Lebensverläufe. Um diesen Begriff zu präzisieren, kann man versuchen, von der Vorstellung auszugehen, daß die Individuen sich in ihren Lebensverläufen an gewissen Regeln orientieren. Man kann sich dann gesellschaftliche Verhältnisse als Bedingungen individueller Lebensverläufe vorstellen, die in gewisser Weise unabhängig von den Individuen existieren. Damit ist nicht Durkheims oder Parsons' Konzeption sozialer Normen gemeint, sondern nur der einfache Sachverhalt, daß die jeweils vorgefundenen Regeln die individuellen Handlungsmöglichkeiten „strukturieren“ und daß niemand als Individuum soziale Regeln verändern kann.[33] Die Individuen können Regeln befolgen oder von ihnen abweichen, in gewisser Weise können sie sich auch dafür einsetzen, daß neue Regeln entstehen. Aber die schließlich resultierenden gesellschaftlichen Verhältnisse sind immer auch, mehr oder weniger, die nicht-intendierten Folgen kontingenter Lebensverläufe.

Formulierungen dieser Art, die für die soziologische Theoriebildung durchaus typisch sind, transzendieren offensichtlich den deskriptiven Gehalt statistischer Beschreibungen von Gesamtheiten von Individuen bzw. Lebensverläufen. Insofern sollte daran festgehalten werden, daß statistische Beschreibungen, die sich zunächst auf Gesamtheiten von Individuen beziehen, einer theoretischen Deutung bedürfen, um dadurch zu einer Darstellung gesellschaftlicher Verhältnisse zu gelangen.

2.3.2 Einzelfallbezogene Wahrscheinlichkeitsaussagen

Eine ganz andere Kategorie von Wahrscheinlichkeitsaussagen bezieht sich auf singuläre Ereignisse oder Sachverhalte; zum Beispiel: morgen wird es wahrscheinlich regnen, oder: die Wahrscheinlichkeit, mit einem gewissen Würfel beim nächsten Wurf eine gerade Zahl zu erzielen, beträgt 0.5. Über die Frage, ob und, wenn ja, in welcher Weise Wahrscheinlichkeitsaussagen dieser Art eine objektivierbare Bedeutung gegeben werden kann, gibt es in der Literatur eine breite Diskussion. Zumindest ein Aspekt von Aussagen dieser Art kann verstanden werden, wenn man sie auf die Tatsache bezieht, daß Menschen offenbar in der Lage sind, über unsichere Ereignisse oder Sachverhalte Erwartungen zu bilden. Man kann insoweit sagen, daß einzelfallbezogene Wahrscheinlichkeitsaussagen Erwartungen zum Ausdruck bringen, die sich ein Subjekt über das Eintreten von nicht sicher voraussagbaren Ereignissen macht, oder über die Beschaffenheit von Sachverhalten, über die nur unzureichende Informationen verfügbar sind.

[33] Eine kritische Diskussion des Verhältnisses von „norms and behavior“ findet sich bei Cancian [1976].

Eine wichtige Eigenart dieser Wahrscheinlichkeitsaussagen liegt darin, daß sie relational sind inbezug auf das *über den jeweiligen Einzelfall* verfügbare Wissen. Man sieht das, wenn man sich überlegt, in welcher Weise deskriptive Wahrscheinlichkeitsaussagen zur Begründung von einzelfallbezogenen Wahrscheinlichkeitsaussagen verwendet werden können. Um die Überlegung zu illustrieren, sei ω^* ein bestimmtes Individuum aus der Grundgesamtheit Ω. Die Frage ist, ob *dieses* Individuum eine gewisse Eigenschaft besitzt, die durch eine Teilmenge $A \subseteq \Omega$ charakterisiert werden kann. Wenn die deskriptive Wahrscheinlichkeit $|\,A\,|\,/\,|\,\Omega\,| = p$ bekannt ist, liegt es nahe, dies als eine Begründung für die Aussage anzusehen, daß ω^* mit der Wahrscheinlichkeit p ein Mitglied der Menge A ist.

Es ist aber durchaus möglich, daß noch weitere Informationen zur Verfügung stehen. Angenommen etwa, wir kennen die bedingte Wahrscheinlichkeit $\mathrm{P}(A \,|\, B) = q$ und wir wissen, daß $\omega^* \in B$. Man würde dann sagen, daß ω^* mit der Wahrscheinlichkeit q (und nicht mit der Wahrscheinlichkeit p) ein Mitglied der Menge A ist. Insofern sind Wahrscheinlichkeitsaussagen über singuläre Individuen (oder Ereignisse oder Sachverhalte) davon abhängig, was wir *jeweils über sie* wissen.

Ein zusätzliches Problem tritt dadurch auf, daß das verfügbare Wissen nicht unbedingt konsistente Schlußfolgerungen erlaubt.[34] Betrachten wir noch einmal das gerade angeführte Beispiel. Man könnte versuchen, die zugrunde liegende Schlußfolgerung folgendermaßen zu beschreiben:

> Es ist bekannt, daß $\mathrm{P}(A \mid B) = p$
>
> Es ist bekannt, daß $\omega^* \in B$
>
> Also gilt mit Wahrscheinlichkeit p, daß $\omega^* \in A$

Die Problematik dieser Schlußweise wird deutlich, wenn man zum Beispiel annimmt, daß folgende Informationen zur Verfügung stehen: $P(A \,|\, B) \leq 0.02$, $P(A \,|\, B') \geq 0.98$ und $\omega^* \in B \cap B'$. Dann gilt (in der oben verwendeten Ausdrucksweise) mit einer Wahrscheinlichkeit ≥ 0.98, daß $\omega^* \in A$, und ebenfalls mit einer Wahrscheinlichkeit ≥ 0.98, daß $\omega^* \notin A$. Dies ist offenbar ein Widerspruch, wenn mit der Formulierung *es gilt mit Wahrscheinlichkeit p, daß ...* eine Behauptung über die Wahrscheinlichkeit von Aussagen gemeint ist.

Soweit es sich um einen formalen Widerspruch handelt, kann er dadurch beseitigt werden, daß zu einer anderen Formulierung übergegangen wird:

> (a) Es ist bekannt, daß $\mathrm{P}(A \mid B) = p$
>
> (b) Es ist bekannt, daß $\omega^* \in B$
>
> (a) und (b) implizieren mit Wahrscheinlichkeit p, daß $\omega^* \in A$

Das angeführte Beispiel führt dann nicht mehr zu einem Widerspruch,

[34] Eine ausführliche Diskussion gibt Stegmüller [1983], Kapitel IX.

denn es ist durchaus möglich, daß ein statistischer Zusammenhang mit hoher Wahrscheinlichkeit impliziert, daß ein bestimmtes Individuum gewisse Eigenschaften besitzt, und daß ein anderer statistischer Zusammenhang mit ebenfalls hoher Wahrscheinlichkeit impliziert, daß das gleiche Individuum diese Eigenschaften nicht besitzt. Wie ausführlich von Stegmüller [1983, Kapitel IX] diskutiert worden ist, beseitigt dieser Trick jedoch nur die formale Seite des Widerspruchs. Eine viel schwieriger zu beantwortende Frage ist, welche Schlußfolgerungen gezogen werden können, wenn die verfügbaren Informationen über statistische Zusammenhänge im Einzelfall zu sich probabilistisch widersprechenden Beurteilungen führen.

Zu überlegen ist jedoch, ob dieses Problem für die empirische Sozialforschung relevant ist. Die Antwort hängt vor allem davon ab, welche Art von Aussagen über gesellschaftliche Verhältnisse angestrebt wird. Zum Beispiel schreiben Mueller, Schuessler und Costner [1970, S. 207]: „Society, as a collectivity of individuals, could not exist without more or less uniform patterns of social behavior which assure a certain fulfillment of mutual expectations. [...] On the other hand, experience has also taught us that the fulfillment of our anticipations is accompanied by considerable uncertainty. This uncertainty gives rise to statements of probability.“ In diesem Kontext handelt es sich um einzelfallbezogene Wahrscheinlichkeitsaussagen, d.h. um Aussagen zur Formulierung von (subjektiven) Erwartungen.[35] Dementsprechend könnte empirische Sozialforschung so konzipiert werden, daß mit ihr intendiert wird, empirisches Wissen zu bilden, mit dem einzelfallbezogene Wahrscheinlichkeitsaussagen über soziale Ereignisse gestützt werden können. Demgegenüber verfolgt die vorliegende Arbeit eine andere Konzeption, bei der, wie ich glaube, von einzelfallbezogenen Wahrscheinlichkeitsaussagen abgesehen werden kann. Ziel ist nicht die Ermöglichung einzelfallbezogener Wahrscheinlichkeitsaussagen, sondern eine Beschreibung gesellschaftlicher Verhältnisse als Bedingungen individueller Lebensverläufe. Dies kann zunächst durch deskriptive Wahrscheinlichkeitsaussagen erreicht werden, die sich nicht auf einzelne, jeweils bestimmte Individuen beziehen, sondern auf Gesamtheiten von Individuen. Deskriptive Wahrscheinlichkeitsaussagen können insofern unabhängig davon verstanden werden, ob bzw. wie mit ihrer Hilfe und ggf. weiteren Informationen einzelfallbezogene Erwartungen gebildet werden können.[36]

[35] Dementsprechend betonen Mueller et al. [1970, S. 209], „that probability is not an absolute condition of nature, but rather a framework imposed on nature by the observer. It is a purely human estimation of the likelihood of an event in the future.“

[36] Allerdings ist es eine offene Frage, ob man nicht dennoch einzelfallbezogene Wahrscheinlichkeitsaussagen benötigt, um deskriptive Wahrscheinlichkeitsaussagen begründen zu können. Denn deskriptive Wahrscheinlichkeitsaussagen beziehen sich in der Regel auf Grundgesamtheiten, über die nur unvollständige Informationen verfügbar sind. Im Kontext ihrer Begründung entsteht infolgedessen ein statistisches Inferenzproblem, und es ist (wie in Kapitel 5 näher ausgeführt wird) unklar, ob man dabei ohne einzelfallbezogene Wahrscheinlichkeitsaussagen auskommen kann.

2.3.3 Aussagen über Zufallsgeneratoren

Viele Autoren sind der Auffassung, daß sich Wahrscheinlichkeitstheorie und Statistik in erster Linie weder mit deskriptiven Wahrscheinlichkeitsaussagen beschäftigen, noch mit Aussagen, mit denen (subjektive) Erwartungen zum Ausdruck gebracht werden, sondern mit (objektivierbaren) Aussagen über Zufallsgeneratoren bzw. Zufallsexperimente. Die für uns wichtige Frage ist wiederum, ob und ggf. auf welche Weise in der empirischen Sozialforschung eine Bezugnahme auf Zufallsgeneratoren erforderlich ist.

Den Begriff „Zufallsgenerator“ verstehe ich so, daß damit Apparate bezeichnet werden, mit denen auf reproduzierbare Weise zufällige Ereignisse erzeugt werden können. „Zufällig“ soll heißen, daß die jeweils individuellen Ereignisse mithilfe des verfügbaren Kausalwissens nicht vorausgesagt werden können.[37] Beispiele für Zufallsgeneratoren sind Glücksspiele oder auch gewisse Algorithmen, die mithilfe eines Computers zur Erzeugung von Zufallszahlen verwendet werden können. Ein in der Wahrscheinlichkeitstheorie häufig verwendetes prototypisches Modell für einen Zufallsgenerator ist eine Urne, die mit Kugeln unterschiedlicher Farbe angefüllt ist; das Zufallsexperiment zur Erzeugung zufälliger Ereignisse besteht darin, aus einer solchen Urne eine Kugel herauszuziehen.[38]

Es erscheint sinnvoll, den Begriff „Zufallsgenerator“ zunächst auf Apparate, d.h. durch Menschen erfundene und praktisch herstellbare Verfahren zur Erzeugung zufälliger Ereignisse zu beziehen. Dies hat mehrere Vorteile. Erstens kann dann praktisch demonstriert werden, was Zufallsgeneratoren sind. Zweitens kann deutlich gemacht werden, daß die übliche Verwendung des Begriffs an reproduzierbare Situationen gebunden ist. Ein Zufallsexperiment kann nur dann durch die Vorstellung eines Zufallsgenerators interpretiert werden, wenn es wiederholbar ist. Drittens kann erklärt werden, in welcher Weise die Charakterisierung eines Apparats oder Verfahrens als Zufallsgenerator vom jeweils verfügbaren Kausalwissen abhängig ist. Zum Beispiel ist festgestellt worden, daß die bei einigen Glücksspielautomaten eingesetzten Zufallsgeneratoren so simpel waren, daß das Verhalten dieser Automaten vorausgesagt werden konnte; d.h. es war in diesen Fällen möglich, ein Kausalwissen zu gewinnen, so daß infolgedessen diese Apprate nicht länger als Zufallsgeneratoren bezeichnet werden konnten.

Die wissenschaftliche Bedeutung von Zufallsgeneratoren liegt natürlich nicht darin, daß mit ihrer Hilfe Glücksspiele geschaffen werden können, sondern in ihrer Eignung als theoretische Modelle zur Repräsentation von Situationen, in denen Ereignisse nicht sicher vorausgesagt werden können. Dabei stellen sich zwei Fragen. Unter welchen Bedingungen kann eine Situation sinnvoll als ein Zufallsgenerator interpretiert werden? Und welche Art von Aussagen können auf der Grundlage einer solchen Interpretation gewonnen werden?

Bezieht man sich auf Zufallsgeneratoren, ist es üblich, von der Wahrscheinlichkeit zu sprechen, mit der ein Ereignis eintritt bzw. eintreten kann. Bezieht man sich auf das Spiel mit einem regulären Würfel, könnte man zum Beispiel sagen: Die Wahrscheinlichkeit, beim nächsten Wurf eine gerade Zahl zu erzielen, beträgt 0.5. Obwohl Wahrscheinlichkeitsaussagen dieser Art üblich sind, ist umstritten, welche Bedeutung ihnen gegeben werden kann. Hauptsächlich zwei Auffassungen stehen sich gegenüber. Auf der einen Seite wird versucht, Wahrscheinlichkeitsaussagen dieser Art als Aussagen über die mit einem Zufallsgenerator erzeugbaren singulären zufälligen Ereignisse zu deuten. Dies entspricht den in Abschnitt 2.3.2 diskutierten einzelfallbezogenen Wahrscheinlichkeitsaussagen, so daß es nicht erforderlich ist, darauf noch einmal einzugehen.

Auf der anderen Seite stehen Versuche, zu objektivierbaren Aussagen über Eigenschaften von Zufallsgeneratoren zu gelangen. Auch im Hinblick auf diese Intention gibt es unterschiedliche Deutungsversuche. Insbesondere haben zahlreiche Autoren versucht, Wahrscheinlichkeitsaussagen als Aussagen über die mit einem Zufallsgenerator realisierbaren Folgen zufälliger Ereignisse zu deuten, genauer: als Aussagen über die relative Häufigkeit, mit der gewisse Ereignisse in solchen Folgen von Ereignissen realisiert werden. Dies wird als „Häufigkeitsinterpretation“ des Wahrscheinlichkeitsbegriffs bezeichnet;[39] in einer Charakterisierung durch Ayer [1972, S. 43]: „The essence of the frequency theory is that it identifies the probability of an event with the proportion of instances in which the property by which the event is identified is in fact distributed among some class, of which the individual to which the property is ascribed is a member.“ Diese Interpretation hat den Vorzug, daß sie den Wahrscheinlichkeitsbegriff auf einen unmittelbar verständlichen Sachverhalt – relative Häufigkeiten – zurückführt. Wie in Abschnitt 2.3.1 ausgeführt wurde, liefert diese Interpretation ein unproblematisches Verständnis deskriptiver Wahrscheinlichkeitsaussagen

[37] Lorenzen [1974, 1978] unterscheidet *Zufallsgeneratoren* und *Zufallsaggregate*: „Ein Gerät heiße ein ‘Zufalsgenerator’, wenn es den folgenden Forderungen genügt: (1) *Eindeutigkeit*: Jede Benutzung des Geräts (jeder ‘Versuch’) ergibt als Resultat genau eine von endlich vielen Aussageformen $E_1, \ldots, E_m$ (‘Elementarereignisse’). (2) *Ununterscheidbarkeit*: Mit keinem Kausalwissen läßt sich ein Grund angeben, der eines der Resultate $E_1, \ldots, E_m$ vor einem anderen auszeichnet. (3) *Wiederholbarkeit*: Nach jedem Versuch ist das Gerät wieder im selben Zustand wie vor dem Versuch.“ Zufallsaggregate entstehen durch verschiedene Formen der Kombination elementarer Zufallsgeneratoren, so daß im Prinzip beliebige Verteilungen konstruiert werden können. Ich verwende in dieser Arbeit das Wort „Zufallsgenerator“ so, daß es auch Zufallsaggregate umfaßt, also als Bezeichnung für Geräte, mit denen für beliebig vorgegebene Wahrscheinlichkeitsverteilungen entsprechende Ereignisfolgen erzeugt werden können.

[38] Eine einführende Übersicht über die vielfältigen Verwendungen solcher Urnenmodelle haben Johnson und Kotz [1977] gegeben.

[39] Sie wurde zum erstenmal systematisch von Venn [1888] ausgeführt. Vertreter dieser Interpretation waren dann u.a. von Mises [1928], Carnap [1945] und Reichenbach [1949], sowie – wichtig vor allem im Hinblick auf die Statistik – R. A. Fisher und J. Neyman; vgl. u.a. Fisher [1956, S 31ff] und Neyman [1950, S. 15f].

über endliche Grundgesamtheiten, da in diesem Fall relative Häufigkeiten eindeutig definiert sind. Sie wird jedoch problematisch, wenn auf Zufallsgeneratoren Bezug genommen wird, mit denen beliebige Folgen zufälliger Ereignisse erzeugt werden können. Der Versuch, Wahrscheinlichkeiten als Grenzwerte relativer Häufigkeiten in beliebigen Folgen zufälliger Ereignisse zu definieren, führt tatsächlich in zahlreiche Schwierigkeiten und wird deshalb von vielen Wissenschaftstheoretikern als aussichtslos angesehen.[40] Eine alternative Interpretation liefert die sog. „Propensity-Theorie", bei der versucht wird, Wahrscheinlichkeitsaussagen nicht als Aussagen über Eigenschaften der mit einem Zufallsgenerator erzeugbaren Folgen von Ereignissen anzusehen, sondern als Aussagen über dispositionale Eigenschaften des Zufallsgenerators.[41] Allerdings ist auch diese Interpretation des Wahrscheinlichkeitsbegriffs umstritten.[42]

Zumindest ein Teil der Probleme, die entstehen, wenn man nach empirisch sinnvollen Deutungen von Wahrscheinlichkeitsaussagen sucht, verdanken sich, wie ich glaube, dem Umstand, daß meistens nach möglichst allgemeinen Interpretationen gesucht wird, die sich auf beliebige Zufallsgeneratoren in beliebigen Kontexten anwenden lassen. Demgegenüber gehe ich in dieser Arbeit von der heuristischen Annahme aus, daß es zweckmäßig ist, die Bedeutung von Wahrscheinlichkeitsaussagen kontextspezifisch zu explizieren. Dementsprechend sollte zunächst überlegt werden, wo – im Kontext der empirischen Sozialforschung – Aussagen über Zufallsgeneratoren überhaupt erforderlich sind.

a) Das erste Anwendungsfeld betrifft den Umstand, daß die in der empirischen Sozialforschung in der Regel verwendeten Daten aus Zufallsstichproben stammen. Vorläufig formuliert: das Interesse zielt auf die Verteilung gewisser Zufallsvariablen, die im Hinblick auf eine Grundgesamtheit Ω definiert sind, zur Verfügung stehen jedoch nur Informationen aus einer Stichprobe S aus dieser Grundgesamtheit. Es ist dann erforderlich, die Verteilung der Zufallsvariablen (oder gewisse Aspekte ihrer Verteilung) mithilfe der Informationen aus S zu schätzen. Einige der üblichen Vorstellungen, um dieses Schätzproblem rationalisierbar zu machen, beruhen darauf, daß die Stichprobe durch die Verwendung eines Zufallsgenerators zustande gekommen ist (darauf wird in Kapitel 5 näher eingegangen). Dieses Schätzproblem betrifft jedoch offensichtlich nur die Frage, wie und mit welcher Sicherheit man zu den schließlich intendierten Aussagen über die Verteilung der Zufallsvariablen in Ω gelangen kann. Die Frage, welche Bedeutung die intendierten Aussagen über die Verteilung der Zufallsvariablen

[40] Vgl. zur Kritik u.a. Ayer [1972, S. 43ff], Stegmüller [1973, S. 32ff].

[41] Diese Interpretation wurde u.a. von Popper [1960], Hacking [1965], Mellor [1971] und Stegmüller [1973, S. 62ff] vertreten. Eine kritische Darstellung der verschiedenen Varianten dieser Interpretation des Wahrscheinlichkeitsbegriffs findet sich bei Kyburg [1974].

[42] Vgl. Sklar [1970], Mackie [1973, S. 179ff].

in Ω haben, ist davon ganz unabhängig. Solange man davon ausgeht, daß es sich um eine endliche Grundgesamtheit handelt, können Wahrscheinlichkeitsaussagen deskriptiv als Aussagen über relative Häufigkeiten verstanden werden.

b) Ein zweiter Problembereich betrifft den Umstand, daß sich die verfügbaren Daten in der Regel nicht nur auf eine Stichprobe beschränken, sondern daß sie darüberhinaus unvollständig sind. Wiederum stellt sich die Frage, ob gleichwohl – und wenn ja, mit welchen Einschränkungen – Aussagen über die im Hinblick auf eine Grundgesamtheit definierte Verteilung von Zufallsvariablen erreicht werden können. Es ist üblich, auch zur Reflexion dieses Aspekts des statistischen Inferenzproblems auf Vorstellungen darüber zurückzugreifen, daß die jeweils verfügbaren Daten zufällig zustandegekommen sind. Insoweit gilt jedoch, was bereits unter (a) gesagt wurde: daß es sich nur um einen Aspekt des statistischen Inferenzproblems handelt, der nicht die Bedeutung der schließlich intendierten Aussagen betrifft.

c) Es bleibt die Frage, ob Zufallsgeneratoren auch zweckmäßig sind, um die Sachverhalte zu interpretieren, mit denen wir uns in der soziologischen Lebensverlaufsforschung beschäftigen, nämlich individuelle Lebensverläufe. Oder anders formuliert: liefern Zufallsgeneratoren sinnvolle Modelle, um den Entwicklungsprozeß von Lebensverläufen zu reflektieren? In gewisser Weise fällt es leicht, darauf eine negative Antwort zu geben. Denn jeder weiß – man braucht nur an den je eigenen Lebensverlauf zu denken –, daß Lebensverläufe nicht durch eine sequentielle Bedienung von Zufallsgeneratoren zustande kommen, vielmehr durch Entscheidungen ihrer Subjekte, die im Rahmen jeweils gegebener Bedingungen und Handlungsmöglichkeiten getroffen werden. Die Frage ist jedoch, ob es gleichwohl sinnvoll sein kann, Lebensverläufe so zu beschreiben, *als ob* sie durch Zufallsgeneratoren entstanden sei könnten. Zwei Überlegungen können dies motivieren.

Erstens kann darauf hingewiesen werden, daß auf diese Weise deutlicher gesagt werden kann, was es heißen soll, gesellschaftliche Verhältnisse durch Chancen zu charakterisieren. Deskriptive Wahrscheinlichkeitsaussagen, wie sie oben eingeführt wurden, liefern nicht unmittelbar eine Einsicht in Chancen. Tatsächlich kann in zwei unterschiedlichen Kontexten über Chancen gesprochen werden. Einerseits kann mit dem Begriff auf Handlungsmöglichkeiten verwiesen werden. In diesem Kontext wird auf Subjekte Bezug genommen, die sich im Hinblick auf die von ihnen wahrnehmbaren Handlungsmöglichkeiten kontingent verhalten können. Andererseits kann man den Chancenbegriff verwenden, um von konkreten Subjekten und von der Kontingenz der von ihnen jeweils individuell realisierten Handlungen zu abstrahieren. Um diesen (statistischen) Chancenbegriff zu verstehen, kann man prototypisch an Zufallsgeneratoren denken. Insofern liefert die Vorstellung, daß individuelle Lebensverläufe durch Zufallsgeneratoren entstanden sein könnten, eine Möglichkeit, um den Sinn einer Beschreibung

gesellschaftlicher Verhältnisse als Chancen (in der statistischen Bedeutung des Wortes) zu verstehen.[43]

Zweitens ist die Vorstellung, daß individuelle Lebensverläufe durch Zufallsgeneratoren entstanden sein könnten, zwar nicht notwendig, aber pragmatisch zweckmäßig, um damit einen Anschluß an die statistische Theoriebildung zu finden. Dies betrifft insbesondere den Begriff eines statistischen Modells. Es ist üblich, diesen Begriff auf die Vorstellung eines „datengenerierenden Prozesses", d.h. auf die Vorstellung eines Zufallsgenerators zu beziehen, der durch das statistische Modell beschrieben werden soll. Die meisten der in der statistischen Theorie zur Modellbildung entwickelten Überlegungen beziehen sich direkt oder indirekt auf diese Vorstellung; und um sie für eine statistische Beschreibung von Lebensverläufen nutzen zu können, ist es insoweit zweckmäßig, sich auf die Vorstellung einzulassen, daß auch individuelle Lebensverläufe so beschrieben werden können, als ob sie durch Zufallsgeneratoren entstanden wären.

Im Hinblick auf soziologische Lebensverlaufsforschung sollte allerdings betont werden, daß statistische Modelle nur ein Hilfsmittel sind, um zu soziologischen Einsichten in die Entwicklung von Lebensverläufen bzw. ihrer gesellschaftlichen Verhältnisse zu gelangen. Ich betrachte die Vorstellung, daß individuelle Lebensverläufe durch Zufallsgeneratoren zustande gekommen sein könnten, als eine Fiktion, die zweckmäßig ist, um statistische Beschreibungen von Lebensverläufen zu verstehen. Tatsächlich kann von dieser Fiktion auf zwei unterschiedliche Weisen Gebrauch gemacht werden. Einerseits als ein theoretisches Hilfsmittel, um die Abstraktion von den Individuen zu charakterisieren, die bei einer statistischen Beschreibung vorgenommen wird. In diesem Kontext kann man dem fiktiven Zufallsgenerator eine klare Bedeutung geben. Gegeben ist eine zeitlich und räumlich fixierte Gesamtheit von Individuen, und den Zufallsgenerator kann man sich als einen Mechanismus vorstellen, mit dem aus dieser Gesamtheit beliebige Individuen mit jeweils der gleichen Wahrscheinlichkeit gewissermaßen als Repräsentanten der Gesamtheit herausgezogen werden können. Diese Vorstellung macht deutlich, wie über „die Individuen", die dieser Gesamtheit angehören, gesprochen werden soll. Zur Beschreibung eignet sich ein deskriptiver Wahrscheinlichkeitsraum, so wie er in Abschnitt 2.3.1 eingeführt wurde. Da es sich stets um endliche Gesamtheiten von Individuen handelt, können auch die intendierten Wahrscheinlichkeitsaussagen problemlos als Aussagen über relative Häufigkeiten verstanden werden. Und da es sich um zeitlich und räumlich fixierte Gesamtheiten von Individuen handelt, ist es nicht erforderlich, eine temporale Stabilität des

[43] Damit hat man allerdings noch keine Begründung dafür, warum es überhaupt sinnvoll sein könnte, gesellschaftliche Verhältnisse durch Chancen (im statistischen Sinne) zu beschreiben. Diese Form der Beschreibung erscheint zwar sinnvoll, um von der Kontingenz individueller Lebensverläufe abstrahieren zu können – und sie entspricht insofern dem Selbstverständnis einer liberalen Gesellschaft –, aber ob sie einen hinreichenden Rahmen zur Reflexion sozialer Ungleichheit bieten kann, ist eine offene Frage.

Zufallsgenerators anzunehmen. Der mit dieser Fiktion eines Zufallsgenerators definierbare Chancenbegriff kann als eine deskriptive Kategorie zur Beschreibung gesellschaftlicher Verhältnisse, die sich historisch verändern können, angesehen werden.

Anders verhält es sich, wenn man glaubt, daß die individuellen Lebensverläufe tatsächlich durch in der Realität existierende Zufallsprozesse zustande kommen. Ich weiß nicht, ob es Soziologen gibt, die diese Auffassung vertreten, es gibt jedoch Formulierungen, die zumindest mißverständlich sind. Die möglichen Mißverständnisse hängen, wie ich glaube, mit der Frage zusammen, in welcher Weise statistische Modelle einer Erklärung individueller Lebensverläufe dienen können. Zum Beispiel sagt Diekmann [1981, S. 319]: „Beobachtete Häufigkeitsverteilungen können zum einen in der Weise ausgewertet werden, daß deskriptive Kennziffern wie Maße der zentralen Tendenz (Mittelwert, Median) oder Streuungsmaße berechnet werden. Diese Vorgehensweise entspricht der üblichen Praxis der Sozialforschung. Darüber hinaus kann zum anderen der Versuch unternommen werden, die beobachtete Verteilung durch einen zugrunde liegenden stochastischen Prozeß zu *erklären*." Die Frage ist, was bei Formulierungen dieser Art mit dem Wort „erklären" gemeint ist, denn statistische Modelle liefern zunächst nur eine Beschreibung von (ggf. bedingten) Verteilungen von Zufallsvariablen. Bezieht man sie, wie in der Einleitung vorgeschlagen wurde, auf Wie-Fragen, können sie als Bestandteile von Erklärungen angesehen werden. Im Kontext der soziologischen Lebensverlaufsforschung erscheint es jedoch nicht sinnvoll, in statistischen Modellen zugleich Antworten auf Warum-Fragen zu sehen. Der Grund liegt darin, daß der „datengenerierende Prozeß" in diesem Fall nicht tatsächlich ein Zufallsgenerator (oder eine zeitliche Sequenz von Zufallsgeneratoren) ist – dies wird nur als Fiktion unterstellt, um einen theoretischen Bezugspunkt für die Formulierung statistischer Beschreibungen zu gewinnen –; er besteht vielmehr darin, daß Menschen in historisch sich ändernden gesellschaftlichen Verhältnissen kontingente Lebensverläufe realisieren. Soziologische Theoriebildung sollte, wie ich glaube, von dieser Eigenart ihres Gegenstandsbereichs nicht abstrahieren, sondern nach Deutungen der statistisch ermittelbaren Regeln bzw. Regelmäßigkeiten suchen, die sie als Beschreibungen gesellschaftlicher Verhältnisse durch deren Subjekte reflektierbar machen.

2.3.4 Bedingte Wahrscheinlichkeitsverteilungen

Statistische Methoden können der soziologischen Lebensverlaufsforschung dienen, indem mit ihrer Hilfe statistische Beschreibungen von Lebensverläufen gewonnen werden können. Darüber hinaus können statistische Methoden verwendet werden, um *Bedingungen* von Lebensverläufen sichtbar zu machen. Wie auf sinnvolle Weise von solchen Bedingungen gesprochen werden kann, ist zwar eine Frage der soziologischen Theoriebildung und hängt davon ab, welches Bild sie sich von den Individuen und ih-

ren gesellschaftlichen Verhältnissen machen will (darauf wird in Kapitel 4 näher eingegangen); aber unabhängig davon kann bereits erklärt werden, wie diese Aufgabe durch statistische Beschreibungen unterstützt werden kann. Das zentrale statistische Konzept ist der Begriff einer bedingten Wahrscheinlichkeitsverteilung.

Im begrifflichen Rahmen eines deskriptiven Wahrscheinlichkeitsraums ist sowohl die Definition als auch die Interpretation einfach. Sei $(\Omega, \mathcal{A}, \mathrm{P})$ der Wahrscheinlichkeitsraum für eine endliche Grundgesamtheit von Individuen. Für beliebige Mengen $A \in \mathcal{A}$ und $B \in \mathcal{A}$ kann, wenn $\mathrm{P}(B) > 0$, definiert werden:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \tag{2.5}$$

Der Ausdruck auf der linken Seite wird als *bedingte Wahrscheinlichkeit von A unter der Bedingung B* bezeichnet. In einem deskriptiven Wahrscheinlichkeitsraum ergibt sich die einfache Interpretation: $\mathrm{P}(A \mid B)$ ist der Anteil von Individuen in der Menge B, die zugleich Mitglieder der Menge A sind. Ist zum Beispiel B die Menge aller Individuen (aus der Grundgesamtheit Ω), die 1960 geboren wurden, und ist A die Menge aller 1990 verheirateten Personen, dann ist $\mathrm{P}(A \mid B)$ der Anteil der 1990 verheirateten Personen in der Menge der 1960 geborenen Personen.[44]

Die Bedeutung der Begriffsbildung resultiert daraus, daß mit ihrer Hilfe auf abstrakte Weise von Bedingungen gesprochen werden kann. Üblicherweise werden solche Aussagen unmittelbar auf Zufallsvariablen bezogen, indem zum Beispiel gesagt wird, daß eine Zufallsvariable X (das Alter) eine Bedingung der Zufallsvariable Y (des Einkommens) ist; und sie legen insofern die Vorstellung eines funktionalen Zusammenhangs zwischen „Variablen“ nahe. Obwohl solche in der „Variablensoziologie“ gängigen Formulierungen den üblichen Sprechweisen der Statistik entsprechen, sollte man sich klar machen, daß es sich nicht um gewöhnliche Funktionen handelt, durch die den Elementen einer gegebenen Menge von Gegenständen

[44] Die Definition bedingter Wahrscheinlichkeiten kann unmittelbar auf Wahrscheinlichkeitsaussagen für Zufallsvariablen übertragen werden. Sind nämlich X und Y zwei Zufallsvariablen mit den Wertebereichen $\mathcal{X}$ bzw. $\mathcal{Y}$, sind außerdem M_x und M_y zwei beliebige Teilmengen aus $\mathcal{X}$ bzw. $\mathcal{Y}$ (da es sich um einen deskriptiven Wahrscheinlichkeitsraum handelt, ist die Meßbarkeitsbedingung automatisch erfüllt), und ist $\mathrm{P}(X \in M_x) > 0$, erhält man durch die in (2.5) angegebene Definition:

$$\begin{aligned} \mathrm{P}\,(Y \in M_y \mid X \in M_x) &= \mathrm{P}(\,\{\omega \mid Y(\omega) \in M_y\} \mid \{\omega \mid X(\omega) \in M_x\}\,) \\ &= \frac{\mathrm{P}(\,\{\omega \mid Y(\omega) \in M_y \text{ und } X(\omega) \in M_x\}\,)}{\mathrm{P}(\,\{\omega \mid X(\omega) \in M_x\}\,)} \end{aligned}$$

Auf der linken Seite steht die Wahrscheinlichkeit, daß Y einen Wert in der Menge M_y annimmt, unter der Bedingung, daß X einen Wert in der Menge M_x annimmt. Die Interpretation in einem deskriptiven Wahrscheinlichkeitsraum ist wiederum: Der Anteil der Individuen mit $Y(\omega) \in M_y$ an der Gesamtheit der Individuen, für die $X(\omega) \in M_x$ gilt.

bestimmte numerische Werte zugeordnet werden Der funktionale Zusammenhang besteht hier vielmehr darin, daß jeder möglichen Menge von Werten einer Zufallsvariable X (genauer: jedem $M_x \subseteq \mathcal{X}$ mit $\mathrm{P}(X \in M_x) > 0$) eine dadurch bedingte Wahrscheinlichkeitsverteilung zugeordnet wird. Folgende Darstellung veranschaulicht diese Betrachtungsweise:

$$M_x \longrightarrow (M_y \longrightarrow \mathrm{P}\,(Y \in M_y \mid X \in M_x))$$

M_x, auf der linken Seite, ist eine Teilmenge aus dem Wertebereich von X, zu interpretieren als die Menge der Individuen aus Ω, bei denen die Zufallsvariable X einen Wert in M_x annimmt. Auf der rechten Seite steht nicht eine bestimmte Wahrscheinlichkeit, sondern eine Wahrscheinlichkeits*verteilung*, eine für die Mengen $M_y \subseteq \mathcal{Y}$ definierte Funktion. M_y ist wiederum zu interpretieren als die Menge der Individuen aus Ω, bei denen die Zufallsvariable Y einen Wert in M_y annimmt.[45]

Ich betone diese Darstellung, weil mit der Vorstellung, daß sich „Variablen“ bedingen, bei der Beschreibung von Lebensverläufen im allgemeinen keine sinnvolle Interpretation erreicht werden kann. In der Regel werden in diesem Zusammenhang bedingte Wahrscheinlichkeitsverteilungen gebildet, um empirische Anhaltspunkte dafür zu gewinnen, wie die Entwicklung von Lebensverläufen von gewissen Merkmalen einer Situation abhängt, die sich in der jeweiligen Vergangenheit herausgebildet hat. Im allgemeinen erscheint es nicht sinnvoll, von der Vorstellung eines unmittelbaren Kausalverhältnisses auszugehen. Es muß nicht nur berücksichtigt werden, daß die Entwicklung von Lebensverläufen auch von Bedingungen abhängig sein kann, die bei der Formulierung einer bedingten Wahrscheinlichkeitsverteilung nicht explizit erfaßt worden sind, sondern es muß auch für die Vorstellung Raum gelassen werden, daß im allgemeinen bei gleichen Bedingungen unterschiedliche Lebensverläufe realisiert werden können.

[45] Wenn die als Bedingung verwendete Zufallsvariable X diskret ist, kann folgende Definition vorgenommen werden:

$$F_{Y|X}(y \mid x) = \mathrm{P}\,(Y \leq y \mid X = x) \quad \text{für} \quad x \in \mathcal{X}$$

Dies wird als *bedingte Verteilungsfunktion* der Zufallsvariable Y in Abhängigkeit von den Werten der Zufallsvariable X bezeichnet. Wenn außerdem Y diskret ist, kann auch eine *bedingte Wahrscheinlichkeitsfunktion* definiert werden:

$$f_{Y|X}(y \mid x) = \mathrm{P}\,(Y = y \mid X = x) \quad \text{für} \quad x \in \mathcal{X}$$

Diese Definitionen sind hier ausreichend, da ein deskriptiver Wahrscheinlichkeitsraum vorausgesetzt wird und infolgedessen alle Zufallsvariablen diskret sind.

2.4 Statistische Beschreibung von Lebensverläufen

Die statistische Beschreibung von Lebensverlaufsdaten beruht auf zwei Voraussetzungen. Erstens muß, wie in den Abschnitten 2.1 und 2.2 ausgeführt wurde, ein formaler Beschreibungsrahmen, insbesondere ein Zustandsraum und eine Zeitachse, definiert werden, in dem dann auf formal eindeutige Weise von Lebensverläufen gesprochen werden kann. Dies impliziert die Festlegung eines Biographieschemas, das die in diesem Rahmen möglichen Lebensverläufe definiert. Zweitens muß ein deskriptiver Wahrscheinlichkeitsraum festgelegt werden, der als Ausgangspunkt zur Definition geeigneter Zufallsvariablen dienen kann. Insbesondere muß also eine endliche Grundgesamtheit von Individuen fixiert werden, auf die sich alle weiteren Definitionen und Aussagen beziehen können.

Sind diese beiden Voraussetzungen gegeben, können Lebensverläufe statistisch beschrieben werden. Das Wort *beschreiben* soll dabei zum Ausdruck bringen, daß wir von der Vorstellung ausgehen, daß die zu beschreibenden Lebensverläufe bereits realisiert worden sind. Es wird also angenommen, daß es eine endliche Grundgesamtheit von Individuen gibt, die wir wie bisher mit dem Symbol Ω bezeichnen, und daß es für jedes Individuum $\omega \in \Omega$ einen realisierten Lebensverlauf $y_\omega(t)$ gibt, der durch Bezugnahme auf einen Zustandsraum $\mathcal{Y}$ und eine Zeitachse $\mathcal{T}$ definierbar ist.

2.4.1 Ausgangszustände und Längsschnittgesamtheiten

Die statistische Beschreibung von Lebensverläufen setzt voraus, daß auf eine Gesamtheit von Individuen Bezug genommen wird. Zu überlegen ist, wie solche Gesamtheiten sinnvoll abgegrenzt werden können. Bei der traditionellen Analyse von Querschnittsdaten ist die Frage weitgehend unproblematisch. Man betrachtet die Gesamtheit der Personen, die zu einem gewissen Zeitpunkt existieren, oder eine daraus nach beliebigen Kriterien abgrenzbare Teilgesamtheit. Bei der statistischen Beschreibung von Lebensverläufen ist die Frage etwas komplizierter, weil dann die Definition einer Grundgesamtheit zeitraumbezogen vorgenommen werden muß. Dabei muß offenbar berücksichtigt werden, daß sich die in einer Gesellschaft lebenden Individuen fortwährend ändern. Zu jedem Zeitpunkt kommen einige Personen durch Geburt oder Migration hinzu, und andere verlassen die Grundgesamtheit, indem sie sterben oder die vorausgesetzte räumliche Abgrenzung verlassen.

Strenggenommen erfordert insofern jede Beschreibung von Lebensverläufen ein demographisches Rahmenmodell. Dies gilt jedenfalls dann, wenn man Lebensverläufe so beschreiben möchte, wie sie sich in einer Gesellschaft in der Kalenderzeit entwickeln. Bei vielen Fragestellungen der soziologischen Lebensverlaufsforschung erscheint es jedoch möglich, auf

ein solches Rahmenmodell zunächst zu verzichten. Die Grundidee ist, daß man zunächst nur gewisse Aspekte der Lebensverläufe bei einer empirisch zugänglichen Gesamtheit von Individuen beschreiben möchte; zum Beispiel das Heiratsverhalten, die Dauer der geschlossenen Ehen, die Dauer der Arbeitslosigkeit, usw.

Um im Hinblick auf solche Aspekte von Lebensverläufen zu sinnvoll vergleichenden Aussagen darüber zu gelangen, wie sich eine Gesamtheit von Lebensverläufen entwickelt, ist es vor allem wichtig, jeweils auf einen vergleichbaren Ausgangszustand bezug zu nehmen, genauer gesagt: auf ein Ereignis, das als Übergang in einen Ausgangszustand interpretiert werden kann. Dies kann zum Beispiel die Geburt sein, grundsätzlich können jedoch auch beliebige andere Ereignisse zur Definition eines Ausgangszustands verwendet werden. Um zum Beispiel die Dauer der Arbeitslosigkeit zu beschreiben, kann der Beginn der Arbeitslosigkeit als ein den Ausgangszustand definierendes Ereignis verwendet werden.

Diese Betrachtungsweise liefert ein einfaches Prinzip zur Definition von Grundgesamtheiten: es werden diejenigen Personen zu einer Gesamtheit zusammengefaßt, die während eines gewissen Kalenderzeitraums in den Ausgangszustand des zu beschreibenden Prozesses geraten sind. Gesamtheiten von Personen, die auf diese Weise definiert worden sind, nenne ich im folgenden *Längsschnittgesamtheiten*. Zum Beispiel die Gesamtheit der Personen, die während eines gewissen Zeitraums geboren wurden, oder die Gesamtheit der Personen, die während eines gewissen Zeitraums geheiratet haben. Allerdings ist zu beachten, daß Definitionen dieser Art bei wiederholbaren Ereignissen präziser gefaßt werden müssen. Zum Beispiel wäre dann genauer zu sagen: alle Personen, die während eines gewissen Zeitraums zum erstenmal geheiratet haben.

Die Betrachtung von Längsschnittgesamtheiten entspricht der wesentlichen Fragestellung der Lebensverlaufsforschung: *Wie entwickeln sich Lebensverläufe?* Die Fragestellung zielt auf einen kontingenten Entwicklungsprozeß. Dies schließt es nicht aus, einen Anfangszustand anzunehmen, der gewissermaßen den Ausgangspunkt für den dann einsetzenden Entwicklungsprozeß bildet. Es ist jedoch wichtig, daß der dann einsetzende Prozeß als eine offene Zukunft betrachtet werden kann, obwohl er erst *beschrieben* werden kann, nachdem er realisiert worden ist. Als Beispiel sei noch einmal auf die Frage des Heiratsalters Bezug genommen. Eine mögliche Formulierung dieser Frage wäre: Wie sieht die Verteilung des Heiratsalters bei denjenigen Personen aus, die während eines gewissen Zeitraums geheiratet haben? Obwohl diese Frage nicht sinnlos ist und ihre Beantwortung möglicherweise eine wichtige Information über die Bedingungen der jeweils folgenden Lebensverläufe liefert, ist sie offenbar ungeeignet, um den Prozeß zu beschreiben, der zur Heirat geführt hat. Um der Tatsache Rechnung zu tragen, daß Lebensverläufe kontingent sind, sollte die Frage vielmehr folgendermaßen gestellt werden: Wie entwickeln sich Lebensverläufe im Hinblick auf das mögliche Ereignis *Heirat*? Bei dieser Betrachtungsweise

wird davon ausgegangen, daß der zu beschreibende Prozeß in einem Zustand beginnt, der dem Ereignis *Heirat* vorausgesetzt werden kann. Dann kann untersucht werden, wie der Prozeß ausgehend von diesem vorausgesetzten Anfangszustand verläuft, ob und ggf. wann und unter welchen Bedingungen eine Heirat erfolgt.

Bei der weiteren Diskussion von Begriffen und Modellen zur statistischen Beschreibung von Lebensverläufen wird stets auf eine Längsschnittgesamtheit Bezug genommen. Die dadurch abgegrenzte Gesamtheit von Lebensverläufen wird als ein *Prozeß* bezeichnet. Es sollte betont werden, daß es sich um eine abstrakte Begriffsbildung handelt. Die Vorstellung eines Prozesses entsteht dadurch, daß wir eine Gesamtheit von Individuen zu einer Längsschnittgesamtheit zusammenfassen und dann, ausgehend von dem die Längsschnittgesamtheit definierenden Ereignis, die weitere Entwicklung der in ihr realisierten Lebensverläufe beschreiben. Die Beschreibung erfolgt auf einer Prozeßzeitachse, die mit dem Übergang in den Ausgangszustand des Prozesses beginnt.

Dieser abstrakte Prozeßbegriff sollte insbesondere von der Vorstellung eines sich in historischer Zeit (Kalenderzeit) entwickelnden „sozialen Prozesses“ unterschieden werden.[46] Auch sollte der Begriff „Längsschnittgesamtheit“ von soziologischen Vorstellungen über „soziale Gruppen“ unterschieden werden. Bei der Definition einer Längsschnittgesamtheit wird eine Menge von Individuen zusammengefaßt, wobei es keine Rolle spielt, ob ihre Lebensverläufe in zeitlicher und räumlicher Hinsicht verbunden sind. Insofern kann in der Regel nicht davon ausgegangen werden, daß es eine soziale Interaktion zwischen den Mitgliedern einer Längsschnittgesamtheit gibt.

2.4.2 Einfache Aspekte von Lebensverläufen

Beschreibungen von Lebensverläufen können unter verschiedenen Aspekten vorgenommen werden. Dies hängt zunächst vom Zustandsraum ab, der für die Beschreibung vorausgesetzt wird; er kann im einfachsten Fall nur aus zwei Zuständen bestehen oder aber sehr komplex sein. Hinzu kommt, daß es nicht immer erforderlich ist, den gesamten Lebensverlauf – von der Geburt bis zum Tod – zu beschreiben, sondern daß es durchaus sinnvoll sein kann, eine Beschreibung auf einzelne Episoden zu beschränken. Schließlich stellt sich auch die Frage, wie die Lebensverläufe einer Gesamtheit von Individuen in zeitlicher Hinsicht vergleichbar gemacht werden sollen. Aus diesen Gründen erscheint es sinnvoll, zunächst einige einfache Formen der statistischen Beschreibung zu behandeln.

a) Ein elementarer Aspekt jedes Lebensverlaufs besteht darin, daß Menschen geboren werden und sterben. Um dies zu beschreiben, benötigt

[46] Wie diese Vorstellung präzisiert werden kann, wird in Abschnitt 2.5 etwas näher diskutiert.

man einen Zustandsraum mit zwei Zuständen: $\mathcal{Y} = \{y_a, y_e\}$. y_a soll den Zustand bezeichnen, in dem eine Person lebt, y_e soll den (Quasi-) Zustand bezeichnen, daß sie gestorben ist. Jeder Lebensverlauf beginnt mit einem Übergang in den Zustand y_a, also mit der Geburt, und endet mit einem Übergang in den Zustand y_e, dem Tod. Wenn wir annehmen, daß die Zeitpunkte für diese beiden Ereignisse auf einer Kalenderzeitachse T definiert sind, kann für die vorausgesetzte Grundgesamtheit Ω folgende zweidimensionale Zufallsvariable definiert werden:

$$(T_1, T_2) : \Omega \longrightarrow \mathcal{T} \times \mathcal{T}$$

Für jedes Individuum $\omega \in \Omega$ gibt es eine Realisierung $(T_1(\omega), T_2(\omega))$ dieser Zufallsvariable, die Zeitpunkte für die Geburt und den Tod dieses Individuums. Aus soziologischer Sicht interessieren jedoch nicht diese individuellen Ereigniszeitpunkte, intendiert ist vielmehr eine Aussage über die Gesamtheit der Personen. Dem entsprechen statistische Aussagen über die Verteilung von Zufallsvariablen. In diesem Beispiel ist die gesamte relevante Information in der Verteilung der Zufallsvariable (T_1, T_2) enthalten.

Daraus können dann weitere Aussagen abgeleitet werden. Zum Beispiel kann man nur die Zufallsvariable T_1 betrachten und erhält dann eine Verteilung der Geburtszeitpunkte auf der zugrundeliegenden Kalenderzeitachse. Insbesondere kann auch die Differenz der beiden Zufallsvariablen betrachtet werden, also $T = T_2 - T_1$. Ihre Verteilung zeigt die Variationen der Lebensdauern bei den Individuen aus der vorausgesetzten Grundgesamtheit.

b) Bildung der Differenz $T = T_2 - T_1$ ist ein einfaches Beispiel für den Übergang von der Kalenderzeit zu einer *Prozeßzeitachse*.[47] Implizit wird dabei angenommen, daß alle Episoden (in diesem Beispiel Lebensverläufe) zum gleichen Zeitpunkt beginnen und daß infolgedessen nur die Dauer der Episode (in diesem Beispiel die Lebensdauer) wissenswert ist. Ob eine solche Abstraktion von der tatsächlichen Lage der Episoden in der Kalenderzeit gerechtfertigt ist, hängt sowohl von der Verteilung der Anfangszeitpunkte als auch von der Zielsetzung der Beschreibung ab. Um Einsichten in sozialen Wandel zu gewinnen, ist es in der Regel sinnvoll, von der Hypothese auszugehen, daß sich Lebensverläufe insbesondere dadurch unterscheiden, in welcher historischen Situation sie realisiert werden.

Eine solche Differenzierung kann auf einfache Weise durch die Bildung von Kohorten erreicht werden. In unserem Beispiel bedeutet dies, daß die Grundgesamtheit in Teilgesamtheiten (Kohorten) aufgeteilt wird, wobei jede Teilgesamtheit diejenigen Personen umfaßt, die „zur gleichen Zeit“ geboren worden sind. Wie die Bildung von Kohorten im Einzelfall vorzunehmen ist, hängt vor allem davon ab, ob sich die zu beschreibenden Sachverhalte in der historischen Entwicklung schnell oder langsam verändern.

[47] Vgl. die Definition in Abschnitt 2.2.

Meistens ist es ausreichend, Zeitintervalle mit einer Dauer von 1 bis 5 Jahren zu verwenden.[48] Formal kann die Bildung von Geburtskohorten dadurch beschrieben werden, daß die Kalenderzeitachse in Zeitintervalle $I_1, I_2, I_3, \ldots$ eingeteilt wird. Jedes Zeitintervall definiert eine Kohorte, in unserem Beispiel:

$$K_j = \{\omega \mid \omega \in \Omega, T_1(\omega) \in I_j\}$$

Die Verteilung der Lebensdauer $T = T_2 - T_1$ kann dann für die verschiedenen Geburtskohorten differenziert werden, indem bedingte Wahrscheinlichkeitsverteilungen betrachtet werden. Für die Kohorte der im Intervall I_j geborenen Personen erhält man die bedingte Verteilung

$$\mathrm{P}\left(T \leq t \mid T_1 \in I_j\right)$$

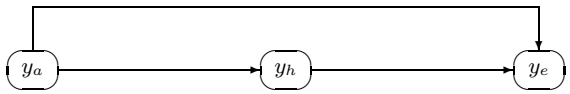
die den Episodenverlauf dieser Kohorte beschreibt.

c) Das bisher betrachtete Beispiel ist insofern einfach, als schließlich stets ein Übergang in den Endzustand y_e stattfinden muß. Bei allen anderen Ereignissen ist das nicht der Fall, es ist stets möglich, daß eine Person stirbt, bevor das betreffende Ereignis eintritt.

Als Beispiel betrachten wir die Episode von der Geburt bis zur ersten Heirat. Man könnte versuchen, zwei Gruppen zu bilden. Einerseits Personen, die (mindestens) einmal geheiratet haben; andererseits diejenigen, die überhaupt nicht heiraten. Für die erste dieser beiden Gruppen kann dann die Zeitdauer bis zur ersten Heirat als eine Zufallsvariable definiert und ihre Verteilung beschrieben werden.

Diese Vorgehensweise setzt jedoch voraus, daß es sich bereits um abgeschlossene Lebensverläufe handelt. Denn nur, wenn eine Person bereits gestorben ist, kann man wissen, daß sie niemals geheiratet hat. Betrachtet man hingegen Lebensverläufe als kontingente Entwicklungsprozesse, muß vorausgesetzt werden, daß jeweils bis zum Tod die Möglichkeit besteht, daß eine Heirat stattfinden könnte. Man kann zwar annehmen, daß es eine Gruppe von Personen gibt, die niemals heiraten werden; aber die Zugehörigkeit einer Person zu dieser Gruppe ist ein kontingenter Sachverhalt, über den immer erst im Nachhinein etwas ausgesagt werden kann. Insofern ist die Annahme, daß es eine solche Gruppe gibt, kein sinnvoller Ausgangspunkt für eine dynamische Beschreibung von Lebensverläufen. Eine Alternative liegt in diesem Beispiel darin, heiraten und sterben als „konkurrierende Risiken“ zu betrachten. Folgende Abbildung kann diese Überlegung veranschaulichen:

[48] In der praktischen Anwendung spielt natürlich auch die Verfügbarkeit von Daten eine entscheidende Rolle. Je kleiner der Stichprobenumfang, desto weniger Kohorten können gebildet werden.



Für alle Individuen beginnt der Lebensverlauf mit einem Übergang in den Zustand y_a, d.h. mit ihrer Geburt; und alle Lebensverläufe enden schließlich mit einem Übergang in den Zustand y_a, d.h. mit dem Tod. Zwischen diesen beiden Ereignissen entwickelt sich der individuelle Lebensverlauf. Bei einigen Personen findet während dieser Zeitspanne ein Übergang in den Zustand y_h statt, d.h. sie heiraten. Ihr Lebensverlauf besteht dann, in diesem einfachen Zustandsraum, aus zwei Episoden: die erste Episode beginnt mit der Geburt und endet mit der Heirat, die zweite Episode beginnt mit der Heirat und endet mit dem Tod. Die anderen Personen haben nur eine Episode, die mit der Geburt beginnt und mit dem Tod endet. Für alle Personen gibt es mindestens eine, die erste Episode, die mit einer Heirat *oder* mit dem Tod endet. Um sie statistisch zu beschreiben, kann man folgende zweidimensionale Zufallsvariable definierten:

$$(T, D) : \Omega \longrightarrow \mathcal{T} \times \mathcal{D}$$

T ist die Zeitdauer der Episode, bis zum erstenmal der Ausgangszustand y_a verlassen wird; T bezeichnet die Prozeßzeitachse, die mit der Geburt beginnt. Die Zufallsvariable D gibt den Zustand an, der am Ende der Episode erreicht wird, der Wertebereich für D ist also die Menge $\mathcal{D} = \{y_h, y_e\}$.

Rein formal können beliebige Aspekte dieser zweidimensionalen Verteilung betrachtet werden. Zum Beispiel könnte man die bedingte Verteilung $\mathrm{P}(T \leq t \mid D = y_h)$ betrachten und als Verteilung der Zeitdauer bis zur ersten Heirat interpretieren. Es muß jedoch überlegt werden, ob eine solche Interpretation sinnvoll ist. Dies ist keine statistische Frage, sie bezieht sich vielmehr darauf, welche Vorstellung wir uns von Lebensverläufen machen wollen. Die Konditionierung auf das Ereignis $(D = y_h)$ erscheint sinnvoll, wenn man annimmt, daß es bereits bei der Geburt eines Menschen feststeht, ob er später einmal heiraten wird. Sie wird jedoch sinnlos, wenn man sich Lebensverläufe als kontingente Prozesse vorstellt, bei denen sich grundsätzlich nur im Nachhinein feststellen läßt, wie sie tatsächlich abgelaufen sind.

2.4.3 Lebensverläufe als kontingente Prozesse

Wie in der Einleitung ausgeführt worden ist, betrachte ich in dieser Arbeit Lebensverläufe als kontingente Prozesse. Lebensverläufe entwickeln sich; zu jedem bereits erreichten Zeitpunkt ist die Vergangenheit abgeschlossen und nicht mehr änderbar, aber die zukünftige Entwicklung des Lebensverlaufs ist mehr oder weniger unbestimmt. Sie unterliegt zwar den Bedingungen,

die sich in der Vergangenheit herausgebildet haben, aber die Vergangenheit determiniert nicht die Zukunft, sondern bildet nur mehr oder weniger einflußreiche Bedingungen für den jeweils weiteren Lebensverlauf.

Es ist verhältnismäßig einfach, diesen Charakter von Lebensverläufen umgangssprachlich zu beschreiben. Die Umgangssprache stellt eine Fülle sehr differenzierter Formulierungen bereit, mit denen die Kontingenz von Lebensverläufen beschrieben und reflektiert werden kann. Darin kommt zum Ausdruck, daß die Annahme kontingenter Lebensverläufe eine Basisannahme des üblichen sozialen Lebens darstellt. Für die soziologische Theoriebildung ist dieses Potential der Umgangssprache jedoch nur begrenzt hilfreich, denn ihr Ziel liegt nicht in der Beschreibung individueller Lebensverläufe. Zumindest ein wesentliches Ziel liegt vielmehr darin, Einsichten in gesellschaftliche Bedingungen individueller Lebensverläufe zu gewinnen. In dieser Arbeit geht es um die Frage, welchen Beitrag die *statistische* Beschreibung von Lebensverläufen zu dieser Aufgabe leisten kann. Es ist also zu überlegen, wie der Kontingenz von Lebensverläufen in ihrer statistischen Beschreibung Rechnung getragen werden kann.

Dies kann, wie ich glaube, durch die Annahme eines einfachen, aber grundlegenden Prinzips erreicht werden: *Die Beschreibung eines zeitlichen Prozesses muß eine sinnvolle Unterscheidung von Vergangenheit, Gegenwart und Zukunft ermöglichen; und Aussagen über die Entwicklung des Prozesses in seiner jeweiligen Gegenwart dürfen nur von Bedingungen abhängig gemacht werden, die sich in der jeweiligen Vergangenheit herausgebildet haben.*

In gewisser Weise handelt es sich bei dieser Basisannahme nur um eine Reformulierung unserer traditionellen Vorstellungen über Kausalität: nur die Vergangenheit kann Einfluß auf die Zukunft nehmen.[49] Ich möchte jedoch an dieser Stelle zunächst offen lassen, ob und ggf. in welcher Weise man sich Lebensverläufe als kausal verursacht vorstellen kann. Die Statistik bietet stattdessen den Begriff einer bedingten Wahrscheinlichkeitsverteilung an, der zunächst deskriptiv interpretiert werden kann und unterschiedliche theoretische Deutungen zuläßt. Die soziologisch wichtige Frage, in welcher Weise davon gesprochen werden kann, daß individuelle Lebensverläufe von gesellschaftlichen Bedingungen abhängen, wird in Abschnitt 4.1 behandelt.

Um eine statistische Beschreibung von Prozessen zu erreichen, die dem oben formulierten Beschreibungsprinzip entspricht, dient der Begriff „Übergangsrate“, zugleich ein zentraler Begriff, um die statistische Modellbildung mit einer soziologischen Theoriebildung zu verknüpfen. Es ist deshalb sinnvoll, diesen Begriff ausführlich zu erörtern.

a) Ich beginne mit dem bereits angeführten Beispiel, in dem es nur zwei mögliche Ereignisse gibt: die Geburt (Übergang in den Zustand y_a) und

[49] Vgl. zur philosophischen Diskussion dieser Basisannahme u.a. Ayer [1956, S. 170ff] und Eells [1991, S. 239ff].

den Tod (Übergang in den Zustand y_e). Zur Vereinfachung setze ich eine Prozeßzeitachse voraus, nehme also an, daß alle Individuen aus der Grundgesamtheit Ω zum gleichen Zeitpunkt ($t = 0$) geboren werden. Zunächst wird außerdem angenommen, daß die Zeitachse diskret ist, so daß jeder Lebensverlauf aus einer *Folge* von Zuständen besteht.

Man kann sich dann die Gesamtheit der Lebensverläufe als einen *Prozeß* vorstellen. Diesen Begriff verwende ich hier und im folgenden in seiner statistischen Bedeutung: als eine zeitliche Folge (oder, im stetigen Fall, als eine zeitlich geordnete „Familie“) von Zufallsvariablen, die für eine gegebene Grundgesamtheit von Individuen definiert sind, in formaler Schreibweise:

$$Y_t : \Omega \longrightarrow \mathcal{Y} = \{y_a, y_e\} \qquad t \in \mathcal{T} \tag{2.6}$$

$\mathcal{T}$ ist eine diskrete Prozeßzeitachse. Zu jedem Zeitpunkt $t \geq 0$ gibt es eine Zufallsvariable Y_t, deren Verteilung angibt, wieviele Personen sich zu diesem Zeitpunkt im Zustand y_a befinden, d.h. noch leben, und wieviele sich im Zustand y_e befinden, d.h. bereits gestorben sind.

Der Prozeß kann dann auf einfache Weise durch eine Übergangsrate beschrieben werden, die folgendermaßen definiert ist.

$$r(t) = \mathrm{P}\left(Y_t = y_e \mid Y_0 = y_a, \ldots, Y_{t-1} = y_a\right) \tag{2.7}$$

Es ist die bedingte Wahrscheinlichkeit, daß zum Zeitpunkt t ein Übergang in den Zustand y_e eintritt, unter der Bedingung, daß der Ausgangszustand y_a bis zum Zeitpunkt $t-1$ (einschließlich) noch nicht verlassen wurde.

In einem deskriptiven Wahrscheinlichkeitsraum ergibt sich eine einfache Interpretation: $r(t)$ ist der Anteil der Personen, an der Gesamtheit der zum Zeitpunkt $t-1$ noch lebenden Personen, die zum Zeitpunkt t sterben. Diese deskriptive Aussage kann offenbar auch als eine Aussage über Chancen bzw. Risiken interpretiert werden. In diesem Fall charakterisiert $r(t)$ das Risiko, daß ein beliebiges Individuum Individuum, das mindestens bis zum Zeitpunkt $t-1$ lebt, zum Zeitpunkt t stirbt.[50] Wichtig ist in jedem Fall, daß die Übergangsrate eine Eigenschaft einer Gesamtheit von Individuen beschreibt, nicht eine Eigenschaft ihrer individuellen Mitglieder. Dies ist eine unmittelbare Folge der statistischen Betrachtungsweise, die von den Individuen abstrahiert, um zu Aussagen über Gesamtheiten von Individuen zu gelangen.

Ersichtlich entspricht der Begriff der Übergangsrate dem oben genannten Beschreibungsprinzip für Prozesse. Erstens kann mit diesem Begriff der Prozeßverlauf *zu jedem Zeitpunkt* beschrieben werden, wodurch unmittelbar die Möglichkeit entsteht, zeitpunktbezogen von Vergangenheit,

[50] Ich betone hier, daß es sich um ein *beliebiges* Individuum handelt, nicht um ein bestimmtes oder um ein „durchschnittliches“ Individuum; insbesondere müßte, um Aussagen über ein bestimmtes Individuum machen zu können, zu einem anderen Typ, nämlich zu einzelfallbezogenen Wahrscheinlichkeitsaussagen übergegangen werden.

Gegenwart und Zukunft zu sprechen. Zweitens wird bei der Bildung einer bedingten Wahrscheinlichkeit nur auf die jeweilige Vergangenheit Bezug genommen; bei der Beschreibung des Prozeßverlaufs zum Zeitpunkt t wird nur auf die jeweils bisherige Entwicklung des Prozesses (im Zeitraum 0 bis $t-1$) Bezug genommen.

Die Definition der Übergangsrate in (2.7) bezieht sich auf die in (2.6) definierten Zufallsvariablen. Der Begriff ist jedoch unabhängig davon, ob ein Prozeß durch eine Folge von Zufallsvariablen oder durch Episoden beschrieben wird. In unserem Beispiel gibt es für jedes Individuum eine Episode, von der Geburt bis zum Tod, und da es nur einen möglichen Zielzustand gibt, kann der Lebensverlauf vollständig durch eine einzige Zufallsvariable

$$T : \Omega \longrightarrow \mathcal{T} \tag{2.8}$$

beschrieben werden, die die Lebensdauer der Individuen repräsentiert. Man erhält dann folgende Definition für die Übergangsrate:

$$r(t) = \mathrm{P}\,(T = t \mid T \geq t)$$

Ersichtlich ist sie äquivalent zur Definition in (2.7); nur die Form der Darstellung hat sich geändert.

Es ist bemerkenswert, daß die Übergangsrate eine vollständige Charakterisierung der Zufallsvariable T liefert. Man sieht dies sehr einfach folgendermaßen. Die Verteilung von T kann zunächst durch eine Verteilungsfunktion

$$F(t) \;=\; \mathrm{P}\,(T \leq t)$$

oder eine Survivorfunktion

$$G(t) \;=\; 1 - F(t) = \mathrm{P}\,(T > t)$$

beschrieben werden. Nun gilt offenbar

$$1 - r(t) \;=\; 1 - \frac{\mathrm{P}\,(T = t)}{\mathrm{P}\,(T \geq t)} \;=\; \frac{\mathrm{P}\,(T > t)}{\mathrm{P}\,(T \geq t)} \tag{2.9}$$

Da ein Zustandswechsel frühestens zum Zeitpunkt $t = 1$ stattfinden kann, gilt außerdem $\mathrm{P}(T \geq 1) = 1$. Aus (2.9) folgt durch Induktion

$$\mathrm{P}\,(T > t) \;=\; \prod_{\tau=1}^{t} (1 - r(\tau))$$

Schließlich ergibt sich

$$\mathrm{P}\,(T = t) \;=\; r(t)\,\mathrm{P}\,(T \geq t) \;=\; r(t) \prod_{\tau=1}^{t-1} (1 - r(t))$$

Kennt man die Übergangsrate $r(t)$, kann die Verteilung von T nach dieser Formel berechnet werden; und umgekehrt kann die Übergangsrate aus einer Kenntnis der Verteilung von T berechnet werden, vgl. (2.9). Diese mathematische Äquivalenz verschleiert jedoch einen wesentlichen inhaltlichen Unterschied. Die üblichen Begriffe zur Beschreibung der Verteilung einer Zufallsvariablen sind statisch; sie setzen voraus, daß der zu beschreibende Sachverhalt sich bereits vollständig realisiert hat. Der Begriff der Übergangsrate erlaubt es dagegen, die *Entwicklung* eines Prozesses zu beschreiben; er erlaubt es, sich auf jeden einzelnen Zeitpunkt in der Entwicklung des Prozesses zu beziehen.

b) Unter der Voraussetzung einer diskreten Zeitachse kann der Begriff der Übergangsrate unmittelbar als eine bedingte Wahrscheinlichkeit eingeführt werden. Geht man stattdessen von einer stetigen Zeitachse aus, muß die Begriffsbildung etwas modifiziert werden. Um dies zu erläutern, gehen wir wieder von der einfachen, in (2.8) definierten Situation aus, betrachten die Lebensdauer T jetzt jedoch als eine stetige Zufallsvariable. Die Übergangsrate kann dann folgendermaßen definiert werden:

$$r(t) \;=\; \lim_{\Delta t \to 0} \frac{1}{\Delta t}\,\mathrm{P}\,(t \leq T < t + \Delta t \mid T \geq t)$$

Die so definierte Übergangsrate ist nicht unmittelbar eine bedingte Wahrscheinlichkeit, sondern eine bedingte Wahrscheinlichkeitsdichte.[51] Der Grund für diese Begriffsbildung liegt darin, daß bei einer stetigen Zeitachse die Wahrscheinlichkeit, daß ein Ereignis zu einem bestimmten Zeitpunkt stattfindet, Null ist. Ein Übergang zu bedingten Wahrscheinlichkeiten kann jedoch, näherungsweise, leicht vollzogen werden. Zunächst ist es zweckmäßig, sich kurz zu vergegenwärtigen, daß auch bei einer stetigen Zeitachse die Übergangsrate eine vollständige Beschreibung der Verteilung der Zufallsvariable T liefert. Ausgangspunkt sei zunächst die Beschreibung durch eine Verteilungsfunktion

$$F(t) = \mathrm{P}\,(T \leq t)$$

Daraus kann unmittelbar die Dichtefunktion

$$f(t) = \frac{\mathrm{d}}{\mathrm{d}t}\,F(t)$$

und die Survivorfunktion

$$G(t) = 1 - F(t) = \mathrm{P}\,(T > t)$$

abgeleitet werden. Unter Verwendung dieser Begriffe erhält man für die Übergangsrate die Formulierung

$$r(t) \;=\; \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t}\,\frac{1}{G(t)} \;=\; \frac{f(t)}{G(t)} \tag{2.10}$$

[51] In der Literatur wird deshalb auch die Bezeichnung *Übergangsintensität* verwendet, oft wird auch von einer *Hazardrate* gesprochen.

d.h. die Übergangsrate kann als Quotient der Dichte- und der Survivorfunktion berechnet werden. Andererseits können aus der Übergangsrate die Dichte-, die Verteilungs- und die Survivorfunktion abgeleitet werden. Denn offenbar gilt

$$-r(t) = \frac{\mathrm{d}}{\mathrm{d}t} \log\left(G(t)\right) \tag{2.11}$$

und durch Integration erhält man

$$\mathrm{P}\left(T \geq t_2 \mid T \geq t_1\right) \;=\; \frac{G(t_2)}{G(t_1)} \;=\; \exp\left\{-\int_{t_1}^{t_2} r(\tau)\,\mathrm{d}\tau\right\}$$

Insbesondere erhält man für $t_1 = 0$ die Beziehung

$$G(t) \;=\; \exp\left\{-\int_{0}^{t} r(\tau)\,\mathrm{d}\tau\right\} \tag{2.12}$$

Durch Integration der Übergangsrate kann also zunächst die Survivorfunktion gewonnen werden, und daraus können dann unmittelbar die Dichte- und die Verteilungsfunktion abgeleitet werden.

Bei einer diskreten Zeitachse entspricht die Übergangsrate einer Übergangswahrscheinlichkeit. Um bei einer stetigen Zeitachse Übergangswahrscheinlichkeiten zu berechnen, kann folgendermaßen vorgegangen werden.

$$\begin{aligned}
\mathrm{P}\left(t \leq T < t + \Delta t \mid T \geq t\right) &= \frac{F(t+\Delta t) - F(t)}{G(t)} \\
&= \frac{G(t) - G(t+\Delta t)}{G(t)} \;=\; 1 - \frac{G(t+\Delta t)}{G(t)} \\
&= 1 - \exp\left\{-\int_{t}^{t+\Delta t} r(\tau)\,\mathrm{d}\tau\right\}
\end{aligned}$$

Oben links steht die gesuchte Übergangswahrscheinlichkeit, d.h. die Wahrscheinlichkeit, daß im Zeitintervall $(t, t+\Delta t)$ ein Ereignis eintritt, unter der Bedingung, daß bis zum Beginn dieses Zeitintervalls der Ausgangszustand noch nicht verlassen wurde. Wie die Ableitung zeigt, kann diese Übergangswahrscheinlichkeit aus der Integration der Übergangsrate in dem entsprechenden Zeitintervall gewonnen werden. Wenn das Zeitintervall Δt klein ist, kann man außerdem eine Näherung vornehmen. Denn für kleine Werte von x gilt näherungsweise: $1 - \exp(-x) \approx x$; Abbildung 2.4.1 illustriert diesen Sachverhalt.

Für kleine Werte von Δt kann man also zunächst das Integral der Übergangsrate durch $r(t)\,\Delta t$ approximieren und erhält dann als Näherung für die Übergangswahrscheinlichkeit die Beziehung

$$\mathrm{P}\left(t \leq T < t + \Delta t \mid T \geq t\right) \;\approx\; r(t)\,\Delta t \tag{2.13}$$

Wenn man insbesondere von einer kleinen Zeiteinheit für die stetige Zeitachse ausgeht, kann man die Übergangsrate näherungsweise mit einer Übergangswahrscheinlichkeit (definiert in der zugrundeliegenden Zeiteinheit) identifizieren.

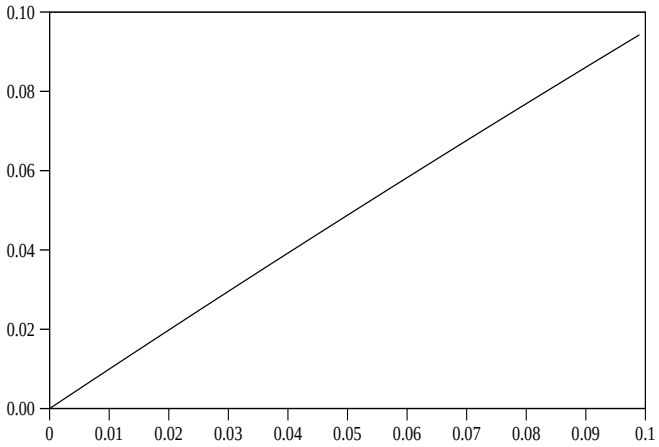


Abbildung 2.4.1 Graph der Funktion $y = 1 - \exp(-x)$ für kleine Werte von x.

Der Übergang von einer diskreten zu einer stetigen Zeitachse impliziert zwar einige Modifikationen in der mathematischen Form der Darstellung, er hat aber im Hinblick auf die Aufgabe, eine geeignete sprachliche Form für die statistische Beschreibung von Lebensverläufen zu finden, keine besondere Bedeutung. Der Grenzübergang zu infinitesimal kleinen Zeitintervallen erlaubt häufig eine einfachere mathematische Darstellung; um eine anschauliche Beschreibung und Interpretation realer Prozesse zu erreichen, müssen jedoch stets Zeitintervalle oder sie repräsentierende Zeitpunkte betrachtet werden.

c) Die Übergangsrate erlaubt eine zeitpunktbezogene Beschreibung eines Prozesses; $r(t)$ informiert darüber, wie sich der Prozeß zum Zeitpunkt t entwickelt. Um eine Beschreibung des gesamten Prozesses zu erreichen, muß man also $r(t)$ für alle möglichen Prozeßzeitpunkte kennen, d.h. $r(t)$ als eine Funktion der Zeit t betrachten.

Im allgemeinen kann sich die Übergangsrate in Abhängigkeit von der Prozeßzeit verändern. Eine zeitkonstante Übergangsrate ist ein Spezialfall, der bei realen Prozessen praktisch nie anzutreffen ist.[52] Bei einer stetigen Zeitachse kann man diesen Spezialfall auf einfache Weise aus Gleichung (2.12) ableiten. Sei nämlich $r(t) = \lambda$, eine Konstante. Dann gewinnt man

[52] Vgl. die ausführliche Diskussion bei Gerchak [1984].

aus (2.12) die Survivorfunktion

$$G(t) = \exp(-\lambda t)$$

Dies ist eine Exponentialverteilung mit den korrespondierenden Verteilungs- und Dichtefunktionen

$$\begin{aligned} F(t) &= 1 - \exp(-\lambda t) \\ f(t) &= \lambda \exp(-\lambda t) \end{aligned}$$

Bei einer stetigen Zeitachse ist dies die einzige Verteilung mit einer zeitunabhängigen (konstanten) Übergangsrate.

Bei der Beschreibung realer Lebensverläufe findet man im allgemeinen, daß die Übergangsrate auf vielfältige Weise mit der Prozeßzeit variiert. Als Beispiel sei die Übergangsrate für die erste Heirat bei den Personen aus der Teilstichprobe A des SOEP angeführt. Eine Abbildung der Survivorfunktion für dieses Beispiel wurde bereits in Abschnitt 2.3.1 gegeben (vgl. Abbildung 2.3.1). Für die Berechnung wurde das Schätzverfahren von Kaplan und Meier verwendet. Leider liefert dies Verfahren nicht unmittelbar auch Schätzungen der Übergangsrate. Eine näherungsweise Berechnung ist jedoch durch numerische Differentiation des Logarithmus der Survivorfunktion möglich; vgl. (2.11). Abbildung 2.4.2 zeigt die auf diese Weise berechnete Übergangsrate.[53]

Die Übergangsrate in Abbildung 2.4.2 liefert eine anschauliche Darstellung der Entwicklung von Lebensverläufen bis zur ersten Heirat. Bis etwa zum 16. Lebensjahr finden fast keine Heiraten statt, dann erfolgt ein rascher Anstieg der Heiratsneigung. Der Höhepunkt wird bei einem Alter von etwa 25 Jahren erreicht, dann flacht die Rate allmählich ab. Zum Beispiel beträgt die Übergangsrate im Alter von 20 Jahren ungefähr 0.06. Unter Verwendung der in (2.13) entwickelten Approximation heißt das, daß etwa 6 % der mit 20 noch ledigen Personen bis zu ihrem 21. Lebensjahr heiraten. Ganz analog findet man, daß etwa 16 % der mit 25 noch ledigen Personen während des folgenden Jahres heiraten.

Bei der Interpretation dieser Zahlen ist darauf zu achten, daß die Übergangsrate eine bedingte Wahrscheinlichkeit (bzw. Dichte) ist, d.h. sie bezieht sich auf diejenigen Personen, für die bis zum jeweiligen Zeitpunkt der zu beschreibende Zustandswechsel *noch nicht* eingetreten ist. Die Verteilung der Anzahl der Ereignisse wird durch die Wahrscheinlichkeitsfunktion oder -dichte der Zufallsvariable T beschrieben, bei einer stetigen Zeitachse

[53] Da das Kaplan-Meier-Verfahren eine Schätzung der Survivorfunktion nur in der Form einer Treppenfunktion zur Verfügung stellt, neigt die Ableitung ihres Logarithmus zu starken Fluktuationen. Um eine einigermaßen glatte Näherung an die Übergangsrate zu erreichen, ist es deshalb sinnvoll, den Logarithmus der Survivorfunktion zunächst zu glätten, bevor eine numerische Differentiation vorgenommen wird. Für die Berechnung der Übergangsrate in Abbildung 2.4.2 wurde die Glättung mit einem von Dierckx [1975] entwickelten Algorithmus vorgenommen.

$f(t) = r(t)\,G(t)$, vgl. (2.10).

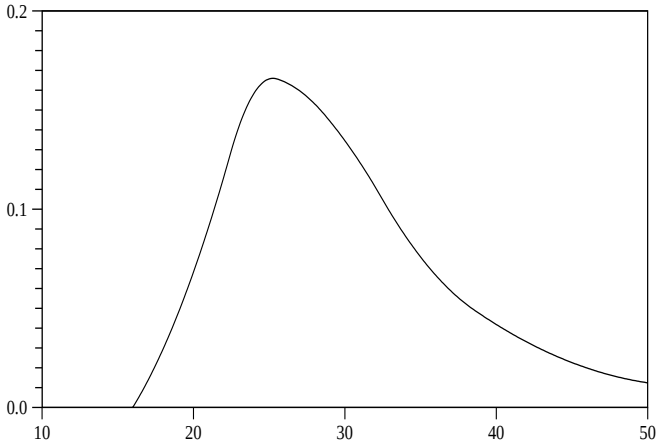


Abbildung 2.4.2 Übergangsrate für das Ereignis *erste Heirat* bei den Personen aus der Teilstichprobe A des SOEP; berechnet durch numerische Differentiation einer geglätteten Kaplan-Meier-Schätzung der Survivorfunktion. Abszisse: Lebensalter in Jahren.

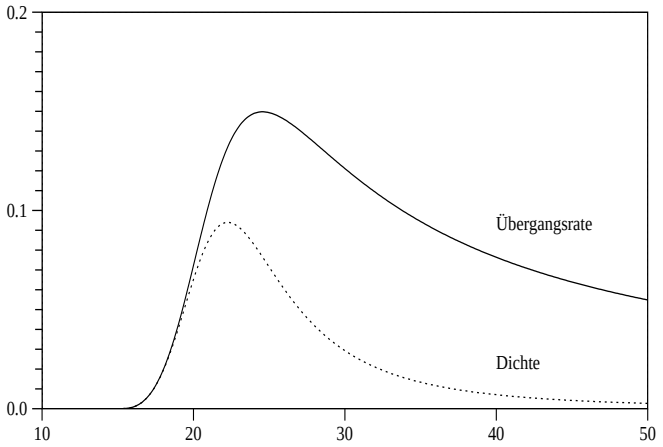


Abbildung 2.4.3 Übergangsrate und Dichtefunktion für das Ereignis *erste Heirat* bei den Personen aus der Teilstichprobe A des SOEP; berechnet aus der Maximum-Likelihood-Schätzung eines erweiterten log-logistischen Modells. Abszisse: Lebensalter in Jahren.

Abbildung 2.4.3 illustriert den Zusammenhang zwischen Übergangsrate und Dichtefunktion, wiederum am Beispiel des Ereignisses *erste Heirat* bei den Personen aus der Teilstichprobe A des Sozio-ökonomischen Panels. Zugrunde liegt eine Maximum-Likelihood-Schätzung eines erweiterten log-logistischen Modells.[54]

d) Die Übergangsrate $r(t)$ ist eine mathematische Funktion der Zeit. Diese Begriffsbildung ist sinnvoll, da sich die Übergangsrate in der Regel, bei den meisten realen Prozessen, während des Prozeßverlaufs verändert. Man kann dann sagen, daß die Übergangsrate „von der Zeit abhängt". Es sollte allerdings betont werden, daß diese Formulierung nicht im Sinne einer kausalen Erklärung verstanden werden kann. Die Übergangsrate ist zunächst ein rein deskriptives Konzept; sie beschreibt die Entwicklung eines Prozesses. Sie liefert Antworten auf Wie-Fragen; über die Frage, warum sich der Prozeß in einer bestimmten Weise entwickelt, wird durch die Feststellung einer zeitabhängigen Übergangsrate zunächst keinerlei Aussage getroffen.

Auf die Frage, wie der statistische Begriff der Übergangsrate für die soziologische Theoriebildung genutzt werden kann, wird in Abschnitt 4.1 näher eingegangen. Hier konzentrieren wir uns zunächst auf den deskriptiven Gehalt des Begriffs. Zwei mögliche Ansatzpunkte für theoretische Deutungen können jedoch bereits in diesem Zusammenhang erwähnt werden.

Der erste Punkt betrifft die Frage, *was* durch die Übergangsrate beschrieben wird. Ausgehend von der Feststellung, daß sich eine deskriptive Interpretation statistischer Begriffe stets auf Gesamtheiten beziehen muß, hatten wir die Übergangsrate als Eigenschaft einer Gesamtheit von Individuen interpretiert. Sie beschreibt den Prozeßverlauf für eine gegebene Gesamtheit von Personen. Dies ist jedenfalls dann sinnvoll, wenn man die Übergangsrate als eine Funktion der Zeit ansieht. Man kann sie jedoch auch für jeden Zeitpunkt gesondert betrachten; $r(t)$ ist dann die Übergangsrate für den Prozeßverlauf *zum Zeitpunkt* t. So betrachtet, bezieht sich die Übergangsrate nur auf diejenigen Personen, die zu diesem Zeitpunkt noch dem Risiko eines Zustandswechsels ausgesetzt sind. Man kann also versuchen, die Übergangsrate nicht nur als eine Eigenschaft der Grundgesamtheit anzusehen, sondern sie als Eigenschaft einer sich im Prozeßablauf verändernden Gesamtheit von Personen zu deuten.

Es ist üblich, die Menge derjenigen Personen, die sich bis zu einem Zeitpunkt t noch im Ausgangszustand der Episode befinden, als *Risikomenge zum Zeitpunkt* t zu bezeichnen, symbolisch: $R(t)$. Der Begriff entspricht einer anschaulichen Vorstellung für die Entwicklung eines Prozesses: Zu

[54] Diese Erweiterung des gewöhnlichen log-logistischen Modells wurde von J. Brüderl vorgeschlagen; vgl. Abschnitt 3.5.4. Im Vergleich zu den einfachen parametrischen Standardmodellen liefert diese Modellvariante eine deutlich bessere Annäherung an die für unser Beispiel verwendeten Daten. Die Modellschätzung wurde mit dem Programm TDA vorgenommen.

Beginn des Prozesses sind noch alle Personen aus der zugrundeliegenden Gesamtheit in dem vorausgesetzten Ausgangszustand der Episode; während des sich entwickelnden Prozesses finden dann Zustandswechsel statt. Zunächst verlassen einige, dann immer mehr Personen ihren Ausgangszustand und geraten in einen der möglichen Folgezustände. Schließlich, nach hinreichend langer Zeit und wenn wir von einem vollständigen Zustandsraum ausgehen, der den Tod als ein irgendwann sicher eintretendes Ereignis umfaßt, haben alle Personen ihren Ausgangszustand verlassen, die Risikomenge ist dann leer.

Diese Überlegung zeigt, daß man die Übergangsrate $r(t)$ als eine Eigenschaft von $R(t)$, der Risikomenge zum Zeitpunkt t ansehen kann. Dabei ist jedoch zu beachten, daß sich die Risikomenge während des Prozesses fortwährend verändert. Welche Personen tatsächlich zur Risikomenge $R(t)$ gehören, steht erst fest, wenn der Prozeß bis zum Zeitpunkt t abgelaufen ist. Betrachten wir unser Heiratsbeispiel. Die Risikomenge zum Zeitpunkt $t = 30$ besteht in diesem Fall aus denjenigen Personen, die bis zu ihrem 30. Lebensjahr noch nicht geheiratet haben. Ob eine Person zu dieser Risikomenge gehört, weiß man frühestens, wenn sie 30 Jahre alt geworden ist. Bis zu diesem Zeitpunkt ist es eine offene Frage, über die man bestenfalls subjektive Erwartungen bilden kann. Die Zugehörigkeit zu einer Risikomenge kann also nicht als statische Eigenschaft einem Individuum zugerechnet werden, sondern sollte als eine kontingente Eigenschaft von Lebensverläufen angesehen werden.

Der zweite Punkt bezieht sich auf die Möglichkeit, die Individuen in einer Grundgesamtheit zu klassifizieren. Zunächst rein formal betrachet, wird dann die Grundgesamtheit Ω in eine Reihe von Teilgesamtheiten $\Omega_1, \ldots, \Omega_m$ zerlegt, so daß der Prozeß für jede dieser Teilgesamtheiten gesondert beschrieben werden kann; $r_j(t)$ sei die Übergangsrate, die die Entwicklung des Prozesses in der Teilgesamtheit Ω_j beschreibt. Man kann dann $r(t)$, die Übergangsrate für den Prozeß in der Grundgesamtheit Ω, als eine „Mischung" aus den Übergangsraten $r_j(t)$ darstellen. Für eine diskrete Zeitachse sieht man dies auf einfache Weise folgendermaßen. Ausgangspunkt ist eine zweidimensionale Zufallsvariable

$$(T, X) : \Omega \longrightarrow \mathcal{T} \times \{1, \ldots, m\}$$

T ist die Verweildauer in der zu beschreibenden Episode, die Zufallsvariable X gibt für jedes Individuum an, zu welcher Teilgesamtheit es gehört. Für

den Zusammenhang der Übergangsraten erhält man

$$\begin{aligned} r(t) &= \mathrm{P}\,(T = t \mid T \geq t) \\ &= \sum_{j=1}^{m} \mathrm{P}\,(T = t, X = j \mid T \geq t) \\ &= \sum_{j=1}^{m} \mathrm{P}\,(T = t \mid T \geq t, X = j)\,\mathrm{P}\,(X = j \mid T \geq t) \\ &= \sum_{j=1}^{m} r_j(t)\,\frac{G_j(t)}{G(t)}\,\mathrm{P}\,(X = j) \\ &= \sum_{j=1}^{m} r_j(t)\,\frac{|R_j(t)|}{|R(t)|} \end{aligned}$$

Die Übergangsrate $r(t)$ ist eine Mischung, in diesem Fall ein gewichteter Durchschnitt, aus den Übergangsraten $r_j(t)$. Die Gewichte sind die Proportionen, in denen sich die Risikomenge $R(t)$ in der Grundgesamtheit aus den Risikomengen $R_j(t)$ in den Teilgesamtheiten zusammensetzt. Unterscheiden sich die Übergangsraten in den Teilgesamtheiten, findet dementsprechend eine unterschiedliche Entwicklung der Risikomengen $R_j(t)$ statt, und es resultiert eine fortwährende Veränderung in den Mischungsgewichten.

Dies kann dazu führen, daß die Übergangsrate in der Grundgesamtheit einen wesentlich anderen Verlauf aufweist als in den einzelnen Teilgesamtheiten.[55] Insbesondere kann die Übergangsrate in der Grundgesamtheit einen fallenden Verlauf aufweisen, obwohl dies in den einzelnen Teilgesamtheiten nicht der Fall ist. Wenn die Übergangsrate in einer Teilgesamtheit überdurchschnittlich groß ist, wird die zugehörige Risikomenge schnell kleiner und ihr Gewicht in der gesamten Risikomenge nimmt dementsprechend ab. Die gesamte Risikomenge besteht zunehmend aus Personen, die solchen Teilgesamtheiten angehören, in denen die Übergangsrate vergleichsweise niedrig ist. – Diese Überlegung zeigt, daß man versuchen kann, den Verlauf der Übergangsrate in einer Grundgesamtheit dadurch *zu erklären*, daß man ihn als eine Mischung aus den Übergangsraten in Teilgesamtheiten darstellt.

2.4.4 Konkurrierende Risiken

Eine gewisse Erweiterung des begrifflichen Rahmens ist erforderlich, um eine angemessene Form der Beschreibung von „konkurrierenden Risiken“ zu

[55] Einige interessante und teilweise überraschende Beispiele werden von Vaupel und Yashin [1985] diskutiert.

erreichen. Wie bereits ausgeführt wurde, ist dies der Regelfall; typischerweise gibt es für jeden Ausgangszustand mindestens zwei, oft zahlreiche mögliche Folgezustände. Insbesondere ist stets die Möglichkeit gegeben, daß ein Mensch stirbt und dadurch seinen Lebensverlauf beendet. Strenggenommen ist die einzige Situation ohne konkurrierende Risiken diejenige, in der nur der Tod als ein alternativenloses Ereignis betrachtet wird.

Eine allgemeine Situation konkurrierender Risiken kann durch eine zweidimensionale Zufallsvariable

$$(T, D) : \Omega \longrightarrow \mathcal{T} \times \mathcal{D}$$

charakterisiert werden. $\mathcal{T}$ ist eine diskrete Prozeßzeitachse, so daß sich alle Personen aus einer vorausgesetzten Gesamtheit Ω zum Zeitpunkt $t = 0$ in einem Anfangszustand y_a befinden. T ist die Zeitdauer der Episode, bis zum erstenmal der Anfangszustand verlassen wird. $\mathcal{D} = \{d_1, \ldots, d_m\}$ ist die Menge der möglichen Folgezustände.

Wir gehen davon aus, daß es sich um einen kontingenten Prozeß handelt, d.h. jedes Individuum kann jederzeit, solange es den Ausgangszustand noch nicht verlassen hat, in einen der möglichen Folgezustände überwechseln. Um einen solchen Prozeß zu beschreiben, liegt es nahe, den Begriff der Übergangsrate folgendermaßen zu erweitern:

$$r_k(t) = \mathrm{P}\,(T = t, D = d_k \mid T \geq t)$$

Man erhält dadurch zustandsspezifische Übergangsraten; $r_k(t)$ ist die bedingte Wahrscheinlichkeit, daß zum Zeitpunkt t ein Übergang in den Zustand d_k erfolgt, unter der Bedingung, daß der Ausgangszustand bis zum Zeitpunkt $t - 1$ noch nicht verlassen wurde. Diese Definition setzt eine diskrete Zeitachse voraus. Ganz analog können zustandsspezifische Übergangsraten bei einer stetigen Zeitachse definiert werden:

$$r_k(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\,\mathrm{P}\,(t \leq T < t + \Delta t, D = d_k \mid T \geq t)$$

Es sei betont, daß die Bildung konditionaler Wahrscheinlichkeiten nur auf die jeweilige Vergangenheit Bezug nimmt. Dies entspricht dem oben, S. 82, eingeführten Prinzip zur Beschreibung kontingenter Prozesse. Für diejenigen Personen, die den Ausgangszustand noch nicht verlassen haben, ist nur die bisherige Verweildauer in diesem Zustand bekannt, nicht bereits, in welchen Folgezustand sie später möglicherweise wechseln werden. Die Begriffsbildung trägt dem Rechnung, indem nur auf diesen Sachverhalt, d.h. auf $T \geq t$, konditioniert wird.

Die zustandsspezifischen Übergangsraten können addiert werden; zum Beispiel ist $r_k(t) + r_l(t)$ die Übergangsrate für einen Wechsel in den Folgezustand d_k *oder* d_l. Insbesondere gilt

$$r(t) = \sum_{d_k \in \mathcal{D}} r_k(t)$$

d.h. die Addition aller zustandsspezifischen Übergangsraten liefert die einfache Abgangsrate, d.h. die Übergangsrate in irgendeinen Folgezustand.

Gelegentlich ist behauptet worden, daß diese grundlegende Eigenschaft der Additivität der zustandsspezifischen Übergangsraten Annahmen über die „Unabhängigkeit“ der konkurrierenden Risiken erfordert. Zum Beispiel sagt Schneider [1990, S. 83]: „Die Zerlegung der allgemeinen bedingten Übergangswahrscheinlichkeit in eine Summe von spezifischen Übergangswahrscheinlichkeiten ist allerdings nur dann zulässig, wenn die Wahrscheinlichkeiten für die Übergänge in spezifische Zielzustände voneinander unabhängig sind.“ Ähnlich haben sich auch einige andere Autoren geäußert.[56] Ich möchte diese Feststellung jedoch bestreiten. Zunächst ist vollständig undefiniert, was es heißen könnte, daß konkurrierende Risiken „unabhängig“ oder „nicht unabhängig“ sind. Die Additivität der zustandsspezifischen Übergangsraten folgt dagegen unmittelbar daraus, daß die möglichen Zielzustände sich ausschließen:

$$\begin{aligned}&\mathrm{P}\left(T = t, (D = d_k \text{ oder } D = d_l)\right) \;= \\ &\qquad \mathrm{P}\left(T = t, D = d_k\right) + \mathrm{P}\left(T = t, D = d_l\right)\end{aligned}$$

Dies gilt ganz allgemein in jedem deskriptiven Wahrscheinlichkeitsraum; auch dann, wenn zusätzlich die Bedingung $(T \geq t)$ eingeführt wird.

Um die Bemerkung Schneiders zu verstehen, ist es jedoch zweckmäßig, kurz darauf einzugehen, daß es zwei wesentlich unterschiedliche Betrachtungsweisen für konkurrierende Risiken gibt. Wir sind bisher davon ausgegangen, daß die konkurrierenden Risiken in einer Menge möglicher Folgezustände bestehen. Die Individuen befinden sich in einem gegebenen Ausgangszustand, und solange dies der Fall ist, besteht die Möglichkeit, daß sie in einen der möglichen Folgezustände wechseln.

Eine ganz andere Betrachtungsweise geht davon aus, daß es *nur einen* möglichen Folgezustand gibt, daß es jedoch eine Reihe kausaler Faktoren $K_1, \ldots, K_m$ gibt, die dazu führen können, daß ein Übergang in diesen Folgezustand stattfindet. Von dieser Betrachtungsweise wird zum Beispiel in der technischen Statistik ausgegangen, wenn man sich für die Frage der Lebensdauer von Maschinen interessiert. Die konkurrierenden Risiken sind dann die Faktoren $K_1, \ldots, K_m$, die dazu führen können, daß eine Maschine funktionsuntüchtig wird.

In diesem Kontext ist es offenbar möglich, ein theoretisches Modell zu konzipieren, in dem sinnvoll über mögliche Abhängigkeiten zwischen den konkurrierenden Risiken gesprochen werden kann. Dann kann zum Beispiel die Frage formuliert werden kann, wie sich die durchschnittliche Lebensdauer der Maschine verändern würde, wenn es gelänge, einige der konkurrierenden Risiken zu beseitigen. Die übliche Modellformulierung geht

[56] Vgl. zum Beispiel die Bemerkungen von Diekmann und Mitter [1990, S. 435], sowie auch Galler [1988] und Klein [1988].

davon aus, für jeden der Faktoren K_j eine latente Verweildauer T_j anzunehmen, interpretierbar als die Lebensdauer von Maschinen (des gleichen Typs), bei denen nur der Faktor K_j wirksam ist. Die Lebensdauer von Maschinen, bei denen alle Faktoren wirksam werden können, ist das Minimum dieser latenten Lebensdauern, also $T = \min(T_1, \ldots, T_m)$. Die Frage ist dann, ob man aus der Beobachtung von T Aussagen über die Verteilung der latenten Lebensdauern T_j gewinnen kann. Insbesondere kann in diesem Kontext die Frage gestellt werden, ob diese latenten Lebensdauern als unabhängige Zufallsvariablen angesehen werden können, denn es ist natürlich vorstellbar, daß sie miteinander korreliert sein könnten.[57]

Wie gezeigt worden ist, ist diese Frage nicht entscheidbar, wenn man nur über Informationen über die Verteilung von T verfügt.[58] Es ist deshalb üblich, bei der Modellbildung von der Annahme auszugehen, daß die konkurrierenden Risiken, d.h. die mit ihnen assoziierten latenten Lebensdauern T_j, unabhängig voneinander sind. Und *diese* Annahme kann natürlich infrage gestellt werden.

Ich möchte jedoch betonen, daß man es bei der Beschreibung von Lebensverläufen mit konkurrierenden Risiken in einem vollständig anderen Sinn zu tun hat. In der technischen Statistik (und ähnlich in der medizinischen Statistik bei der Beschäftigung mit Mortalitätsrisiken) geht es um die Frage, wie kausal unterscheidbare Risikofaktoren die Lebensdauerverteilung beeinflussen. Bei der soziologischen Beschreibung kontingenter Lebensverläufe geht es darum, daß es für jeden gegebenen Ausgangszustand eine Vielzahl möglicher Folgezustände gibt. Die Aufgabe besteht zunächst darin, zu beschreiben, wie diese Übergänge vollzogen werden. Man kann dann auch versuchen, diese Übergänge zu erklären; aber es wäre offenbar sinnlos, sich die *potentiellen* Zielzustände als Ursachen der (jeweils noch zu realisierenden) Zustandsänderungen vorzustellen.[59]

Dementsprechend sind auch die konjunktiven Fragestellungen der technischen und medizinischen Statistik bei der soziologischen Beschreibung von Lebensverläufen in der Regel nicht angemessen bzw. nicht sinnvoll beantwortbar.[60] Menschen entwickeln ihre Lebensverläufe in einem jeweils gegebenen Spektrum von (subjektiv wahrgenommenen) Hand-

[57] Als Einführung in diese Betrachtungsweise konkurrierender Risiken vgl. man u.a. Namboodiri und Suchindran [1987, Kap. 8].

[58] Tsiatis [1975]; vgl. auch die Diskussion bei Gail [1975].

[59] In der Literatur wird häufig versäumt, diese beiden Betrachtungsweisen klar zu unterscheiden. Das aus der technischen Statistik bekannte Modell konkurrierender Risiken wird umstandslos so verwendet, als ob es auch für die Beschreibung kontingenter Lebensverläufe verwendet werden könnte. Zum Beispiel bezieht sich Teachman [1983, S. 287] zunächst auf eine Situation mit „possible events that an individual can experience“, dann führt er zielzustandsspezifische Übergangsraten ein und bezeichnet sie schließlich als „cause-specific hazard functions“.

[60] Vgl. dazu auch die kritischen Bemerkungen von Aalen [1987, S. 178] und Lancaster [1990, S. 107].

lungsmöglichkeiten. Wenn sich diese Handlungsmöglichkeiten verändern, verändert sich mit großer Wahrscheinlichkeit auch das Verhalten. In diesem Sinne sind die Übergangsraten sicherlich nicht „unabhängig“ von den jeweils vorhandenen konkurrierenden Risiken. Daß Menschen sich an den jeweils vorhandenen Chancen und Risiken (Handlungsmöglichkeiten) orientieren, hat jedoch mit der Vermutung, daß konkurrierende Risiken (genauer: die ihnen assoziierbaren latenten Verweildauern) in einem statistischen Sinne korreliert sein könnten, überhaupt nichts zu tun. Insofern ist es auch nicht sinnvoll, zu sagen, daß das Additivitätsprinzip für zustandsspezifische Übergangsraten auf irgendeiner statistischen Unabhängigkeitsannahme *beruht*. Wenn es an dieser Stelle ein Problem gibt, besteht es darin, daß sowohl unsere Beschreibungen als auch die schließlich zu konstruierenden statistischen Modelle von der Annahme eines jeweils gegebenen Zustandsraums abhängig sind, dieser sich jedoch auf eine nicht antizipierbare Weise verändern kann.

Darin liegt ein wesentlicher Grund für die Grenzen der Prognostizierbarkeit von Lebensverläufen. Wenn sich während der Entwicklung von Lebensverläufen die möglichen Zielzustände verändern, entsteht eine jeweils qualitativ neue Situation, für die man mit einem Modell, dessen Konstruktion und Berechnung auf dem Vorhandensein einer ganz anderen Situation beruht, kaum Prognosen machen kann. Man denke zum Beispiel an ein Modell, bei dem drei Zustände unterschieden werden: vollzeit, teilzeit und nicht erwerbstätig. Es wäre kaum sinnvoll, aus einem Modell für diesen Zustandsraum Prognosen für eine Situation abzuleiten, in der es die Möglichkeit zur Teilzeitarbeit nicht mehr gibt.

2.4.5 Lebensverläufe als Folgen von Episoden

Bisher haben wir einzelne Episoden betrachtet. Die Grundgesamtheit bestand aus allen Personen, die sich zu einem gewissen Zeitpunkt in einem vorausgesetzten Anfangszustand befinden. Es wurde dargestellt, wie mithilfe des Begriffs zustandsspezifischer Übergangsraten Episodenverläufe als kontingente Prozesse beschrieben werden können.

Diese Überlegungen können verallgemeinert werden, um komplexe Lebensverläufe zu beschreiben. Die Grundidee besteht darin, Lebensverläufe als Folgen von Episoden zu betrachten. Jeder Lebensverlauf beginnt mit der Geburt, mit dem Übergang in einen Anfangszustand y_a. Dann entwickelt sich der Lebensverlauf als eine Sequenz von Episoden. Die erste Episode, die mit dem Übergang in den Anfangszustand y_a beginnt, endet, wenn zum erstenmal dieser Zustand verlassen und ein neuer Zustand eingenommen wird. Dieser neue Zustand ist dann der Ausgangspunkt für die zweite Episode, und so weiter. Schließlich endet jeder Lebensverlauf mit dem Tod, mit dem Übergang in einen absorbierenden Endzustand y_e.

Welche Episoden möglich sind, hängt vom Zustandsraum ab, der jeder statistischen Beschreibung von Lebensverläufen vorausgesetzt werden muß. Im einfachsten Fall, wenn der Zustandsraum nur die beiden Zustände y_a und y_e umfaßt, ist nur eine Episode möglich, die mit der Geburt beginnt und mit dem Tod endet. Sobald weitere Zustände berücksichtigt werden, wird eine komplexere Beschreibung von Lebensverläufen möglich. Nicht nur kann jeder neue Zustand zum Ausgangspunkt eines neuen Episodentyps werden, sondern es werden auch immer vielfältigere Folgen von Episoden möglich.

Die Grundidee, Lebensverläufe als Folgen von Episoden zu betrachten, liefert einen einfachen Zugang zu ihrer statistischen Beschreibung. Denn es kann dann auf einfache Weise dem auf S. 82 formulierten Beschreibungsprinzip für kontingente Prozesse Rechnung getragen werden: jede Episode wird gesondert beschrieben, dabei jedoch die mögliche Abhängigkeit des Episodenverlaufs von der jeweils erreichten Vorgeschichte berücksichtigt.

a) Um die statistische Beschreibung von Lebensverläufen als Folgen von Episoden darzustellen, setzen wir, wie bisher, einen deskriptiven Wahrscheinlichkeitsraum für eine endliche Grundgesamtheit (Ω) von Individuen voraus; außerdem einen Zustandsraum $\mathcal{Y}$, der insbesondere den Anfangszustand y_a und einen absorbierenden Endzustand y_e umfaßt, und eine Zeitachse $\mathcal{T}$. Um die Einführung der Grundbegriffe etwas zu vereinfachen, nehmen wir an, daß es sich um eine diskrete Zeitachse handelt.[61] Jeder Lebensverlauf besteht dann aus einer *Folge* von Zuständen, und der Prozeß, d.h. die Gesamtheit der Lebensverläufe der Individuen aus Ω, kann durch eine Folge von Zufallsvariablen repräsentiert werden.

b) Jedes Individuum beginnt seinen Lebensverlauf mit einem kontingenten Ereignis, seiner Geburt. Im allgemeinen ist also davon auszugehen, daß die Individuen aus Ω ihre Lebensverläufe zu jeweils unterschiedlichen Kalenderzeitpunkten beginnen. Um dies angemessen beschreiben zu können, ist es erforderlich, die individuellen Lebensverläufe in einen historischen Prozeß einzubetten. Um von den damit verbundenen Problemen hier absehen zu können, setzen wir eine Prozeßzeitachse voraus, die für jedes Individuum mit dem Zeitpunkt seiner Geburt beginnt. Zeitpunkte auf dieser Zeitachse können also als Lebensalter der Individuen interpretiert werden.

Als Ausgangspunkt für die statistische Beschreibung ergibt sich somit eine Folge von Zufallsvariablen

$$Y_t : \Omega \longrightarrow \mathcal{Y} \qquad t \in \mathcal{T} = \{0, 1, 2, 3, \ldots\} \tag{2.14}$$

Zu jedem Zeitpunkt $t = 0, 1, 2, 3, \ldots$ gibt es eine Zufallsvariable, die die Verteilung der Individuen aus Ω auf die durch den vorausgesetzten Zustandsraum möglichen Zustände erfaßt. Zum Zeitpunkt $t = 0$ findet die Geburt statt, alle Individuen befinden sich dann im gleichen Ausgangszustand. Dann beginnt für jedes Individuum sein jeweils individueller Le-

[61] Alle Begriffsbildungen können jedoch analog auch für eine stetige Zeitachse eingeführt werden, vgl. Hamerle [1989].

bensverlauf.

c) Ausgehend von den in (2.14) definierten Zufallsvariablen können jetzt Episoden konstruiert werden. Dabei ist zu berücksichtigen, daß die Anzahl der Episoden zwischen den Individuen variieren kann; es handelt sich also um eine Zufallsvariable:

$$L : \Omega \longrightarrow \{1, 2, 3, \ldots\}$$

$L(\omega)$ ist die Anzahl der Episoden für das Individuum $\omega \in \Omega$. Da der Zustandsraum endlich ist, gibt es eine maximal mögliche Anzahl von Episoden; sie wird mit $L_{\max}$ bezeichnet.

Es können jetzt Zufallsvariablen zur Beschreibung der Episoden definiert werden:

$$(T_l, D_l) : \Omega \longrightarrow \mathcal{T} \times \mathcal{Y} \qquad l = 0, 1, \ldots, L_{\max}$$

Die Definition erfolgt rekursiv. T_l ist der Zeitpunkt, zu dem die l.te Episode endet, und D_l ist ihr Endzustand. Dies liefert zugleich die Anfangsbedingungen für die $l+1$.te Episode, wenn es eine solche Episode gibt. Um eine vollständige Definition für $l = 1, \ldots, L_{\max}$ zu erreichen, wird als Konvention festgesetzt:

$$T_l(\omega) = T_{l-1}(\omega) \quad \text{und} \quad D_l(\omega) = D_{l-1}(\omega) \quad \text{wenn} \quad l > L_{\max}$$

Der Episodenverlauf bis zur l.ten Episode kann dann durch bedingte Wahrscheinlichkeiten beschrieben werden:

$$\mathrm{P}\left(D_1 = d_1, T_1 = t_1, \ldots, D_l = d_l, T_l = t_l \mid L \geq l\right) \tag{2.15}$$

Die Konditionierung ist erforderlich, da die Wahrscheinlichkeit für einen bestimmten Episodenverlauf bis zur l.ten Episode nur für diejenigen Individuen sinnvoll verstanden werden kann, die mindestens l Episoden erreichen. Darüberhinaus wird hier und im folgenden stets angenommen, daß eine Konditionierung auf den fest vorgegebenen Anfangszustand $(D_0 = y_a,\ T_0 = 0)$ erfolgt.

d) Die in (2.15) formulierte Wahrscheinlichkeitsverteilung liefert eine statistische Beschreibung des Prozesses bis zum Ende der l.ten Episode. Die eingangs genannte Idee besteht nun darin, diese Wahrscheinlichkeitsverteilung in ein Produkt bedingter Wahrscheinlichkeitsverteilungen zu zerlegen, die jeweils eine Episode beschreiben, jedoch unter Berücksichtigung der bis zum Episodenbeginn realisierten Vorgeschichte.

Um hierfür die Notation zu vereinfachen, ist es zweckmäßig, folgende Abkürzung für die Geschichte des Prozesses bis zum Ende der l.ten Episode zu verwenden:

$$H_l = \mathrm{P}\left(D_1, T_1, \ldots, D_l, T_l\right)$$

Die möglichen Realisierungen dieser Zufallsvariablen werden mit h_l bezeichnet. Folgende Gleichung liefert dann die angestrebte zeitlich rekursive Zerlegung in eine Folge bedingter Wahrscheinlichkeiten:

$$\mathrm{P}\left(H_l = h_l \mid L \geq l\right) = \prod_{k=1}^{l} \mathrm{P}\left(D_k = d_k, T_k = t_k \mid H_{k-1} = h_{k-1}, L \geq l\right) \tag{2.16}$$

Auf der linken Seite steht eine statistische Beschreibung für den Prozeßverlauf bis zum Ende der l.ten Episode. Auf der rechten Seite steht ein Produkt von Beschreibungen jeweils einer Episode, in der Form bedingter Wahrscheinlichkeiten, wobei die Bedingungen im Prozeßverlauf bis zum Beginn der jeweiligen Episode besteht.

e) Es ist wichtig, die Bedingung $L \geq l$ zu beachten. Sie steht offenbar im Widerspruch zu dem grundlegenden Prinzip, daß sich bei der Bildung bedingter Wahrscheinlichkeitsverteilungen zur Beschreibung von Prozessen die Bedingungen nur auf die jeweilige Vergangenheit beziehen sollten. Denn während die Individuen ihre Lebensverläufe realisieren, steht noch nicht fest, wieviele Episoden sie schließlich durchlaufen werden. Es erscheint deshalb sinnvoll, folgende zusätzliche Annahme einzuführen:

$$\mathrm{P}\left(H_l = h_l \mid L \geq l\right) = \mathrm{P}\left(H_l = h_l \mid L \geq k\right) \qquad \text{für alle } \ k \geq l \tag{2.17}$$

Sie besagt, daß der Verlauf der l.ten Episode unabhängig davon ist, wieviele Episoden ihr noch folgen werden. Sie entspricht insofern einer für die Lebensverlaufsanalyse typischen Betrachtungsweise: Der Lebensverlauf resultiert aus einem zeitlich sequentiellen Durchlaufen einzelner Episoden; was nach dem Ende einer Episode folgt, ist eine offene Zukunft, die auf den jeweils gegenwärtigen Episodenverlauf keinen Einfluß hat.[62]

Setzt man das in (2.17) angegebene Beschreibungsprinzip voraus, kann die Basisgleichung (2.16) folgendermaßen umgeschrieben werden:

$$\mathrm{P}\left(H_l = h_l \mid L \geq l\right) = \prod_{k=1}^{l} \mathrm{P}\left(D_k = d_k, T_k = t_k \mid H_{k-1} = h_{k-1}, L \geq k\right) \tag{2.18}$$

das heißt der Episodenverlauf bis zum Ende der l.ten Episode ergibt sich aus den bedingten Wahrscheinlichkeiten für alle bis dahin durchlaufenen Episoden, wobei sich jede einzelne dieser bedingten Wahrscheinlichkeiten auf jeweils alle Individuen bezieht, die den Beginn (und also auch das Ende) einer Episode noch erreichen.

[62] Diese Formulierung zeigt, daß es sich eigentlich nicht um eine Annahme handelt, die empirisch geprüft werden könnte, sondern um ein Prinzip, das die Art der Beschreibungen definiert, die wir erreichen möchten.

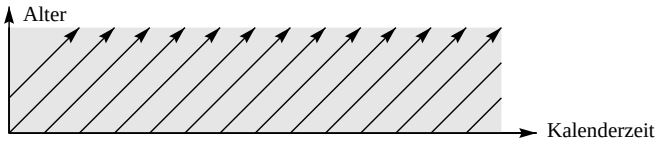
Die in (2.18) angegebene Formulierung zeigt, daß eine statistische Beschreibung einer Folge von Episoden dadurch erreicht werden kann, daß für jede Episode eine separate Beschreibung gegeben wird, wobei jeweils die bis zum Beginn der Episode durchlaufene Vorgeschichte als Bedingung des Episodenverlaufs dient. Jede einzelne Episode kann auf der Grundlage einer Prozeßzeitachse beschrieben werden, d.h. durch Verwendung von $T_l - T_{l-1}$ als Verweildauer für die l.te Episode.[63]

[63] Vgl. Hamerle [1989].

2.5 Soziale Prozesse und sozialer Wandel

Bisher sind wir von einem abstrakten Prozeßbegriff ausgegangen. Ein Prozeß wurde definiert als eine Gesamtheit von Lebensverläufen, deren Subjekte einer Längsschnittgesamtheit angehören (vgl. Abschnitt 2.4.1). Für die statistische Beschreibung wurde dementsprechend eine Prozeßzeitachse zugrunde gelegt, ohne einen expliziten Bezug zur Kalenderzeit herzustellen. Da die statistische Beschreibung schließlich der soziologischen Theoriebildung dienen soll, ist dies unzureichend. Es muß überlegt werden, wie gesellschaftliche Verhältnisse als Bedingungen individueller Lebensverläufe sichtbar gemacht werden können. Das Problem hat zwei Aspekte. Erstens muß überlegt werden, wie überhaupt von *Bedingungen* kontingenter Lebensverläufe gesprochen werden kann. Dieses Problem wird in Kapitel 4 näher diskutiert. Zweitens gibt es ein in gewisser Weise vorgelagertes Problem: Es muß überlegt werden, wie dargestellt werden kann, daß Lebensverläufe in einer sozial definierten Kalenderzeit ablaufen, also im Rahmen eines sozialen Prozesses, auf den bezogen dann auch von einem sozialen Wandel gesprochen werden kann.

Einen zweckmäßigen Ausgangspunkt bildet die Unterscheidung von Geburtskohorten. Eine Gesamtheit von Lebensverläufen kann dann in einem *Lexis-Diagramm* dargestellt werden, das Kalenderzeit und Lebensalter als zwei Dimensionen verknüpft,[64] etwa folgendermaßen:[65]



Jeder der diagonalen Pfeile repräsentiert eine Geburtskohorte. Die kohortenspezifischen Lebensverläufe entwickeln sich gewissermaßen simultan in zwei Zeitdimensionen: einerseits parallel zur Kalenderzeit, andererseits parallel zu einer kohortenspezifischen Lebenszeit.

Einige Autoren haben versucht, dieses Lexis-Diagramm als einen formalen Rahmen für eine makro-soziologische Theoriebildung zu verwenden.[66] Darauf soll hier nicht näher eingegangen werden. Das Lexis-Diagramm soll hier nur dazu dienen, ein anschauliches Bild der Problemstellung zu vermitteln: wie individuelle Lebensverläufe mit der Vorstellung eines sozialen Prozesses, d.h. sich in der Kalenderzeit entwickelnder gesellschaftlicher Verhältnisse, verknüpft werden können.

[64] Die Bezeichnung *Lexis-Diagramm* übernehme ich von Mayer und Huinink [1990a].

[65] Eine ähnliche Darstellung wurde bereits in Abschnitt 2.2 verwendet, vgl. Abb. 2.2.4 auf S. 55.

[66] Vgl. zum Beispiel das *age-stratification model* von Riley [1986].

Um diese Problemstellung diskutieren zu können, muß zunächst präzisiert werden, in welcher Weise von sozialen Prozessen gesprochen werden soll. Da dieser Begriff mit einer Vielzahl theoretischer Konnotationen belastet ist, erscheint es sinnvoll, mit einer sehr einfachen, statistischen Definition zu beginnen, nämlich den Begriff *sozialer Prozeß* zur Bezeichnung stochastischer Prozesse der folgenden Art zu verwenden:

$$Y_t^* : \Omega \longrightarrow \mathcal{Y}^* \qquad t \in \mathcal{T}^* \tag{2.19}$$

$\mathcal{T}^*$ ist eine Kalenderzeitachse, Ω eine Gesamtheit von Individuen, deren Lebensverläufe sich ganz oder partiell mit dieser Kalenderzeitachse überschneiden.[67] $\mathcal{Y}^*$ ist der Zustandsraum. Um berücksichtigen zu können, daß Individuen während der Kalenderzeit $\mathcal{T}^*$ neu geboren werden und sterben können, enthält dieser Zustandsraum insbesondere die Quasi-Zustände *noch nicht geboren* und *bereits gestorben*. Schließlich bildet die Folge der Zufallsvariablen Y_t^* die statistische Repräsentation des sozialen Prozesses. Zu jedem Kalenderzeitpunkt t beschreibt die Zufallsvariable Y_t^* die Verteilung der Individuen aus Ω auf die in $\mathcal{Y}^*$ definierten möglichen Zustände.

Zweck dieser Begriffsbildung ist, einen formalen Rahmen zu gewinnen, in dem zumindest einige Aspekte der gesellschaftlichen Entwicklung – sozialen Wandels – *beschrieben* werden können.[68] Wie in der Einleitung ausgeführt worden ist, können zwei elementare Formen der Beschreibung gesellschaftlicher Verhältnisse und mithin sozialen Wandels unterschieden

[67] Um Spekulationen über Anfang und Ende sozialer Prozesse zu vermeiden, nehmen wir an, daß die Kalenderzeitachse aus einem beschränkten Zeitintervall besteht, das für deskriptive Zwecke willkürlich festgelegt wird. Außerdem wird angenommen, daß es sich um eine diskrete Zeitachse handelt, um auf einfache Weise von *Folgen* von Zufallsvariablen sprechen zu können.

[68] Ich betone dies, da in der Literatur gelegentlich Formulierungen verwendet werden, die den Gedanken nahelegen, daß soziale Prozesse Sachverhalte sein könnten, die individuelle Verhaltensweisen, Lebensverläufe oder irgendetwas anderes „hervorbringen“. Zum Beispiel sagen Blossfeld et al. [1991, S. 213], „that the observed occupational attainment of an individual at a specific time is the result of a process of change.“ Ich verstehe dies so, daß damit gesagt werden soll, daß der Zustand, den ein Individuum zu einem gewissen Zeitpunkt einnimmt, als transitorischer Endpunkt einer vorausgegangenen Lebensgeschichte beschrieben und erklärt werden sollte. Schwer verständlich und problematisch erscheint demgegenüber folgende Formulierung von Bergman et al. [1991, S. 1]: „The ultimate goal for developmental research is to understand and explain the developmental process underlying an individual's way of thinking, feeling, acting and reacting at a certain stage of the life process.“ Oder auch folgende Formulierung von Willekens [1988, S. 91]: „Live events and the combination of life events in careers are viewed as the outcome of substantive and random processes. [...] The explanation of life events in terms of the substantive processes causing the events to occur remains a major challenge in life course analysis.“ Es scheint gemeint zu sein, daß durch die Bezugnahme auf einen „underlying process“ die Phänomene erklärt werden können, durch die wir den Prozeß empirisch beschreiben. Ob dieser Vorstellung ein nachvollziehbarer Sinn gegeben werden kann, sei dahingestellt. In dieser Arbeit gehe ich jedenfalls davon aus, daß der Prozeßbegriff eine deskriptive Kategorie ist; er dient dazu, beobachtbare Phänomene begrifflich zusammenzufassen, um sie in ihrer zeitlichen Abfolge darstellen zu können.

werden. Dem entsprechen zwei unterschiedliche Formen der Bezugnahme auf soziale Prozesse.

a) Erstens kann sozialer Wandel auf der Ebene der in (2.19) definierten Zufallsvariablen Y_t^* definiert werden. Sozialer Wandel besteht dann einfach darin, daß sich im Zeitablauf die Verteilung dieser Zufallsvariablen verändert. Zumindest einige Aspekte sozialen Wandels können in diesem Rahmen sinnvoll beschrieben werden, zum Beispiel: die Entwicklung der durchschnittlichen Einkommen und ihrer Verteilung, die Anzahl der Personen, die einer Teilzeitbeschäftigung nachgehen, das Ausmaß nichtehelicher Lebensgemeinschaften, die Entwicklung des durchschnittlichen Heiratsalters, usw. In allen Fällen handelt es sich um Sachverhalte, für die eine kalenderzeitpunktbezogene Repräsentation gefunden werden kann. Ist dies der Fall, können die in der Kalenderzeit stattfindenden Veränderungen beschrieben werden.

b) Eine zweite elementare Form der Beschreibung gesellschaftlicher Verhältnisse geht von der Vorstellung aus, daß die Individuen in ihren Lebensverläufen zumindest partiell gewissen Regeln folgen, die empirisch als Regelmäßigkeiten ermittelt werden können. Sozialer Wandel besteht dann darin, daß sich diese Regeln bzw. Regelmäßigkeiten im Zeitablauf verändern. Bei dieser Betrachtungsweise ist es offensichtlich wesentlich schwieriger, eine angemessene statistische Form der Beschreibung zu finden. Die Schwierigkeit liegt darin, daß es nicht möglich ist, zeitpunktbezogen von sozialen Regeln zu reden. Zu sagen, daß es eine soziale Regel gibt, impliziert, daß sie *für einen gewissen Zeitraum* gilt. Aber Anfang und Ende dieses Zeitraums können in den meisten Fällen nicht genau bestimmt werden. Dies macht es zugleich schwierig, davon zu sprechen, daß sich Regeln im Zeitablauf verändern. Soweit ich sehen kann, gibt es keine geeignete mathematische Form, um diese Vorstellung auszudrücken. Man ist gezwungen, auf mehr oder weniger willkürlich abgegrenzte Zeitperioden zurückzugreifen.

Geht man von der unter (a) genannten Konzeption aus, scheint es auf einfache Weise möglich zu sein, sozialen Wandel zu beschreiben, indem man von folgendem Schema ausgeht:

$$\cdots\ Y_{t-1}^* \longrightarrow Y_t^* \longrightarrow Y_{t+1}^* \cdots \tag{2.20}$$

Da es sich um einen in der Zeit ablaufenden Prozeß handelt, liegt es nahe, die statistische Beschreibung und Modellbildung auf die Übergangswahrscheinlichkeiten

$$\mathrm{P}\left(Y_t^* = y_t^* \mid Y_{t-1}^* = y_{t-1}^*, Y_{t-2}^* = y_{t-2}^*, \ldots\right)$$

zu konzentrieren. Tatsächlich verfolgen viele Modellansätze in der Literatur diesen Ansatz. Er ist insbesondere typisch für die sog. Panel-Modelle, bei denen die Modellbildung unmittelbar von einer diskreten Kalender-

zeitachse ausgeht.[69]

Dieser Ansatz der statistischen Modellbildung ist jedoch im Hinblick auf die Zielsetzung, soziologische Einsichten in sozialen Wandel zu gewinnen, unbefriedigend. Denn hinter den Veränderungen in den Verteilungen der Zustandsvariablen Y_t^* verbergen sich zahlreiche unterschiedliche Vorgänge. Man sieht dies bereits anhand des oben skizzierten Lexis-Diagramms. Ein Modellansatz, der unmittelbar dem Schema (2.20) folgt, vergleicht gewissermassen eine Serie von Querschnittsverteilungen auf der Kalenderzeitachse. Die auf dieser Ebene resultierenden Veränderungen sind insbesondere das Ergebnis demographischer Prozesse und dadurch bedingter Veränderungen in der Altersverteilung der jeweils lebenden Querschnittsbevölkerung. Diese demographischen Prozesse vermischen sich dann mit Veränderungen in den individuellen Lebensverläufen, die ihrerseits wiederum auch vom Lebensalter ihrer Subjekte abhängen. Für die soziologische Theoriebildung sind diese Unterscheidungen offensichtlich wichtig; und damit die statistische Modellbildung der soziologischen Theoriebildung dienen kann, sollten sie darin explizit berücksichtigt werden.[70]

Wie kann dies erreicht werden? Eine Möglichkeit besteht darin, von der Frage auszugehen, wer die *Träger* des jeweils beobachtbaren sozialen Wandels sind.[71] Es gibt zwei unterschiedliche Perspektiven, um sich einer Antwort zu nähern.

a) Einerseits kann man von den Veränderungen in den Verteilungen der Zustandsvariablen Y_t^* ausgehen und nach einer informativen Dekomposition suchen. Als ein Beispiel betrachten wir folgenden einfachen Zustandsraum:

$$\mathcal{Y} = \begin{cases} 1 & \text{noch nicht geboren} \\ 2 & \text{unverheiratet} \\ 3 & \text{verheiratet} \\ 4 & \text{gestorben} \end{cases}$$

Die Veränderungen zwischen zwei Kalenderzeitpunkten t und t' können

[69] Eine Einführung in diesen Typ statistischer Modelle, die insbesondere in der Ökonometrie eine breite Verwendung gefunden haben, findet sich bei Maddala [1987].

[70] Eine der wesentlichen in der Lebensverlaufsforschung erzielten Einsichten liegt darin, daß ein unmittelbarer Vergleich von Querschnitten (Perioden), wie in (2.20) angedeutet, ohne die zugrundeliegenden Veränderungen (oder Stabilitäten) in den individuellen Lebensverläufen zu berücksichtigen, zu weitgehend irreführenden Vorstellungen über sozialen Wandel führen kann. Am Beispiel der Entwicklung periodenspezifischer Scheidungsraten in den USA wurde dies exemplarisch von Ferriss [1970] gezeigt.

[71] Hinter dieser Fragestellung steht natürlich die Vorstellung, daß sozialer Wandel durch Veränderungen in individuellen Lebensverläufen „hervorgebracht“ wird. Diese Formulierung soll jedoch zunächst rein deskriptiv verstanden werden und keinerlei Vermutung darüber implizieren, ob bzw. in welcher Weise sozialer Wandel durch Veränderungen in individuellen Lebensverläufen *erklärt* werden kann.

dann in Form einer Kontingenztabelle dargestellt werden:

$Y_t^* \backslash Y_{t'}^*$	1	2	3	4
1	h_{11}	h_{12}	–	–
2	–	h_{22}	h_{22}	h_{24}
3	–	h_{32}	h_{32}	h_{34}
4	–	–	–	h_{44}

h_{jk} ist die Anzahl der Personen, die sich zum Zeitpunkt t im Zustand j und zum Zeitpunkt t' im Zustand k befinden. In der Hauptdiagonale befinden sich diejenigen Personen, die ihren Zustand nicht wechseln, insbesondere die (bis t') noch nicht geborenen und die (bis t) bereits gestorbenen Personen. Außerhalb der Hauptdiagonalen befinden sich die Personen, die ihren Zustand verändern und dadurch den sozialen Wandel zwischen den Kalenderzeitpunkten t und t' hervorbringen. Um herauszufinden, wer die Träger des sozialen Wandels sind, ist es schließlich nur erforderlich, die Merkmale der Personen zu beschreiben, die die Aggregate h_{jk} bilden. Eine elementare Form dieser Dekomposition liefert bereits das Lexis-Diagramm, bei dem eine Zerlegung nach dem Alter vorgenommen wird.[72]

Diese Methode der Dekomposition (auch in den etwas raffinierteren Formen der sog. APK-Analyse) hat jedoch eine Reihe von Nachteilen.[73] Das Hauptproblem kann darin gesehen werden, daß die Bezugnahme auf individuelle Lebensverläufe implizit und unanschaulich bleibt. Man erfährt zum Beispiel etwas über die Altersverteilung derjenigen Personen, die im Zeitraum $[t, t']$ geheiratet bzw. nicht geheiratet haben; aber es ist praktisch unmöglich, daraus eine Einsicht in die Lebensverläufe der beteiligten Personen zu gewinnen.

b) Eine alternative Perspektive ergibt sich, wenn man explizit von den individuellen Lebensverläufen ausgeht und dann versucht, sie so zu beschreiben, daß sich daraus Einsichten in sozialen Wandel gewinnen lassen. Diese Betrachtungsweise kann unmittelbar mit der Überlegung verknüpft werden, daß sozialer Wandel durch die Entwicklung individueller Lebensverläufe hervorgebracht wird, sich als in der Regel nicht-intendierte Folge kontingenter Lebensverläufe herausbildet. Um theoretische Deutungen sozialen Wandels mit einer empirischen Grundlage zu versehen, ist sie deshalb geeigneter als schematische Dekompositionsverfahren.

Die übliche Methode, um diese Betrachtungsweise empirisch umzusetzen, besteht darin, die statistische Beschreibung individueller Lebensverläufe durch eine Differenzierung nach Kohorten zu ergänzen. Es wird,

[72] Vgl. als ein Beispiel die Darstellung altersspezifischer Geburtsraten bei Mayer und Huinink [1990a].

[73] Eine ausführliche Diskussion dieser Probleme findet sich bei Mayer und Huinink [1990a, 1990b].

wie in den vorangegangenen Abschnitten, auf eine Längsschnittgesamtheit Bezug genommen. Diese Längsschnittgesamtheit wird dann in Kohorten eingeteilt, wobei als kohortendefinierendes Ereignis der Beginn der Lebensverläufe verwendet wird.[74] Auf diese Weise erhält man kohortenspezifische Beschreibungen von Lebensverläufen, die, wie in einem verallgemeinerten Lexis-Diagramm, als eine in der Kalenderzeit geordnete Folge von Prozessen interpretiert werden können. Sozialer Wandel wird dann als Resultat einer Überlagerung kohortenspezifischer Prozesse von Lebensverläufen sichtbar.

[74] Genauer gesagt, der Beginn derjenigen Episode oder Folge von Episoden, auf die sich die intendierte Beschreibung richtet.

2.6 Exkurs: Bezüge zur Ungleichheitsforschung

Soziologische Lebensverlaufsforschung steht in einem engen Zusammenhang zu soziologischer Ungleichheitsforschung.[75] Die wechselseitigen Bezüge sind sehr vielseitig,[76] und in der vorliegenden Arbeit kann darauf nicht im einzelnen eingegangen werden. Es erscheint jedoch sinnvoll, zumindest den Perspektivenwechsel anzudeuten, den die Lebensverlaufsforschung zur Reflexion sozialer Ungleichheit vorschlägt.

Aus dem grundsätzlichen Ansatz der Lebensverlaufsforschung ergeben sich zwei gleichermaßen wichtige Ausgangspunkte. (1) Soziale Ungleichheit wird als Ungleichheit zwischen Individuen betrachtet. Dies schließt zwar nicht aus, Gesamtheiten (Gruppen) von Individuen zu bilden und zu vergleichen, aber dies ist bereits ein Resultat soziologischer Theoriebildung. Gruppen, Schichten, Klassen, Lebenslagen, Lebensstile, usw. können nicht unmittelbar beobachtet werden, es sind theoretische Begriffe, deren empirischer Sinn nur durch Bezugnahme auf individuelle Lebensverläufe deutlich gemacht werden kann. (2) Individuen werden durch Lebensverläufe definiert. Soziale Ungleichheit wird infolgedessen als Ungleichheit von Lebensverläufen betrachtet.

Ein wesentlicher Beitrag der Lebensverlaufsforschung zur Diskussion sozialer Ungleichheit kann bereits darin gesehen werden, daß sich aus ihrer Sicht eine weitreichende Kritik an traditionellen *statischen* Konzeptionen sozialer Ungleichheit formulieren läßt. Aus ihrer Sicht sind alle Merkmale, die als Indikatoren sozialer Ungleichheit herangezogen werden können, als grundsätzlich transitorische Zustände im Ablauf individueller Lebensverläufe aufzufassen. Dies impliziert selbstverständlich nicht, daß es sich um „kurzfristige“ Zustände handelt. Gerade aus der Kritik an statischen Konzeptionen sozialer Ungleichheit folgt, daß die Dauerhaftigkeit von Ungleichheitszuständen als eine empirische Frage betrachtet werden muß.

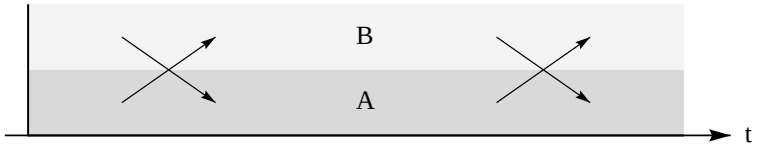
Ein zweiter wesentlicher Beitrag der Lebensverlaufsforschung kann darin gesehen werden, daß sie einen *empirischen* Zugang zur Erforschung sozialer Ungleichheit – verstanden als Ungleichheit individueller Lebensverläufe – erschließt. Unter Rückgriff auf die in den vorangegangenen Abschnitten dieses Kapitels angestellten Überlegungen zur Beschreibung von Lebensverläufen will ich versuchen, einige Aspekte dieses empirischen Zugangs zur Beschreibung sozialer Ungleichheit anzudeuten.

a) Eine naheliegende Möglichkeit, soziale Ungleichheit im Zeitablauf zu untersuchen, scheint darin zu liegen, eine (in der Kalenderzeit definierte) Folge von Zustandsvariablen zu betrachten. Diese Vorgehensweise ist weit verbreitet, um Ungleichheiten im Zeitablauf sichtbar zu machen; zum Beispiel: die Entwicklung der Arbeitslosenquote, die Entwicklung des

[75] Einen instruktiven Einblick in den gegenwärtigen Diskussionsstand in der BRD vermittelt ein von Berger und Hradil [1990] herausgegebener Sammelband.

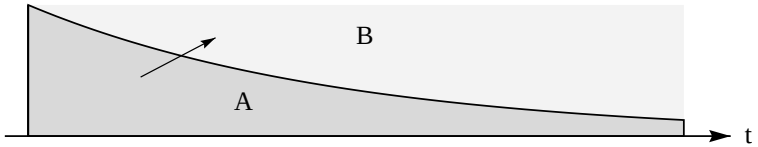
[76] Vgl. exemplarisch den Diskussionsbeitrag von Mayer und Blossfeld [1990].

Anteils von Sozialhilfeempfängern, usw. Aus der Perspektive der Lebensverlaufsforschung ist diese Vorgehensweise jedoch irreführend, da mit ihr nicht sichtbar gemacht werden kann, auf welche Weise und in welchem Ausmaß die Individuen *in ihren Lebensverläufen* von dem jeweiligen Sachverhalt sozialer Ungleichheit betroffen sind. Folgende Abbildung illustriert das Problem.



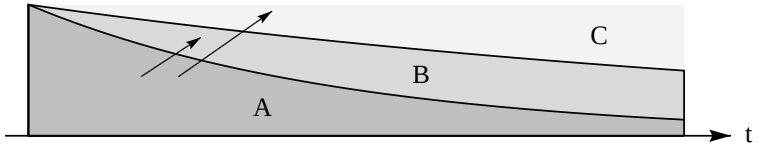
In diesem Beispiel gibt es zwei Gruppen, A und B. Das Bild zeigt, wie sich eine Menge von Individuen auf die beiden Gruppen verteilt und liefert insofern ein Bild sozialer Ungleichheit. Aber das Bild zeigt nicht, ob und, wenn ja, in welchem Ausmaß die Individuen im Zeitablauf ihre Zugehörigkeit zu den beiden Gruppen wechseln. Obwohl die Zeitdimension berücksichtigt wird, erhält man keinen Einblick in die Dynamik sozialer Ungleichheit.

b) Mit den im Rahmen der Lebensverlaufsforschung entwickelten Modellen wird demgegenüber versucht, einen expliziten Bezug zu individuellen Lebensverläufen herzustellen. Das Grundprinzip liegt darin, von einer gegebenen Menge von Anfangszuständen auszugehen und dann zu untersuchen, wie sich die Lebensverläufe im weiteren entwickeln. Im einfachsten Fall gibt es nur einen Ausgangszustand A und einen Folgezustand B. Man erhält dann zum Beispiel folgendes Bild.



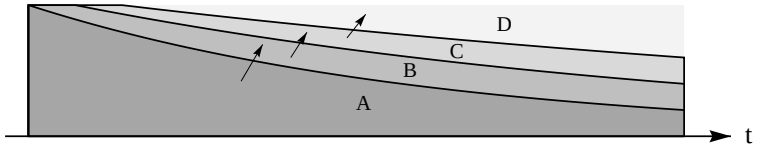
In statistischer Sprechweise handelt es sich um eine einfache Survivorfunktion. Alle Individuen beginnen im Zustand A, im Zeitablauf wechseln dann einige, in diesem Beispiel immer mehr Individuen in den Zustand B. Offensichtlich vermittelt dieses Bild einen Einblick in die Dynamik sozialer Ungleichheit auf der Ebene individueller Lebensverläufe. Zunächst gibt es keinerlei Ungleichheit, alle Personen befinden sich im gleichen Zustand. Dann entwickelt sich Ungleichheit, zunächst wird sie größer, dann wieder kleiner.

c) Eine einfache Verallgemeinerung führt zu einer Situation konkurrierender Risiken. Die folgende Abbildung illustriert zwei konkurrierende Risiken.



Alle Individuen beginnen wieder im gleichen Ausgangszustand A, dann kann ein Übegang in einen von zwei Folgezuständen, B oder C, stattfinden. Wiederum erhält man einen anschaulichen Einblick in den zeitlichen Prozeß, in dem sich Ungleichheit herausbildet.

d) Komplizierter wird es, wenn man Folgen von Episoden, insbesondere wiederholbare Zustände, untersuchen will. Die nächste Abbildung illustriert eine einfache Situation mit drei Episoden, ohne konkurrierende Risiken. Alle Individuen beginnen im Zustand A, dann können sie in den Zustand B, dann in den Zustand C, und schließlich in den Zustand D wechseln. Zum Beispiel können die Zustände A und C Episoden der Arbeitslosigkeit, die Zustände B und D Episoden der Beschäftigung repräsentieren.



Diese Beispiele zeigen, in welcher Weise eine statistische Beschreibung individueller Lebensverläufe Einsichten in die Entwicklung sozialer Ungleichheit vermitteln kann. Das Hauptproblem liegt darin, daß mit zunehmender Anzahl möglicher Zustände die resultierenden Beschreibungen sehr komplex werden.

Um dies zu illustrieren, betrachte ich einige Informationen, die im SOEP über arbeitslos gewordene Personen enthalten sind. Es wird von folgender Längsschnittstichprobe ausgegangen: alle Stammpersonen, die zur Teilstichprobe A gehören, an mindestens den ersten drei Wellen teilgenommen und eine Angabe über ihren Berufsbeginn gemacht haben; insgesamt $N = 6015$ Personen. Aus dieser Längsschnittstichprobe werden dann diejenigen 178 Personen betrachtet, die *während* des Jahres 1983 arbeitslos geworden sind. Die Frage ist, wie sich ihre weiteren Lebensverläufe entwickelt haben, und zwar im Hinblick auf die möglichen Folgezustände: *vollzeit beschäftigt*, *teilzeit beschäftigt* und *sonstiges*. Insoweit handelt es sich um eine einfache Situation mit drei konkurrierenden Risiken. Darüber hinaus soll jedoch berücksichtigt werden, daß Arbeitslosigkeit ein wiederholbarer Zustand ist. Eine Person kann den Zustand der Arbeitslosigkeit verlassen und zum Beispiel eine neue Beschäftigung finden, dann jedoch erneut arbeitslos werden, und so weiter.

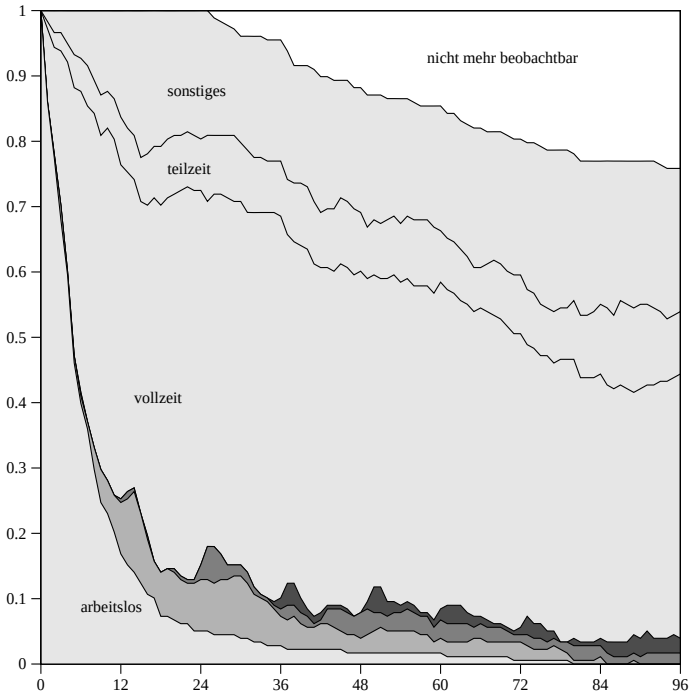


Abbildung 2.6.1 Überlagerungen von Arbeitslosigkeitsepisoden in den Erwerbsbiographien von 178 Personen aus der Teilstichprobe A des SOEP, die während des Jahres 1983 arbeitslos geworden sind. Darstellung auf einer Prozeßzeitachse (in Monaten), die mit dem Eintritt in die Arbeitslosigkeit während des Jahres 1983 beginnt.

Abbildung 2.6.1 zeigt, wie sich die Erwerbsbiographien dieser 178 Personen in den folgenden 8 Jahren entwickelt haben. Bei dieser Darstellung werden bis zu drei weitere Arbeitslosigkeitsepisoden erfaßt. Die Darstellung erfolgt durch sich überlagernde Survivorfunktionen. Die unteren vier Bereiche, durch unterschiedliche Grautöne gekennzeichnet, repräsentieren die vier möglichen Arbeitslosigkeitsepisoden. Alle Personen beginnen im gleichen Ausgangszustand, der ersten Arbeitslosigkeitsepisode. Die Ausgangsepisode wird in einen der möglichen Folgezustände verlassen: Vollzeit- oder Teilzeitbeschäftigung oder Sonstiges. Ist dieser Übergang vollzogen, kann es zu einer zweiten Arbeitslosigkeitsepisode kommen. Abbildung 2.6.1 zeigt zum Beispiel, daß nach zwei Jahren für etwa 95 % der Personen die erste Arbeitslosigkeitsepisode beendet ist, daß sich jedoch zu diesem Zeitpunkt fast 10 % der Personen bereits in einer zweiten oder dritten Arbeitslosigkeitsepisode befinden.

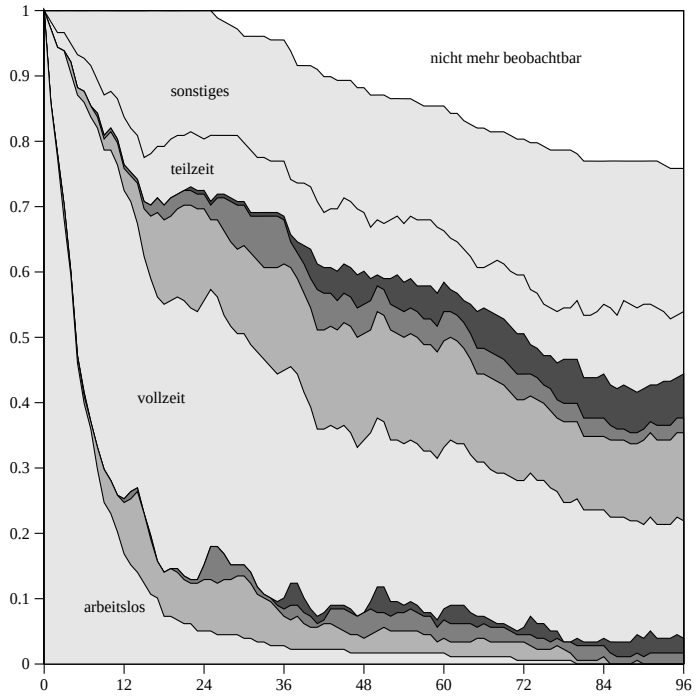


Abbildung 2.6.2 Überlagerungen von Arbeitslosigkeits- und Vollzeiterwerbstätigkeitsepisoden in den Erwerbsbiographien von 178 Personen aus der Teilstichprobe A des SOEP, die während des Jahres 1983 arbeitslos geworden sind. Darstellung auf einer Prozeßzeitachse (in Monaten), die mit dem Eintritt in die Arbeitslosigkeit während des Jahres 1983 beginnt.

Entsprechende Differenzierungen können bei den anderen Zuständen vorgenommen werden. Abbildung 2.6.2 liefert ein Bild, bei dem zusätzlich bis zu drei Wiederholungen von Vollzeitbeschäftigungen unterschieden werden.

Die Abbildungen illustrieren, wie mit Hilfe von Lebensverlaufsdaten Einsichten in die Entwicklung sozialer Ungleichheit gewonnen werden können. Sie zeigen vor allem, daß das resultierende Bild sozialer Ungleichheit sehr komplex wird, sobald man mehrere, insbesondere wiederholbare Zustände berücksichtigt. Darin kommt ein generelles Problem zum Ausdruck. Die statistischen Methoden der Ereignisanalyse sind sehr gut geeignet, um Zustandswechsel zu beschreiben und zu analysieren, gewissermaßen die Knotenpunkte, an denen sich die individuellen Lebensverläufe sukzessive verzweigen. Die Anwendung dieser Methoden stößt jedoch an Grenzen, wenn beabsichtigt wird, die schließlich resultierende Vielfalt in den realisierten Lebensverläufen angemessen darzustellen.

Ungleichheitsmaße und das Komplexitätsproblem

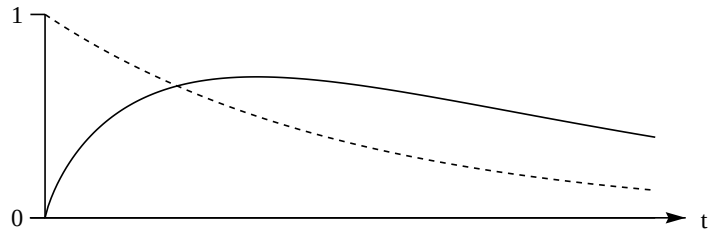
Es liegt nahe, statistische Ungleichheitsmaße zu verwenden, um die Entwicklung sozialer Ungleichheit sichtbar zu machen. Bei quantifizierbaren Zustandsvariablen, wie zum Beispiel beim Einkommen, können unmittelbar einfache Streuungsmaße verwendet werden. Zum Beispiel haben Mayer und Blossfeld [1990] Standardabweichungen verwendet, um nicht nur die durchschnittliche Entwicklung, sondern auch die Streuung von Prestige-Werten in der Entwicklung von Berufsverläufen sichtbar zu machen. In der Regel beziehen wir uns jedoch auf einen diskreten Zustandsraum, so daß Ungleichheitsmaße für die Verteilung von Personen auf eine Menge diskreter Zustände verwendet werden müssen.

In der statistischen Methodenlehre sind zahlreiche Maße entwickelt worden, mit deren Hilfe die Streuung einer diskreten Verteilung ausgedrückt werden kann. Ein einfach zu interpretierendes Ungleichheitsmaß ist die sog. *Entropie*.[77] Wenn es die Zustände $1, \ldots, m$ gibt, und wenn h_j den Anteil der Personen im Zustand j bezeichnet, ist die Entropie dieser Verteilung folgendermaßen definiert:

$$E_m = \sum_{j=1}^{m} h_j \log(h_j) \tag{2.21}$$

Es gilt $0 \leq E_m \leq \log(m)$. Die Entropie nimmt ihren minimalen Wert an, wenn sich alle Personen im gleichen Zustand aufhalten; sie nimmt ihren maximalen Wert an, wenn sich in allen Zuständen gleich viele Individuen aufhalten. Die Entropie kann daher unmittelbar als ein Ungleichheitsmaß interpretiert werden. Daß der maximale Wert der Entropie von der Anzahl der Zustände abhängt, erscheint in diesem Zusammenhang durchaus sinnvoll, denn dadurch kann berücksichtigt werden, daß soziale Ungleichheit immer diffuser wird, je mehr unterschiedliche Aspekte einbezogen werden.

Folgende Abbildung illustriert das Entropiemaß anhand einer einfachen Situation, in der es nur zwei Zustände gibt, einen Ausgangszustand A und einen möglichen Folgezustand B.



[77] Einführungen in den statistischen Entropiebegriff und seine vielfältigen Anwendungen geben u.a. Theil [1972] und Kapur [1989].

Die gestrichelte Linie zeigt den Verlauf einer Survivorfunktion für eine Standard-Exponentialverteilung (konstante Rate $r = 1$). Die durchgezogene Linie zeigt den Verlauf der korrespondierenden Entropie. Am Anfang befinden sich alle Personen im gleichen Ausgangszustand (A), so daß die Entropie den Wert 0 annimmt. Dann wechseln zunehmend Personen in den Folgezustand (B), und die Entropie wird dementsprechend größer. Ihr Maximum nimmt sie beim Median der Verteildauer an, wenn sich jeweils 50 % der Personen in den Zuständen A und B befinden. Schließlich wird die Entropie wieder geringer, wenn sich mehr und mehr Personen im Folgezustand B konzentrieren.

In dieser einfachen Situation liefert die Entropie natürlich keine Einsichten, die nicht bereits unmittelbar aus einer Betrachtung der Survivorfunktion gewonnen werden könnten. Sie verschafft jedoch zusätzliche Einsichten, wenn zunehmend komplexere Zustandsräume betrachtet werden. Um dies zu illustrieren, beziehe ich mich noch einmal auf die in den Abbildungen 2.6.1 und 2.6.2 dargestellten Lebensverläufe von 1983 arbeitslos gewordenen Personen.

Abbildung 2.6.3 zeigt drei verschiedene Varianten der Berechnung von Entropiemaßen zur Erfassung der Ungleichheit, die sich mit dem Eintritt in die Arbeitslosigkeit herausbildet. Bei der Variante A werden nur vier Zustände unterschieden: arbeitslos, vollzeit erwerbstätig, teilzeit erwerbstätig und sonstiges. Das Entropiemaß erfaßt die sich im Zeitablauf wandelnde Verteilung der Personen auf diese vier Zustände. Bei der Variante B werden drei weitere Zustände berücksichtigt, nämlich die Möglichkeit von drei weiteren Arbeitslosigkeitsepisoden; dies entspricht Abbildung 2.6.1. Bei der Variante C werden darüberhinaus bis zu vier Vollzeit-Erwerbstätigkeitsepisoden unterschieden; dies entspricht Abbildung 2.6.2.[78]

[78] Bei der Berechnung der Entropiemaße für Abbildung 2.6.3 wurde von den nicht mehr beobachtbaren Individuen abgesehen.

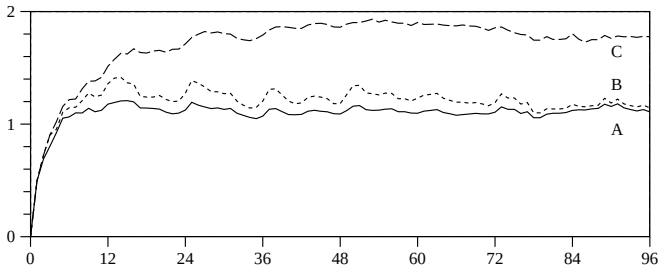


Abbildung 2.6.3 Entropiemaße für die Entwicklung der Zustandsverteilungen von 178 Personen aus der Teilstichprobe A des SOEP, die während des Jahres 1983 arbeitslos geworden sind. (A) Nur drei mögliche Folgezustände: vollzeit und teilzeiterwerbstätig sowie sonstiges; (B) zusätzliche Unterscheidung von bis zu drei weiteren Arbeitslosigkeitsepisoden; (C) zusätzliche Unterscheidung von bis zu drei weiteren Vollzeit-Erwerbstätigkeitsepisoden. Darstellung auf einer Prozeßzeitachse (in Monaten), die mit dem Eintritt in die Arbeitslosigkeit während des Jahres 1983 beginnt.

Die Abbildung zeigt, daß die durch die Entropie erfaßte Art der Ungleichheit zunächst schnell größer wird; dies entspricht dem relativ raschen Verlassen der ersten Arbeitslosigkeitsepisode. Danach wird ein relativ konstanter Verlauf erreicht. Dies ist jedoch in Wirklichkeit nur ein Reflex der Tatsache, daß bei der Berechnung nur eine kleine Anzahl von sukzessiven Zustandswechseln berücksichtigt worden ist. Das wird bei der Variante A besonders deutlich. Sie erfaßt nur die jeweils ersten Übergänge, die zum Verlassen des Ausgangszustands führen. Es ist intuitiv einleuchtend, daß die Bedeutung der durch sie erzeugten Ungleichheit in der weiteren Entwicklung gewissermaßen verschwindet. Sobald jedoch neue, sich anschließende Zustandswechsel berücksichtigt werden, wird die resultierende Verteilung zunehmend diffus und die Entropie nimmt immer größere Werte an.

Das Beispiel illustriert, in welcher Weise das Ausmaß der erfaßbaren sozialen Ungleichheit davon abhängt, wie das Biographieschema beschaffen ist, in dessen Rahmen Lebensverläufe betrachtet ist. Je differenzierter die Beschreibung wird, desto größer wird die durch sie wahrnehmbare Ungleichheit. Diese Sichtweise macht deutlich, daß soziale Ungleichheit kein absolut definierbarer Sachverhalt ist; und sie verweist auf das Erfordernis, eine – schließlich nur sozialpolitisch (normativ) begründbare – Unterscheidung zwischen der individuellen Vielfalt von Lebensverläufen und sozialer Ungleichheit vorzunehmen.[79]

[79] Vgl. Nussbaum und Sen [1993] als eine Einführung in die gegenwärgige Diskussion.

Kapitel 3

Deskriptive Modelle für Lebensverläufe

Unsere bisherigen Überlegungen zielten auf die Frage, wie Gesamtheiten vergleichbarer Lebensverläufe statistisch *beschrieben* werden können. Diese Fragestellung ist grundlegend, um den Gegenstand unseres theoretischen Interesses begrifflich fassen zu können. Aus soziologischer Sicht interessieren zwar nicht die jeweils individuellen Lebensverläufe, sondern die gesellschaftlichen Verhältnisse, in denen sie sich entwickeln. Aber die individuellen Lebensverläufe bilden nicht nur den empirischen Ausgangspunkt für soziologische Einsichten in gesellschaftliche Verhältnisse; für die hier verfolgte soziologische Perspektive bilden sie auch das schließlich relevante Bezugsproblem: wir interessieren uns für gesellschaftliche Verhältnisse *als Bedingungen individueller Lebensverläufe*. In diesem und im folgenden Kapitel wird diskutiert, wie die Konstruktion statistischer Modelle diesem Erkenntnisinteresse dienen kann. In diesem Kapitel wird die Modellbildung unter deskriptiven Gesichtspunkten behandelt. Der Grundgedanke ist, wie in dem auf S. 39 angeführten Zitat Fishers, daß statistische Modelle eine vereinfachende Darstellung komplexer Daten ermöglichen sollen. In einer Formulierung von Cox und Snell [1981, S. 28]: „The objective of statistical analysis is to discover what conclusions can be drawn from data and to present these conclusions in as simple and lucid a form as is consistent with accuracy.“[1] Im Zentrum der Modellbildung steht das Konzept zustandsspezifischer Übergangsraten, das in Abschnitt 2.4.4 eingeführt worden ist. Da Lebensverlaufsdaten in der Regel nicht nur komplex, sondern auch unvollständig sind, nimmt das Problem, wie dies bei der Modellbildung berücksichtigt werden kann, einen verhältnismäßig breiten Raum ein (Abschnitt 3.4). Eine systematische Diskussion der Frage, wie statistische Modelle auch als Instrumente verstanden werden können, um gesellschaftliche Bedingungen individueller Lebensverläufe sichtbar zu machen, erfolgt erst in Kapitel 4.

3.1 Statistische Modelle und theoretische Deutungen

Der Übergang von der statistischen Beschreibung individueller Lebensverläufe zu soziologischen Vorstellungen über gesellschaftliche Verhältnisse

[1] Zu ähnliche Auffassungen über den Sinn der statistischen Modellbildung vgl. u.a. Nelder [1984].

(als Bedingungen individueller Lebensverläufe) kann allgemein als *Modellbildung* betrachtet werden. Vor allem zwei Aspekte sind wesentlich.

Erstens ist jede Modellbildung eine Vereinfachung der komplexen Realität. Dies gilt insbesondere, wenn man versucht, ein soziologisches Verständnis von Lebensverläufen zu gewinnen. An die Stelle des Versuchs, die Vielfalt und Komplexität der in einer Gesellschaft sich entwickelnden Lebensverläufe wahrzunehmen, muß eine vereinfachende, abstrahierende, schematisierende Betrachtungsweise treten. Es gibt drei wesentliche Schritte: Die Definition eines Biographieschemas, um von den jeweils möglichen Lebensverläufen sprechen zu können, die statistische Betrachtungsweise, die auf Verteilungen von Merkmalen in Gesamtheiten zielt und dadurch von allen Besonderheiten der die Gesamtheiten bildenden Individuen abstrahiert, und schließlich noch ein dritter Abstraktionsschritt, der als *statistische Modellbildung* bezeichnet werden kann: das Ziel liegt darin, empirisch ermittelbare Verteilungen von Zufallsvariablen durch vereinfachende theoretische Modelle zu approximieren, um ihren wesentlichen Informationsgehalt für eine sich anschließende theoretische Reflexion zugänglich zu machen.

Zweitens zielt Modellbildung auf eine Verknüpfung empirischen (deskriptiven) Wissens mit theoretischen Vorstellungen über den jeweils interessierenden Gegenstandsbereich. In unserem Fall handelt es sich darum, das empirische Wissen, das mithilfe statistischer Modelle über Lebensverläufe gewonnen werden kann, mit theoretischen Vorstellungen über gesellschaftliche Verhältnisse und deren Bedeutung für die Entwicklung individueller Lebensverläufe zu verknüpfen.

Über das Verhältnis von „Theorie“ und „Empirie“ gibt es unterschiedliche Ansichten.[2] In dieser Arbeit gehe ich davon aus, daß es nicht erforderlich (und vermutlich auch nicht möglich) ist, beide „Bereiche“ strikt abzugrenzen. Soziologisches Wissen entsteht durch direkte oder indirekte Beobachtungen *und* durch die theoretische Reflexion dieser Beobachtungen. Gleichwohl erscheint es sinnvoll, an einigen Unterscheidungen festzuhalten.

a) Der Informationsgehalt von Daten ist zwar davon abhängig, was man wissen möchte, von einer vorausgesetzten theoretischen Fragestellung. Soziologische Theoriebildung kann jedoch in den seltensten Fällen unmittelbar in ein statistisches Modell übersetzt werden. Es ist – auch deshalb – sinnvoll, statistische Modellbildung und soziologische Theoriebildung als zwei unterschiedliche Aspekte der empirischen Sozialforschung zu trennen. Die statistische Modellbildung kann dann als eine Art von Vermittlung betrachtet werden, die zwischen den Daten und der Theoriebildung steht. Ihre primäre Funktion besteht darin, die Theoriebildung mit empirischem

[2] Über die Entwicklung des Begriffs „Empirie“ in der Geschichte der empirischen Sozialforschung, insbesondere seine unterschiedlichen Prägungen durch sich wandelnde wissenschaftstheoretische Vorstellungen, vgl. Bonß [1982].

Wissen zu versorgen; aber damit sie dies tun kann, muß die Modellkonstruktion den theoretischen Fragestellungen folgen.

b) Das Verhältnis von Daten und statistischen Modellen kann auf unterschiedliche Weisen betrachtet werden. Eine in der statistischen Literatur verbreitete Auffassung betrachtet statistische Modelle als Formulierungen für Hypothesen, die mithilfe von Daten „getestet“ werden können. Mit Hypothesen sind dabei „allgemeine Gesetzmäßigkeiten“ gemeint, deren vermutete Geltung mithilfe der verfügbaren Daten bestätigt oder infrage gestellt werden soll. Wie ich in der Einleitung auszuführen versucht habe, ist diese Auffassung problematisch, insofern dabei implizit von einer temporalen Stabilität der durch Gesetzmäßigkeiten zu beschreibenden Sachverhalte ausgegangen wird. In dieser Arbeit gehe ich deshalb davon aus, daß die empirische Aufgabe zunächst darin besteht, Regeln bzw. soziale Regelmäßigkeiten in Erfahrung zu bringen, denen die Menschen in einer Gesellschaft folgen; daß dabei berücksichtigt werden sollte, daß sich diese Regeln verändern können; und daß infolgedessen eine besondere Aufgabe auch darin liegt, den Prozeß der Veränderung dieser Regeln zu beschreiben.

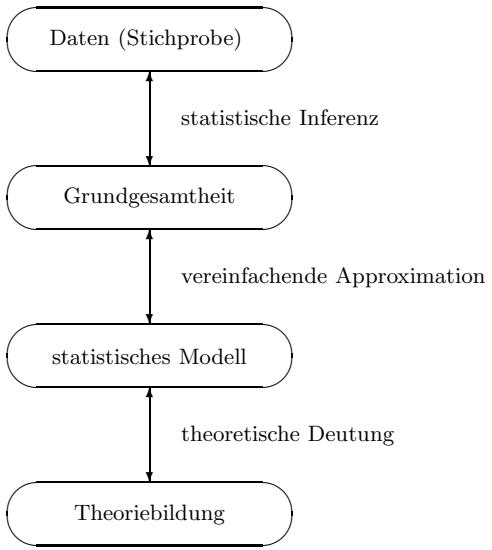
Daraus ergeben sich Implikationen sowohl für unsere Auffassung statistischer Modelle als auch für unser Verständnis des Verhältnisses von Daten und statistischen Modellen:[3] Statistische Modelle werden als formale Hilfsmitttel zur Beschreibung gesellschaftlicher Verhältnisse betrachtet. Um diese Intention für die Modellbildung formal präzisieren zu können, wird die Vorstellung einer räumlich und zeitlich abgrenzbaren, endlichen Grundgesamtheit von Individuen gebildet, so daß man sagen kann, daß *deren* gesellschaftliche Verhältnisse beschrieben werden sollen. Das Verhältnis zwischen Daten und statistischen Modellen ist dann zunächst dadurch charakterisierbar, daß die verfügbaren Daten – gemessen an der deskriptiven Intention der Modellbildung – in der Regel unvollständig sind. Typischerweise sind wir nur in der Lage, die interessierenden Sachverhalte für eine vergleichsweise kleine Stichprobe aus der eigentlich interessierenden Grundgesamtheit zu ermitteln; außerdem sind die verfügbaren Stichprobendaten in der Regel unvollständig, gemessen an der für eine vollständige Prozeßbeschreibung eigentlich erforderlichen Information. Daraus resultiert das statistische Inferenzproblem: wie mithilfe der jeweils vorliegenden Daten Einsichten in die Entwicklung der Lebensverläufe in einer Längsschnittgesamtheit von Individuen gewonnen werden können.

c) Geht man davon aus, daß es in der empirischen Lebensverlaufsforschung zunächst darum geht, deskriptive Modelle für räumlich und zeitlich abgrenzbare gesellschaftliche Verhältnisse – formal: für endliche Grundgesamtheiten von Individuen – zu bilden, können im Verhältnis von Daten und Modellen sinnvoll zwei Aspekte unterschieden werden. Erstens die Frage, wie mithilfe der Daten aus einer Stichprobe Einsichten in die *em-*

[3] Ich beziehe mich natürlich stets, ohne das immer wieder zu betonen, auf statistische Modellbildung als ein Werkzeug der soziologischen Lebensverlaufsforschung.

pirische Verteilung der interessierenden Zufallsvariablen in der Grundgesamtheit gewonnen werden können. Dies kann als das statistische Inferenzproblem (im engeren Sinne des Wortes) bezeichnet werden. Zweitens die Frage, wie für die mit den verfügbaren Daten *geschätzte* empirische Verteilung in der Grundgesamtheit eine sinnvoll vereinfachende Deskription gefunden werden kann. Im folgenden gehe ich von dieser Unterscheidung aus. Mit der Formulierung *deskriptive Modellbildung* ist also stets gemeint, daß eine sinnvoll vereinfachende Darstellung für die empirische Verteilung einer Reihe von Zufallsvariablen in einer endlichen Grundgesamtheit von Individuen angestrebt wird.

Das folgende Bild veranschaulicht die wesentlichen Punkte dieser Auffassung statistischer Modelle und ihres Verhältnisses einerseits zu Daten und andererseits zur soziologischen Theoriebildung.



Ausgangspunkt sind Daten, typischerweise handelt es sich um eine Stichprobe aus einer endlichen Grundgesamtheit von Individuen. Dadurch entsteht ein statistisches Inferenzproblem. Das statistische Modell bezieht sich auf die Grundgesamtheit, es soll die gesellschaftlichen Verhältnisse in dieser Grundgesamtheit beschreiben. Diese Zielsetzung verlangt, daß die für die Grundgesamtheit ermittelbaren bzw. schätzbaren Merkmalsverteilungen in eine vereinfachende Darstellungsform gebracht werden. Dabei gibt es zwei sich tendenziell widersprechende Anforderungen. Einerseits soll das statistische Modell den Bedürfnissen der Theoriebildung dienen, gewissermaßen als ein Werkzeug der theoretischen Spekulation. Andererseits soll es der Theoriebildung eine empirische Grundlage verschaffen, gewährleisten, daß die Theoriebildung sich auf die jeweils realen gesellschaftlichen Verhältnisse beziehen kann.

3.1.1 Hypothesen und theoretische Deutungen

Ein problematischer Aspekt dieser Betrachtungsweise liegt sicherlich in der Formulierung *theoretische Deutung*, um das Verhältnis zwischen Theoriebildung und deskriptiven statistischen Modellen zu charakterisieren. Ich verwende diese Formulierung, um noch einmal zum Ausdruck zu bringen, daß mir die im Kontext der „Analytischen Wissenschaftstheorie“ verbreitete Auffassung, die das Wesen der Theoriebildung in der Formulierung und Überprüfung von Hypothesen über allgemeine Gesetzmäßigkeiten sieht, für die empirische Sozialforschung zu eng erscheint. Die Frage hängt natürlich damit zusammen, worin man den Sinn der Theoriebildung sehen will.

Die zeitweilig verbreitete Orientierung an einem deduktiv-nomologischen Erklärungsschema hat dazu geführt, daß der wesentliche Sinn der Theoriebildung darin gesehen wurde, erfolgreiche Prognosen zu ermöglichen.[4] Theoretische Hypothesen erscheinen dann gleichbedeutend mit Annahmen über Gesetzmäßigkeiten, mit denen sich Ereignisse voraussagen lassen. Ich glaube, daß diese Betrachtungsweise in zweifacher Hinsicht zu eng ist. Zunächst schon deshalb, weil bereits für die Aufgabe einer Beschreibung gesellschaftlicher Verhältnisse Hypothesen, d.h. Bezugnahmen auf theoretische Vorstellungen über den zu beschreibenden Gegenstand, erforderlich sind. Hauptsächlich erscheint mir jedoch diese Betrachtungsweise deshalb zu eng, weil sie den Bereich sinnvoller Hypothesenbildung auf solche Aspekte der gesellschaftlichen Verhältnisse einschränkt, die eine hinreichende temporale Stabilität aufweisen, um Prognosen zu erlauben. Ob bzw. inwieweit es eine temporale Stabilität gesellschaftlicher Verhältnisse – und infolgedessen und insoweit auch eine Möglichkeit für konditionale Prognosen – gibt, muß jedoch als eine empirische Frage betrachtet werden. Es wäre nicht sinnvoll, diese für die Soziologie substantiell wichtige, also empirisch zu behandelnde Frage durch ein *methodologisches* Kriterium auszublenden (was man tut, wenn man die Möglichkeit verläßlicher Hypothesenbildung als eine Sinnvoraussetzung in die theoretische Sprache einbaut).

Ohne die Berechtigung des Interesses an konditionalen Prognosen bestreiten zu wollen, glaube ich – ähnlich wie Hacking und andere Autoren[5] –,

[4] Wenn hier und im folgenden von Prognosen gesprochen wird, sind stets *konditionale* Prognosen gemeint, nicht die von Popper [1965] kritisierten „Prophezeiungen“.

[5] Vgl. v.a. Hacking [1965, Kap. III]. Bereits Fisher [1955] hatte sich gegen die im Anschluß an Neyman, Pearson und Wald verbreitete Auffassung gewandt, Fragen der Wissensbildung mit praktischen Entscheidungsproblemen zu vermengen. Ähnliche Vorbehalte gegen eine entscheidungstheoretische Konzeption wissenschaftlicher Hypothesenbildung finden sich zum Beispiel bei Rozeboom [1960], Barnard und Smith in Savage [1962, S. 40 und S. 59], Anderson [1965, S. 234], Kempthorne [1971, S. 471f], Stegmüller

daß eine sinnvolle Unterscheidung zwischen Problemen der Theoriebildung und Entscheidungsproblemen, die aus der Verfolgung praktischer Zwecke resultieren, getroffen werden kann und daß sich daraus ein unterschiedlicher Sinn für die Formulierung von Hypothesen ergibt.[6]

Im Kontext der Theoriebildung und der auf sie bezogenen empirischen Sozialforschung hat die Formulierung von Hypothesen zunächst eine heuristische Bedeutung für das Verständnis des interessierenden Gegenstandsbereichs. Hypothesen dienen der Beschreibung, Interpretation und Erklärung sozialer Phänomene und können ganz unterschiedlicher Art sein. Dem oft zitierten Kriterium, daß es möglich sein sollte, die Hypothesen einer Theorie durch Beobachtungen zu bestätigen bzw. zu falsifizieren, kann selten eine klare Bedeutung gegeben werden.[7] Bestenfalls gelingt dies dort, wo es sich um räumlich und zeitlich eng begrenzte Annahmen über empirisch identifizierbare Aspekte einer Gesellschaft handelt. Schwierig wird es bereits, wenn eine Theorie Hypothesen über die Regeln formuliert, an denen sich die Menschen in einer Gesellschaft orientieren, denn solche Regeln verändern sich und ihre Geltung besteht nicht darin, daß sie stets ausnahmslos befolgt werden. Noch schwieriger wird es bei Hypothesen allgemeinerer Art, zum Beispiel: soziale Akteure handeln aufgrund von Erwartungen in bezug auf ihre soziale Umwelt. Bei allgemeinen Hypothesen dieser Art ist es kaum möglich, sie von den semantischen Regeln zu unterscheiden, die die Theoriesprache konstituieren. Schließlich können Hypothesen auch ohne einen identifizierbaren empirischen Anspruch in Gestalt eines hypothetischen Modells formuliert werden, um die Implikationen der durch das Modell formulierten Regeln zu untersuchen; von dieser Art sind zum Beispiel zahlreiche Rational-Choice-Modelle. Insofern glaube ich, daß man sagen kann, daß Hypothesen in der soziologischen Theoriebildung hauptsächlich eine *heuristische* Bedeutung haben. Alternative Hypothesen können reflektiert und zur Diskussion gestellt werden, ohne eine definitive Entscheidung treffen zu müssen.[8]

[1973, S. 81, 176f], Henkel [1976, S. 37f].

[6] Eine parallellaufende Unterscheidung findet sich in Stegmüllers [1971] Diskussion des Induktionsproblems. Er betont dort (S. 13f), „daß man zwei vollkommen verschiedene Arten von Gesichtspunkten scharf auseinanderhalten sollte: die *theoretische* Stellungnahme und die *praktisch-menschliche* Stellungnahme zu einer Hypothese. Der zweite Gesichtspunkt kann zwar auch systematisch untersucht werden. (Dies geschieht im Rahmen der rationalen Entscheidungstheorie.) Er sollte aber mit dem ersten nicht vermengt werden." Stegmüller fügt als Fußnote hinzu: „Leider ist diese Vermengung heute große Mode. Sie findet sich nicht erst bei Carnap. Man trifft sie auf Schritt und Tritt in der mathematischen Statistik, insbesondere in der statistischen Test- und Schätzungstheorie, an – dieser seltsamen Kombination von mathematischen Virtuosenleistungen und teils unausgegorenen, teils konfusen wissenschaftstheoretischen Vorstellungen."

[7] Das Problem ist in der wissenschaftstheoretischen Literatur ausführlich diskutiert worden, ohne daß ein definitives Ergebnis erzielt werden konnte; vgl. als Einführung Harding [1976].

[8] Dies rechtfertigt es, wissenschaftliche Hypothesen grundsätzlich als heuristisch an-

Wesentlich anders verhält es sich bei der Verfolgung praktischer Zwecke durch individuelle oder kollektive Akteure. Sie müssen Entscheidungen treffen, in der Regel unter Unsicherheit. Die Hypothesenbildung dient in diesem Kontext zur Rationalisierung von Entscheidungen unter Unsicherheit; die Brauchbarkeit von Hypothesen hängt also davon ab, daß sie zur Formulierung konditionaler Prognosen verwendet werden können. Die Frage, ob verläßliche konditionale Prognosen *möglich* sind, spielt dabei in gewisser Weise keine Rolle, denn wie Reichenbach [1949, S. 492] einmal bemerkt hat: „Wenn einer handeln will, so braucht er an den Erfolg nicht zu glauben." Der *Sinn* der Hypothesenbildung liegt nicht darin, deskriptiv angemessene Einsichten in die Verfassung gesellschaftlicher Verhältnisse zu erreichen, sondern darin, daß sie einen Beitrag zur Rationalisierung von Entscheidungen unter Unsicherheit liefern können.[9] Es ist deshalb verständlich, daß in diesem Kontext auch noch die Kriterien zur Prüfung und Beurteilung von Hypothesen ihrem Verwendungszweck unterworfen werden, formal dadurch, daß diese Kriterien durch Rückgriff auf eine dem Entscheidungsträger unterstellte Nutzenfunktion definiert werden; verständlich – als konzeptionelle Idee, abgesehen davon, ob sie forschungspraktisch realisierbar ist – zum Beispiel dann, wenn in der medizinischen Forschung Hypothesen über die Wirksamkeit von Medikamenten formuliert und geprüft werden sollen. Es erscheint mir jedoch offensichtlich, daß diese Orientierung an Entscheidungsproblemen unter Unsicherheit keinen geeigneten, zumindest keinen allgemein verbindlichen Rahmen für die Wissensbildung in der empirischen Sozialforschung liefern kann.[10]

Es geht jedoch nicht nur um die Frage, inwieweit ein Interesse an der Ermöglichung konditionaler Prognosen als Erkenntnisinteresse der empirischen Sozialforschung akzeptiert werden sollte. Überlegt werden sollte

zusehen. Lehnt man diese Betrachtungsweise ab, entfällt zugleich die wissenschaftstheoretische Bedeutung der Unterscheidung von Wissensbildung und praktischen Entscheidungen, und man könnte – wie z.B. Birnbaum [1977] – von zwei Varianten von Entscheidungen sprechen: einerseits praktische Entscheidungen, andererseits Entscheidungen, die mit dem „Akzeptieren" oder „Verwerfen" von Hypothesen verbunden sind.

[9] Wie Hacking [1965, S. 28] festgestellt hat, gibt es keinen unmittelbaren, logisch notwendigen Zusammenhang zwischen Hypothesen und Entscheidungen: „If one hypothesis is better supported than another, it would usually be, I believe, right to call it the more reasonable. But of course it need not be reasonable positively to believe the best supported hypothesis, nor the most reasonable one. Nor need it be reasonable to act as if one knew the best supported hypothesis were true."

[10] Damit wird nicht ausgeschlossen, daß empirische Sozialforschung Wissen zur Verfügung stellen kann, auf das soziale Akteure bei Entscheidungsproblemen unter Unsicherheit zurückgreifen können, d.h. zur Bildung einzelfallbezogener (subjektiver) Wahrscheinlichkeitsaussagen; vgl. Abschnitt 2.3.2. Die hier diskutierte Unterscheidung zielt auf eine Konzeption empirischer Sozialforschung, die sich nicht unmittelbar an der Ermöglichung solcher einzelfallbezogenen Wahrscheinlichkeitsaussagen orientiert, sondern an einer Vorstellung objektivierbarer Einsichten in die Verfassung und Entwicklung gesellschaftlicher Verhältnisse, zunächst also an deskriptiven Wahrscheinlichkeitsaussagen.

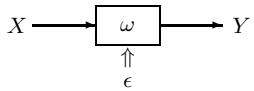
auch der potentielle wissenschaftliche Sinn einer Unterscheidung von spekulativ orientierter Theoriebildung und deskriptiv orientierter statistischer Modellbildung. Ein mir wichtig erscheinender Aspekt des Problems besteht darin, daß die Ansicht, daß empirische Daten nur als Material zur Überprüfung theoretischer Hypothesen von Belang sind, dazu verleitet, die Grenze zwischen theoretischer Spekulation und dem, was wir empirisch wissen (können), zu verwischen. Es ist zwar richtig, daß jeder Versuch einer solchen Grenzziehung problematisch ist, insofern empirische Wissensbildung nicht ohne eine theoretische Sprache auskommen kann und infolgedessen der Sinn empirischer Aussagen von der zu ihrer Formulierung gewählten Sprache abhängt. Dies gilt insbesondere für die empirische Sozialforschung, denn ihr Gegenstand – gesellschaftliche Verhältnisse und deren Wandel – kann nicht unmittelbar beobachtet werden. Beschreibungen gesellschaftlicher Verhältnisse und ihrer Veränderungen müssen vielmehr aus einer Vielzahl einzelner Beobachtungen konstruiert werden; und diese Konstruktionen erfordern in der Regel theoretische Annahmen, die nicht oder nur eingeschränkt durch die verfügbaren Daten begründet werden können. Insofern besteht bei der Modellkonstruktion die Möglichkeit, von *beliebigen* Annahmen auszugehen. Die Gefahr, daß eine theoretische Beliebigkeit entsteht, kann schließlich nur durch Konventionen eingeschränkt werden, die von einer *scientific community* für den Umgang mit ihrer theoretischen Sprache akzeptiert werden. Gleichwohl glaube ich, daß zumindest in zweierlei Hinsicht eine Unterscheidung zwischen theoretischer Spekulation und empirischem Wissen getroffen werden kann. Der erste Aspekt betrifft die Bezugnahme auf Raum und Zeit. So wie es selbstverständlich ist, daß jede Beobachtung an räumliche und zeitliche Koordinaten gebunden ist, gilt dies auch für den *empirischen* Anspruch eines mithilfe von Beobachtungen konstruierten statistischen Modells.[11] Der zweite Aspekt betrifft Annahmen über Merkmale von Individuen, für die keine Beobachtungen verfügbar sind. Es sollte davon ausgegangen werden, daß die Nichtverfügbarkeit von Beobachtungen eine Grenze empirischer Wissensbildung darstellt, die nur durch zusätzliche Beobachtungen, nicht jedoch durch theoretische Annahmen überwunden werden kann.[12]

[11] „Regularities, trends and laws can be observed only at a *local* and/or *partial* level, in the *past*. They should not be extended to the future. They should not be made more general than they are. Statements concerning the future can be but *conjectures*. Why, after all, would it be more interesting to look for general laws of development than to understand why there can be no such laws – what I have tried to do here – and to understand why the development of Columbia in the twentieth century was different, say, from the development of England in the eighteenth century." (Boudon [1983, S. 16]).

[12] Vgl. dazu die Diskussion „unbeobachteter Heterogenität" in Abschnitt 4.1.7.

3.1.2 Deskriptive statistische Modelle

Es gibt unterschiedliche Auffassungen über die Bedeutung und den Sinn statistischer Modelle. Häufig erfolgt eine unmittelbare Orientierung an der Aufgabenstellung, statistische Modelle für konditionale Prognosen über *individuelle* Verhaltensweisen zu entwickeln. Folgt man dieser Auffassung, kann unmittelbar an ein im Hinblick auf naturwissenschaftliche Anwendungen entwickeltes Verständnis statistischer Modelle angeknüpft werden. Ausgangspunkt ist die Annahme, daß Lebensverläufe Vorgänge sind, die zumindest partiell gewissen Gesetzmäßigkeiten unterliegen, die mithilfe statistischer Methoden sichtbar gemacht werden können.[13] Damit ist nicht unbedingt die Behauptung verbunden, daß sich Lebensverläufe von anderen „natürlichen" Vorgängen nicht unterscheiden. Solche Unterschiede mag es geben, aber von ihnen wird – wenn man der hier zunächst zu beschreibenden Auffassung folgt – für die statistische Modellbildung und die mit ihr verbundene (soziologische) Theoriebildung abstrahiert. Das dieser Auffassung zugrunde liegende Basismodell, ich nenne es im folgenden das ND-Modell, kann infolgedessen unabhängig von irgendwelchen Besonderheiten menschlicher Lebensverläufe formuliert werden. Folgende Abbildung kann zur Illustration dienen.



Die Variable Y repräsentiert das Verhalten eines Objekts ω. Dieses Verhalten hängt davon ab, wie das Objekt beschaffen ist und welchen Einflüssen es unterliegt. Ein Teil dieser Einflußfaktoren kann systematisch erfaßt und in einer Variable X repräsentiert werden. Ein anderer Teil der Bedingungen, denen das Verhalten des Objekts unterliegt, kann nicht systematisch erfaßt werden und wird durch eine unspezifische Variable ϵ repräsentiert. Die intendierte Gesetzmäßigkeit, um das Verhalten von ω zu erklären, kann dann als eine mathematische Funktion $Y = g(X, \epsilon)$ vorgestellt werden; und die Aufgabe besteht darin, Funktionen dieser Art zu finden, mit denen das beobachtbare Verhalten realer Objekte erfolgreich erklärt werden kann.

Das ND-Modell wird zu einem *statistischen* Modell dadurch, daß ϵ als eine Zufallsvariable interpretiert wird, als ein Sachverhalt, dessen mögliche individuelle Realisierungen „zufällig", d.h. mit dem gegenwärtig verfügbaren Kausalwissen nicht prognostizierbar sind. Erklärungen des Verhaltens Y, das infolgedessen ebenfalls zu einer Zufallsvariablen wird, durch die Ursachen X werden dann probabilistisch: kennt man die Ursachen X, kann das Verhalten Y nicht sicher vorausgesagt werden, sondern nur mit einer gewissen Wahrscheinlichkeit, abhängig von dem Ausmaß, in dem die

[13] Besser erscheint es, hier nicht von einer „Annahme" zu sprechen, d.h. von einer Vermutung, die empirisch infrage gestellt werden könnte, sondern von einer „regulativen Idee", die das intendierte Wissen definiert.

Zufallseinflüsse ϵ wirksam sind. Da – im Kontext der hier zunächst zu beschreibenden Auffassung – das Ziel darin liegt, möglichst vollständige Erklärungen (d.h. hier: möglichst gute konditionale Prognosen) zu erreichen, wird ϵ auch als „unerklärte Varianz“ bezeichnet. Ihr Ausmaß erscheint abhängig vom jeweils erreichten Wissensstand; und es erscheint nicht ausgeschlossen, daß immer mehr systematisch wirksame Einflußfaktoren gefunden werden können und auf diese Weise das jeweils verfügbare Kausalwissen erweitert werden kann.

Offensichtlich liefert das ND-Modell (in seiner statistischen Variante) auch einen möglichen formalen Rahmen zur Erklärung menschlicher Lebensverläufe. Es ist nicht schwer, sich vorzustellen, daß Lebensverläufe von Bedingungen abhängig sind, die zumindest teilweise systematisch erfaßt werden können. Unterschiede zu anderen, „natürlichen“ Vorgängen können darin gesehen werden, daß das Ausmaß der zufälligen Einflüsse vergleichsweise groß ist und daß die Gesetzmäßigkeiten zur Erklärung von Lebensverläufen einem historischen Wandel unterliegen. Aber diese Besonderheiten machen das Basismodell nicht hinfällig oder sinnlos. Wenn sich das Erkenntnisinteresse darauf richtet, Gesetzmäßigkeiten zu finden, mit denen individuelles Verhalten bzw. individuelle Lebensverläufe erklärt werden können, bietet das ND-Modell einen zweckmäßigen Rahmen, um dieses Erkenntnisinteresse zu verfolgen. Die häufig, zum Beispiel von Lieberson [1985] beklagte Tatsache, daß die auf der Grundlage dieses Modells ermittelbaren Gesetzmäßigkeiten nur eine sehr begrenzte temporale Stabilität aufweisen und daß typischerweise ein hohes Ausmaß an nicht erklärter Varianz vorhanden ist, kann nicht dem Modell angelastet werden, sondern charakterisiert den zu erklärenden Sachverhalt: kontingente Lebensverläufe.

Ein wesentliches Merkmal des ND-Modells kann darin gesehen werden, daß sich das intendierte Wissen unmittelbar auf *individuelle* Ereignisse richtet. Wird es im Kontext der Lebensverlaufsforschung verwendet, liegt das Ziel darin, Gesetzmäßigkeiten zu finden, mit denen *individuelle* Lebensverläufe erklärt werden können. Als ein alternatives Bezugsproblem für die soziologische Lebensverlaufsforschung kann man die Aufgabe betrachten, gesellschaftliche Verhältnisse als Rahmenbedingungen individueller Lebensverläufe zu beschreiben und zu erklären. Wie ich in der Einleitung auszuführen versucht habe, liegt ein wesentlicher theoretischer Vorteil dieser Konzeption darin, daß sie mit der Kontingenz individueller Lebensverläufe vereinbar ist. Die Frage ist, ob sich daraus Folgen für ein Verständnis der Bedeutung und des Sinns statistischer Modelle (zur Beschreibung von Lebensverläufen) ergeben. Auf den ersten Blick scheint dies nicht der Fall zu sein, denn auch das ND-Modell kann so interpretiert werden, daß es der Kontingenz individueller Lebensverläufe nicht widerspricht; die Zufallsvariable ϵ läßt dafür genügend Raum. Allerdings gibt es offensichtlich ein Spannungsverhältnis, wenn man auf das üblicherweise mit dem ND-Modell verfolgte Erkenntnisinteresse achtet. Denn dieses Erkenntnisinteresse läßt

die Kontingenz individueller Lebensverläufe gewissermaßen nur als eine Grenze erfolgreicher (nomologischer) Erklärungen erscheinen. Kontingenz erscheint nicht als ein wesentliches Merkmal von Lebensverläufen, sondern nur als Ausdruck unseres mangelhaften Kausalwissens.

Eine Alternative besteht darin, Kontingenz nicht als mangelhaftes Kausalwissen, sondern als ein wesentliches Merkmal der zu erforschenden Sachverhalte zu betrachten.[14] Damit verändert sich zugleich die Aufgabe der Theoriebildung. Ihre Problemstellung richtet sich nicht mehr unmittelbar auf die kontingenten Lebensverläufe der Individuen, sondern auf die sozialen Bedingungen, unter denen sich diese Kontingenz entwickelt.

Versucht man, dieser alternativen Betrachtungsweise zu folgen, verändert sich schließlich auch das theoretische Bezugsproblem zum Verständnis statistischer Modelle (im Kontext der soziologischen Lebensverlaufsforschung). Ihr primärer Sinn liegt dann darin, Beschreibungen gesellschaftlicher Verhältnisse als Bedingungen individueller Lebensverläufe zu ermöglichen. Daraus ergeben sich, wie ich glaube, zwei Unterschiede zum ND-Modell. Der erste Unterschied betrifft die Residuen. Das Ziel liegt nicht länger darin, die in diesen Residuen erscheinende „unerklärte Varianz“ möglichst klein zu machen. Stattdessen wird versucht, diese Residuen als Ausdruck einer für die jeweiligen gesellschaftlichen Verhältnisse typischen Vielfalt möglicher individueller Verhaltensweisen zu interpretieren. Diese Formulierung ist zweifellos vage, sie liefert jedoch einen ersten Leitfaden zum Verständnis von Übergangsratenmodellen, denn Übergangsraten beschreiben einen Sachverhalt, der aus der Perspektive des ND-Modells nur als eine Verteilung von Residuen wahrgenommen wird.

Der zweite Unterschied betrifft die Interpretation der zur Modellbildung verwendeten bedingten Wahrscheinlichkeitsverteilungen. Im Kontext des ND-Modells werden sie als isolierbare kausale Faktoren interpretiert, die – gemeinsam mit einer Reihe (noch) nicht erfaßter Faktoren – das jeweils individuell realisierte Verhalten determinieren. In einer Formulierung von Mueller, Schuessler und Costner [1970, S. 209]: „We conceive of any event as a resultant of innumerable factors, which may be classified into two broad categories: *determining* factors and *chance* factors. The determining factors are those which have been isolated and to which we attribute the explanation or causation of the event. [...] Chance factors are the remaining or unknown factors affecting an event.“ Zu einer anderen theoretischen Deutung gelangt man, wenn man sich die Entwicklung individueller Lebensverläufe als Resultat situationsbedingter Handlungen vorstellt:

[14] Ich folge teilweise Liebersons [1985] kritischen Überlegungen zur Verwendung des ND-Modells in der empirischen Sozialforschung, insbesondere seiner Kritik an der Auffassung, daß das Ausmaß „erklärter Varianz“ ein sinnvolls Kriterium für erfolgreiche soziologische Theoriebildung sei.

$$S \xrightarrow[\text{soziale Regeln}]{\text{individuelles Verhalten}}$$

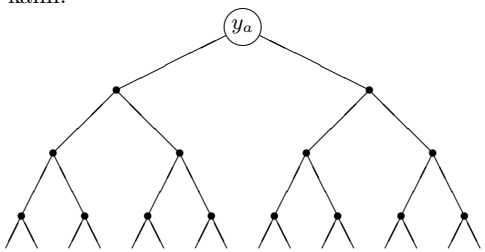
Der wesentliche Punkt ist, daß nicht einfach „die Vielfalt“ individueller Verhaltensweisen beschrieben wird, sondern daß diese Beschreibungen situationsbezogen – aus soziologischer Sicht: bezogen auf typische soziale Situationen – vorgenommen wird. Diese Betrachtungsweise erlaubt es, auf eine angemessene Weise von Bedingungen von Lebensverläufen zu sprechen; insbesondere wird dann die Darstellung vereinbar mit der üblichen Reflexion von Bedingungen von Lebensverläufen durch die sozialen Akteure selbst. Darüber hinaus erlaubt diese Betrachtungsweise eine Verknüpfung der statistischen Modellbildung mit der soziologischen Vorstellung, daß die Individuen in der Entwicklung ihrer Lebensverläufe sozialen Regeln bzw. Regelmäßigkeiten folgen. Der wichtige Punkt liegt wiederum darin, daß individuelle Verhaltensweisen auf spezifische Situationen bezogen werden, denn soziale Regeln sind stets auf spezifische Situationen bezogen. Der theoretische Sinn der bedingten Wahrscheinlichkeitsverteilungen, die im Mittelpunkt der statistischen Modellbildung stehen, liegt also bei dieser Betrachtungsweise darin, daß mit ihrer Hilfe ein Situationsbezug individuellen Verhaltens hergestellt werden kann. Man kann dann von Bedingungen individuellen Verhaltens sprechen, ohne dies mit der Annahme verknüpfen zu müssen, daß das individuelle Verhalten durch die jeweilige Situation, in der es stattfindet, determiniert wird.

Diese hier zunächst nur angedeutete Interpretation statistischer Modelle für die Lebensverlaufsforschung wird in den folgenden Abschnitten dieses und des nächsten Kapitels hoffentlich etwas genauer formuliert werden können. Um die angedeuteten Unterschiede zum ND-Modell zu betonen (ohne zugleich auf ein Alternativmodell verweisen zu müssen), spreche ich unspezifisch von *deskriptiven* statistischen Modellen. Der elementare Sinn liegt darin, empirisches Wissen zur Beschreibung gesellschaftlicher Verhältnisse zu vermitteln und, durch geeignete Bildung bedingter Wahrscheinlichkeitsverteilungen, Antworten auf Wie-Fragen (*wie* sich Lebensverläufe situationsspezifisch entwickeln) zu ermöglichen.

3.1.3 Ansatzpunkte für die Modellbildung

Lebensverläufe sind komplex, in zweierlei Hinsicht. Erstens gibt es eine Vielzahl unterschiedlicher Lebensverläufe, jedes Individuum entwickelt schließlich seinen eigenen, einmaligen Lebensverlauf. Zweitens ist jeder individuelle Lebensverlauf selbst ein hochgradig komplexer Prozeß. Der wesentliche methodische Schritt, um Einsichten in diese komplexen Sachverhalte zu gewinnen, besteht (aus soziologischer Sicht) darin, von den Individuen zu abstrahieren und stattdessen Gesamtheiten von Individuen zu betrachten und einen Begriff ihrer gesellschaftlichen Verhältnisse zu entwickeln. Diesem Perspektivenwechsel dient die statistische Modellbildung, die insofern ein wesentliches Hilfsmittel der soziologischen Lebensverlaufsforschung ist.

Voraussetzung dafür ist, daß die individuellen Lebensverläufe *vergleichbar* gemacht werden. Dies kann, wie in Abschnitt 2.2 dargestellt worden ist, durch die Konzeption eines Biographieschemas erreicht werden. Die Definition eines Biographieschemas fixiert jeweils diejenigen Aspekte der individuellen Lebensverläufe, die vergleichend beschrieben werden sollen. Aber selbst dann, wenn ein Biographieschema fixiert worden ist, bleibt noch eine beträchtliche Komplexität, wie man sich an folgendem Bild verdeutlichen kann.



Dieses Bild soll dazu dienen, die Entwicklung einer Gesamtheit von Lebensverläufen schematisch vorstellbar zu machen. Alle Lebensverläufe beginnen mit der Geburt, in einem gemeinsamen Ausgangszustand y_a. Dann verzweigen sich die Lebensverläufe. Jeder Knoten markiert eine Alternative, eine Situation konkurrierender Risiken.[15] Die Entwicklung führt gewissermaßen durch eine Serie von Alternativen, und bei jeder Alternative ergeben sich neue Verzweigungen. Schließlich entsteht eine komplexe Vielfalt unterschiedlicher individueller Lebensverläufe, die kaum noch vergleichbar sind.

Das Bild zeigt jedoch auch, daß sich die Modellbildung auf zwei unterschiedliche Aspekte in der Entwicklung von Lebensverläufen beziehen kann. Man kann sich darauf konzentrieren, die Vielfalt der schließlich realisierten Lebensverläufe sichtbar zu machen. Gegenstand der Darstellung sind dann die „vollständigen“ Lebensverläufe, die in einem Ausgangszustand beginnen und schließlich in einem „absorbierenden“ Endzustand aufhören. Stattdessen kann man sich zunächst auf die einzelnen Verzweigungen konzentrieren. Die Fragestellung ist dann, wie sich die Lebensverläufe verzweigen, wie die Übergänge verlaufen und wovon dies abhängt.

Wichtig erscheint mir, den an dieser Stelle einsetzenden Perspektivenwechsel zu verstehen. Bezieht man sich auf die einzelnen Verzweigungssituationen, besteht das Ziel zunächst nicht darin, die Vielfalt der unterschiedlichen Lebensverläufe wahrzunehmen, die tatsächlich realisiert wer-

[15] Um eine einfache graphische Darstellung zu erreichen, zeigt die Abbildung nur binäre Alternativen. Die in Abschnitt 2.4.4 eingeführte Terminologie konkurrierender Risiken kann dagegen beliebige Verzweigungen berücksichtigen.

den; vielmehr geht es darum, die Regeln zu verstehen, nach denen sich die Lebensverläufe in jeweils spezifischen, der Modellbildung vorausgesetzten Situationen entwickeln. Ein einfaches Modell für eheliche und nichteheliche Lebensgemeinschaften kann helfen, die Unterscheidung zu verdeutlichen. Abbildung 3.1.1 veranschaulicht das hierfür gewählte Biographieschema, den Zustandsraum und die möglichen Übergänge.

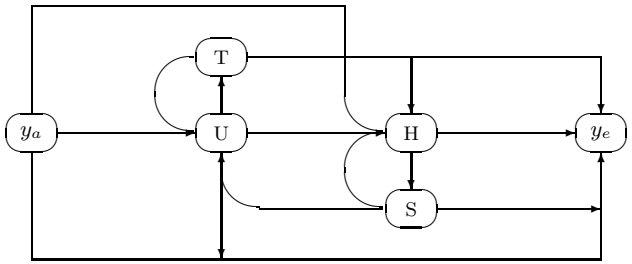


Abbildung 3.1.1 Biographieschema für die Bildung und Auflösung von Lebensgemeinschaften.

Alle Lebensverläufe beginnen im Anfangszustand y_a und enden im Endzustand y_e. Vom Anfangszustand ausgehend sind drei Übergänge möglich: Heirat (H), Übergang in eine nichteheliche Lebensgemeinschaft (U), oder ein direkter Übergang in den Endzustand. Ausgehend vom Zustand H sind zwei Übergänge möglich: Scheidung (S) oder Übergang in den Endzustand; und ausgehend vom Zustand U sind drei Übergänge möglich: Heirat (H), Trennung (T), oder Übergang in den Endzustand. Schließlich berücksichtigt das Biographieschema, daß sowohl eheliche als auch nichteheliche Lebensgemeinschaften mehrfach wiederholt werden können.

Innerhalb dieses Biographieschemas sind zahlreiche unterschiedliche Lebensverläufe möglich. Bereits ohne Berücksichtigung wiederholbarer Ereignisse gibt es 10 mögliche Verläufe. Zieht man in Betracht, daß alle internen Zustände mehrfach durchlaufen werden können und daß sich Lebensverläufe auch dadurch unterscheiden, wie lange sich die Individuen in den einzelnen Zuständen aufhalten, wird die Zahl der möglichen Verläufe *sehr* groß. Jeder Versuch, ein Bild der realen Vielfalt von Lebensverläufen zu gewinnen, stößt infolgedessen sehr schnell an Grenzen.

Andererseits können die im Biographieschema von Abbildung 3.1.1 möglichen Lebensverläufe durch ein System zustandsspezifischer Übergangsraten beschrieben werden. Man benötigt insgesamt 12 Übergangsraten, entsprechend der Anzahl möglicher Zustandswechsel. Das System dieser Übergangsraten gibt dann eine erste Einsicht in die Regeln, nach denen sich die Lebensverläufe als eine Abfolge kontingenter Episoden entwickeln.

Dieser Perspektivenwechsel, der sich zunächst auf die einzelnen Verzweigungssituationen in der Entwicklung von Lebensverläufen konzentriert, erscheint als ein sinnvoller Ausgangspunkt sowohl für die statistische Modellbildung als auch für die soziologische Theoriebildung. Tatsächlich beschränkt sich die soziologische Lebensverlaufsforschung, die sich der statistischen Methoden der Ereignisanalyse bedient, bisher weitgehend auf diesen Ansatzpunkt für die Modell- und Theoriebildung. Es sollte jedoch nicht vergessen werden, daß auf diesem Wege nicht ohne weiteres ein hinreichendes Bild der tatsächlichen Vielfalt der in einer Gesellschaft realisierten Lebensverläufe gewonnen werden kann.[16]

[16] Vgl. die Diskussion von „transitions“ und „trajectories“ bei Hagestad [1991].

3.2 Modellkonstruktion und Schätzverfahren

Als zentraler Begriff zur Beschreibung der Entwicklung von Lebensverläufen (stets in einer endlichen Grundgesamtheit von Individuen) wurde in Kapitel 2 das Konzept der Übergangsrate dargestellt. Auf diesen Begriff konzentriert sich deshalb auch die Modellbildung. Es geht darum, mithilfe statistischer Modelle Einsichten in den Verlauf und in mögliche Bedingungen für den Verlauf von Übergangsraten zu gewinnen.

Um die Grundgedanken der Modellkonstruktion darzustellen, soll zunächst eine einfache Situation konkurrierender Risiken betrachtet werden. Die Grundgesamtheit Ω besteht aus einer Menge von Individuen, die zum gleichen Zeitpunkt in einen gegebenen Ausgangszustand geraten sind. Hiervon ausgehend können sie Übergänge in einen von mehreren möglichen Folgezuständen vollziehen. Die Situation kann formal durch eine zweidimensionale Zufallsvariable

$$(T, D) : \Omega \longrightarrow \mathcal{T} \times \mathcal{D}$$

charakterisiert werden. T ist die Zeitdauer der Episode, bis zum erstenmal der Anfangszustand verlassen wird, $\mathcal{T}$ ist eine Prozeßzeitachse. $\mathcal{D}$ ist die Menge der möglichen Folgezustände: $\mathcal{D} = \{1, \ldots, m\}$.

Kennt man für alle Individuen aus der Grundgesamtheit die von ihnen realisierten Werte dieser zweidimensionalen Zufallsvariable, können die zustandsspezifischen Übergangsraten $r_k(t)$ (für $k = 1, \ldots, m$) berechnet werden. Sie beschreiben den Prozeß so, wie er bei diesen Individuen tatsächlich abgelaufen ist.

Das Motiv für die Konstruktion eines statistischen Modells liegt zunächst darin, eine einfachere mathematische Form zur Darstellung dieser empirischen Übergangsraten zu finden. Der übliche Ausgangspunkt für die Modellbildung ist deshalb die Annahme einer parametrischen Klasse möglicher Modelle,[17] in formaler Darstellung:

$$r_k(t) \;\approx\; \tilde{r}_k(t;\theta) \qquad \theta \in \Theta$$

Auf der linken Seite steht die empirische Übergangsrate; auf der rechten Seite steht eine Klasse theoretischer Übergangsraten, die sich nur durch einen Parametervektor θ unterscheiden, der in einer Menge möglicher Werte (Θ) variieren kann. Das Zeichen $\approx$ soll zum Ausdruck bringen, daß die theoretischen Übergangsraten einer approximativen Darstellung der empirischen Übergangsraten dienen sollen. (Es sei angemerkt, daß wir in dieser Arbeit stets das Tilde-Zeichen verwenden, um auf Modelle hinzuweisen.)

[17] Dies gilt auch für die sog. semi-parametrischen Modelle, die im wesentlichen dadurch charakterisiert werden können, daß sie eine vergleichsweise große Menge an freien Parametern enthalten. Zu nennen sind insbesondere die von Cox [1972] eingeführte Klasse von Modellen sowie die von Wu und Tuma [1990] vorgeschlagenen „lokalen" Ratenmodelle. Auf beide Typen von Modellen wird jedoch in dieser Arbeit nicht näher Bezug genommen.

a) Es stellt sich die Frage, wie man geeignete Modelle, d.h. Klassen theoretischer Übergangsraten findet. Ein Teil dieser Frage ist leicht zu beantworten. Die theoretische Statistik stellt eine große Anzahl unterschiedlicher Modellklassen zur Verfügung; es gibt eine inzwischen sehr breite Literatur, in der diese Modelle dargestellt und ihre Eigenschaften diskutiert werden.[18] Es bleibt allerdings die Frage, wie man eine *geeignete* Modellklasse findet. Auf diese Frage gibt es keine einfache Antwort, da – wie bereits weiter oben festgestellt wurde – statistische Modelle zwei unterschiedlichen, sich tendenziell widersprechenden Anforderungen genügen sollen. Sie sollen einerseits eine *vereinfachende* Beschreibung einer gegebenen Menge an Daten im Hinblick auf eine zugrundeliegende theoretische Fragestellung liefern; aber andererseits soll der empirische Informationsgehalt der Daten gewahrt bleiben.[19] Schließlich gelangt man immer wieder zu der Frage, wie eine *sinnvolle* Vereinfachung gefunden werden kann, und hierfür gibt es keine objektiven Kriterien.

b) Ein zweites Problem betrifft die Frage, wie aus der vorgegebenen Klasse möglicher Modelle dasjenige gefunden werden kann, das am besten zu den verfügbaren Daten paßt. Dafür benötigt man ein Maß, mit dem die Ähnlichkeit statistischer Verteilungen beurteilt werden kann. Zahlreiche unterschiedliche Distanzmaße sind vorstellbar und in der statistischen Literatur diskutiert worden. Wir gehen in dieser Arbeit von einem von Kullback und Laibler [1951] eingeführten Distanzmaß aus, das einen unmittelbaren Anschluß an die Theorie der Maximum-Likelihood-Schätzung ermöglicht. Eine Darstellung dieses Distanzmaßes erfolgt weiter unten.

c) Die Formulierung *Schätzung eines statistischen Modells* soll bedeuten, daß aus einer gegebenen Klasse möglicher Modelle ein Modell ausgesucht wird, das – bezogen auf ein vorausgesetztes Distanzmaß – am besten zu den verfügbaren Daten paßt. In diesem Sinne kann man sagen, daß das schließlich ausgewählte Modell eine optimale vereinfachende Beschreibung der Daten liefert. Vereinfachung ist zwar grundlegend, es gibt jedoch noch einen anderen, gleichermaßen wesentlichen Aspekt der Modellschätzung.

[18] Vgl. zum Beispiel Lawless [1982], Andreß [1985, 1989], Blossfeld, Hamerle und Mayer [1986, 1989], Schneider [1991], Courgeau und Lelievre [1992].

[19] Der latente Widerspruch kommt auch in der Unterscheidung von parametrischen und semi- und nicht-parametrischen Modellen zum Ausdruck. Auf der einen Seite stehen die „stark vereinfachenden" parametrischen Modelle, auf der anderen Seite die „flexibleren" semi- und nicht-parametrischen Modelle, die in der Regel bessere Approximationen an die empirischen Verteilungsfunktionen erlauben. Es ist jedoch kaum mögich, eine grundsätzliche Unterscheidung zu treffen. Auch bei der Verfolgung semi- und nicht-parametrischer Modellansätze kommt man nicht umhin, zunächst eine Klasse möglicher Modelle zu spezifizieren. In der Regel liegt der Unterschied nur in der Anzahl der Parameter, die zur Anpassung des Modells an die empirische Verteilung variiert werden können. Je größer die Anzahl der Parameter, desto mehr „Freiheitsgrade" gibt es, um das Modell den Daten anzupassen; aber desto schwerer wird auch eine theoretische Interpretation des Modells.

Er besteht darin, daß Modelle auch mit *unvollständigen* Daten geschätzt werden können. Dies ist gleichermaßen grundlegend, denn in der Regel stehen uns für die Modellschätzung nur unvollständige Daten zur Verfügung.

Daß Daten unvollständig sind, kann wiederum unter zwei unterschiedlichen Aspekten betrachtet werden. Erstens ist damit gemeint, daß uns in der Regel nur eine Stichprobe für die eigentlich interessierende Grundgesamtheit (Längsschnittgesamtheit) zur Verfügung steht. Von dem damit verbundenen statistischen Inferenzproblem wird hier zunächst abgesehen, einige Aspekte werden in Kapitel 5 diskutiert. Eine wesentlich andere Art von Unvollständigkeit resultiert daraus, daß die uns interessierenden Modelle den Verlauf eines Prozesses darstellen sollen. Unvollständigkeit bedeutet in diesem Zusammenhang, daß wir nur unvollständige Informationen darüber haben, wie der Prozeß abgelaufen ist.

Die Unvollständigkeit der jeweils verfügbaren Daten ist offensichtlich ein zentrales Problem für die statistische Modellbildung. Es gibt im wesentlichen zwei Möglichkeiten, um mit diesem Problem umzugehen. Die erste Möglichkeit besteht darin, daß man sich auf eine Beschreibung der jeweils vorhandenen Daten beschränkt, d.h. daß man ein Modell für die verfügbaren Daten konstruiert. Diese Vorgehensweise ist jedoch unbefriedigend, da wir in der Regel nicht primär an den jeweils verfügbaren Daten interessiert sind, sondern mithilfe dieser Daten Einsichten in die durch sie nur partiell beobachteten gesellschaftlichen Verhältnisse gewinnen möchten. Dann benötigt man allerdings Annahmen, und dies führt zur zweiten Möglichkeit, um mit unvollständigen Daten umzugehen. Die Basisannahme besteht darin, daß sich das, was man nicht beobachtet hat, „so ähnlich“ verhält wie das, was man beobachtet hat. Es ist jedoch evident, daß diese Basisannahme nicht generell richtig sein kann. Es ist deshalb in jedem Anwendungsfall zu überlegen, unter welchen Bedingungen diese Annahme getroffen werden kann. Einige der damit verbundenen Probleme werden in Abschitt 3.4 diskutiert.

Das KL-Distanzmaß zum Vergleich von Verteilungen

Im Mittelpunkt der Modellschätzung steht ein Vergleich der empirischen Verteilung einer Zufallsvariable in einer gegebenen Menge von Daten und einer Klasse theoretischer Verteilungen für die gleiche Zufallsvariable. Man benötigt also ein Maß, um Aussagen über die Ähnlichkeit von Verteilungen machen zu können. Wie bereits erwähnt, wird in dieser Arbeit ein von Kullback und Laibler [1951] eingeführtes Distanzmaß verwendet, weil es sich unmittelbar mit der in der Praxis hauptsächlich verwendeten Maximum-Likelihood-Schätzung verbinden läßt. Ich nenne es im folgenden abkürzend das *KL-Distanzmaß*.

Das KL-Distanzmaß kann in verschiedenen Varianten dargestellt werden, je nachdem ob stetige und / oder diskrete Verteilungen miteinander verglichen werden sollen. Zur Erläuterung beschränke ich mich zunächst

auf eine Situation, in der für eine diskrete Zufallsvariable ein diskretes Modell gefunden werden soll. Ausgangspunkt ist also eine diskrete Zufallsvariable

$$X : \Omega \longrightarrow \mathcal{X}$$

mit einer empirischen Verteilungsfunktion $F_X(x)$; $\mathcal{X} = \{a_1, \ldots, a_m\}$ sei der Wertebereich. Die Verteilung von X kann also vollständig durch die Wahrscheinlichkeiten

$$P = (p_1, \ldots, p_m) \quad \text{mit} \quad p_j = \mathrm{P}\left(X = a_j\right) > 0$$

ausgedrückt werden. Eine allgemeine Klasse möglicher diskreter Modelle kann in diesem Fall durch

$$\mathcal{K}_m = \left\{ (\tilde{p}_1, \ldots, \tilde{p}_m) \;\middle|\; \sum_{j=1}^{m} \tilde{p}_j = 1,\, \tilde{p}_j > 0 \right\}$$

definiert werden.[20] Werden keine zusätzlichen Restriktionen vorausgesetzt, enthält sie die Verteilung P als einen Spezialfall. Die Definition des KL-Distanzmaßes sieht dann folgendermaßen aus:

$$D[P, \tilde{P}] = \sum_{j=1}^{m} p_j \, \log \left\{ \frac{p_j}{\tilde{p}_j} \right\}$$

Die auf diese Weise definierte Distanz kann als gewichteter Mittelwert (Erwartungswert bezüglich P) der Abweichungen $\log(p_j/\tilde{p}_j)$ zwischen den Verteilungen P und $\tilde{P}$ interpretiert werden. Je größer die durchschnittliche Abweichung der beiden Verteilungen, desto größer ist der Wert des Distanzmaßes. Seine grundlegenden Eigenschaften sind:[21]

$$\begin{aligned} D[P, \tilde{P}] &\geq 0 \\ D[P, \tilde{P}] &= 0 \quad \text{genau dann, wenn } P = \tilde{P} \\ D[P, \tilde{P}] &\quad \text{ist eine in den Variablen } \tilde{p}_j \text{ streng konvexe Funktion} \end{aligned}$$

Abbildung 3.2.1 illustriert das KL-Distanzmaß für eine Binomialverteilung $P = (0.4, 0.6)$. Da $\tilde{p}_1 + \tilde{p}_2 = 1$, genügt es, $D[P, \tilde{P}]$ als Funktion von $\tilde{p}_1$ zu betrachten. Man erkennt anschaulich die drei genannten Eigenschaften.

Da $D[P, \tilde{P}]$ streng konvex ist, kann stets genau ein optimales Modell $\tilde{P}^\circ$ gefunden werden, das die Distanz zur empirischen (vorgegebenen) Verteilung P minimiert. Ist P in der Klasse von Modellen $\tilde{P}$ enthalten, ist

[20] Bei dieser Formulierung wird davon ausgegangen, daß es für alle möglichen Werte aus dem Wertebereich von X eine positive Wahrscheinlichkeit gibt. Diese Restriktion ist jedoch überflüssig, wenn man von der Konvention $0 \log(0) = 0$ ausgeht.

[21] Beweise findet man zum Beispiel bei Kapur [1989].

$\tilde{P}^\circ = P$. Hiervon kann jedoch im allgemeinen nicht ausgegangen werden, denn die Modellbildung zielt darauf, die empirische Verteilung durch eine einfachere theoretische Verteilung zu approximieren. Im allgemeinen hängt dann die Existenz und Eindeutigkeit optimaler Modell von der vorausgesetzten Modellklasse ab. Diese Frage ist infolgedessen bei der Modellkonstruktion stets gesondert zu überlegen.

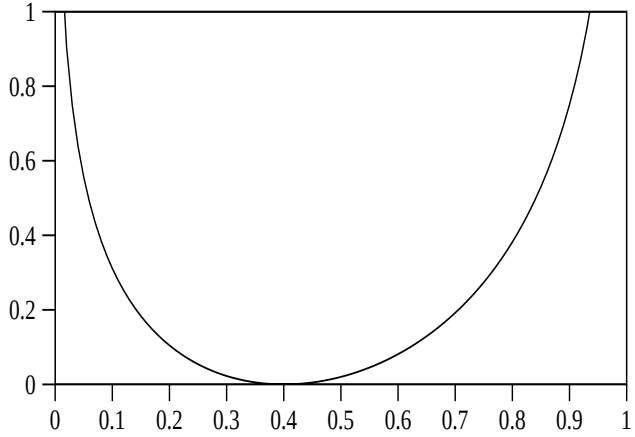


Abbildung 3.2.1. Darstellung von $D[P, \tilde{P}]$ mit $P = (0.4, 0.6)$ und $\tilde{P} = (\tilde{p}_1, \tilde{p}_2)$ als Funktion von $\tilde{p}_1$.

Maximum-Likelihood-Schätzung

In formaler Hinsicht ist die Verwendung des KL-Distanzmaßes äquivalent zum Maximum-Likelihood-Schätzverfahren. Man sieht dies sehr leicht folgendermaßen. Sei $f_X(x)$ die empirische Wahrscheinlichkeitsfunktion für die diskrete Zufallsvariable X mit dem Wertebereich $\mathcal{X}$, und sei $\tilde{f}_X(x;\theta)$ ($\theta \in \Theta$) eine Klasse möglicher Modelle. Das KL-Distanzmaß ist dann

$$D(\theta) = \sum_{x\in\mathcal{X}} f_X(x)\, \log\left\{\frac{f_X(x)}{\tilde{f}_X(x;\theta)}\right\} = \tag{3.1}$$

$$\sum_{x\in\mathcal{X}} f_X(x)\, \log\{f_X(x)\} - \sum_{x\in\mathcal{X}} f_X(x)\, \log\left\{\tilde{f}_X(x;\theta)\right\}$$

Betrachtet man die Distanz als eine Funktion des zu schätzenden Parametervektors θ, sieht man, daß ihre Minimierung äquivalent ist zur Maximierung des Ausdrucks, der in der untersten Zeile von (3.1) ganz rechts steht. Dieser Ausdruck ist jedoch proportional zur Log-Likelihood. Denn

$$\sum_{x\in\mathcal{X}} f_X(x)\, \log\left\{\tilde{f}_X(x;\theta)\right\} = \frac{1}{N}\sum_{i=1}^{N} \log\left\{\tilde{f}_X(x_i;\theta)\right\}$$

wobei $x_1, \ldots, x_N$ die in der Grundgesamtheit realisierten Werte der Zufallsvariable X sind. Die Minimierung des KL-Distanzmaßes ist also äquivalent zur Maximierung der Likelihood bzw. Log-Likelihood

Die Interpretation der ML-Methode mit dem KL-Distanzmaß liefert eine einfache und anschauliche Deutung: Die Modellschätzung besteht darin, aus einer vorgegebenen Klasse möglicher Modelle dasjenige auszuwählen, das – im Sinne des KL-Distanzmaßes – von der empirischen Verteilung der Daten am wenigsten abweicht. Diese Interpretation setzt jedoch eine vollständige Kenntnis der empirischen Verteilung in der Grundgesamtheit voraus. Wenn dies nicht der Fall ist, bedarf es zusätzlicher Überlegungen, um das Verfahren der Modellschätzung zu rechtfertigen. Dies betrifft einerseits das statistische Inferenzproblem, wenn nur Daten aus einer Stichprobe verfügbar sind, und andererseits das Problem, wie eine geeignete Modellschätzung bei unvollständigen Daten erreicht werden kann. Darauf wird in Abschnitt 3.4 näher eingegangen.

3.3 Modelle für eine diskrete Zeitachse

Um die Konstruktion und Schätzung von Übergangsratenmodellen zu illustrieren, beginne ich mit einer sehr einfachen Situation: es gibt nur eine Episode mit einem möglichen Zielzustand, und die Modellbildung soll auf einer diskreten Prozeßzeitachse erfolgen. Alle Individuen aus der vorausgesetzten Grundgesamtheit Ω beginnen die Episode zum gleichen Zeitpunkt $t = 0$ im Anfangszustand y_a, und sie verlassen ihn nach einer mehr oder weniger langen Verweildauer in den Endzustand y_e. Der Episodenverlauf kann also vollständig durch eine diskrete Zufallsvariable T beschrieben werden. Der Wertebereich für T besteht aus der Prozeßzeitachse $\mathcal{T} = \{1, 2, 3, \ldots\}$.

Zusätzlich nehmen wir an, daß die Individuen zum Beginn des Episodenverlaufs in m unterschiedliche Gruppen eingeteilt werden können. Um dies formal ausdrücken zu können, wird eine Zufallsvariable X mit dem Wertebereich $\mathcal{X} = \{1, \ldots, m\}$ verwendet. Der Ausgangspunkt für die Modellbildung besteht also in einem deskriptiven Wahrscheinlichkeitsraum mit der zweidimensionalen Zufallsvariable

$$(T, X) : \Omega \longrightarrow \mathcal{T} \times \mathcal{X}$$

Hiervon ausgehend können die empirischen Übergangsraten

$$r(t \mid x) \;=\; \mathrm{P}\,(T = t \mid T \geq t, X = x)$$

definiert werden. Für jeden Wert von X erhält man eine spezielle Übergangsrate, die den Prozeßverlauf für die Individuen aus der korrespondierenden Teilgesamtheit beschreibt.

Der nächste Schritt besteht darin, eine Klasse möglicher Modelle zu definieren. Dafür gibt es zahlreiche unterschiedliche Möglichkeiten.[22] Im einfachsten Fall könnte man zum Beispiel annehmen, daß es in jeder der m Teilgesamtheiten einen konstanten Verlauf der Übergangsrate gibt. Der Modellansatz kann dann folgendermaßen formuliert werden:

$$r(t \mid x) \;\approx\; \tilde{r}(t \mid x) = \theta_x \quad \text{wobei} \quad 0 \leq \theta_x \leq 1 \tag{3.2}$$

Hier ist die theoretische Übergangsrate mit dem zu schätzenden Parameter identisch, und da Übergangsraten bei einer diskreten Zeitachse bedingte Wahrscheinlichkeiten sind, muß der zu schätzende Parameter ein Wert aus dem Intervall $[0, 1]$ sein.

Eine äquivalente Formulierung wird durch das logistische Regressionsmodell erreicht. Die übliche Modellformulierung kann am einfachsten dargestellt werden, wenn man die Zufallsvariable X durch m Variablen

[22] Außer der bereits in Anmerkung 18 auf Seite 131 genannten Literatur vgl. als Arbeiten, die sich insbesondere mit diskreten Übergangsratenmodellen beschäftigen, Allison [1982] und Hamerle und Tutz [1989].

$X_1, \ldots, X_m$ repräsentiert, die nur die Werte 0 oder 1 annehmen können, also

$$X_j \;=\; \begin{cases} 1 & \text{wenn } X = j \\ 0 & \text{andernfalls} \end{cases} \tag{3.3}$$

Als Standardformulierung für das logistische Regressionsmodell erhält man dann

$$\tilde{r}(t \mid x; \beta_1, \ldots, \beta_m) \;=\; \frac{\exp(X_1\beta_1 + \cdots + X_m\beta_m)}{1 + \exp(X_1\beta_1 + \cdots + X_m\beta_m)} \tag{3.4}$$

Dies ist wiederum eine Klasse von Modellen, der Parametervektor ist in diesem Fall $\theta = (\beta_1, \ldots, \beta_m)$. Die Modellformulierung hat offensichtlich die Eigenschaft, daß für beliebige Parameterwerte die Bedingung $\tilde{r}(t \mid x; \theta) \in [0, 1]$ erüllt ist.

Bei der in (3.2) bzw. (3.4) definierten Modellklasse können sich die Übergangsraten zwar zwischen den Teilgesamtheiten unterscheiden, innerhalb jeder Teilgesamtheit gibt es jedoch nur einen zeitunabhängigen konstanten Wert. Einen flexibleren Modellansatz, der auch zeitabhängige Verläufe der Übergangsrate erfassen kann, erhält man zum Beispiel durch folgende einfache Verallgemeinerung des logistischen Regressionsmodells:

$$\tilde{r}(t \mid x; \theta') \;=\; \frac{\exp(\alpha_t + X_2\beta_2 + \cdots + X_m\beta_m)}{1 + \exp(\alpha_t + X_2\beta_2 + \cdots + X_m\beta_m)}$$

Eine Alternative und einfacher zu interpretieren wäre ein proportionales Übergangsratenmodell, das eine beliebige, jedoch für alle Teilgesamtheiten identische Basisrate annimmt. Die Übergangsraten in den Teilgesamtheiten verlaufen dann proportional. Eine mögliche Modellformulierung wäre

$$\tilde{r}(t \mid x; \theta'') \;=\; \alpha_t \, \frac{\exp(X_2\beta_2 + \cdots + X_m\beta_m)}{1 + \exp(X_2\beta_2 + \cdots + X_m\beta_m)}$$

Modellschätzung mit vollständigen Daten

Wenn vollständige Daten vorliegen, können Übergangsratenmodelle sehr leicht mithilfes des KL-Distanzmaßes bzw. der ML-Methode geschätzt werden. Man kennt dann für jedes Individuum der Grundgesamtheit die realisierte Verweildauer und seine Gruppenzugehörigkeit. Die verfügbaren Daten sehen in diesem Fall folgendermaßen aus:

Individuum	Verweildauer	Kovariable
ω_1	$T(\omega_1)$	$X(\omega_1)$
ω_2	$T(\omega_2)$	$X(\omega_2)$
ω_3	$T(\omega_3)$	$X(\omega_3)$
$\vdots$	$\vdots$	$\vdots$
ω_N	$T(\omega_N)$	$X(\omega_N)$

(3.5)

$T(\omega_i)$ ist der beim Individuum ω_i realisierte Wert der Verweildauer T, $X(\omega_i)$ ist der realisierte Wert der Kovariable X, die die Gruppenzugehörigkeit angibt.

Das KL-Distanzmaß vergleicht die empirische Verteilung der Daten mit den für sie theoretisch angenommenen Verteilungen. Die empirische Verteilung ist in Form der zweidimensionalen Wahrscheinlichkeitsfunktion

$$f(t,x) = Pr(T = t, X = x)$$

bekannt. Da die möglichen Modelle durch Klassen von Übergangsraten definiert sind, müssen zunächst die korrespondierenden Wahrscheinlichkeitsverteilungen berechnet werden. Wie in Abschnitt 2.4.3 bereits gezeigt worden ist, kann dies immer erreicht werden, da die Übergangsrate eine vollständige Beschreibung des Episodenverlaufs liefert. Eine einfache Verallgemeinerung der dort angestellten Überlegungen liefert

$$\tilde{f}(t \mid x;\theta) \;=\; \tilde{r}(t \mid x;\theta) \prod_{\tau=1}^{t-1} (1 - \tilde{r}(\tau \mid x;\theta)) \tag{3.6}$$

Auf der linken Seite steht die durch den Modellansatz $\tilde{r}(t \mid x;\theta)$ implizierte theoretische Wahrscheinlichkeitsfunktion für die Verteilung der Verweildauern in den Teilgesamtheiten. Es ist eine bedingte Wahrscheinlichkeitsfunktion, da von der Verteilung der Individuen auf die einzelnen Teilgesamtheiten abstrahiert wird. Für die Modellschätzung ist dies jedoch ausreichend, denn wir können einfach annehmen, daß

$$f(t,x) \;\approx\; \tilde{f}(t \mid x;\theta)\,\mathrm{P}\,(X = x) \tag{3.7}$$

Oder anders gesagt: es ist für die Modellschätzung (zunächst) nicht erforderlich, auch für die Verteilung der Kovariablen ein theoretisches Modell anzunehmen. Geht man nämlich für die Formulierung des KL-Distanzmaßes von (3.7) aus, erhält man

$$D(\theta) \;=\; \sum_{t\in\mathcal{T}} \sum_{x\in\mathcal{X}} \mathrm{P}\,(T = t, X = x)\, \log\left\{ \frac{\mathrm{P}\,(T = t, X = x)}{\tilde{f}(t \mid x;\theta)\,\mathrm{P}(X = x)} \right\}$$

Dies ist jedoch identisch mit

$$D(\theta) \;=\; \sum_{t\in\mathcal{T}} \sum_{x\in\mathcal{X}} f(t,x)\, \log\left\{ \frac{f(t \mid x)}{\tilde{f}(t \mid x;\theta)} \right\}$$

Um ein optimales Modell zu finden, muß dieser Ausdruck als eine Funktion des Parametervektors θ minimiert werden. D.h. es ist ein Parametervektor $\hat{\theta}$ zu suchen, so daß $\tilde{f}(t \mid x;\hat{\theta})$ eine im Rahmen der vorgegebenen Modellklasse optimale Approximation an die empirische *bedingte* Verteilung $f(t \mid x)$ liefert.

Wie bereits gezeigt worden ist, kann dies auch durch die Maximierung der entsprechenden Log-Likelihood erreicht werden, also durch Maximierung des Ausdrucks

$$\ell(\theta) \;=\; \sum_{t\in\mathcal{T}} \sum_{x\in\mathcal{X}} f(t,x)\, \log\left\{ \tilde{f}(t \mid x;\theta) \right\}$$

Bezeichnet man die in der Grundgesamtheit realisierten Werte der Zufallsvariable (T, X) mit (t_i, x_i) $(i = 1, \ldots, N)$, kann man schließlich noch eine äquivalente Umformung in

$$\ell(\theta) \;=\; \sum_{i=1}^{N} \log\left\{ \tilde{f}(t_i \mid x_i;\theta) \right\} \tag{3.8}$$

vornehmen. Die Maximierung dieser Log-Likelihood ergibt eine Schätzung $\hat{\theta}$ für den Parametervektor, also auch eine Schätzung des gesuchten Modells $\tilde{f}(t \mid x;\hat{\theta})$.

Ein einfaches Beispiel ist das in (3.4) angegebene logistische Regressionsmodell. Anwendung von (3.6) liefert die bedingte Wahrscheinlichkeitsfunktion

$$\tilde{f}(t \mid x;\theta) \;=\; \frac{\exp(X_1\beta_1 + \cdots + X_m\beta_m)}{(1 + \exp(X_1\beta_1 + \cdots + X_m\beta_m))^t}$$

wobei $\theta = (\beta_1, \ldots, \beta_m)$. Durch Einsetzen in (3.8) erhält man die Log-Likelihood für die Modellschätzung. Bezeichnet man die realisierten Werte für die in (3.3) definierten 0/1-Variablen mit $x_{i1}, \ldots, x_{im}$ (für die Individuen $i = 1, \ldots, N$ aus der Grundgesamtheit), erhält man die Formulierung

$$\ell(\theta) \;=\; \sum_{i=1}^{N} \log\left\{ \frac{\exp(x_{i1}\beta_1 + \cdots + x_{im}\beta_m)}{(1 + \exp(x_{i1}\beta_1 + \cdots + x_{im}\beta_m))^{t_i}} \right\}$$

Die Maximierung dieser Log-Likelihood, als eine Funktion des Parametervektors θ, liefert die ML-Schätzung $\hat{\theta}$. Da in diesem Fall die Log-Likelihood global streng konkav ist, existiert ein eindeutiges Maximum.

3.4 Modellschätzung mit unvollständigen Daten

Wir sind bisher davon ausgegangen, daß vollständige Daten verfügbar sind, vgl. (3.5). Bei der Verwendung realer Lebensverlaufsdaten ist diese Bedingung in der Regel nicht erfüllt. Typischerweise kann der interessierende Prozeß nicht bis zu einem Zeitpunkt beobachtet werden, zu dem alle Individuen einen der möglichen Zielzustände erreicht haben; dies ist insbesondere stets dann der Fall, wenn ein „in der Gegenwart“ ablaufender Prozeß beobachtet wird. In vielen Fällen gibt es außerdem für den Beginn eines Prozesses nur unvollständige Informationen.

Zur formalen Darstellung dieses Aspekts unvollständiger Daten ist es zweckmäßig, zwei Zufallsvariablen zu definieren:

$$(T_s, T_f) : \Omega \longrightarrow \mathcal{T} \times \mathcal{T}$$

mit denen Beginn und Ende der Beobachtung erfaßt werden. Wir nehmen an, daß der Lebensverlauf jedes Individuums $\omega \in \Omega$ innerhalb des Zeitraums von $T_s(\omega)$ bis $T_f(\omega)$ beobachtet werden kann. $\mathcal{T}$ ist eine Kalenderzeitachse.

Ob die verfügbaren Daten unvollständig sind, hängt natürlich nicht nur von den Werten der Zufallsvariablen T_s und T_f ab, sondern auch von der Definition des zu beschreibenden Prozesses. Eine vollständige Definition verlangt die Angabe eines Anfangszustands, dessen Eintritt den Beginn des Prozesses fixiert, sowie die Angabe eines oder mehrerer möglicher Endzustände, mit denen der Prozeßverlauf abgeschlossen wird. Darauf bezugnehmend kann dann gesagt werden, ob bzw. in welcher Weise Beobachtungen unvollständig sind. Zwei sich nicht ausschließende Haupttypen unvollständiger Beobachtungen können unterschieden werden:

1. Man sagt, daß die Beobachtung eines Individuums *links zensiert* ist, wenn die Beobachtung seines Lebensverlaufs erst nach dem Zeitpunkt beginnt, zu dem es den Anfangszustand des Prozesses erreicht hat.

2. Entsprechend spricht man von einer *rechts zensierten* Beobachtung, wenn die Beobachtung vor dem Zeitpunkt endet, zu dem das Individuum einen der möglichen Endzustände des Prozesses erreicht hat.

Wenn die Begriffe *links* und / oder *rechts zensierte* Beobachtungen verwendet werden, wird üblicherweise unterstellt, daß gleichwohl Informationen über alle Individuen aus der Grundgesamtheit oder aus einer repräsentativen Stichprobe verfügbar sind, daß der Informationsmangel nur darin besteht, daß bei einigen Individuen der Anfang und / oder das Ende des Prozesses nicht beobachtet werden kann. Ein anderer Aspekt unvollständiger Daten ergibt sich daraus, daß gelegentlich nur Daten über eine Teilgesamtheit vorliegen, deren Auswahl mit dem zu beschreibenden Sachverhalt korreliert ist. Dann liegt ein Selektionsproblem vor, daß dazu führt, daß

mit den verfügbaren Daten nicht ohne weiteres sinnvolle Aussagen über die Grundgesamtheit getroffen werden können. Ein für die statistische Beschreibung wichtiger Spezialfall dieses Selektionsproblems besteht in einer Situation mit sog. *links abgeschnittenen* Beobachtungen. Darauf wird in Abschnitt 3.4.2 etwas näher eingegangen. Zuvor soll jedoch diskutiert werden, wie mit rechts zensierten Daten umgegangen werden kann, denn dies ist der Regelfall bei aus Umfragen gewonnenen Lebensverlaufsdaten. Dabei gehen wir wie bisher von einer einfachen Episode aus, bei der es nur einen möglichen Endzustand gibt. Alle wesentlichen Probleme zeigen sich bereits in dieser einfachen Situation.

3.4.1 Rechts zensierte Beobachtungen

Ausgangspunkt ist, wie in Abschnitt (3.3), eine zweidimensionale Zufallsvariable

$$(T, X) : \Omega \longrightarrow \mathcal{T} \times \mathcal{X}$$

T ist die Verweildauer bis zum Erreichen des Endzustands der Episode, gemessen auf einer diskreten Prozeßzeitachse, X ist eine Variable, mit der die Individuen zum Beginn der Episode in m Teilgesamtheiten klassifiziert werden können.

Wir nehmen jetzt an, daß der Episodenverlauf nur unvollständig beobachtet werden kann, so daß die Beobachtungen für einen Teil der Individuen rechts zensiert sind. Die verfügbaren Daten können dann durch folgende Zufallsvariablen repräsentiert werden:

$$(T', X, \delta) : \Omega \longrightarrow \mathcal{T} \times \mathcal{X} \times \{0, 1\} \tag{3.9}$$

T' ist die beobachtete Verweildauer, δ ist eine Indikatorvariable, die die Bedeutung der beobachteten Verweildauer charakterisiert: wenn $\delta(\omega) = 1$, für ein Individuum $\omega \in \Omega$, wissen wir, daß dieses Individuum zum Zeitpunkt $T'(\omega)$ einen Übergang in den Endzustand der Episode vollzogen hat; wenn $\delta(\omega) = 0$, wissen wir nur, daß bis zum Zeitpunkt $T'(\omega)$ noch kein Übergang in den Endzustand stattgefunden hat.[23] Also

$$T'(\omega) \quad \begin{cases} = T(\omega) & \text{wenn} \quad \delta(\omega) = 1 \\ < T(\omega) & \text{wenn} \quad \delta(\omega) = 0 \end{cases}$$

[23] Wir nehmen an, daß der Zustand jedes Individuums ω mindestens bis zum Zeitpunkt $T'(\omega)$ beobachtet werden kann, und zwar den Zeitpunkt $T'(\omega)$ mit eingeschlossen. Bei einer zensierten Beobachtung wissen wir dann, daß ein Ereignis erst *nach* diesem Zeitpunkt eintreten kann.

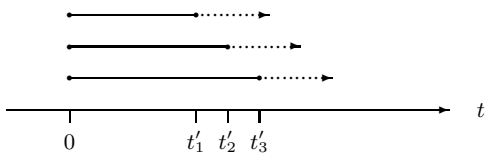
Das Problem besteht darin, aus einer Kenntnis der Verteilung der Zufallsvariablen (T', X, δ) Einsichten in die Verteilung der Zufallsvariablen (T, X) zu gewinnen. Es gibt zwei sich ergänzende, jedoch zu unterscheidende Betrachtungsweisen.

1. Man kann, gestützt auf die verfügbaren Daten, nach einer möglichst plausiblen *Schätzung* für den tatsächlichen, jedoch unbekannten Prozeßverlauf suchen. Es ist klar, daß dies ohne zusätzliche Annahmen nicht erreicht werden kann. Die Frage ist dann, wie solche Annahmen formuliert und gerechtfertigt werden können.

2. Man kann versuchen, gestützt auf die verfügbaren Daten, einen *Bereich möglicher Prozeßverläufe* zu finden, so daß sich der tatsächliche, jedoch unbekannte Prozeßverlauf mit Sicherheit oder mit einer gewissen (subjektiven) Wahrscheinlichkeit innerhalb dieses Bereichs bewegt.

Im folgenden werden diese beiden Betrachtungsweisen des Schätzproblems zunächst anhand einiger unterschiedlicher Situationen mit rechts zensierten Daten diskutiert. Die Frage ist, was jeweils über den Prozeßverlauf, über den nur unvollständige Informationen vorliegen, ausgesagt werden kann. Im Anschluß wird dann die Frage behandelt, wie Übergangsratenmodelle geschätzt werden können, wenn ein Teil der verfügbaren Daten rechts zensiert ist.

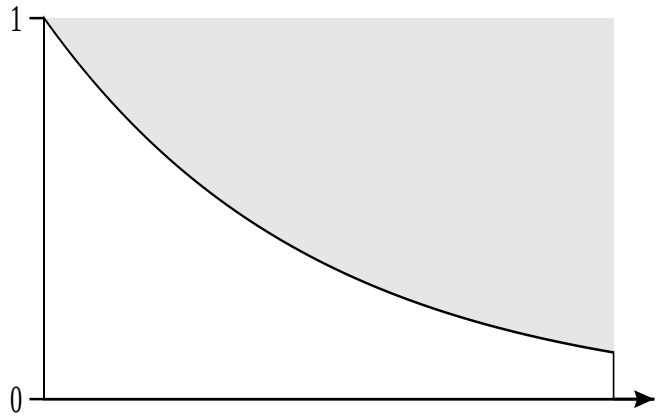
a) Ein Extremfall liegt vor, wenn *alle* Beobachtungen rechts zensiert sind. Folgendes Bild veranschaulicht diese Situation für die Beobachtung von drei Indivdiuen.



Was kann in einer solchen Situation über den Episodenverlauf ausgesagt werden? Etwas mehr als nichts. Wir können zwar nicht den genauen Verlauf der Übergangsraten angeben, jedoch eine untere Grenze für die Survivorfunktion. Denn man kann T, die Zufallsvariable für die tatsächlichen, aber unbekannten Ereigniszeitpunkte als eine Summe $T = T' + T_u$ auffassen. Der erste Summand, T', repräsentiert die beobachteten, aber zensierten Verweildauern, der zweite Summand repräsentiert die jeweils noch verbleibenden Verweildauern. Offenbar gilt $T_u \geq 0$, also

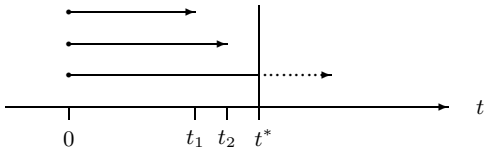
$$G(t) = \mathrm{P}\,(T > t) \geq \mathrm{P}\,(T' > t) = G'(t)$$

$G'(t)$ ist die Survivorfunktion für die beobachtete Zufallsvariable T'. Die tatsächliche Survivorfunktion $G(t)$ verläuft irgendwo oberhalb von $G'(t)$. Folgende Abbildung illustriert diese Situation.



Aus den Daten läßt sich $G'(t)$, eingezeichnet als durchgezogene Linie, berechnen. $G(t)$ verläuft irgendwo im grau eingezeichneten Bereich oberhalb von $G'(t)$. Jede beliebige monoton fallende Funktion, die innerhalb dieses Bereichs verläuft ist mit den Daten verträglich.

b) Eine andere, in gewisser Weise ebenfalls extreme Situation liegt vor, wenn der Prozeßverlauf für alle Individuen bis zu einem festen Zeitpunkt beobachtet werden kann.[24] Folgende Abbildung veranschaulicht diese Situation.

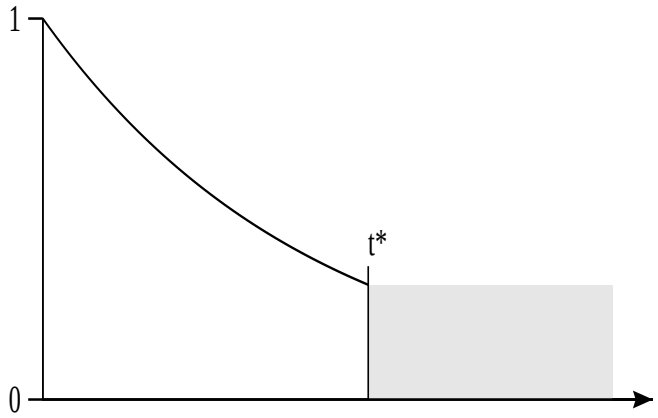


Man hat vollständige Kenntnis über den Prozeßverlauf bis zum Zeitpunkt t^*. Alle Ereigniss, die bis zu diesem Zeitpunkt stattfinden, können beobachtet werden. Jedoch darüber, wie sich der Prozeß nach t^* weiter entwickelt, hat man *keinerlei* Information. Diese Situation ist infolgedessen einfach zu handhaben. Man kann und sollte sich sowohl bei der Beschreibung als auch bei der Modellierung des Prozesses auf den Zeitraum bis t^* beschränken.

Die Beschreibung des Prozesses bis zum Zeitpunkt t^* kann folgendermaßen vorgenommen werden. Es sei $R(t)$ die Risikomenge zum Zeitpunkt t, R_t sei die Anzahl der Individuen in der Risikomenge, und E_t sei die Anzahl der Ereignisse, die zum Zeitpunkt t stattfinden. Also kann man für

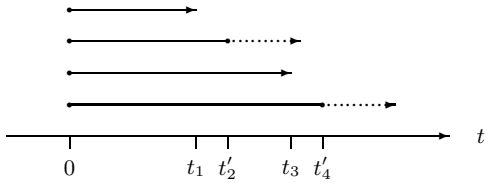
[24] Eine wesentlich kompliziertere Situation tritt ein, wenn nur über diejenigen Individuen eine Information verfügbar ist, die bis zum Ende des Beobachtungszeitraum ein Ereignis hatten. Diese Situation wird hier nicht näher behandelt, einige Überlegungen dazu finden sich bei Kalbfleisch und Lawless [1988].

die vorausgesetzte diskrete Zeitachse die Übergangswahrscheinlichkeiten $h(t) = E_t/R_t$ und schließlich auch die Survivorfunktion berechnen.[25] Im Hinblick auf die Survivorfunktion kann die Situation durch folgene Abbildung veranschaulicht werden:



Bis zum Zeitpunkt t^* kennt man den tatsächlichen Verlauf der Survivorfunktion. Danach kann die Survivorfunktion einen beliebigen (monoton fallenden) Verlauf innerhalb der grau eingezeichneten Fläche haben.

c) Eine andere Situation liegt vor, wenn wir annehmen, daß wir den Prozeß zwar beliebig lange beobachten können, daß jedoch – aus irgendwelchen Gründen – die Beobachtbarkeit einiger Indivdiuen während des Prozeßverlaufs aufhört. Folgende Abbildung veranschaulicht diese Situation.



Die erste Frage, die sich in dieser Situation stellt, ist: Bis zu welchem Zeitpunkt kann der Prozeß beschrieben werden? Die Antwort muß sich offensichtlich auf das Zeitintervall $[t_a, t_b]$ beziehen, wobei t_a der Zeitpunkt ist, bei dem zum erstenmal die Beobachtung eines Individuums aufhört (d.h. t_a ist der kleinste Zensierungszeitpunkt), und t_b der Zeitpunkt, bei dem zum letztenmal ein Ereignis beobachtet werden kann (d.h. t_b ist der

[25] Bei vollständig beobachteten Verläufen ist die Risikomenge proportional zur Survivorfunktion.

größte Ereigniszeitpunkt). Bis zum Zeitpunkt t_a kann der Prozeßverlauf vollständig beobachtet werden; ab dem Zeitpunkt t_b hat man keinerlei Informationen über den weiteren Prozeßverlauf.

Welcher Zeitpunkt innerhalb dieses Intervalls gewählt werden sollte, hängt von den Daten ab, vom Ausmaß der zensierten Beobachtungen und von der Lage der Zensierungs- und Ereigniszeitpunkte. Als Fragestellung sollte zunächst davon ausgegangen werden, wie der Prozeßverlauf bis zum Zeitpunkt t_b beschrieben werden kann.

Es erscheint sinnvoll, mit den tatsächlich verfügbaren Beobachtungen zu beginnen. Zu jedem Zeitpunkt gibt es eine *beobachtete* Risikomenge $R'(t)$, die aus R'_t Individuen besteht, und eine *beobachtete* Anzahl von Ereignissen E'_t. Dies liefert die beobachteten Übergangswahrscheinlichkeiten $h'(t) = E'_t/R'_t$, woraus dann eine Survivorfunktion berechnet werden kann:

$$G'(t) = \prod_{\tau=1}^{t} (1 - h'(\tau))$$

Auf dieser Vorgehensweise beruht das üblicherweise verwendete Kaplan-Meier-Schätzverfahren.[26] Zu überlegen ist, in welcher Weise man dadurch eine Schätzung der tatsächlichen Survivorfunktion $G(t)$ erhält. Das Problem kann folgendermaßen sichtbar gemacht werden. Die tatsächliche Survivorfunktion resultiert aus den Übergangswahrscheinlichkeiten

$$h(t) = \frac{E_t}{R_t} = \frac{E'_t + E''_t}{R'_t + R''_t}$$

Wir beobachten jedoch nur einen Teil der tatsächlichen Ereignisse E_t und der tatsächlichen Risikomenge R_t. E''_t ist die Anzahl der nicht-beobachteten Ereignisse zum Zeitpunkt t, und R''_t ist die Anzahl der Individuen in der Risikomenge zum Zeitpunkt t, die nicht beobachtet werden. Um $G'(t)$ als eine Schätzung für $G(t)$ ansehen zu können, muß in irgendeiner Weise eine Annahme über diese nicht-beobachteten Sachverhalte gemacht werden. Die einfachste Annahme ist natürlich, daß sich das Nicht-beobachtete so ähnlich verhält wie das Beobachtete, d.h. in diesem Fall: daß die Übergangswahrscheinlichkeiten bei den nicht (mehr) beobachteten Individuen näherungsweise gleich sind zu den Übergangswahrscheinlichkeiten bei den beobachteten Individuen. Es wird also angenommen, daß

$$\frac{E'_t}{R'_t} \approx \frac{E''_t}{R''_t} \tag{3.10}$$

woraus dann folgen würde, daß $G'(t)$ näherungsweise gleich $G(t)$ ist.

[26] Eine detaillierte Diskussion findet sich bei Cox und Oakes [1984, Kap. 4], Einführungen geben u.a. Diekmann und Mitter [1984, S. 76ff], Blossfeld et al. [1989, S. 44f], Andreß [1992, S. 147ff].

Wenn die Annahme (3.10) erfüllt ist, führt das Kaplan-Meier-Verfahren zu sinnvollen, näherungsweise gültigen Schätzungen der Survivorfunktion $G(t)$. Das Problem ist, daß diese Annahme nicht erfüllt sein kann und daß es in der Regel nicht möglich ist, diese Annahme auf der Grundlage der jeweils verfügbaren Daten zu prüfen. Man könnte sagen: Wenn es keine besonderen Gründe gibt, an der Annahme (3.10) zu zweifeln (wenn der Zensierungsmechanismus „nicht informativ“ ist), gibt es auch keine Gründe, daran zu zweifeln, daß das Kaplan-Meier-Verfahren zu einer näherungsweise zutreffenden Schätzung der tatsächlichen Survivorfunktion führt.

Immer dann, wenn eine Überlegung auf Annahmen beruht, die nicht oder nur begrenzt gerechtfertigt werden können, sollte jedoch auch die Frage gestellt werden, welche Konsequenzen es hätte, wenn die Annahmen unzutreffend wären. Diese Frage führt zur zweiten der oben unterschiedenen Betrachtungsweisen. Gesucht ist dann ein Bereich möglicher Prozeßverläufe, so daß sich der tatsächliche Prozeßverlauf mit Sicherheit oder mit einer gewissen subjektiven Wahrscheinlichkeit innerhalb dieses Bereichs bewegt.

Im Hinblick auf die Survivorfunktion $G(t)$ kann dieses Problem auf einfache Weise gelöst werden. Zunächst findet man sichere Unter- und Obergrenzen auf folgende Weise. Eine sichere Untergrenze ergibt sich daraus, daß bei allen zensierten Beobachtungen angenommen wird, daß ein Ereignis dem Zensierungszeitpunkt unmittelbar folgt.[27] Ein sichere Obergrenze entsteht entsprechend durch die Annahme, daß bei allen zensierten Beobachtungen ein Ereignis frühestens nach dem letzten beobachteten Ereignis eintritt.

Abbildung 3.4.1 illustriert diese Unter- und Obergrenzen für die Schätzung einer Survivorfunktion für das Alter bei der ersten Heirat.[28] Die tatsächliche Survivorfunktion $G(t)$ bewegt sich irgendwo innerhalb des grau eingezeichneten Bereichs; jede monoton fallende Funktion innerhalb dieses Bereichs ist mit den verfügbaren Daten vereinbar.[29] Die Breite des Bereichs möglicher Prozeßverläufe hängt sowohl von der Zeitstruktur als auch von der Anzahl zensierter Beobachtungen ab; bei jedem Zensierungszeitpunkt wird das Intervall breiter. In diesem Beispiel gibt es insgesamt 6416 Individuen, darunter 1171 (etwa 18 %) zensierte Beobachtungen.

Das Kaplan-Meier-Verfahren beruht auf der Annahme, daß der Prozeßverlauf bei den nicht (mehr) beobachteten Individuen im wesentlichen

[27] D.h. die Untergrenze ist proportional zur beobachteten Risikomenge.

[28] Die Berechnung wurde mit Daten aus dem SOEP vorgenommen. Verwendet wurden Informationen über alle Stammpersonen aus der Teilstichprobe A, die im Zeitraum 1918 – 1868 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben, vgl. Abschnitt 2.2.

[29] Wie zu Beginn dieses Kapitels ausgeführt wurde, abstrahieren wir hier davon, daß es sich bei den verwendeten Daten um eine Stichprobe handelt. Außerdem wird von der Problematik der Meßfehler abgesehen.

identisch ist mit dem Prozeßverlauf bei den beobachteten Individuen. Die Konstruktion sicherer Unter- und Obergrenzen beruht auf der Annahme, daß der Prozeß bei den nicht (mehr) beobachteten Individuen vollständig anders verlaufen könnte. Dazwischen liegt ein breites Spektrum anderer Möglichkeiten.

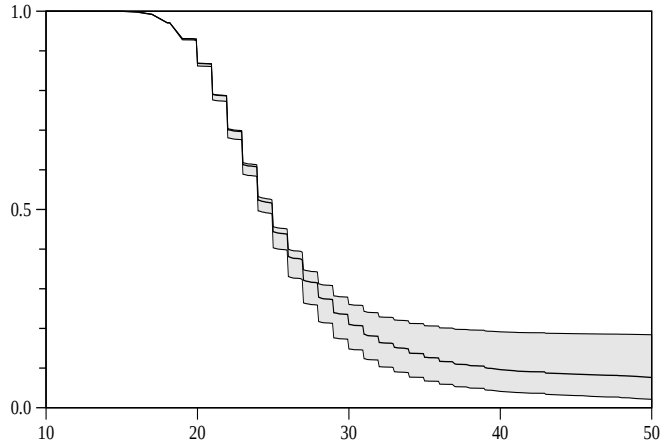


Abbildung 3.4.1 Kaplan-Meier-Schätzung der Survivorfunktion für das Alter bei der ersten Heirat sowie sichere Unter- und Obergrenzen. Teilstichprobe A des SOEP. Abszisse in Jahren.

Um sich diese Möglichkeiten zu veranschaulichen, kann man folgendermaßen vorgehen. Es sei $Z(t)$ die Menge der Individuen, deren Beobachtung vor dem Zeitpunkt t zensiert ist. Für jedes Individuum $i \in Z(t)$ gibt es also einen Zensierungszeitpunkt $t'_i < t$. Die Individuen aus $Z(t)$ sind in der beobachteten Risikomenge $R'(t)$ nicht enthalten, möglicherweise jedoch in der tatsächlichen Risikomenge $R(t) = R'(t) \cup R^*(t)$. Bei der Berechnung der Übergangswahrscheinlichkeiten beim Kaplan-Meier-Verfahren wird von diesen Individuen abstrahiert; implizit wird angenommen, daß bei diesen Individuen der unbeobachtete Prozeßverlauf während der Zeitpunkte $t'_i+1, \ldots, t$ durch die für diesen Zeitraum geschätzten Übergangswahrscheinlichkeiten $\hat{h}(t)$ angemessen beschrieben werden kann. Diese Annahme kann richtig sein oder falsch. Nehmen wir an, daß der unbeobachtete Prozeßverlauf tatsächlich den Übergangswahrscheinlichkeiten $h^*(t)$ folgt, die von $\hat{h}(t)$ verschieden sein können. Dann kann berechnet werden, welcher Anteil der Individuen aus $Z(t)$ in der tatsächlichen Risikomenge $R(t)$ enthalten ist, nämlich

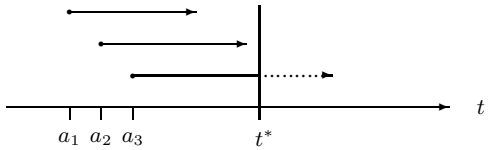
$$R_t^* = \sum_{i \in Z(t)} \prod_{\tau=t'_i+1}^{t-1} (1 - h^*(\tau))$$

Außerdem kann der Anteil der Individuen aus $R^*(t)$ berechnet werden, bei dem zum Zeitpunkt t ein Ereignis stattfindet, nämlich $h^*(t)\,R_t^*$. Daraus ergibt sich als neue Schätzung für die Übergangswahrscheinlichkeit zum Zeitpunkt t:

$$\bar{h}(t) = \frac{E_t' + h^*(t)\,R_t^*}{R_t' + R_t^*} \tag{3.11}$$

Wird angenommen, daß $h^*(t) = \hat{h}(t)$, ist auch $\bar{h}(t) = \hat{h}(t)$, und die resultierende Schätzung der Survivorfunktion ist mit ihrer Kaplan-Meier-Schätzung identisch. Formel (3.11) bietet jedoch die Möglichkeit, Abweichungen von dieser Annahme zu reflektieren. Zum Beispiel kann man annehmen, daß $h^*(t) = \lambda\,\hat{h}(t)$ ist, wobei λ in einem gewissen Intervall variieren kann.

d) Meistens werden Lebensverlaufsdaten aus einmaligen Retrospektiverhebungen gewonnen, ggf. verknüpft mit Folgeerhebungen in der Form eines Panels. Ein wesentlicher Vorteil besteht darin, daß es bei einem solchen Erhebungsdesign möglich ist, links zensierte Beobachtungen zu vermeiden. Man kann bis zum Erhebungszeitpunkt vollständige Informationen über alle Personen gewinnen, die während eines gewissen vergangenen Zeitraums in den Anfangszustand des zu untersuchenden Prozesses eingetreten sind (bei den üblichen Befragungen natürlich nur für Personen, die zum Erhebungszeitpunkt noch leben). Folgende Abbildung illustriert diese Situation, wobei von einer Kalenderzeitachse ausgegangen wird.



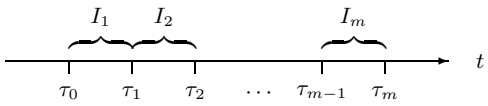
a_i ist der Zeitpunkt, zu dem das i.te Individuum in den Anfangszustand des zu beschreibenden Prozesses gerät, t^* ist der Zeitpunkt, zu dem die Retrospektiverhebung stattfindet. Man erhält für jedes Individuum eine vollständige Prozeßbeschreibung für den Zeitraum von a_i bis t^*.

Allerdings hängt die Dauer der Beobachtung davon ab, wann – auf der zugrundeliegenden Kalenderzeitachse – der Übergang in den Anfangszustand des Prozesses stattgefunden hat. Je später dieser Übergang erfolgt, desto kürzer ist die Beobachtungsdauer und desto größer ist infolgedessen die Möglichkeit, daß die Beobachtung rechts zensiert ist.

Wenn der Prozeß unabhängig davon verläuft, zu welchem Kalenderzeitpunkt er begonnen hat, kann diese Korrelation zwischen Anfangszeitpunkt und Zensierungswahrscheinlichkeit ignoriert werden. Bei soziologischen Untersuchungen von Lebensverläufen ist diese Unabhängigkeitsannahme jedoch fast immer falsch. Es muß vielmehr in der Regel davon ausgegangen werden, daß Lebensverläufe auch davon abhängen, zu welchem

historischen Zeitpunkt sie begonnen haben.

Eine geeignete Methode, um diesem Sachverhalt Rechnung zu tragen, besteht in der Bildung von Kohorten. Die Kalenderzeitachse wird in Intervalle eingeteilt, etwa in folgender Form:



so daß alle Anfangszeitpunkte a_i in einem der Intervalle $I_1, \ldots, I_m$ enthalten sind. Die j.te Kohorte besteht aus allen Individuen, die den Prozeß im Intervall I_j begonnen haben. Für alle Individuen, die der gleichen Kohorte angehören, ergibt sich eine (abhängig von der Intervallbreite) ungefähr gleiche Beobachtungsdauer. Beschreibt man den Prozeßverlauf gesondert für jede Kohorte, ergibt sich also im Hinblick auf das Zensierungsproblem im wesentlichen eine Situation, wie sie oben in Abschnitt (b) diskutiert worden ist.

Zur Illustration betrachten wir wiederum den Übergang in die erste Heirat bei den Stammpersonen aus der Teilstichprobe A des SOEP. Zum Beispiel können folgende Geburtskohorten unterschieden werden.

Kohorte	Personen	zensiert	%
$1918-27$	908	46	5.1
$1928-37$	1209	54	4.5
$1938-47$	1400	81	5.8
$1948-57$	1415	185	13.1
$1958-68$	1484	805	54.2
Insgesamt	6416	1171	18.3

(3.12)

Man erkennt deutlich, wie das Ausmaß der rechts zensierten Beobachtungen mit dem Geburtsjahr zusammenhängt. Nicht nur, aber auch aus diesem Grund ist es sinnvoll, kohortenspezifischen Survivorfunktionen zu schätzen und darauf zu achten, daß die Zeitspannen, für die diese Schätzungen sinnvoll interpretiert werden können, von der jeweiligen Kohorte abhängen. Abbildung 3.4.2 illustriert den Sachverhalt für die in Tabelle (3.12) unterschiedenen Geburtskohorten.[30]

[30] Man erkennt deutlich einige Unterschiede: die älteste Geburtskohorte nimmt eine mittlere Stellung ein, in den dann folgenden Kohorten wurde zunächst in zunehmend jüngerem Alter geheiratet, schließlich wird in der jüngsten Kohorte wieder deutlich später geheiratet.

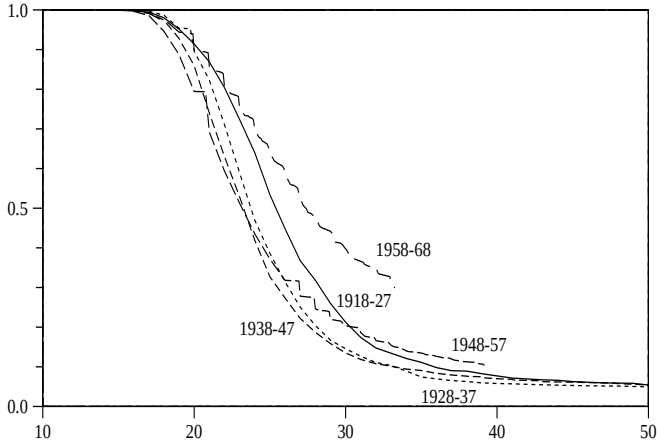


Abbildung 3.4.2 Kaplan-Meier-Schätzungen der Survivorfunktionen für das Alter bei der ersten Heirat, differenziert nach den in (3.12) definierten Geburtskohorten. Teilstichprobe A des SOEP. Abszisse in Jahren.

Eine Datenanalyse, die auf eine Kohortendifferenzierung verzichtet, kann tatsächlich leicht zu irreführenden Ergebnissen führen. Abbildung 3.4.3 zeigt zum Beispiel (oben links) einen Vergleich der Verteilungen für das Alter bei der ersten Heirat in den beiden Teilstichproben A (überwiegend Deutsche) und B (Ausländer). Es scheint so, daß die in Deutschland lebenden Ausländer durchweg früher geheiratet haben. Daß diese Feststellung irreführend ist, zeigt sich jedoch, sobald man Geburtskohorten unterscheidet. Es wird dann deutlich, daß der Unterschied im wesentlichen nur eine Folge dessen ist, daß die Mitglieder der jüngeren deutschen Geburtskohorten angefangen haben, später zu heiraten.[31]

e) Häufig werden Retrospektiverhebungen mit einer oder mehreren Folgeerhebungen verbunden. Ein typisches Beispiel ist das Sozio-ökonomische Panel (SOEP). Die erste Welle wurde 1984 durchgeführt und lieferte für eine Basisstichprobe von Individuen und Haushalten sowohl Informationen über ihren gegenwärtigen Zustand (zum Befragungszeitpunkt) als auch eine Fülle von Retrospektivinformationen über den jeweils vergangenen Lebensverlauf. Die gleichen Individuen bzw. Haushalte wurden dann in jedem der folgenden Jahre erneut befragt, wodurch zusätzliche Informationen über die Lebensverläufe seit dem ersten Befragungszeitpunkt gewonnen werden konnten. Gegenwärtig (Mitte 1994) liegen die Ergebnisse

[31] Ein aus soziologischer Sicht besonders wichtiger Grund kann nicht nur im bei den jüngeren Geburtskohorten generell gestiegenen Bildungsniveau gesehen werden, sondern vor allem in den damit verbundenen längeren Ausbildungszeiten, während der in der Regel noch nicht geheiratet wird. Vgl. zum Beispiel Blossfeld und Jaenichen [1990] sowie mehrere Aufsätze in Diekmann und Weick [1993].

aus neun Erhebungswellen (1984 – 1992) vor.

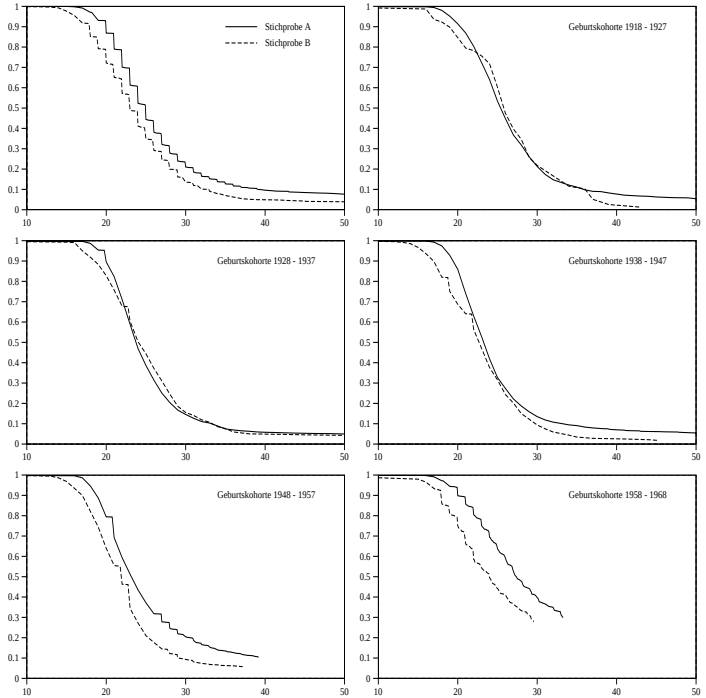
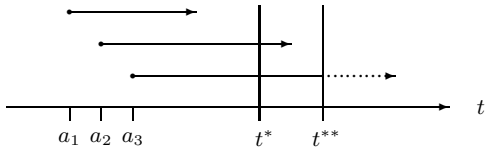


Abbildung 3.4.3 Vergleich der Survivorfunktionen für das Alter bei der ersten Heirat in den Teilstichproben A und B des SOEP; für jeweils alle Personen (oben links) sowie differenziert nach Geburtskohorten. Kaplan-Meier-Schätzungen. Abszisse in Jahren.

Im Hinblick auf das Problem rechts zensierter Beobachtungen bei Retrospektiverhebungen, die mit einer oder mehreren Folgeerhebungen verbunden werden, ergeben sich zunächst die gleichen Probleme, wie oben in Abschnitt (d) behandelt. Es kommt jedoch ein zusätzliches Problem hinzu. Folgende Abbildung veranschaulicht die Situation für eine Folgeerhebung.[32]

[32] Ich gehe davon aus, daß die relevante Grundgesamtheit diejenige ist, aus der die Basisstichprobe in der ersten Erhebungswelle gezogen wird.



Verwendet wird wieder eine Kalenderzeitachse. Die erste Retrospektiverhebung findet zum Zeitpunkt t^*, die Folgeerhebung findet zum Zeitpunkt t^{**} statt.

Im Unterschied zu einmaligen Retrospektiverhebungen kommt jetzt das Problem hinzu, daß in der Regel nicht alle Ereignisse, die zwischen den beiden Erhebungszeitpunkten stattfinden, beobachtet werden können. Ereignisse können nur dann festgestellt werden, wenn die Individuen auch noch in der Folgeerhebung befragt werden können; wenn sie zum Zeitpunkt der Folgeerhebung bereits gestorben sind oder wenn es in der Folgeerhebung nicht mehr gelingt, sie zu befragen, können zwischenzeitlich stattgefundene Ereignisse nicht ermittelt werden, liegt also eine zum Zeitpunkt t^* rechts zensierte Beobachtung vor.[33]

Für die Zensierungsproblematik ist vor allem wichtig, ob das zu untersuchende Ereignis die Wahrscheinlichkeit für das Auftreten einer zensierten Beobachtung beeinflußt. Damit muß in vielen Fällen gerechnet werden. Auch in unserem Beispiel, dem Übergang in die erste Heirat, ist ein solcher Effekt vorstellbar. Man kann vermuten, daß eine Heirat die Wahrscheinlichkeit erhöht, daß in der folgenen Erhebungswelle keine Teilnahme mehr erfolgt. Das würde bedeuten, daß die Ereigniswahrscheinlichkeit in der jeweils unbeobachteten Risikomenge größer ist als in der beobachteten Risikomenge. Eine formale Bedingung, unter der gleichwohl näherungsweise verläßliche Schätzungen erzielt werden können, wird im Kontext der jetzt zu besprechenden Modellschätzung behandelt.

Modellschätzung mit rechts zensierten Daten

Wie in den beiden ersten Abschnitten dieses Kapitels besprochen worden ist, sehen wir den primären Sinn der statistischen Modellbildung darin, für einen empirisch erfaßbaren Prozeß eine vereinfachende Beschreibung zu erreichen. Wie die Modellschätzung vorgenommen werden kann, wenn alle erforderlichen Daten verfügbar sind, wurde in Abschnitt 3.3 bereits für einfache Übergangsratenmodelle behandelt. Der Grundgedanke ist, aus einer vorgegebenen Klasse möglicher Modelle dasjenige auszuwählen, das im Sinne des KL-Distanzmaßes am besten zur gegebenen empirischen Verteilung der Daten paßt. Wenn einige der für die Modellschätzung zu verwendenden

[33] Ein Teil dieses Selektionsproblems stellt sich natürlich bereits bei einmaligen Retrospektiverhebungen, denn in der Regel sind nicht alle für eine Stichprobe ausgewählten Individuen bereit, an der Befragung teilzunehmen.

Beobachtungen rechts zensiert sind, tritt eine etwas andere Situation ein: ein Teil der für die Modellschätzung eigentlich erforderlichen Information fehlt. Es muß dann überlegt werden, ob und ggf. wie ein sinnvoller Vergleich zwischen den Beobachtungen und der Klasse möglicher theoretischer Modelle gleichwohl vorgenommen werden kann.

Das theoretische Modell soll sich auf die Verteilung der Zufallsvariablen

$$(T, X) : \Omega \longrightarrow \mathcal{T} \times \mathcal{X}$$

beziehen, genauer gesagt: auf die durch X bedingte Verteilung von T. Der Modellansatz ist also

$$\mathrm{P}\,(T = t \mid X = x) \approx \tilde{f}(t \mid x; \theta) \qquad \theta \in \Theta \tag{3.13}$$

Die Verteilung von (T, X) kann jedoch nicht unmittelbar beobachtet werden. Die verfügbaren Beobachtungen liefern vielmehr empirische Wahrscheinlichkeiten für die Verteilung der Zufallsvariablen (T', X, δ), vgl. (3.9). Die Frage ist, wie man mithilfe dieser empirischen Wahrscheinlichkeiten Einsichten in die Verteilung von (T, X) gewinnen kann.

Betrachten wir zunächst die nicht zensierten Beobachtungen. Der Zusammenhang mit der Verteilung von (T, X) kann folgendermaßen sichtbar gemacht werden:[34]

$$\begin{aligned} &\mathrm{P}\,(T' = t, \delta = 1 \mid X = x) = \\ &\quad \mathrm{P}\,(T' = t, \delta = 1, T = t \mid X = x) = \\ &\quad \mathrm{P}\,(T' = t, \delta = 1 \mid T = t, X = x)\,\mathrm{P}\,(T = t \mid X = x) \end{aligned} \tag{3.14}$$

Die Wahrscheinlichkeit $\mathrm{P}(T = t \,|\, X = x)$ (bzw. das für diese Wahrscheinlichkeit zu bildende Modell) kann also nicht unmittelbar mit der empirischen Wahrscheinlichkeit $\mathrm{P}(T' = t, \delta = 1 \,|\, X = x)$ verglichen werden. Dazwischen steht die Wahrscheinlichkeit für eine nicht zensierte Beobachtung unter der Bedingung, daß ein Ereignis stattgefunden hat.

Die Zerlegung in (3.14) liefert jedoch einen Hinweis darauf, unter welchen Bedingungen ein Vergleich dennoch sinnvoll vorgenommen werden kann. Eine hinreichende Bedingung kann folgendermaßen formuliert werden:

$$\mathrm{P}\,(T' = t, \delta = 1 \mid T = t, X = x) \;\propto\; \mathrm{P}\,(T' = t, \delta = 1 \mid X = x) \tag{3.15}$$

Auf der rechten Seite steht die Wahrscheinlichkeit für eine nicht zensierte Beobachtung in der Teilgesamtheit $\Omega(x)$, d.i. die Menge der Individuen, die zur Klasse x gehören. Auf der linken Seite steht die Wahrscheinlichkeit für eine nicht zensierte Beobachtung in der Teilgesamtheit $\Omega(t, x)$, d.i. die Menge der Individuen, die zur Klasse x gehören und zum Zeitpunkt

[34] Da sich die Modellbildung auf die durch X bedingte Verteilung von T beziehen soll, werden im folgenden stets nur die konditionalen Wahrscheinlichkeiten betrachtet.

t ein Ereignis haben. Beide Wahrscheinlichkeiten sollen – unabhängig von t – proportional zueinander sein, d.h. die Zeitpunkte, zu denen eine nicht zensierte Beobachtung erreicht werden kann, sollen nicht von den Ereigniszeitpunkten abhängen.

Wenn die in (3.15) formulierte Annahme (näherungsweise) erfüllt ist, kann die Wahrscheinlichkeit $\mathrm{P}(T' = t, \delta = 1 \,|\, T = t, X = x)$ durch eine (unbekannte) Funktion $\rho_1(x)$ repräsentiert werden, und der in (3.14) formulierte Zusammenhang zwischen den empirischen Wahrscheinlichkeiten für nicht zensierte Beobachtungen und der Verteilung von T kann folgendermaßen ausgedrückt werden:

$$\mathrm{P}\left(T' = t, \delta = 1 \mid X = x\right) \approx \rho_1(x)\, \mathrm{P}\left(T = t \mid X = x\right) \tag{3.16}$$

Wenn dieser Zwischenschritt gerechtfertig werden kann, kann schließlich auch ein Vergleich der empirischen Wahrscheinlichkeiten mit den potentiellen Modellen vorgenommen werden. Unter Verwendung von (3.13) erhält man

$$\mathrm{P}\left(T' = t, \delta = 1 \mid X = x\right) \approx \rho_1(x)\, \tilde{f}(t \mid x; \theta) \tag{3.17}$$

Ganz analog kann man sich den Zusammenhang zwischen den zensierten Beobachtungen und der Verteilung von T überlegen. Jetzt benötigt man die korrespondierende Annahme, daß die zensierten Beobachtungen von den Ereigniszeitpunkten unabhängig sind, also

$$\mathrm{P}\left(T' = t, \delta = 0 \mid T > t, X = x\right) \propto \mathrm{P}\left(T' = t, \delta = 0 \mid X = x\right) \tag{3.18}$$

Dann kann man die Wahrscheinlichkeit $\mathrm{P}(T' = t, \delta = 0 \,|\, T > t, X = x)$ durch eine unbekannte, nur von x abhängige Funktion $\rho_2(x)$ repräsentieren und den Zusammenhang zwischen den empirischen Wahrscheinlichkeiten für zensierte Beobachtungen mit der Verteilung von T folgendermaßen herstellen:

$$\begin{aligned}
&\mathrm{P}\left(T' = t, \delta = 0 \mid X = x\right) = \\
&\quad \mathrm{P}\left(T' = t, \delta = 0, T > t \mid X = x\right) = \\
&\quad \mathrm{P}\left(T' = t, \delta = 0 \mid T > t, X = x\right) \mathrm{P}\left(T > t \mid X = x\right) = \\
&\quad \rho_2(x)\, \mathrm{P}\left(T > t \mid X = x\right)
\end{aligned}$$

Diese Darstellung zeigt, wie zensierte Beobachtungen mit einem Modell für die Verteilung von T verglichen werden können. Nämlich, analog zu (3.17), durch

$$\mathrm{P}\left(T' = t, \delta = 0 \mid X = x\right) \approx \rho_2(x)\, \tilde{G}(t \mid x; \theta) \tag{3.19}$$

also durch die durch das Modell vermutete Wahrscheinlichkeit, daß bis zum Ende der Beobachtung kein Ereignis eingetreten ist.

Zusammengenommen zeigen (3.17) und (3.19), wie die empirisch ermittelbaren Wahrscheinlichkeiten mit den potentiellen Modellen für die Verweildauer T verglichen werden können. Um den Vergleich praktisch auszuführen, kann das KL-Distanzmaß verwendet werden. Ein optimales Modell (aus der vorausgesetzten Modellklasse) erhält man dann durch die Minimierung folgender Distanz:

$$\begin{aligned}
\sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}} \mathrm{P}\left(T' = t, \delta = 1, X = x\right) \log \left\{ \frac{\mathrm{P}\left(T' = t, \delta = 1 \mid X = x\right)}{\rho_1(x)\, \tilde{f}(t \mid x; \theta)} \right\} + \\
\mathrm{P}\left(T' = t, \delta = 0, X = x\right) \log \left\{ \frac{\mathrm{P}\left(T' = t, \delta = 0 \mid X = x\right)}{\rho_2(x)\, \tilde{G}(t \mid x; \theta)} \right\}
\end{aligned} \tag{3.20}$$

Wichtig ist, daß man, um einen optimalen Parametervektor $\hat{\theta}$ zu finden, die Funktionen $\rho_1(x)$ und $\rho_2(x)$ nicht zu kennen braucht, da sie von θ unabhängig sind. Man sieht das, wenn man die zu (3.20) korrespondierende Log-Likelihood betrachtet. Mit den Bezeichnungen $t'_i = T'(\omega_i)$, $\delta_i = \delta(\omega_i)$ und $x_i = X(\omega_i)$ kann sie folgendermaßen geschrieben werden:

$$\ell(\theta) = \sum_{i=1}^{N} \delta_i \log \left\{ \tilde{f}(t'_i \mid x_i; \theta) \right\} + (1 - \delta_i) \log \left\{ \tilde{G}(t'_i \mid x_i; \theta) \right\} \tag{3.21}$$

Die Maximierung dieser Log-Likelihood ist äquivalent zur Minimierung der in (3.20) angegebenen Distanz und liefert die gleiche Parameterschätzung $\hat{\theta}$.

Es ist bemerkenswert, daß die in (3.15) und (3.18) angegebenen Bedingungen nur mögliche Korrelationen der Zensierungszeitpunkte mit den Ereigniszeitpunkten betreffen, ein unterschiedliches Ausmaß zensierter Beobachtungen in den durch die Kovariable X definierten Klassen spielt für die Modellschätzung keine Rolle.[35] Wie in Abschnitt 3.4.1.d dargestellt worden ist, tritt diese Situation insbesondere dann auf, wenn eine Klassifizierung in Kohorten vorgenommen wird. Obwohl die Tatsache, daß das Ausmaß zensierter Beobachtungen in den Kohorten typischerweise stark variiert, für die Modellschätzung keine Rolle spielt, sollte jedoch beachtet werden, daß daraus gewisse Grenzen für die empirische Interpretierbarkeit des geschätzten Modells resultieren. Ein deskriptiver Anspruch kann nur für diejenigen Bereiche der Zeitachse formuliert werden, für die hinreichend viele Ereignisse beobachtet werden können.

Um den Sachverhalt zu illustrieren, werden simulierte Daten verwendet; Beispiele auf der Grundlage der SOEP-Daten für das Alter bei der ersten Heirat werden in Abschnitt 3.5.4 gegeben.

[35] Dabei wird natürlich vorausgesetzt, daß es sich um Variablen handelt, deren Werte zum Episodenbeginn feststehen.

Beispiele mit simulierten Daten

Zur Illustration des Problems der Modellschätzung mit rechts zensierten Beobachtungen soll eine mit Zufallszahlen simulierte Grundgesamtheit betrachtet werden, die aus zwei Gruppen (jeweils 1000 Individuen) besteht.[36] Zwischen den Gruppen wird mit der Variable X (0 in Gruppe A, 1 in Gruppe B) unterschieden. Die Verweildauer soll in beiden Gruppen einem logistischen Regressionsmodell folgen (vgl. Abschnitt 3.3), jedoch mit einer unterschiedlichen Übergangsrate. Für die Gruppe A ($X = 0$) wird eine Rate $r_A(t) = 0.1$, für die Gruppe B ($X = 1$) wird eine Rate $r_B(t) = 0.3$ angenommen. Mithilfe eines Generators für Zufallszahlen kann dann ein Datensatz erzeugt werden, der diesen Annahmen entspricht. Abbildung 3.4.4 zeigt Survivorfunktionen für die unzensierten Verweildauern, die mit dem im folgenden verwendeten Datensatz berechnet worden sind.

Ausgehend von dieser simulierten Grundgesamtheit wird jetzt ein logistisches Regressionsmodell der Form

$$\tilde{r}(t \mid x; \gamma, \beta) = \frac{\exp(\gamma + x\beta)}{1 + \exp(\gamma + x\beta)} \qquad x \in \{0, 1\}$$

geschätzt, bei unterschiedlichen Annahmen über zensierte Beobachtungen. Für die Modellschätzung wird das Maximum-Likelihood-Prinzip verwendet.

[36] Es sei betont, daß es sich nur um eine Illustration, nicht um eine systematische Untersuchung handelt, bei der die Datenerzeugung viele Male wiederholt und dann die Verteilung der Schätzergebnisse betrachtet werden müßte. Eine etwas extensivere Monte-Carlo-Studie rechts zensierter Daten wurde u.a. von Tuma und Hannan [1979] vorgelegt; vgl. auch Tuma und Hannan [1984, S. 140ff].

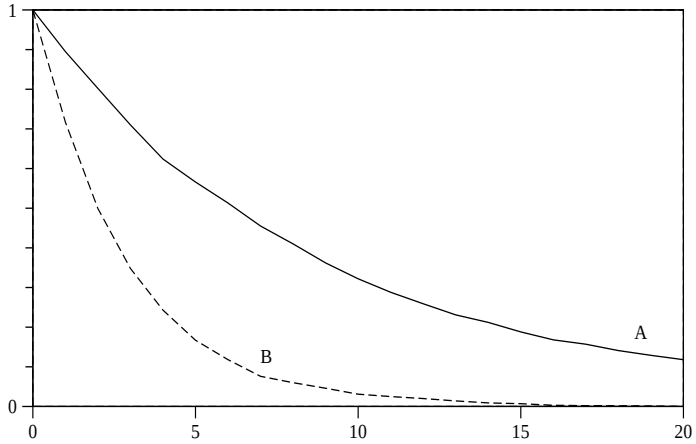


Abbildung 3.4.4. Survivorfunktionen für die nicht-zensierten Verweildauern in Gruppe A, $r(t) = 0.1$, und Gruppe B, $r(t) = 0.3$. Jeweils 1000 Individuen.

a) Zunächst wird ein Modell unter der Annahme geschätzt, daß für alle Verweildauern unzensierte Beobachtungen vorliegen. In diesem Fall liefert eine Modellschätzung mit dem KL-Distanzmaß die Parameterschätzungen $\hat{\gamma} = -2.2412$ und $\hat{\beta} = 1.3362$. Dem entsprechen $\hat{r}_A = 0.104$ und $\hat{r}_B = 0.295$ als Schätzwerte für die Übergangsraten. Sie weichen ersichtlich nur wenig von den zur Datenerzeugung verwendeten Raten ab.

b) Jetzt wird angenommen, daß die Beobachtungen auf zufällige Weise zensiert sind. Zweierlei ist dafür erforderlich. Erstens muß die Auswahl der Individuen, deren Beobachtung rechts zensiert ist, zufällig erfolgen; dies kann durch eine einfache Zufallsauswahl aus der Gesamtheit der 2000 Individuen erfolgen. Zweitens müssen auch die Zensierungszeitpunkte zufällig bestimmt werden. Um dies zu simulieren, wird zunächst ein Bereich $[t_1, t_2]$ auf der Zeitachse festgelegt, dann wird für jedes der zuvor zufällig ausgewählten Individuen eine in diesem Bereich gleichverteilte Zufallszahl erzeugt. Wenn sie kleiner ist als die Verweildauer des Individuums, wird sie als Zensierungszeitpunkt verwendet.

Tabelle (3.22) zeigt Schätzergebnisse für vier unterschiedliche Intervalle und dementsprechend unterschiedliche Prozentsätze zensierter Beobachtungen. Man erkennt, daß in diesem Fall selbst bei einem hohen Anteil zensierter Beobachtungen noch gute Schätzergebnisse erzielt werden.

Intervall	Zensiert	$\hat{\gamma}$	$\hat{\beta}$	$\hat{r}_A$	$\hat{r}_B$
$1-40$	15%	-2.1594	1.2967	0.103	0.297
$1-20$	25%	-2.1301	1.2825	0.106	0.300
$1-10$	40%	-2.1605	1.3010	0.103	0.297
$1-\ \ 5$	55%	-2.1268	1.2677	0.107	0.298

(3.22)

c) Um eine Unabhängigkeit der Zensierungszeitpunkte von den Verweildauern zu erreichen, ist es nicht unbedingt erforderlich, daß sie zufällig verteilt sind. Eine andere Möglichkeit besteht darin, feste Zensierungszeitpunkte anzunehmen. Um dies zu illustrieren, gehe ich wieder davon aus, daß die Individuen, deren Beobachtung ggf. zensiert ist, zunächst zufällig aus der Gesamtheit ausgewählt werden. Ihre Beobachtung wird dann als zensiert angenommen, wenn ihre Verweildauer größer ist als der willkürlich gewählte Zeitpunkt $t = 3$, vgl. Abbildung 3.4.4, der dann als Zensierungszeitpunkt angenommen wird. Tabelle (3.23) zeigt die Schätzergebnisse für diese Situation.

Zensiert	$\hat{\gamma}$	$\hat{\beta}$	$\hat{r}_A$	$\hat{r}_B$
10%	-2.1697	1.3130	0.103	0.298
20%	-2.1655	1.3178	0.103	0.300
30%	-2.1435	1.2757	0.105	0.296
40%	-2.1354	1.2589	0.106	0.293
53%	-2.1163	1.2413	0.107	0.294

(3.23)

Die Angaben in der ersten Spalte beziehen sich auf den Prozentsatz zensierter Fälle im gesamten Datensatz. Bei 53 % zensierten Fällen sind tatsächlich alle Beobachtungen zensiert, bei denen die Verteildauer größer ist als $t = 3$. Man erkennt, daß auch bei diesem Zensierungsmechanismus brauchbare Schätzergebnisse erzielt werden können.

d) Schließlich sei noch ein Beispiel angeführt, bei dem die Zensierungszeitpunkte systematisch mit der Verweildauer korreliert sind. Wiederum werden die Individuen, deren Verweildauer als zensiert angenommen wird, zufällig aus der Gesamtheit ausgewählt; aber der Zensierungszeitpunkt wird dann jeweils bei der halben tatsächlichen Verweildauer festgesetzt.[37] Tabelle (3.24) zeigt die Schätzergebnisse.

[37] Da in diesem Beispiel nur ganzzahlige Verweildauern möglich sind, wird ggf. aufgerundet.

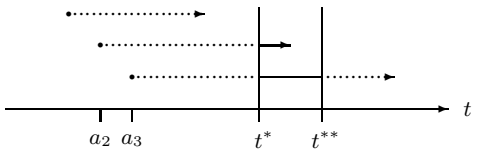
Zensiert	$\hat{\gamma}$	$\hat{\beta}$	$\hat{r}_A$	$\hat{r}_B$
10%	-2.2219	1.2696	0.098	0.278
20%	-2.3101	1.2639	0.090	0.260
30%	-2.4079	1.2346	0.083	0.236
40%	-2.5128	1.2088	0.075	0.213
50%	-2.6201	1.2056	0.068	0.196
60%	-2.7630	1.1535	0.059	0.167
70%	-2.9795	1.1005	0.048	0.133
80%	-3.2739	0.9735	0.036	0.091
90%	-3.7947	0.8289	0.022	0.049

(3.24)

Man erkennt an diesem Beispiel deutlich, daß in einer Situation, in der die Zensierungszeitpunkte mit den tatsächlichen Verweildauern korreliert sind, mit durchaus erheblichen Verzerrungen in der Modellschätzung gerechnet werden muß. Wie bereits erwähnt worden ist, ist eine solche Korrelation bei einmalig retrospektiv erhobenen Lebensverlaufsdaten im allgemeinen nicht zu erwarten. Wenn sie jedoch mit einer Serie von Panelerhebungen verknüpft wird, besteht durchaus die Möglichkeit, daß das Ausscheiden aus der Stichprobe im Verlaufe des Panels auch durch das zu untersuchende Ereignis bedingt wird, so daß auf diese Weise eine Korrelation zwischen Verweildauer und Zensierungszeitpunkten entsteht.

3.4.2 Links abgeschnittene Beobachtungen

Von links abgeschnittenen Beobachtungen spricht man, wenn die (retrospektive) Beobachtung eines Prozesses erst einsetzt, nachdem der Prozeß für einige oder alle Personen bereits begonnen hat, und wenn die Beobachtbarkeit der Personen davon abhängt, daß sie den Ausgangszustand des Prozesses zum Beginn der Beobachtungsperiode noch nicht verlassen haben.[38] Folgende Abbildung illustriert eine Situation links abgeschnittener Beobachtungen, wobei von einer Kalenderzeitachse ausgegangen wird.



Der Prozeß wird während des Zeitraums von t^* bis t^{**} beobachtet. Erfaßt

[38] In der englischsprachigen Literatur wird von *left truncated observations* gesprochen. Diese Situation unvollständiger Beobachtungen ist in der Literatur bereits häufig behandelt worden, vgl. u.a. Lancaster [1979], Ridder [1984], Helsen [1990], Schneider [1992], Guo [1993].

werden können alle Personen, die sich zum Zeitpunkt t^* noch im Ausgangszustand des Prozesses befinden. Außerdem kann für alle diese Personen festgestellt werden, wann sie in den Anfangszustand des Prozesses eingetreten sind.

Das Problem besteht offenbar darin, daß man keine Informationen über diejenigen Personen hat, die den Prozeß vor t^* begonnen *und auch bereits beendet* haben. In der Abbildung ist dies die erste Person. Die zweite Person beendet den Prozeß innerhalb des Beobachtungsfensters, und infolgedessen kann der Zeitpunkt a_1, zu dem sie den Prozeß begonnen hat, ermittelt werden. Dasselbe gilt für die dritte Person, die jedoch den Prozeß erst nach dem Ende des Beobachtungszeitraums beendet; die Beobachtung ist in diesem Fall links abgeschnitten und rechts zensiert.[39]

Das Problem links abgeschnittener Beobachtungen stellt sich immer dann, wenn die verfügbare Stichprobe nur Personen umfaßt, die sich zum Beginn der Beobachtungsperiode bereits im Anfangszustand des zu untersuchenden Prozesses befinden. Ein typisches Beispiel wäre etwa der Versuch, Aussagen über die Dauer der Arbeitslosigkeit mithilfe einer Stichprobe zu gewinnen, die nur aus Personen besteht, die bereits arbeitslos sind; oder in unserem Heiratsbeispiel, wenn man versuchen würde, Aussagen über das Heiratsalter aus einer Befragung von bereits verheirateten Personen zu gewinnen.

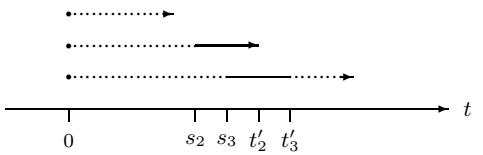
Ob bzw. in welcher Weise eine Situation links abgeschnittener Beobachtungen vorliegt, hängt u.a. davon ab, wie die Grundgesamtheit für den zu untersuchenden Prozeß definiert wird. Würde man die Grundgesamtheit als die Gesamtheit derjenigen Personen definieren, die sich zum Anfangszeitpunkt der Beobachtung in einem gewissen Zustand (dem Anfangszustand des interessierenden Prozesses) befinden, gäbe es keine links abgeschnittenen Beobachtungen. Diese Vorgehensweise ist jedoch im allgemeinen nicht sinnvoll, wir interessieren uns vielmehr für den Prozeßverlauf in einer geeignet definierten Längsschnittgesamtheit (vgl. Abschnitt 2.4.1). Die Frage ist dann, welche Beobachtungen über den Prozeßverlauf in der intendierten Längsschnittgesamtheit verfügbar sind, und wie mit ihrer Hilfe der Prozeßverlauf *in der Längsschnittgesamtheit* beschrieben werden kann. Auf diese Frage bezogen, können zwei Situationen unterschieden werden.

(1) Zunächst eine Situation, in der der zu beschreibende Prozeß zumindest während eines gewissen Zeitraums beobachtet werden kann; in der obigen Abbildung ist dies der Zeitraum von t^* bis t^{**}. Im allgemeinen kann natürlich nicht vorausgesetzt werden, daß es sich um einen festen Zeitabschnitt auf der Kalenderzeitachse handelt. Entscheidend ist jedoch, daß eine Risikomenge beobachtet werden kann und daß für die Personen in dieser Risikomenge empirische Wahrscheinlichkeiten für einen Zustandswechsel ermittelt werden können.

[39] Wir betrachten hier also den allgemeinen Fall, in dem es sowohl links abgeschnittene als auch rechts zensierte Beobachtungen geben kann.

(2) Eine andere Situation liegt vor, wenn nur eine einmalige Erhebung durchgeführt wird. In der obigen Abbildung wäre das eine Datenerhebung, die nur zum Zeitpunkt t^* stattfindet. Das Resultat wäre eine Menge links abgeschnittener Beobachtungen, die zugleich in allen Fällen auch zum Erhebungszeitpunkt rechts zensiert sind. Für keines der Individuen hat man dann eine Information über einen Zustandswechsel.

Im folgenden wird nur die erste dieser beiden Situationen etwas näher betrachtet.[40] Zumindest für einige Individuen aus der intendierten Längsschnittgesamtheit kann dann der Prozeßverlauf für einen gewissen Zeitraum beobachtet werden. Die folgende Abbildung illustriert diese Situation für eine Prozeßzeitachse, die mit dem Eintritt in den Anfangszustand des Prozesses beginnt.



In dieser Illustration gibt es drei Individuen, alle beginnen den Prozeß zum Zeitpunkt $t = 0$ auf der Prozeßzeitachse. Über das erste Individuum gibt es keinerlei Information. Das zweite Individuum kann vom Zeitpunkt s_2 bis zum Zeitpunkt t'_2 beobachtet werden, und es kann ermittelt werden, daß zum Zeitpunkt t'_2 ein Ereignis eingetreten ist. Das dritte Individuum kann vom Zeitpunkt s_3 bis zum Zeitpunkt t'_3 beobachtet werden; die Beobachtung ist zum Zeitpunkt t'_3 rechts zensiert.

Wiederum können zwei Fälle unterschieden werden. Erstens, man hat ausschließlich Informationen über diejenigen Individuen, die zumindest für einen gewissen Zeitraum beobachtet werden können (die zweite und dritte Person in der obigen Abbildung). In diesem Fall kennt man nur N', die Anzahl der beobachteten Individuen, man hat keine Information über N, die Anzahl der Personen in der intendierten Längsschnittgesamtheit. Zweitens, man kennt nicht nur N', sondern auch N. In der Regel kennt man N nicht, und wir werden uns deshalb hauptsächlich auf diesen Fall beziehen.

Um das Schätzproblem in dieser Situation links abgeschnittener Beobachtungen formal präzisieren zu können, ist es erforderlich, die verfügbaren Informationen als Zufallsvariablen zu definieren. Ich beziehe mich zunächst, wie in (3.9), auf eine in diskreter Prozeßzeit definierte Episode mit einem einfachen Endzustand. Das verfügbare Beobachtungswissen

[40] Bei der zweiten Situation wird in der englischsprachigen Literatur häufig von „backward recurrence times“ gesprochen. Vgl. als Beiträge, die diese Situation diskutieren, u.a. Sørensen [1977], Allison [1985], Baydar und White [1988].

kann dann durch folgende Zufallsvariablen repräsentiert werden:

$$(S, T', X, \delta) : \Omega' \longrightarrow \mathcal{T} \times \mathcal{T} \times \mathcal{X} \times \{0, 1\}$$

Ω' ist die Menge der Individuen aus der Längsschnittgesamtheit Ω, bei denen der Prozeß für einen gewissen Zeitraum beobachtet werden kann. Für jedes Individuum $\omega \in \Omega'$ gibt es folgende Informationen: $S(\omega)$ ist der Zeitpunkt, zu dem die Beobachtung beginnt, und $T'(\omega)$ ist der Zeitpunkt, zu dem die Beobachtung endet. δ ist eine Indikatorvariable, die die Bedeutung von $T'(\omega)$ charakterisiert: wenn $\delta(\omega) = 1$, wissen wir, daß dies Individuum zum Zeitpunkt $T'(\omega)$ einen Übergang in den Endzustand der Episode vollzogen hat; wenn $\delta(\omega) = 0$, wissen wir nur, daß bis zum Zeitpunkt $T'(\omega)$ noch kein Übergang in den Endzustand stattgefunden hat. Schließlich ist $X(\omega)$ ein Index für die Teilklasse (in Ω), der das Individuum zum Episodenbeginn angehört.[41] Die Frage ist, welche Aussagen mithilfe von (S, T', X, δ) über die Zufallsvariable

$$(T, X) : \Omega \longrightarrow \mathcal{T} \times \mathcal{X}$$

gemacht werden können. Die Grundidee zur Lösung dieses Problems liefert wiederum die Vorstellung einer Risikomenge $R(t)$, die Gesamtheit der Personen, die sich zum Zeitpunkt t noch im Ausgangszustand befinden und für die infolgedessen ein Zustandswechsel noch möglich ist. Zwar kennen wir $R(t)$ nicht, jedoch gibt es zu jedem Zeitpunkt eine *beobachtete* Risikomenge $R'(t)$. Jedes Individuum $\omega \in \Omega'$ gehört für den Zeitraum von $S(\omega)$ bis $T'(\omega)$ zur beobachteten Risikomenge:

$$R'(t) = \{\omega \mid \omega \in \Omega', S(\omega) < t \leq T'(\omega)\}$$

Gestützt auf dieses Konzept einer beobachteten Risikomenge kann, wie in einer Situation nur rechts zensierter Beobachtungen, das Kaplan-Meier-Verfahren zur Schätzung einer Survivorfunktion entwickelt werden. Zunächst lassen sich empirische Übergangswahrscheinlichkeiten definieren: $q'(t) = E'_t/R'_t$, wobei E'_t die Anzahl beobachteter Ereignisse und R'_t die Anzahl der Personen in der beobachteten Risikomenge bezeichnet. Dann erhält man eine Schätzung der Survivorfunktion durch

$$G'(t) = \prod_{\tau=1}^{t} (1 - q'(t)) \tag{3.25}$$

Zu überlegen ist, unter welchen Bedingungen dies eine sinnvolle Schätzung der Survivorfunktion $G(t)$ in der Längsschnittgesamtheit Ω liefert. Eine Bedingung, die den Sachverhalt rechts zensierter Beobachtungen betrifft,

[41] Es wird, wie bisher, angenommen, daß sich die Klassenzugehörigkeit während des Episodenverlaufs nicht verändern kann.

wurde bereits in Abschnitt 3.4.1 dargestellt: die Zensierungszeitpunkte müssen in gewisser Weise unabhängig von den Ereigniszeitpunkten sein. Eine analoge Bedingung läßt sich auch für links abgeschnittene Beobachtungen formulieren: die durch die Zufallsvariable S repräsentierten Anfangszeitpunkte der jeweiligen Prozeßbeobachtung müssen unabhängig von den Ereigniszeitpunkten sein.

Bevor ich versuche, diese Bedingung im Kontext der Modellschätzung formal zu präzisieren, soll sie zunächst anhand der SOEP-Daten über das Alter bei der ersten Heirat illustriert werden. Ich betrachte drei unterschiedliche Situationen.

a) Ausgangspunkt sind die Daten, die das SOEP über den Prozeß bis zur ersten Heirat zur Verfügung stellt. Der Stichprobenumfang beträgt, wie in den bisherigen Beispielen, $N = 6416$.[42] Für jedes Individuum kennen wir beobachtete Werte für die Zufallsvariablen (T', X, δ). Die verfügbaren Daten sind nicht links abgeschnitten, d.h. $S = 0$.

Um eine Situation links abgeschnittener Daten zu simulieren, wird angenommen, daß der Prozeß bei diesen Personen nicht von Anfang an (seit ihrer Geburt) beobachtet werden kann, sondern daß die Prozeßbeobachtung bei jedem Individuum $i = 1, \ldots, N$ erst zu einem Zeitpunkt $s_i \geq 0$ beginnt. s_i wird für jedes Individuum als eine gleichverteilte Zufallszahl im Altersbereich von 1 bis K Jahren bestimmt ($K = 30, 40, 50, 60$). Dies hat zur Folge, daß für einen Teil der Individuen überhaupt keine Beobachtungen mehr vorliegen, wenn nämlich $s_i \geq t'_i$. Die restlichen Beobachtungen bilden die beobachtete Teilgesamtheit Ω' mit zu den Zeitpunkten s_i links abgeschnittenen Beobachtungen.

Bei dieser Vorgehensweise ist die Bedingung einer statistischen Unabhängigkeit von Anfangszeiten der Beobachtung und Ereigniszeitpunkten offensichtlich erfüllt, und das in (3.25) definierte Schätzverfahren liefert, wie durch Abbildung 3.4.5 illustriert wird, tatsächlich gute Ergebnisse.[43]

[42] Alle Stammpersonen aus der Teilstichprobe A, die im Zeitraum 1918 – 1968 geboren wurden und an mindestens den ersten drei Wellen des SOEP teilgenommen haben.

[43] Die Schätzung der Survivorfunktionen erfolgt mit dem Kaplan-Meier-Verfahren, wobei allerdings berücksichtigt werden muß, daß die Risikomenge während der Prozeßzeit nicht nur kleiner, sondern auch größer werden kann. Diese Möglichkeit bietet die im Programmpaket TDA implementierte Form des Kaplan-Meier-Verfahrens.

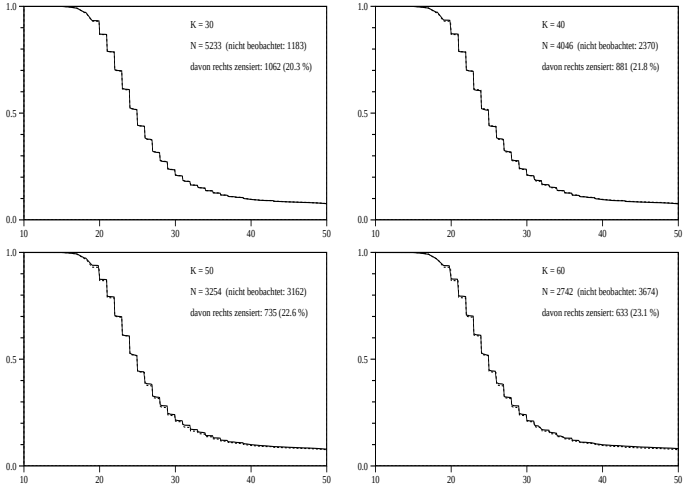


Abbildung 3.4.5. Survivorfunktionen für das Ereignis *erste Heirat* bei den Personen aus der Teilstichprobe A des SOEP. Als durchgezogene Linien sind Kaplan-Meier-Schätzungen für vier unterschiedlich simulierte Formen links abgeschnittener Daten eingezeichnet. Jede Abbildung zeigt außerdem als gestrichelte Linie eine Kaplan-Meier-Schätzung mit der vollen, nicht links abgeschnittenen Stichprobe. Abszisse in Jahren.

b) Eine andere Variante statistischer Unabhängigkeit von Anfangszeitpunkten der Beobachtung und Ereigniszeitpunkten liegt vor, wenn die Beobachtung bei allen Individuen zum gleichen Zeitpunkt beginnt. Zur Illustration dieser Situation nehmen wir an, daß die Beobachtung bei allen Individuen erst mit dem Abschluß des 20. Lebensjahr beginnt. Alle Individuen, die vor ihrem 20. Lebensjahr geheiratet haben, können nicht beobachtet werden. In unserer Stichprobe verbleiben dann $N' = 5530$ Personen (davon 1127 rechts zensiert), 886 Personen können nicht beobachtet werden.

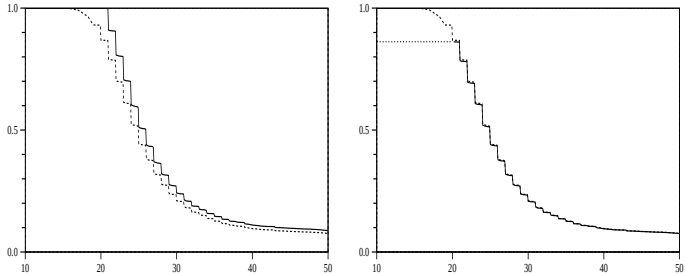


Abbildung 3.4.6. Kaplan-Meier-Schätzungen der Survivorfunktion für das Ereignis *erste Heirat* bei den Personen aus der Teilstichprobe A des SOEP. Durchgezogene Linien: beim Alter von 20 Jahren links abgeschnittene Daten, gestrichelte Linien: Schätzung mit nicht links abgeschnittenen Daten. Rechte Abbildung: Multiplikation der geschätzten Survivorfunktion mit dem Faktor 5530/6416.

Abbildung 3.4.6 zeigt die Schätzergebnisse für die so konstruierten Daten. Die linke Abbildung zeigt die geschätzte Survivorfunktion in der üblichen Form: sie hat bis zum Auftreten des ersten Ereignisses den Wert 1. Dies entspricht allerdings nicht dem Verlauf der Survivorfunktion in der Grundgesamtheit. Dort sind bereits all diejenigen Personen ausgeschieden, die vor dem Abschluß ihres 20. Lebensjahres geheiratet haben. In unserer Stichprobe sind dies 886 von 6416 Personen; der Wert der Survivorfunktion zum Zeitpunkt $t = 20$ ist infolgedessen nicht 1, sondern $5530/6416 \approx 86.2\,\%$.

Kennt man die Anzahl der nicht beobachteten Individuen, kann man die geschätzte Survivorfunktion durch Multiplikation mit dem Faktor N'/N korrigieren, so daß sie für den Beobachtungszeitraum ein zutreffendes Bild der Survivorfunktion in der Grundgesamtheit liefert (vgl. die rechte Abbildung). Kennt man diese Anzahl nicht, läßt sich über die Survivorfunktion in der Grundgesamtheit tatsächlich keine angemessene Aussage bilden. Dies gilt generell, auch in der unter (a) beschriebenen Situation. Eine mit dem Kaplan-Meier-Verfahren geschätzte Survivorfunktion liefert (bestenfalls) eine angemessene Beschreibung für den Zeitraum $[t_a, t_b]$, wobei t_a der Zeitpunkt ist, bei dem zum erstenmal eine Risikomenge beobachtet werden kann (das Minimum der Zufallsvariable S), und t_b der Zeitpunkt, bei dem zum letzten Mal ein Ereignis beobachtet werden kann. Die fehlende Information über die bis zum Zeitpunkt t_a bereits ausgeschiedenen Personen kann nicht kompensiert werden. In der unter (a) beschriebenen Situation ist diese Tatsache nur deshalb nicht sichtbar geworden, weil der Zufallsmechanismus zur Auswahl von Anfangszeitpunkten für die Beobachtung dazu geführt hat, daß t_a vor dem ersten Ereignis liegt, das in der gesamten Stichprobe vorkommt.

Es sollte jedoch betont werden, daß dieses Problem nur die Schätzung der Survivorfunktion in der Grundgesamtheit betrifft. Die mit dem

Kaplan-Meier-Verfahren ermittelbare Survivorfunktion $G'(t)$ ist zu einer sinnvollen Schätzung der Survivorfunktion in der Grundgesamtheit, $\hat{G}(t)$, proportional. Daraus folgt jedoch, daß die durch $G'(t)$ bzw. $\hat{G}(t)$ definierbaren Übergangsraten identisch sind.[44] Konzentriert man sich zur Prozeßbeschreibung auf eine Verwendung von Übergangsraten, wie dies typischerweise bei der Modellbildung der Fall ist, ist die Tatsachse, daß man bei links abgeschnittenen Daten die Anzahl der nicht beobachteten Individuen nicht kennt, ohne Bedeutung. Es bleibt natürlich die Restriktion, daß empirische Aussagen über den Prozeßverlauf nur ab einem Zeitpunkt möglich sind, bei dem zum erstenmal eine Risikomenge beobachtet werden kann.

c) Schließlich betrachten wir noch eine Situation, in der die Anfangszeitpunkte der Beobachtung (s_i) mit den Ereigniszeitpunkten korreliert sind. Eine einfache Möglichkeit zur Simulation einer solchen Situation besteht in der Annahme, daß die Beobachtung jeweils 10 oder 20 Jahre vor dem Ende der Beobachtung (t'_i) beginnt. Abbildung 3.4.7 zeigt die Schätzergebnisse unter dieser Annahme. Man erkennt, daß die Korrelation von S und T' bzw. T zu deutlichen Verzerrungen in den Schätzergebnissen führt.

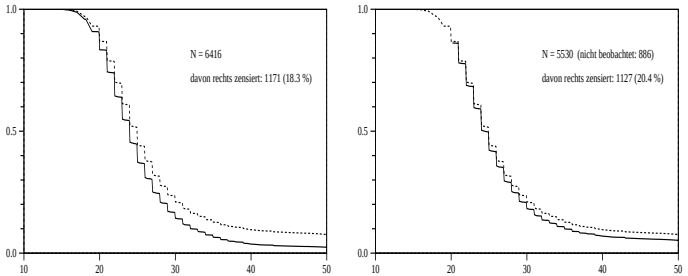


Abbildung 3.4.7. Kaplan-Meier-Schätzungen der Survivorfunktion für das Ereignis *erste Heirat* bei den Personen aus der Teilstichprobe A des SOEP. Durchgezogene Linien: links abgeschnitten bei 10 (linke Abbildung) bzw. 20 (rechte Abbildung) Jahren vor dem Ende der Beobachtung; gestrichelte Linien: Schätzung mit nicht links abgeschnittenen Daten.

Modellschätzung mit links abgeschnittenen Beobachtungen

Ziel der Modellschätzung ist ein Modell für die Zufallsvariable (T, X) bzw. für die durch X bedingte Verteilung von T. Im allgemeinen können die verfügbaren Beobachtungen sowohl links abgeschnitten als auch rechts zensiert sein. Zur Darstellung des Problems der Modellschätzung ist es jedoch zweckmäßig, von der Möglichkeit rechts zensierter Beobachtungen

[44] Dies folgt unmittelbar daraus, daß die Übergangsrate als negative logarithmische Ableitung der Survivorfunktion formuliert werden kann, vgl. (2.11) in Abschnitt 2.4.

zunächst abzusehen. Das Beobachtungswissen kann dann durch die Zufallsvariablen

$$(S, T', X) : \Omega' \longrightarrow \mathcal{T} \times \mathcal{T} \times \mathcal{X}$$

repräsentiert werden. S liefert die Anfangszeitpunkte der Beobachtung, T' die Ereigniszeitpunkte, und X ist eine Klassifizierungsvariable. Alle Zufallsvariablen beziehen sich auf die Gesamtheit Ω', d.h. diejenigen Individuen aus der Längsschnittgesamtheit Ω, die zumindest partiell beobachtet werden können.

Aus der gemeinsamen Verteilung von (S, T', X) könnte nun unmittelbar die Randverteilung von (T', X) gebildet werden. Aber es ist klar, daß diese Verteilung keine unverzerrte Darstellung für die Verteilung von (T, X) in der Grundgesamtheit Ω liefert. Um den für die Modellbildung wesentlichen Zugang zu dieser Verteilung zu finden, muß deshalb anders vorgegangen werden. Einen möglichen Zugang liefert folgende Zerlegung der empirisch verfügbaren Wahrscheinlichkeiten:

$$\begin{aligned} &\mathrm{P}\,(T' = t, S = s \mid X = x) = \\ &\quad \mathrm{P}\,(T' = t \mid S = s, X = x)\,\mathrm{P}\,(S = s \mid X = s) \end{aligned}$$

Diese Zerlegung zeigt, wo man ansetzen kann, um aus der Verteilung von (T', S, X) Informationen über die Verteilung von (T, X) zu gewinnen, nämlich bei der bedingten Wahrscheinlichkeit für das Auftreten eines Ereignisses, wobei die Bedingung darin besteht, daß die Individuen einer durch X definierten Teilgesamtheit angehören *und* daß ihre Beobachtung zu einem gewissen Zeitpunkt begonnen hat. Es liegt also nahe, folgende Annahme zu formulieren:

$$\mathrm{P}\,(T' = t \mid S = s, X = x) \approx \mathrm{P}\,(T = t \mid T > s, X = x) \tag{3.26}$$

Auf der rechten Seite steht die Wahrscheinlichkeit, daß bei den Individuen in der Teilgesamtheit $\Omega(s, x) = \{\omega \mid T(\omega) > s, X(\omega) = x)$ zum Zeitpunkt t ein Ereignis stattfindet. Die linke Seite bezieht sich auf diejenigen Individuen aus $\Omega(s, x)$, die tatsächlich beobachtet worden sind. Die Bedingung fordert also, daß die Wahrscheinlichkeit eines Ereignisses bei den beobachteten Individuen näherungsweise gleich ist zur Wahrscheinlichkeit eines Ereignisses bei allen Individuen aus $\Omega(s, x)$. Diese Annahme kann gerechtfertigt werden, wenn die Auswahl der Werte für die Zufallsvariable S, d.h. der Anfangszeitpunkte für die Beobachtung, unabhängig von der Verteilung der durch X bedingten Zufallsvariable T, d.h. unabhängig von den Ereignispunkten in der Grundgesamtheit Ω erfolgt. Im wesentlichen handelt es sich wiederum um eine Repräsentativitätsannahme: es wird angenommen, daß – *während des Beobachtungszeitraums* – die beobachteten Prozeßverläufe der Individuen aus Ω' für die Gesamtheit der

Prozeßverläufe bei den Individuen aus Ω repräsentativ sind.[45]

Wenn die Annahme (3.26) gerechtfertigt erscheint, kann sie zur Modellschätzung genutzt werden. Im Mittelpunkt steht stets ein Vergleich: auf der einen Seite eine aus den verfügbaren Daten ableitbare empirische Verteilung, auf der anderen Seite eine Klasse möglicher theoretischer Verteilungen zur Approximation dieser empirischen Verteilung. Im vorliegenden Fall besteht der Ausgangspunkt in einer Klasse möglicher Modelle für die durch X bedingte Verteilung von T, also

$$\mathrm{P}\,(T = t \mid X = x) \;\approx\; \tilde{f}(t \mid x; \theta) \qquad \theta \in \Theta$$

Daraus folgt eine Approximation für die rechte Seite von (3.26) und, wenn die dort formulierte Annahme zutreffend ist, auch für die empirische Wahrscheinlichkeit auf der linken Seite von (3.26), also:

$$\mathrm{P}\,(T' = t \mid S = s, X = x) \;\approx\; \frac{\tilde{f}(t \mid x; \theta)}{\tilde{G}(s \mid x; \theta)}$$

Im Mittelpunkt der Modellschätzung steht also dieser Vergleich. Für die praktische Durchführung kann wiederum das KL-Distanzmaß verwendet werden. Ein optimales Modell (aus der vorgegebenen Modellklasse) erhält man dann durch die Minimierung der Distanz

$$\sum_{s,t \in \mathcal{T}} \sum_{x \in \mathcal{X}} \mathrm{P}\,(T' = t, S = s, X = x) \, \log \left\{ \frac{\mathrm{P}\,(T' = t \mid S = s, X = x)}{\tilde{f}(t \mid x; \theta) / \tilde{G}(s \mid x; \theta)} \right\}$$

Ein äquivalentes Kriterium für die Modellschätzung liefert die Log-Likelihood

$$\ell(\theta) = \sum_{i=1}^{N'} \log \left\{ \frac{\tilde{f}(t'_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} \right\}$$

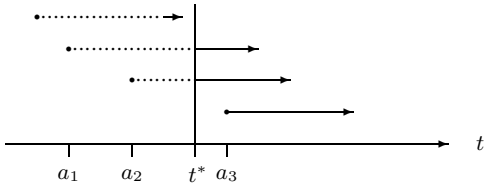
Um zu berücksichtigen, daß einige der links abgeschnittenen Beobachtungen zugleich rechts zensiert sein können, kann auf die Ausführungen in Abschnitt 3.4.1 zurückgegriffen werden. Die dort angegebene Log-Likelihood nimmt dann folgende Form an:

$$\ell(\theta) = \sum_{i=1}^{N'} \delta_i \, \log \left\{ \frac{\tilde{f}(t'_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} \right\} + (1 - \delta_i) \, \log \left\{ \frac{\tilde{G}(t'_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} \right\} \tag{3.27}$$

[45] Problematische Selektionen entstehen, wenn eine solche Repräsentativitätsannahme nicht gerechtfertigt werden kann. Vgl. Berk [1983] als eine allgemeine Einführung in das Selektionsproblem in der empirischen Sozialforschung sowie Heckman [1990] für eine Diskussion aus ökonometrischer Sicht.

3.4.3 Links zensierte Beobachtungen

Bei links abgeschnittenen Beobachtungen kann nur ein Teil des Prozeßverlaufs beobachtet werden, aber man kennt die Zeitpunkte, zu denen die Episoden begonnen haben; der beobachtete Teil des Episodenverlaufs kann infolgedessen in den Gesamtverlauf der Episode eingeordnet werden. Bei links zensierten Beobachtungen ist der Anfangszeitpunkt der Episode unbekannt, und es ist infolgedessen ebenfalls unbekannt, zu welchem Zeitpunkt *auf der Prozeßzeitachse* die Beobachtung beginnt. Folgende Abbildung illustriert diese Situation.



Die Abbildung zeigt eine Kalenderzeitachse; die Prozeßbeobachtung beginnt zum Zeitpunkt t^*. Über die erste Episode hat man keinerlei Information. Die beiden folgenden Beobachtungen sind links abgeschnitten oder links zensiert. Sie sind links abgeschnitten, wenn die Anfangszeitpunkte a_1 bzw. a_2 ermittelt werden können, andernfalls handelt es sich um links zensierte Beobachtungen. Die vierte Beobachtung (ganz unten) ist vollständig.

Im allgemeinen können Beobachtungen im Hinblick auf den Anfangsverlauf eines Prozesses vollständig, links abgeschnitten oder links zensiert sein. Außerdem können einige Beobachtungen rechts zensiert sein, davon wird im folgenden jedoch abgesehen.

Zur formalen Beschreibung gehen wir wie bisher von einer Längsschnittgesamtheit Ω aus, die sich aus Individuen mit den Indizes $i = 1, \ldots, N$ zusammensetzt. Für jedes Individuum kennen wir ein Intervall $[s'_i, s''_i]$, in dem sich der Anfangszeitpunkt der Beobachtung, s_i, befindet. Außerdem kennen wir die Zeitdauer d_i, während der das Individuum beginnend mit dem Zeitpunkt s_i beobachtet worden ist. Je nach der Beschaffenheit des Intervalls $[s'_i, s''_i]$ gibt es drei Möglichkeiten:

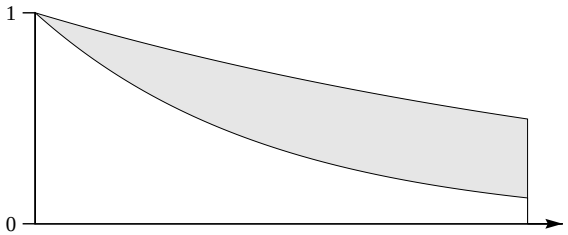
a) Wenn $s'_i = s''_i = 0$, ist die Beobachtung vollständig. Man weiß dann, daß der Prozeßverlauf für dieses Individuum beginnend mit dem Zeitpunkt $s_i = 0$ beobachtet worden ist.

b) Wenn $0 < s'_i = s''_i$, handelt es sich um eine links abgeschnittene Beobachtung. Man weiß dann, daß das Individuum beginnend mit dem Zeitpunkt $s_i = s'_i$ beobachtet worden ist.

c) Wenn $0 \leq s'_i < s''_i$, handelt es sich um eine links zensierte Beobachtung. Man weiß dann nur, daß der Anfangszeitpunkt für die Beobachtung des Individuums *irgendwo innerhalb* des Intervalls $[s'_i, s''_i]$ liegt. Die Breite dieses Intervalls kann als ein Indikator für das Ausmaß der fehlenden Information angesehen werden. Im allgemeinen weiß man nur, daß $s'_i \geq 0$

und daß $s_i'' - s_i'$ kleiner oder gleich dem Lebensalter der Person ist, bei dem ihre Beobachtung begonnen hat. Ggf. sind natürlich zusätzliche Annahmen möglich. Will man zum Beispiel den Übergang in die erste Heirat untersuchen, erscheint es sinnvoll, die Prozeßzeitachse erst mit einem Lebensalter beginnen zu lassen, ab dem das „Risiko“ einer Heirat besteht.[46] Verwendet man hierfür das 15. Lebensjahr und wird dann eine Person erst ab dem 20. Lebensjahr beobachtet, weiß man, daß $s_i'' - s_i'$ kleiner oder gleich 5 Jahre ist.

Es ist zweckmäßig, zunächst mit der Frage zu beginnen, wie sich in einer Situation, in der einige Beobachtungen links zensiert sind, Grenzen für den Verlauf der Survivorfunktion bestimmen lassen.[47] Analog zu den Überlegungen in Abschnitt 3.4.1 findet man eine Untergrenze dadurch, daß man bei allen links zensierten Beobachtungen annimmt, daß $s_i = s_i'$; entsprechend erhält man durch die Annahme $s_i = s_i''$ eine Obergrenze. Der Bereich möglicher Survivorfunktionen sieht dann etwa folgendermaßen aus.



Wie breit der Bereich möglicher Survivorfunktionen ist, hängt hauptsächlich von zwei Eigenschaften der verfügbaren Daten ab. Erstens vom Anteil der links zensierten Beobachtungen, zweitens von der Breite der Intervalle $[s_i', s_i'']$. Um dies zu veranschaulichen, verwenden wir wieder unser Standardbeispiel, also das Ereignis *erste Heirat* bei den Personen aus der Teilstichprobe A des SOEP. Zur Simulation links zensierter Daten wird folgendermaßen vorgegangen.

1. Ausgangspunkt ist die Stichprobe mit $N = 6416$ Personen. Mit einem Zufallsgenerator wird ein gewisser Prozentsatz von Personen ausgewählt, der als links zensiert angenommen wird.

2. Für jede Person kennen wir t_i'', das Alter bei der ersten Heirat bzw. das Alter beim Ende der Beobachtung, wenn es sich um eine rechts zensierte Beobachtung handelt. Wir verwenden eine Prozeßzeitachse, die mit dem 15. Lebensjahr beginnt. Die tatsächliche Beobachtungsdauer ist dann $t_i' = t_i'' - 15$.

[46] Vgl. Anmerkung 8 auf Seite 46.

[47] Die Möglichkeit, daß es gleichzeitig links abgeschnittene Beobachtungen geben kann, wird hierbei außer Acht gelassen.

3. Wenn es sich um Personen handelt, deren Beobachtung als nicht links zensiert angenommen wird, wird $s_i = 0$ und die hypothetisch angenommene Beobachtungsdauer $d_i' = t_i'$ gesetzt.

4. Bei links zensierten Beobachtungen wird angenommen, daß der Anfangszeitpunkt der Beobachtung sich zufällig (gleichverteilt) in einem Intervall $[0, K]$ bewegt; s_i wird also als eine gleichverteilte Zufallszahl aus diesem Intervall gezogen. Als Beobachtungsdauer wird $d_i' = t' - s_i$ angenommen. Fälle, bei denen $d_i' \leq 0$ ist, werden ausgeschlossen.

Für die nicht links zensierten Beobachtungen kennen wir s_i und d_i', bei den links zensierten Beobachtungen kennen wir nur das Intervall $[0, K]$, in dem sich der Anfangszeitpunkt der Beobachtung befindet, sowie die sich an diesen Anfangszeitpunkt anschließende Beobachtungsdauer d_i'.

Zunächst wird ein festes Zensierungsintervall mit einer Breite von $K = 5$ Jahren angenommen, außerdem vier unterschiedliche Prozentsätze für den Anteil links zensierter Beobachtungen: 10, 25, 50 und 100 %. Es werden jeweils eine Untergrenze und eine Obergrenze für die Survivorfunktion geschätzt, verwendet wird das Kaplan-Meier-Verfahren. Für die Untergrenze wird angenommen, daß (bei allen links zensierten Beobachtungen) $s_i = 0$, für die Obergrenze wird angenommen, daß $s_i = K$. Abbildung 3.4.8 zeigt die Schätzergebnisse.

Das Beispiel zeigt, daß auch bei Vorliegen links zensierter Beobachtungen sinnvolle Aussagen über die Survivorfunktion gewonnen werden können. Die Unsicherheit wächst natürlich mit dem Anteil links zensierter Beobachtungen und mit der Breite des Zensierungsintervalls.

Leider können diese Überlegungen nicht ohne weiteres auf das Problem der Schätzung von Übergangsraten übertragen werden. Man sieht das bereits anhand von Abbildung 3.4.8. Ein Bereich möglicher Survivorfunktionen impliziert zwar einen Bereich möglicher Übergangsraten; aber praktisch sind fast alle Übergangsraten möglich. Bei links abgeschnittenen Beobachtungen kann dies Problem vermieden werden, weil dann eine beobachtete Risikomenge verfügbar ist; darauf gestützt können dann empirische Übergangswahrscheinlichkeiten berechnet werden. Genau diese Möglichkeit entfällt jedoch bei links zensierten Beobachtungen. Liegt eine links zensierte Beobachtung mit dem Zensierungsintervall $[s_i', s_i'']$ und der Beobachtungsdauer d_i' vor, kann man zwar sagen, daß sie bis zum Zeitpunkt d_i' sicher zur Risikomenge gehört und daß sie ab dem Zeitpunkt $s_i'' + d_i'$ sicher nicht mehr zur Risikomenge gehört. Aber für den Zeitraum $d_i' < t < s_i'' + d_i'$ gibt es keinerlei Information darüber, ob sie zur Risikomenge gehört oder nicht. Eine Modellschätzung in einer Situation, in der ein Teil der verfügbaren Beobachtungen links zensiert ist, ist infolgedessen ohne zusätzliche Annahmen nicht möglich.[48]

[48] Vgl. dazu u.a. Tuma und Hannan [1984, S. 128ff], Hamerle [1991].

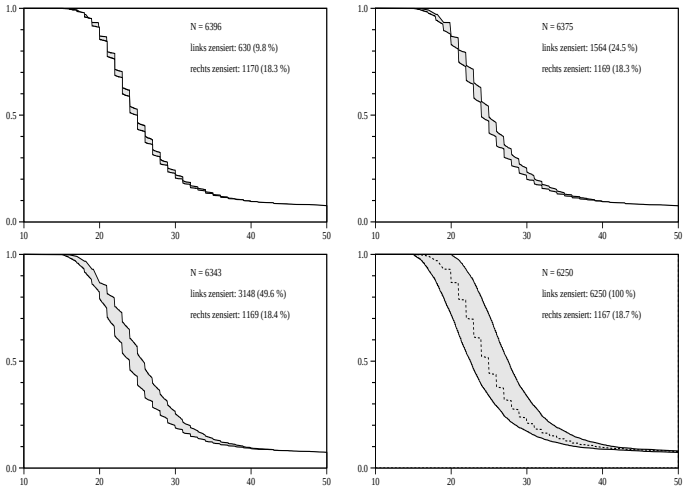


Abbildung 3.4.8. Unter- und Obergrenzen der Survivorfunktion für das Ereignis *erste Heirat* bei den Personen aus der Teilstichprobe A des SOEP, bei unterschiedlichen Prozentsätzen links zensierter Beobachtungen. Abszisse in Jahren.

3.5 Modelle für eine stetige Zeitachse

In den vorangegangenen Abschnitten dieses Kapitels sind wir davon ausgegangen, daß nicht nur der zu beschreibende Prozeß, sondern auch das zu seiner vereinfachenden Beschreibung intendierte Modell auf der Grundlage einer diskreten Zeitachse definiert sind. In diesem Abschnitt soll dargestellt werden, wie die Modellschätzung bei einer stetigen Zeitachse vorgenommen werden kann.

Die Modellbildung soll eine vereinfachende Beschreibung eines empirisch in einer Längsschnittgesamtheit Ω erfaßbaren Prozesses liefern. Wie bisher gehen wir davon aus, daß dieser Prozeß, um empirisch erfaßbar zu sein, zunächst auf einer diskreten Zeitachse definiert werden muß. Wird eine einfache Episode mit einem möglichen Zielzustand betrachtet, besteht der Ausgangspunkt wie bisher in den Zufallsvariablen

$$(T, X) : \Omega \longrightarrow \mathbf{T} \times \mathcal{X}$$

T ist die Verweildauer, definiert auf einer diskreten Prozeßzeitachse $\mathcal{T}$, und X ist eine Zufallsvariable, mit der die Individuen aus Ω in unabhängig vom Prozeßverlauf definierbare Teilgesamtheiten klassifiziert werden können.

Die Modellbildung beginnt mit der Spezifikation einer Klasse vereinfachender Darstellungen für die durch X bedingte Verteilung von T. Um dies formal zu repräsentieren, schreiben wir wie bisher $\tilde{f}(t \mid x; \theta)$, $\theta \in \Theta$. Im Unterschied zur bisher behandelten Situation diskreter Modelle handelt es sich jedoch nicht um eine Klasse von Wahrscheinlichkeitsfunktionen, sondern um eine Klasse von Dichtefunktionen.[49]

Im Mittelpunkt der Modellbildung steht ein Vergleich der empirischen Wahrscheinlichkeiten für den Prozeßverlauf in der Grundgesamtheit mit ihrer vereinfachenden Beschreibung durch eine Klasse möglicher Modelle. Um diesen Vergleich formal sichtbar zu machen, haben wir bisher die Formulierung

$$\mathrm{P}\left(T = t \mid X = x\right) \approx \tilde{f}(t \mid x; \theta) \tag{3.28}$$

verwendet. Wenn die Modellbildung auf der Grundlage einer diskreten Zeitachse erfolgen soll, stehen auf beiden Seiten Wahrscheinlichkeitsfunktionen, und es kann ein unmittelbarer Vergleich vorgenommen werden. Soll die Modellbildung jedoch auf einer stetigen Zeitachse erfolgen, ist die in (3.28) gegebene Formulierung zunächst unklar. Es stellt sich die Frage, wie die empirische Wahrscheinlichkeitsfunktion (auf der linken Seite) mit einer Klasse von Dichtefunktionen (auf der rechten Seite) sinnvoll verglichen werden kann.

[49] Wir nehmen im folgenden stets an, daß diese Dichtefunktionen für alle nicht negativen Werte von t definiert, ggf. $= 0$ sind. Die stetige Prozeßzeitachse wird mit $\mathcal{T}^*$ bezeichnet.

Um diese Frage zu erörtern, ist es zweckmäßig, zwei Aspekte zu unterscheiden. Folgende Formulierung kann helfen, sie sichtbar zu machen:

$$\mathrm{P}\,(T = t \mid X = x) \;\approx\; f^*(t \mid x) \;\approx\; \tilde{f}(t \mid x; \theta) \tag{3.29}$$

Auf der linken Seite steht die empirische Wahrscheinlichkeitsfunktion, die den Prozeß in der endlichen Grundgesamtheit Ω beschreibt. In der Mitte steht eine Dichtefunktion, die den Prozeß so, wie er in der Grundgesamtheit verläuft, theoretisch repräsentieren soll. Auf der rechten Seite steht eine Klasse von Dichtefunktionen, mit deren Hilfe ein vereinfachendes Modell für $f^*(t\,|\,x)$ erreicht werden soll. Da sowohl $f^*(t\,|\,x)$ als auch $\tilde{f}(t\,|\,x;\theta)$ Dichtefunktionen sind, können sie mit einer stetigen Variante des KL-Distanzmaßes verglichen werden. Dieser Vergleich kann konzeptionell vollständig analog zur Modellbildung auf einer diskreten Zeitachse vorgenommen werden.

Schwieriger ist die Frage, wie man sich das Verhältnis zwischen der den Prozeßverlauf empirisch beschreibenden Wahrscheinlichkeitsfunktion $\mathrm{P}(T = x\,|\,X = x)$ und der Dichtefunktion $f^*(t\,|\,x)$, die den Prozeßverlauf „theoretisch repräsentieren“ soll, vorstellen kann. Zwei unterschiedliche Betrachtungsweisen sind möglich.

1. Zunächst die übliche Betrachtungsweise, die auf der Annahme beruht, daß sich die Theoriebildung auf eine *unendliche* Grundgesamtheit bezieht, die durch stetige Verteilungsmodelle der Form $f^*(t\,|\,x)$ angemessen repräsentiert werden kann.[50] Bei dieser Betrachtungsweise wird $f^*(t\,|\,x)$ nicht als eine Approximation zur Beschreibung einer diskreten Wahrscheinlichkeitsverteilung in einer endlichen Grundgesamtheit aufgefaßt, vielmehr wird angenommen, daß die real existierende endliche Grundgesamtheit gemäß $f^*(t\,|\,x)$ (und einer analog konzipierten Wahrscheinlichkeitsverteilung für den Kovariablenvektor X) *erzeugt* worden ist, also eine Zufallsstichprobe aus einem hypothetisch unterstellten Verteilungsmodell ist.

   Diese Basisannahme wird üblicherweise mit einer zweiten Annahme verbunden, daß die empirisch verfügbaren Daten exakt gemessene Realisationen der als stetig unterstellten Zufallsvariablen sind.

2. Eine alternative Betrachtungsweise des Schätzproblems ergibt sich daraus, daß $f^*(t\,|\,x)$ als ein theoretisches Modell zur *Approximation* einer diskreten Verteilung in einer endlichen Grundgesamtheit aufgefaßt wird. Die statistische Modellbildung besteht dann gewissermaßen in einer zweistufigen Approximation. Zunächst wird eine als gegeben vorausgesetzte diskrete Wahrscheinlichkeitsverteilung durch ein stetiges Verteilungsmodell approximiert, dann wird für dieses Verteilungsmodell eine (ebenfalls stetige) vereinfachende Darstellung gesucht.

[50] Diese Annahme wurde zum erstenmal von Venn [1888, S. 109] systematisch formuliert, dann vor allem von R. A. Fisher in die statistische Theorie eingeführt, vgl. zum Beispiel Fisher [1925] und [1970, S.4f]. Sie erscheint plausibel, wenn sich die statistische Modellbildung auf Zufallsexperimente beziehen läßt, die hinreichend oft wiederholbar sind, um die hypothetische Bezugnahme auf die Vorstellung einer beliebigen Wiederholbarkeit sinnvoll zu machen. Im Hinblick auf statistische Modelle zur Beschreibung nicht wiederholbarer Lebensverläufe ist diese Annahme natürlich problematisch.

Wie sich zeigen wird, führen beide Betrachtungsweisen zu approximativ ähnlichen Schätzverfahren. Insbesondere liefert die unter (2) genannte Betrachtungsweise auch einen Zugang zum Problem, daß die praktisch verfügbaren Informationen nicht der theoretisch angemessenen Meßgenauigkeit entsprechen.

3.5.1 Exakt erfaßte Ereigniszeitpunkte

Ich beginne mit der ersten der beiden oben unterschiedenen Betrachtungsweisen. Ausgehend von (3.29) besteht die Aufgabe darin, einen optimalen Parametervektor $\hat{\theta}$ zu finden, so daß $\tilde{f}(t\,|\,x;\hat{\theta})$ eine im Rahmen der gegebenen Modellklasse optimale Approximation für $f^*(t\,|\,x)$ ist. Das dafür verfügbare empirische Wissen besteht in den Beobachtungen (t_i, x_i) für die Individuen aus einer endlichen Gesamtheit $\Omega = \{\omega_i, \ldots, \omega_N\}$.[51]

Hätte man eine vollständige Information über $f^*(t\,|\,x)$, könnte diese Dichtefunktion unmittelbar mit den korrespondierenden Dichtefunktionen aus der vorgegebenen Modellklasse verglichen werden. Das KL-Distanzmaß könnte dann in folgender Modifikation verwendet werden:

$$D^*(\theta) = \sum_{x \in \mathcal{X}} \mathrm{P}\,(X = x) \int_0^\infty f^*(t \mid x)\, \log\left\{ \frac{f^*(t \mid x)}{\tilde{f}(t \mid x; \theta)} \right\} \mathrm{d}t$$

Wie im diskreten Fall ist eine Minimierung dieser Distanz äquivalent zur Maximierung der Log-Likelihood

$$\ell^*(\theta) = \sum_{x \in \mathcal{X}} \mathrm{P}\,(X = x) \int_0^\infty f^*(t \mid x)\, \log\left\{ \tilde{f}(t \mid x; \theta) \right\} \mathrm{d}t \tag{3.30}$$

Allerdings sind nur endlich viele Beobachtungen (t_i, x_i) verfügbar. Es muß also überlegt werden, wie mithilfe dieser Beobachtungen das Integral in (3.30) approximiert werden kann.

Die Approximation kann auf unterschiedliche Weisen dargestellt werden. Bezieht man sich auf die Theorie der Riemann-Stieltjes-Integrale, kann man das Integral in (3.30) als einen Grenzwert darstellen. Ausgangspunkt ist eine im Prinzip beliebige Einteilung der Zeitachse in Zeitintervalle, zum Beispiel abgegrenzt durch Zeitpunkte

$$\tau_j = j\,\Delta \qquad j = 0, 1, 2, \ldots$$

[51] Ich verwende die Bezeichnungen $t_i = T(\omega_i)$ und $x_i = X(\omega_i)$ für $i = 1, \ldots, N$.

In diesem Fall erhält man für jedes $\Delta > 0$ eine Zerlegung der Zeitachse in Intervalle der Länge Δ, und das Integral kann auf folgende Weise als ein Grenzwert dargestellt werden:[52]

$$\int_0^\infty f^*(t \mid x) \log\left\{\tilde{f}(t \mid x;\theta)\right\} \mathrm{d}t = \tag{3.31}$$
$$\lim_{\Delta\to 0} \sum_{j=1}^{\infty} \left(F^*(\tau_j \mid x) - F^*(\tau_{j-1} \mid x)\right) \log\left\{\tilde{f}(\tau_j \mid x;\theta)\right\}$$

Diese Darstellung kann verwendet werden, um das Integral durch die empirisch verfügbaren Beobachtungen (t_i, x_i) zu approximieren. Zunächst erhält man eine Approximation der theoretischen Verteilungsfunktion $F^*(t \mid x)$ durch die aus den Daten ermittelbare empirische Verteilungsfunktion

$$F(t \mid x) \equiv \text{ Anteil der Individuen in } \Omega \text{ mit } t_i \le t \text{ und } x_i = x$$

Dies ist eine Treppenfunktion, die sich nur bei den empirisch beobachteten Ereigniszeitpunkten verändert. Werden diese Ereigniszeitpunkte durch t_j $(j = 1, \ldots, N)$ bezeichnet, erhält man als Approximation des Integrals (3.31) die Formulierung

$$\int_0^\infty f^*(t \mid x) \log\left\{\tilde{f}(t \mid x;\theta)\right\} \mathrm{d}t \approx$$
$$\sum_{j=1}^{N} \left(F(t_j \mid x) - F(t_{j-1} \mid x)\right) \log\left\{\tilde{f}(t_j \mid x;\theta)\right\}$$

$F(t_j \,|\, x) - F(t_{j-1} \,|\, x)$ ist der Anteil der Individuen in Ω mit einer Verweildauer t_j. Also kann man die Approximation (3.32) auch folgendermaßen schreiben:

$$\int_0^\infty f^*(t \mid x) \log\left\{\tilde{f}(t \mid x;\theta)\right\} \mathrm{d}t \approx \sum_{i=1}^{N} \log\left\{\tilde{f}(t_i \mid x_i;\theta)\right\}$$

Schließlich kann man diese Approximation in (3.30) einsetzen und erhält als Kriterium für die Modellschätzung die Log-Likelihood

$$\ell(\theta) = \sum_{i=1}^{N} \log\left\{\tilde{f}(t_i \mid x_i;\theta)\right\}$$

Dies ist die übliche Formulierung für die Log-Likelihood des Modells $\tilde{f}(t \,|\, x;\theta)$ im Hinblick auf die verfügbaren Daten. Sie kann auf die gleiche Weise zur Modellschätzung verwendet werden, wie es in Abschnitt 3.3 für diskrete Modelle beschrieben worden ist. Analog gelten auch alle Bemerkungen über die erforderlichen Modifikationen der Log-Likelihood zur Berücksichtigung unvollständiger Beobachtungen.

[52] $F^*(t)$ bezeichnet die zu $f^*(t)$ korrespondierende Verteilungsfunktion.

3.5.2 Ungenau erfaßte Beobachtungszeitpunkte

Folgt man der oben unter (2) genannten Betrachtungsweise, ist das theoretische Modell $f^*(t \,|\, x)$ eine idealisierende Approximation für die empirische Verteilung der Verweildauern in der endlichen Grundgesamtheit, also

$$\mathrm{P}\left(T = t \mid X = x\right) \approx f^*(t \mid x)\,\mathrm{d}t \qquad t \in \mathcal{T}^*$$

Wie kann diese Approximation begrifflich gefaßt werden? Als erste Möglichkeit kann man eine Approximation auf der Ebene der Verteilungsfunktionen in Betracht ziehen, also

$$F(t \mid x) \approx F^*(t \mid x)$$

Die theoretische Verteilungsfunktion $F^*(t \,|\, x)$, die der stetigen Dichte $f^*(t \,|\, x)$ entspricht, wird dann als Approximation der in der endlichen Grundgesamtheit empirisch gegebenen diskreten Verteilungsfunktion $F(t \,|\, x)$, eine Treppenfunktion mit endlich vielen Sprungstellen, angesehen.

Diese Interpretation führt offensichtlich zu genau dem gleichen Schätzverfahren, wie es oben dargestellt worden ist. Dies ist zwar mathematisch trivial, jedoch unter theoretischen Gesichtspunkten wichtig. Es zeigt nämlich, daß das Standard-Schätzverfahren (ausgehend vom KL-Distanzmaß oder direkt vom Maximum-Likelihood-Prinzip) nicht unbedingt die metaphysische Annahme einer unendlichen Grundgesamtheit (von Individuen) und eines für sie geltenden Verteilungsmodells voraussetzt. Es kann gleichermaßen als ein Schätzverfahren für eine stetige Approximation an eine empirisch gegebene diskrete Verteilung aufgefaßt werden.

Ein alternatives Schätzverfahren entsteht erst dann, wenn man die Interpretation der beobachteten Ereigniszeitpunkte bzw. Verweildauern ändert. Bisher haben wir angenommen, daß sich die Ereigniszeitpunkte durch Zahlenangaben auf einer stetigen Zeitachse exakt repräsentieren lassen. Als eine alternative Betrachtungsweise kann man sich vorstellen, daß ein Ereigniszeitpunkt nur durch die Angabe eines Zeitintervalls beschrieben werden kann, in dem das fragliche Ereignis stattgefunden hat. Wie bereits in Abschnitt 2.1 bemerkt worden ist, kann diese Betrachtungsweise durch zwei Überlegungen begründet werden. Einerseits kann man daran denken, daß Ereigniszeitpunkte nur ungenau erfaßt werden können; andererseits kann man sagen, daß soziologisch relevante Ereignisse eine inhärente zeitliche Ausdehnung haben. Beide Überlegungen schließen sich nicht aus, insofern kann das im folgenden darzustellende Schätzverfahren insbesondere bei ungenau erfaßten Ereigniszeitpunkten verwendet werden.[53]

[53] In der Literatur wird diese Problemstellung üblicherweise unter dem Gesichtspunkt diskutiert, daß nur ungenaue Beobachtungen über die zeitliche Dauer von Episoden verfügbar sind, d.h. man kennt nur gewisse zeitliche Intervalle, in denen Ereignisse

Wichtig ist, daß sich durch diese alternative Betrachtungsweise die Bedeutung des theoretischen Modells verändert. Seine Bedeutung liegt dann nicht mehr darin, daß es exakt erfaßbare Ereigniszeitpunkte generieren kann, vielmehr darin, daß es als ein theoretisches Hilfsmittel angesehen werden kann, um die Vorstellung begrifflich zu fassen, daß Ereignisse nur in gewissen zeitlichen Intervallen stattfinden bzw. beobachtet werden können.

Um diese Überlegung formal zu präzisieren, wird der empirische Prozeß durch folgende Zufallsvariablen definiert:

$$(T', T'', X) : \Omega \longrightarrow \mathcal{T} \times \mathcal{T} \times \mathcal{X}$$

$\mathcal{T}$ ist wahlweise eine diskrete oder stetige Prozeßzeitachse. Die Zufallsvariablen T' und T'' sind die Anfangs- bzw. Endpunkte für die Ereignisse, X ist wie bisher eine prozeßunabhängig definierbare Klassifizierungsvariable. Die empirischen bedingten Ereigniswahrscheinlichkeiten sind durch

$$\mathrm{P}\left(T \in [t', t''] \mid X = x\right)$$

gegeben, also Wahrscheinlichkeiten dafür, daß unter der Bedingung $X = x$ ein Ereignis im Intervall $[t', t'']$ stattfindet. Wie gesagt, sind zwei unterschiedliche, sich jedoch nicht ausschließende Interpretationen möglich. Man kann annehmen, daß das Ereignis *während* des gesamten Intervalls $[t', t'']$ stattfindet, oder man kann annehmen, daß bei einer genaueren Beobachtung ein Zeitpunkt oder ein Subintervall innerhalb von $[t', t'']$ bestimmt werden könnte, zu dem bzw. in dem das Ereignis stattfindet, daß jedoch infolge von Beobachtungsungenauigkeiten nur gesagt werden kann, daß das Ereignis *irgendwann* innerhalb des Intervalls $[t', t'']$ stattfindet.

Die Modellschätzung kann unabhängig von der jeweils für sinnvoll erachteten Interpretation konzipiert werden. Im Mittelpunkt steht jetzt der Vergleich

$$\mathrm{P}\left(T \in [t', t''] \mid X = x\right) \approx \int_{t'}^{t''} \tilde{f}(\tau \mid x; \theta)\, \mathrm{d}\tau$$

Wird der Vergleich mit dem KL-Distanzmaß vorgenommen, erhält man die Log-Likelihood

$$\bar{\ell}(\theta) = \sum_{i=1}^{N} \log \left\{ \int_{t_i'}^{t_i''} \tilde{f}(\tau \mid x_i; \theta)\, \mathrm{d}\tau \right\} \tag{3.32}$$

Das Integral kann auch durch Survivor- oder Verteilungsfunktionen dargestellt werden. Bezeichnet $\tilde{G}(t \mid x; \theta)$ die durch das Modell angenommene

eingetreten oder Beobachtungen zensiert sind. Dazu liegen bereits zahlreiche Arbeiten vor, u.a. Thompson [1977], Bartlett [1978], Prentice und Gloeckler [1978], Pierce et al. [1979], Lawless [1982, S. 259ff], Aranda-Ordaz [1983], Hutchison [1987].

Survivorfunktion, liefert

$$\bar{\ell}(\theta) = \sum_{i=1}^{N} \log \left\{ \tilde{G}(t_i' \mid x_i; \theta) - \tilde{G}(t_i'' \mid x_i; \theta) \right\} \tag{3.33}$$

ein zu (3.32) äquivalentes Kriterium für die Modellschätzung.

a) Um rechts zensierte Beobachtungen berücksichtigen zu können, muß die in (3.33) angegebene Log-Likelihood modifiziert werden. Grundsätzlich kann so vorgegangen werden, wie in Abschnitt 3.4.1 beschrieben worden ist, d.h. die empirische Wahrscheinlichkeit, daß bis zum Ende des Beobachtungszeitraums kein Ereignis eingetreten ist, wird mit der durch die Modellannahme gegebenen Survivorfunktion verglichen. Allerdings muß dann entschieden werden, zu welchem Zeitpunkt die theoretisch angenommene Survivorfunktion berechnet werden soll. Es gibt verschiedene Möglichkeiten: man kann den Anfangspunkt des Intervalls, den Endpunkt oder einen mittleren Wert verwenden; ein objektives Kriterium gibt es nicht.[54] Wenn es sich um ungenau erfaßte Zensierungszeitpunkte handelt, erscheint es am sinnvollsten, alle drei Möglichkeiten zu verfolgen, um auf diese Weise Anhaltspunkte zur Beurteilung der Auswirkungen der Ungenauigkeit der Beobachtungen auf die Schätzergebnisse zu bekommen.

b) Bei der Formulierung der Log-Likelihood (3.33) wurde davon ausgegangen, daß für jede individuelle Beobachtung ein spezielles Intervall $[t_i', t_i'']$ verwendet werden kann. Als einen Spezialfall erhält man einen Schätzansatz für eine Situation, in der die Prozeßzeitachse in feste Zeitintervalle eingeteilt ist und die verfügbaren Beobachtungen sich auf diese Zeitintervalle beziehen, d.h. man weiß für jedes Individuum, in welchem dieser Zeitintervalle ein Ereignis stattgefunden hat oder die Beobachtung rechts zensiert ist. Wenn alle Zeitintervalle die gleiche Länge haben, kann man sie formal mit den Zeitpunkten einer diskreten Zeitachse identifizieren.

c) Wir haben angenommen, daß nur die Ereignis- bzw. Zensierungszeitpunkte ungenau beobachtet werden können. Wesentlich komplizierter wird es, einen geeigneten Schätzansatz zu finden, wenn auch die Anfangszeitpunkte für den Episodenverlauf nur ungenau ermittelt werden können. Wenn die Intervalle sehr kurz sind, kann man ad-hoc-Näherungen verwenden oder, wenn alle Intervalle die gleiche Länge haben, ein diskretes Modell verwenden. Wenn jedoch die Intervalle vergleichsweise groß sind, entsteht eine Situation links zensierter Beobachtungen (vgl. Abschnitt 3.4.3) und man kommt ohne explizite Verteilungsannahmen nicht weiter.[55]

[54] Vgl. die Diskussion dieser Frage bei Thompson [1977, S. 465], Lawless [1982, S. 261].

[55] Vgl. die Diskussion bei Galler [1985], Wurzel [1988a, 1988b].

3.5.3 Bemerkungen zur Modellidentifikation

Zur Modellbildung für empirisch beobachtbare Prozeßverläufe werden fast immer Klassen theoretischer Modelle verwendet, die für die gesamte positive Zeitachse definiert sind. Würde man diese Modelle ohne Vorbehalt akzeptieren, gäbe es für jeden beliebigen Zeitpunkt eine positive Wahrscheinlichkeit dafür, daß einige Individuen bis zu diesem Zeitpunkt den Ausgangszustand noch nicht verlassen haben. Im Hinblick auf menschliche Lebensverläufe ist das natürlich unsinnig. Man sieht jedoch auf einfache Weise, daß dieses Problem nur die Modellinterpretation betrifft.

Für jede Grundgesamtheit (oder Stichprobe) Ω können nämlich zwei Zeitpunkte definiert werden: t_a ist der Zeitpunkt bei dem zum erstenmal ein Individuum aus Ω den Ausgangszustand verläßt, t_b ist der Zeitpunkt, bei dem zum letztenmal ein Individuum den Ausgangszustand verläßt. Zu jeder Klasse theoretischer Modelle $\tilde{f}(t \mid x; \theta)$ kann dann eine korrespondierende Klasse *gestutzter*, auf das Intervall $[t_a, t_b]$ eingeschränkter Verteilungen definiert werden:

$$\tilde{f}_{t_a,t_b}(t \mid x; \theta) \quad = \quad \begin{cases} \tilde{f}(t \mid x; \theta) \big/ \int_{t_a}^{t_b} \tilde{f}(\tau \mid x; \theta)\, \mathrm{d}\tau & \text{wenn } t_a \le t \le t_b \\ 0 & \text{andernfalls} \end{cases}$$

Wie man sich leicht überzeugt, kann jedoch mithilfe der verfügbaren Daten zwischen den Modellen $\tilde{f}(t \mid x; \theta)$ und $\tilde{f}_{t_a,t_b}(t \mid x; \theta)$ nicht unterschieden werden. Gleichgültig, von welcher der beiden Modellklassen man ausgeht, man erhält den gleichen Schätzwert $\hat{\theta}$. Dies bedeutet, daß das geschätzte Modell außerhalb des Intervalls $[t_a, t_b]$ jedenfalls nicht deskriptiv interpretiert werden kann. Im Hinblick darauf, daß das Modell der Beschreibung eines empirisch erfaßbaren Prozesses dienen soll, ist es infolgedessen zweckmäßig, die Interpretation auf dieses Zeitintervall einzuschränken.

Die Tatsache, daß ein Modell geschätzt werden kann, bedeutet also nicht, daß zugleich alle Aspekte des Modells sinnvoll interpretiert werden können. Zum Beispiel sagen Wu und Tuma [1991a, S. 358]: „An often neglected aspect of a hazard model is its implication for the probability of *ever* experiencing the event of interest.“ Als Beispiel beziehen sie sich auf das Ereignis *erste Heirat*, und sie stellen fest, daß die üblicherweise verwendeten parametrischen Übergangsratenmodelle implizieren, daß schließlich alle Personen heiraten (was vermutlich falsch ist). Ich glaube jedoch, daß die angeführte Bemerkung irreführend ist, denn in der Regel ist die Wahrscheinlichkeit dafür, ob ein gewisser Zustand schließlich erreicht wird, mit den verfügbaren Daten nicht ermittelbar und mithin auch aus einer Modellschätzung auf der Grundlage dieser Daten nicht ableitbar. Sogenannte „defekte“ Verteilungen, bei denen die Survivorfunktion schließlich nicht gegen Null konvergiert, bieten zur Lösung dieses Problems keine Vorteile.[56]

[56] Zum Beispiel schreibt Diekmann [1990, S. 177]: „As in the Hernes model and unlike in the log-logistic model, the sickle model has a defective distribution, in that it allows for a proportion of ‘immune’ cases in the population (Hernes, 1972). This property is sometimes desirable, particularly for the analysis of divorce data, where a large proportion of the married population does not experience an event.“ Diese Aussage sollte jedoch relativiert werden, wenn man daran denkt, daß jede gestutzte Verteilung zu einer defekten Verteilung fortgesetzt werden kann. Das eigentlich interessante Problem an dieser Stelle besteht nicht darin, ob gewöhnliche oder gestutzte oder defekte Verteilungen angemessener sind, um den Verlauf einer Episode zu beschreiben, sondern darin, daß bei Vorhandensein rechts zensierter Daten keine verläßlichen Aussagen über den Anteil derjenigen Personen getroffen werden können, die nicht in den vorgegebenen Zielzustand wechseln. Dies ist jedoch kein Schätzproblem, sondern nur ein Reflex der Tatsache, daß in einer Situation konkurrierender Risiken weder zu Beginn noch während des Verlaufs determiniert ist, welcher der möglichen Übergänge schließlich realisiert wird.

Ein theoretisch angemessenes Modell für eine Situation, in der nicht von vornherein erwartet werden kann, daß alle Individuen schließlich in einen bestimmten Endzustand überwechseln, erhält man durch die Betrachtung konkurrierender Risiken. Zumindest die Alternative, daß ein Mensch sterben kann, bevor er den jeweils zu beschreibenden Zustandswechsel vollzogen hat, sollte berücksichtigt werden. Allerdings gilt die Feststellung, daß mit den üblicherweise verfügbaren Lebensverlaufsdaten (d.h. wenn es rechts zensierte Beobachtungen gibt) nur begrenzte Schlußfolgerungen gezogen werden können, auch für Modelle mit konkurrierenden Risiken.

3.5.4 Beispiele

Um die in den vorangegangenen Abschnitten besprochenen Methoden der Modellschätzung zu illustrieren, werden als Beispiel noch einmal die SOEP-Daten für den Lebensverlauf bis zur ersten Heirat betrachtet. Wie bisher verwenden wir Informationen über Personen, die der Teilstichprobe A angehören, im Zeitraum 1918 – 1968 geboren wurden und an mindestens den ersten drei Wellen teilgenommen haben. Als Klassifizierungsvariable werden die in Abschnitt 3.4 (Tabelle 3.12) definierten Geburtskohorten verwendet.

a) Als erstes Beispiel wird eine von J. Brüderl vorgeschlagene Erweiterung des herkömmlichen log-logistischen Modell betrachtet, die hinreichend flexibel erscheint, um die Daten angemessen beschreiben zu können.[57] Die Übergangsrate, zunächst ohne Berücksichtigung von Kovariablen, sieht bei dieser Modellklasse folgendermaßen aus:

$$\tilde{r}(t; \alpha, \beta, \gamma) = \gamma \, \frac{\beta \, (\alpha \, t)^{\beta - 1}}{1 + (\alpha \, t)^{\beta}} \tag{3.34}$$

Das Modell hat drei eindimensionale Parameter: α, β und γ. Jeder dieser Parameter kann über eine Link-Funktion mit Kovariablen verknüpft werden. In unserem Beispiel haben wir jedoch nur eine einfache Klassifizierung entsprechend den fünf unterschiedlichen Geburtskohorten. Es liegt also nahe, für jede dieser Kohorten ein eigenes Modell zu schätzen. Die

[57] Vgl. Brüderl und Diekmann [1994a, 1994b].

Prozeßzeit wird in Jahren seit Beginn des 15. Lebensjahrs gemessen. Man erhält dann mit einer Maximum-Likelihood-Schätzung unter Berücksichtigung rechts zensierter Beobachtungen die in Tabelle (3.35) angegebenen Schätzergebnisse:[58]

Kohorte	$\hat{\alpha}$	SE($\hat{\alpha}$)	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\gamma}$	SE($\hat{\gamma}$)
$1918-27$	-2.1483	0.0342	1.4087	0.0485	-2.7796	0.0597
$1928-37$	-1.9628	0.0255	1.6226	0.0469	-2.7920	0.0581
$1938-47$	-1.8707	0.0263	1.5445	0.0444	-2.6983	0.0539
$1948-57$	-1.8070	0.0388	1.3204	0.0493	-2.6052	0.0572
$1958-68$	-2.2084	0.1002	1.1083	0.0630	-2.8434	0.1062

(3.35)

Wie sind diese Schätzergebnisse zu interpretieren? Die geschätzten Standardfehler (SE) sollen zunächst ignoriert werden. Aus den geschätzten Parametern ($\hat{\alpha}$, $\hat{\beta}$ und $\hat{\gamma}$) lassen sich, indem man sie in (3.34) einsetzt, Formulierungen für die kohortenspezifischen Übergangsraten und Survivorfunktionen berechnen. Abbildung 3.5.1 zeigt die auf diese Weise geschätzten Survivorfunktionen, Abbildung 3.5.2 zeigt die entsprechenden Übergangsraten.

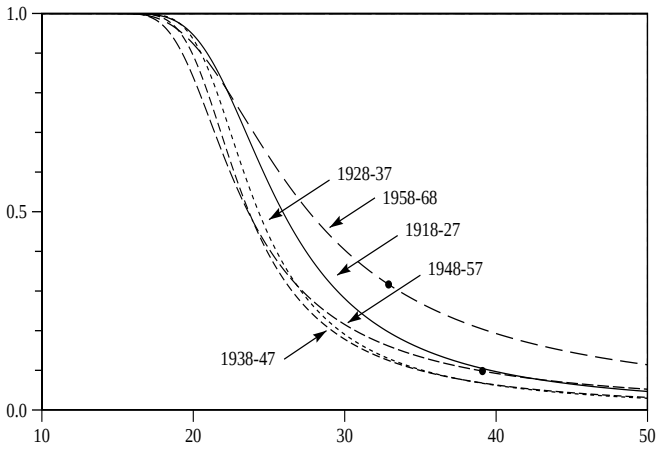


Abbildung 3.5.1 Schätzung kohortenspezifischer Survivorfunktionen für das Ereignis *erste Heirat* mit einem erweiterten log-logistischen Modell. Teilstichprobe A des SOEP. Abszisse in Jahren.

[58] Die Schätzung wurde mit dem Programm TDA durchgeführt. Die in Tabelle 3.35 angegebenen Schätzwerte beziehen sich auf die Anti-Logarithmen der in (3.34) verwendeten Modellparameter.

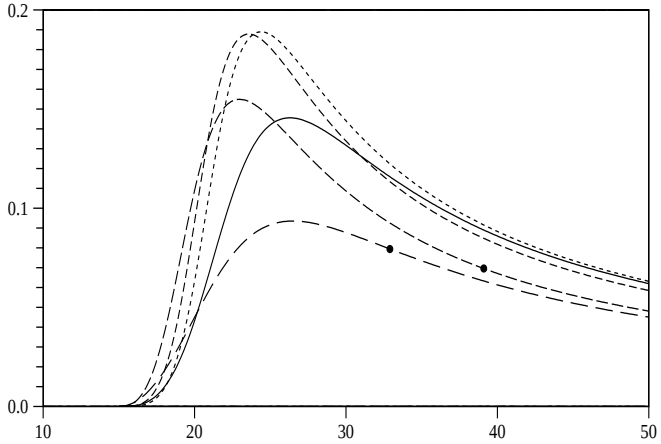


Abbildung 3.5.2 Schätzung kohortenspezifischer Übergangsraten für das Ereignis *erste Heirat* mit einem erweiterten log-logistischen Modell. Teilstichprobe A des SOEP. Abszisse in Jahren. Linientypen wie in Abbildung 3.5.1.

Es stellt sich natürlich die Frage, ob die geschätzten Modelle eine sinnvolle vereinfachende Beschreibung der über den Prozeß vorliegenden Daten liefern. Auf den ersten Blick erscheinen die geschätzten Survivorfunktionen sehr ähnlich zu ihren Kaplan-Meier-Schätzungen in Abbildung 3.4.2. Diese Frage bedarf jedoch einer näheren Prüfung; darauf wird in Abschnitt 3.7 etwas näher eingegangen.

Ein zweiter Punkt betrifft das in Abschnitt 3.5.3 angesprochene Identifikationsproblem. Die Abbildungen 3.5.1 und 3.5.2 präsentieren die Schätzergebnisse für eine Prozeßzeitachse vom 15. bis zum 50. Lebensjahr, und natürlich könnten die Kurven auch noch weiter fortgesetzt werden. Tatsächlich werden sie jedoch nur in einem beschränkten Altersintervall durch die verfügbaren Daten gestützt. Für die zwei jünsten Geburtskohorten ist dies in den Abbildungen durch Punkte markiert worden.

Kohorte	Personen	zensiert	t_a	t_b
$1918-27$	908	46	16	61
$1928-37$	1209	54	17	55
$1938-47$	1400	81	16	51
$1948-57$	1415	185	16	39
$1958-68$	1484	805	16	33

(3.36)

Tabelle (3.36) zeigt für die fünf Kohorten die Zeitpunkte t_a und t_b: t_a ist das Alter, bei dem zum erstenmal, t_b ist das Alter, bei dem zum letztenmal eine Heirat beobachtet werden kann. Wenn mit der Modellschätzung

ein deskriptiver Anspruch vertreten wird, sollte sich die Interpretation auf diese Zeitspannen beschränken. Natürlich ist es möglich, die Ergebnisse der Modellschätzung zur Formulierung von *Vermutungen* über das Heiratsverhalten der Kohorten auch außerhalb der jeweils verfügbaren Daten zu verwenden. Ein möglicher Sinn der Modellbildung liegt sicherlich darin, die jeweils verfügbaren Daten zur Formulierung solcher Vermutungen zu nutzen. Aber die Beschreibung eines beobachteten Sachverhalts sollte von Vermutungen über nicht beobachtete Sachverhalte unterschieden werden.

b) Eine wichtige Aufgabe der Modellschätzung liegt darin, Einsichten in den zeitlichen Verlauf eines Prozesses zu vermitteln, d.h. in die Zeitabhängigkeit der Übergangsraten. Wie in Abschnitt 2.4.3 diskutiert worden ist, verstehen wir diese Abhängigkeit der Übergangsraten von der Zeit zunächst rein deskriptiv. Abbildung 3.5.2 zeigt anhand unseres Beispiels die mit einem erweiterten log-logistischen Modell geschätzten Verläufe der Übergangsraten. Diese Ratenverläufe sind jedoch, bis zu einem gewissen Grad, durch die verwendete Modellklasse bedingt. Denn jedes parametrische Modell kann nur eine beschränkte Menge von Ratenverläufen zum Ausdruck bringen.

Einen einfachen Ausweg aus dieser Schwierigkeit liefert eine Modellklasse, bei der die Prozeßzeitachse in Zeitintervalle eingeteilt und angenommen wird, daß es in jedem Zeitintervall eine unterschiedliche, jedoch konstante Übergangsrate geben kann. Ich nenne dies im folgenden ein *Exponentialmodell mit Zeitperioden*.[59] Ausgangspunkt für die Modellbildung ist eine Einteilung der Prozeßzeitachse in Zeitperioden:

$$0 = \tau_1 < \tau_2 < \ldots < \tau_K < \tau_{K+1} = \infty$$

Dies liefert K Zeitintervalle

$$I_k = \{t \mid \tau_k \leq t < \tau_{k+1}\} \qquad k = 1, \ldots, K$$

Die Bildung solcher Zeitintervalle kann grundsätzlich beliebig vorgenommen werden. Je mehr Intervalle gebildet werden, desto besser wird die Approximation an den empirischen Prozeß. Allerdings wächst mit der Anzahl der Zeitintervalle die Anzahl der zu schätzenden Modellparameter, so daß es bei praktischen Anwendungen stets Grenzen gibt. Insbesondere müssen in jedem Zeitintervall hinreichend viele Ereigniszeitpunkte liegen, damit das Modell noch geschätzt werden kann.[60]

[59] In der englischsprachigen Literatur wird häufig von einem „piecewise constant exponential model“ gesprochen. Das Modell ist bereits oft dargestellt worden, vgl. u.a. Brown [1975], Holford [1976], Friedman [1982], Blossfeld et al. [1989]. Es ist auch gezeigt worden, daß dieses Übergangsratenmodell als ein log-lineares Modell für Kontingenztafeln interpretiert und geschätzt werden kann, vgl. Holford [1980], Laird und Olivier [1981].

[60] Vgl. die Diskussion dieser Frage bei Brown [1975].

Hat man diese Zeitintervalle definiert, kann für jedes Intervall ein eigener Verlauf der Übergangsrate angenommen werden, im einfachsten Fall eine konstante Übergangsrate,[61] also

$$\tilde{r}(t; \theta_k) = \theta_k \quad \text{wenn} \quad t \in I_k$$

Dieser Modellansatz kann erweitert werden, um die Abhängigkeit von Kovariablen (Klassifizierungsvariablen) berücksichtigen zu können. Da Übergangsraten keine negativen Werte annehmen können, ist es üblich, eine exponentielle Linkfunktion zu verwenden. Der Modellansatz ist dann

$$\tilde{r}(t; \alpha_k, \beta) = \exp(\alpha_k + x\beta) \quad \text{wenn} \quad t \in I_k$$

Dies ist die Übergangsrate im Zeitintervall I_k. Durch die Parameter $\alpha_1, \ldots, \alpha_K$, im folgenden zu einem Parametervektor α zusammengefaßt, kann diese Rate in jedem Zeitintervall einen unterschiedlichen Verlauf annehmen. Außerdem hängt sie durch den Parametervektor β von der (ggf. mehrdimensionalen) Klassifizierungsvariablen X ab.[62]

Um den Modellansatz zu vervollständigen, muß noch die Survivorfunktion für die Verweildauer im Ausgangszustand berechnet werden. Mithilfe des Begriffs der bedingten Survivorfunktion kann dies einfach erreicht werden. Zur Vereinfachung der Notation soll folgende Definition für die bei der Rechnung zu berücksichtigenden Zeitintervalle verwendet werden:

$$\Delta_k(t) \quad = \quad \begin{cases} \tau_{k+1} - \tau_k & \text{wenn} \quad t \geq \tau_{k+1} \\ t - \tau_k & \text{wenn} \quad \tau_k < t \leq \tau_{k+1} \\ 0 & \text{wenn} \quad t < \tau_k \end{cases}$$

Für die Zeitintervalle, bei denen $t \geq \tau_{k+1}$ ist, erhält man folgende Formulierung für die bedingte Survivorfunktion:

$$\tilde{G}(\tau_{k+1} \mid \tau_k, x) = \exp\left\{ - \int_{\tau_k}^{\tau_{k+1}} \tilde{r}(\tau \mid x; \alpha_k, \beta)\, \mathrm{d}\tau \right\}$$

Für das Zeitintervall von τ_{k+1} bis t sind die Integrationsgrenzen entsprechend zu ändern. Da sich die Übergangsraten in den Zeitintervallen nicht ändern, können die Integrale leicht ausgerechnet werden:

$$\int_{\tau_k}^{\tau_{k+1}} \tilde{r}(\tau \mid x; \alpha_k, \beta)\, \mathrm{d}\tau = \exp(\alpha_k + x\beta)\ \Delta_k(t)$$

[61] Tatsächlich ist es nicht unbedingt erforderlich, in jedem Zeitintervall eine *konstante* Übergangsrate anzunehmen. Auch andere parametrische Übergangsratenmodelle können für eine in Intervalle zerlegte Zeitachse reformuliert werden. Noura und Read [1990] zeigen dies z.B. für ein Weibull-Modell.

[62] Hier nehmen wir an, daß die Abhängigkeit der Übergangsrate von der Klassifizierungsvariable unabhängig von den Zeitintervallen besteht. Das Modell kann jedoch verallgemeinert werden, so daß der Parametervektor β zwischen den Intervallen variieren kann; vgl. Blossfeld et al. [1989, S. 211ff].

und die Survivorfunktion für die gesamte Verweildauer im Ausgangszustand kann als ein Produkt der bedingten Survivorfunktionen folgendermaßen dargestellt werden:

$$\tilde{G}(t \mid x; \alpha, \beta) = \prod_{k=1}^{K} \exp\{-\exp(\alpha_k + x\beta)\ \Delta_k(t)\}$$

Schließlich kann eine Log-Likelihoodfunktion gebildet werden, durch deren Maximierung das Modell geschätzt werden kann. Für die Formulierung der Log-Likelihood wird angenommen, daß für N Individuen die Werte (t_i, δ_i, x_i) $(i = 1, \ldots, N)$ beobachtet werden können. t_i ist die beobachtete Verweildauer, δ_i ist ein Zensierungsindikator, der den Wert 0 bei rechts zensierten, den Wert 1 bei nicht rechts zensierten Beobachtungen annimmt, und x_i ist der beobachtete Wert der Klassifizierungsvariablen. Außerdem sei $k(t)$ der Index desjenigen Zeitintervalls, daß den Zeitpunkt t enthält. Da die für die Likelihoodformulierung erforderliche Dichtefunktion als Produkt der Übergangsrate und der Survivorfunktion geschrieben werden kann, erhält man zunächst

$$\ell(\alpha, \beta) = \sum_{i=1}^{N} \delta_i \log\left\{\tilde{r}(t_i \mid x_i; \alpha_{k(t_i)}, \beta)\right\} + \sum_{i=1}^{N} \log\left\{\tilde{G}(t_i \mid x_i; \alpha, \beta)\right\}$$

Durch Einsetzen der oben entwickelten Formulierungen für die Übergangsrate und die Survivorfunktion erhält man schließlich

$$\ell(\alpha, \beta) = \sum_{i=1}^{N} \delta_i \left(\alpha_{k(t_i)} + x_i\beta\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} \Delta_k(t_i) \exp(\alpha_k + x_i\beta)$$

Die Modellschätzung erfolgt, indem diese Log-Likelihood maximiert wird. Vorausgesetzt wird allerdings eine mehr oder weniger willkürliche Einteilung der Prozeßzeitachse in Zeitperioden. Zur Illustration der Modellschätzung gehen wir zunächst von Zeitperioden (Altersgruppen) aus, die jeweils eine Länge von 5 Jahren haben, beginnend mit dem 15. Lebensjahr. Tabelle (3.37) zeigt die Schätzergebnisse.[63] Jede Zeile entspricht einer Altersgruppe (Zeitperiode), jede Spalte einer Geburtskohorte. Die Angaben sind geschätzte Übergangsraten, berechnet durch $\exp(\hat{\alpha}_k)$.

[63] Für die praktische Durchführung der Modellschätzung wurde das Programm TDA verwendet.

	Kohorte				
Alter	1918 - 27	1928 - 37	1938 - 47	1948 - 57	1958 - 68
$15-20$	0.0098	0.0095	0.0147	0.0235	0.0123
$20-25$	0.0789	0.1385	0.1601	0.1492	0.0735
$25-30$	0.1970	0.2289	0.2231	0.1545	0.1014
$30-35$	0.1757	0.1351	0.1154	0.1002	0.0739
$35-40$	0.0818	0.0898	0.0571	0.0468	
$40-45$	0.0470	0.0206	0.0242		
$45-50$	0.0255	0.0155			
$50-$	0.0081				

(3.37)

Diese Tabelle liefert einen Einblick in die kohortenspezifischen Heiratsmuster. Die geschätzten Übergangsraten sind allerdings von den jeweils gewählten Zeitintervallen (Altersgruppen) abhängig. Es liegt nahe, diese Zeitintervalle möglichst klein zu machen, um eine immer bessere Approximation des tatsächlichen, durch die Daten gegebenen Verlaufs der Übergangsraten zu erreichen. Dies stößt jedoch bei den gewöhnlich verfügbaren Stichproben schnell an Grenzen. Abilldung 3.5.3 zeigt Schätzergebnisse, nur für die Kohorte der 1918 – 1927 geborenen Personen, bei drei unterschiedlichen Einteilungen der Zeitachse in Intervalle: 5 Jahre, 2 Jahre und 1 Jahr. Man erkennt, daß zunehmend stichprobenbedingte Fluktuationen auftreten, die kaum noch verläßlich interpretiert werden können.

Abbildung 3.5.3 zeigt allerdings auch, daß das erweiterte log-logistische Modell nur eine recht grobe Approximation des Verlaufs der Übergangsraten liefert. Es ist ersichtlich nicht in der Lage, die vergleichsweise starke Konzentration der Heiraten im Alter von Mitte 20 bis Anfang 30 angemessen zu repräsentieren. Dies ist eine generelle Schwäche einfacher parametrischer Übergangsratenmodelle. Für die Forschungspraxis ist das Exponentialmodell mit Zeitperioden in der Regel vorzuziehen.

c) Bisher haben wir für jede der fünf Geburtskohorten ein separates Modell geschätzt. Bei der praktischen Verwendung von Übergangsratenmodellen ist es demgegenüber üblich, nur ein Modell zu schätzen und dabei wichtig erscheinende Klassifikationen als Kovariablen einzubeziehen. Für unser Beispiel ist es naheliegend, fünf Indikatorvariablen (0/1-Variablen) zu definieren, um die Zugehörigkeit zu den Geburtskohorten zu erfassen; sie seien mit $X_1, \ldots, X_5$ bezeichnet. Vier dieser Indikatorvariablen können dann als Kovariablen in die Modellspezifikation aufgenommen werden, etwa folgendermaßen:

$$\tilde{r}(t; \alpha_k, \beta_2, \ldots, \beta_5) = \exp(\alpha_k + x_{2,i}\beta_2 + \cdots x_{5,i}\beta_5) \quad \text{wenn} \quad t \in I_k \tag{3.38}$$

Ein Modell dieser Art kann auf einfache Weise geschätzt und interpre-

tiert werden. Allerdings liegt diesem Modellansatz eine häufig problematische Vereinfachung zugrunde. Es wird implizit angenommen, daß die Übergangsraten in allen durch die Kovariablen unterschiedenen Teilgesamtheiten proportional verlaufen. Man sieht das unmittelbar anhand der Modellformulierung in (3.38). Wenn alle Kovariablen den Wert Null haben, liefert das Modell den Verlauf der Übergangsrate für die als Referenz dienende Geburtskohorte, hier die erste Kohorte, deren Mitglieder im Zeitraum 1918 – 1927 geboren wurden. Für alle anderen Geburtskohorten ergeben sich – und zwar durch die Modellspezifikation – dazu proportionale Verläufe der Übergangsraten. Wie jedoch Abbildung 3.5.2 zeigt, ist dies jedenfalls in unserem Beispiel nicht zutreffend.

Dies ist eine allgemeine Eigenschaft sog. proportionaler Übergangsratenmodelle. Diese Modelle können auf folgende Weise charakterisiert werden:

$$\tilde{r}(t \mid x; \alpha, \beta) = \tilde{r}_b(t; \alpha)\, g(x; \beta)$$

Die Übergangsrate kann in zwei Faktoren zerlegt werden. Der erste Faktor liefert eine Basis-Übergangsrate, die von einem Parametervektor α, jedoch nicht von den Kovariablen (x) abhängig ist. Der zweite Faktor ist eine beliebige Funktion der Kovariablen, jedoch unabhängig von der Zeit (t). Mit Modellen dieser Art kann offensichtlich nicht erfaßt werden, daß der Verlauf der Übergangsraten in einzelnen, durch die Kovariablen unterscheidbaren Teilgesamtheiten unterschiedlich sein kann.

Eine Möglichkeit, um dieses Problem im Rahmen des Exponentialmodells mit Zeitperioden zu lösen, besteht darin, den Modellansatz so zu erweitern, daß die Wirkung der Kovariablen in den einzelnen Zeitperioden unterschiedlich sein kann. In formaler Schreibweise, anknüpfend an die Formulierung in (3.38):

$$\tilde{r}(t; \alpha_k, \beta_{k,2}, \cdots, \beta_{k,5}) = \exp\left(\alpha_k + x_{2,i}\beta_{k,2} + \cdots x_{5,i}\beta_{k,5}\right) \quad \text{wenn} \quad t \in I_k \tag{3.39}$$

In diesem Fall erhielte man die gleichen Schätzergebnisse wie bei der Schätzung separater Modelle für jede der fünf Geburtskohorten. Gleichwohl bietet der in (3.39) formulierte Modellansatz Vorteile. Denn man kann auf einfache Weise zusätzliche Bedingungen formulieren, zum Beispiel die Annahme, daß gewisse Modellparameter identisch sind; und man kann bei Modellen, die eine große Anzahl von Kovariablen berücksichtigen sollen, annehmen und in gewissen Grenzen auch prüfen, daß nur ein Teil dieser Kovariablen nicht-proportionale Verläufe der Übergangsraten bedingt.

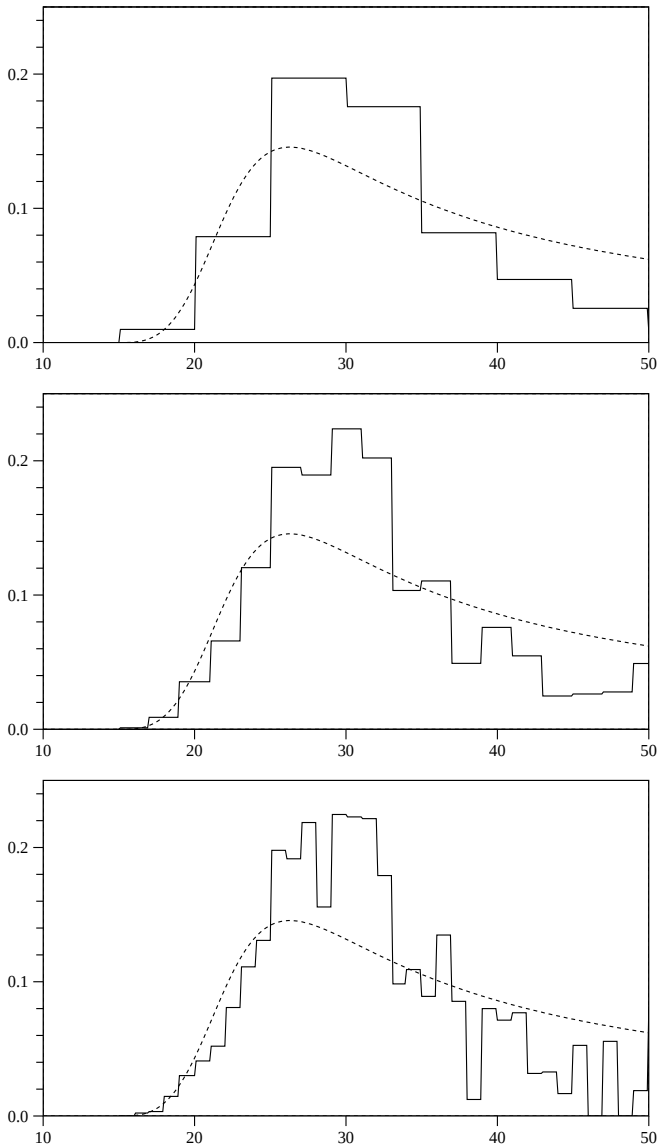


Abbildung 3.5.3 Übergangsraten für das Ereignis *erste Heirat* bei der Geburtskohorte 1918 – 1927, geschätzt mit einem erweiterten log-logistischen Modell und mit einem Exponentialmodell mit Zeitperioden; von oben nach unten: Periodenbreite 5 Jahre, 2 Jahre, 1 Jahr.

3.6 Episoden mit alternativen Zielzuständen

In den vorangegangenen Abschnitten wurde angenommen, daß die zu beschreibende Episode nur einen möglichen Endzustand hat. Strenggenommen ist diese Annahme nur dann sinnvoll, wenn als Zielzustand der Tod eines Individuums betrachtet wird, denn dies ist das einzige Ereignis, zu dem es keine Alternative gibt. Im allgemeinen ist davon auszugehen, daß es mehrere mögliche Zielzustände gibt, in die der Anfangszustand einer Episode verlassen werden kann, also von einer Situation konkurrierender Risiken.

Wie bereits in Abschnitt 2.4.4 dargestellt worden ist, kann dann der zu beschreibende Prozeß durch folgende Zufallsvariablen repräsentiert werden:

$$(T, D, X) : \Omega \longrightarrow \mathcal{T} \times \mathcal{D} \times \mathcal{X}$$

T ist die Verweildauer im Ausgangszustand der Episode, gemessen auf einer diskreten Prozeßzeitachse $\mathcal{T}$. D, mit Werten in der Menge $\mathcal{D}$, beschreibt den am Ende der Episode erreichten neuen Zustand. X ist eine Variable, mit der die Individuen unabhängig vom Episodenverlauf in Teilgesamtheiten klassifiziert werden können.

Als empirische Konzeption des Prozeßverlaufs hat man dann die bedingten Wahrscheinlichkeiten $\mathrm{P}(T = t, D = d \,|\, X = x)$; diese Wahrscheinlichkeiten sind additiv. Wir setzen außerdem einen vollständigen Zustandsraum voraus, d.h. jedes Individuum muß den Anfangszustand der Episode nach einer endlichen Verweildauer in einen der möglichen Endzustände verlassen.[64] Es gilt dann

$$\sum_{t \in \mathcal{T}} \sum_{d \in \mathcal{D}} \mathrm{P}\left(T = t, D = d \mid X = x\right) = 1$$

Der Prozeß kann auch durch zustandsspezifische Übergangsraten beschrieben werden. Auch dies wurde bereits in Abschnitt 2.4.4 behandelt. Um die Verweise zu erleichtern, werden hier noch einmal die wichtigsten Formeln zusammengestellt. Die zustandsspezifischen Übergangsraten werden durch

$$r_d(t \mid x) = \mathrm{P}\left(T = t, D = d \mid T \geq t, X = x\right)$$

definiert. Die unspezifische Abgangsrate ergibt sich aus der Summierung der zustandsspezifischen Übergangsraten:

$$r(t \mid x) = \sum_{d \in \mathcal{D}} r_d(t \mid x)$$

[64] Diese Bedingung ist insbesondere stets dann erfüllt, wenn der Tod als einer der möglichen Endzustände in der Definition von $\mathcal{D}$ berücksichtigt wird.

Die Survivorfunktion für das Verbleiben im Ausgangszustand der Episode ist

$$G(t \mid x) = \mathrm{P}\left(T > t \mid x\right) = \prod_{\tau=1}^{t} \left(1 - r(\tau \mid x)\right)$$

Und schließlich kann das Eintreten von Ereignissen durch eine zweidimensionale Wahrscheinlichkeitsfunktion beschrieben werden:

$$f(t, d \mid x) = \mathrm{P}\left(T = t, D = d \mid X = x\right) = r_d(t \mid x)\, G(t - 1 \mid x) \quad (3.40)$$

Die Modellbildung zielt darauf, eine vereinfachende Beschreibung des auf der diskreten Prozeßzeitachse $\mathcal{T}$ definierten empirischen Prozesses zu ermöglichen. Die Definition potentieller Modelle kann auf einer diskreten oder stetigen Zeitachse erfolgen. In beiden Fällen besteht der Ausgangspunkt darin, eine Klasse zustandsspezifischer Übergangsraten zu definieren, die zur vereinfachten Prozeßbeschreibung in Betracht gezogen werden sollen:[65]

$$\tilde{r}_d(t \mid x; \theta_d) \qquad \theta_d \in \Theta_d$$

a) Soll die Modellbildung auf einer diskreten Zeitachse erfolgen, ist $\tilde{r}_d(t \,|\, x; \theta)$ eine diskrete Übergangsrate, so daß daraus analog zu (3.40) eine zweidimensionale Wahrscheinlichkeitsfunktion

$$\tilde{f}(t, d \mid x; \theta) = \tilde{r}_d(t \mid x; \theta_d)\, \tilde{G}(t - 1 \mid x; \theta) \quad (3.41)$$

gebildet werden kann. Man beachte, daß für die zustandsspezifischen Übergangsraten ein jeweils spezifischer Parametervektor θ_d angenommen wird; die Wahrscheinlichkeitsfunktion $\tilde{f}$ und die Survivorfunktion $\tilde{G}$ hängen jedoch von allen zustandsspezifischen Übergangsraten ab, d.h. von einem Parametervektor θ, der als Komponenten die Parametervektoren θ_d enthält.[66] Der Parameterraum für θ wird mit Θ bezeichnet.

Im Mittelpunkt der Modellbildung steht dann der Vergleich

$$\mathrm{P}\left(T = t, D = d \mid X = x\right) = f(t, d \mid x) \;\approx\; \tilde{f}(t, d \mid x; \theta)$$

Auf der linken Seite steht die den Prozeß empirisch beschreibende Wahrscheinlichkeitsfunktion, auf der rechten Seite steht eine Klasse potentieller Modelle, ebenfalls diskrete Wahrscheinlichkeitsfunktionen. Der Vergleich

[65] Wie bisher verwenden wir das Tilde-Zeichen, um anzudeuten, daß es sich um Modelle handelt.

[66] Ggf. kann durch zusätzliche Bedingungen festgelegt werden, daß einige der Modellparameter identisch sein sollen. Modelle, die solchen „constraints“ unterliegen, können zum Beispiel mit dem Programm TDA geschätzt werden.

kann mit dem KL-Distanzmaß vorgenommen werden:

$$D(\theta) = \sum_{t\in\mathcal{T}} \sum_{d\in\mathcal{D}} \sum_{x\in\mathcal{X}} \mathrm{P}\,(T=t, D=d, X=x) \log\left\{\frac{\mathrm{P}\,(T=t, D=d \mid X=x)}{\tilde{f}(t,d \mid x;\theta)}\right\} \tag{3.42}$$

Die Minimierung dieser Distanz, als eine Funktion von θ, liefert (falls eindeutig möglich) einen Parametervektor $\hat{\theta}$, d.h. ein optimales Modell aus der vorgegebenen Klasse potentieller Modelle. Als ein äquivalentes Kriterium für die Modellschätzung kann eine Log-Likelihood definiert werden. Gibt es Beobachtungen für $i = 1, \ldots, N$ Individuen und bezeichnet t_i die beobachtete Verweildauer im Ausgangszustand der Episode und d_i den schließlich erreichten Zielzustand, erhält man eine zu (3.42) äquivalente Log-Likelihood durch

$$\ell(\theta) = \sum_{i=1}^{N} \log\left\{\tilde{f}(t_i, d_i \mid x_i;\theta)\right\} \tag{3.43}$$

Die Maximierung dieser Log-Likelihood liefert den gleichen Schätzwert $\hat{\theta}$ wie die Minimierung der Distanz (3.42).

b) Wenn die Modellbildung auf einer stetigen Zeitachse erfolgen soll, ist $\tilde{r}_d(t \,|\, x;\theta)$ eine Klasse stetiger Übergangsraten. Wie für einfache Episoden bereits beschrieben worden ist, muß daraus zunächst eine Formulierung gewonnen werden, die sinnvoll mit den empirischen Wahrscheinlichkeiten $\mathrm{P}(T = t, D = d \,|\, X = x)$ verglichen werden kann. Dies kann folgendermaßen erreicht werden. Zunächst gilt die Additivität auch bei stetigen Übergangsraten, also

$$\tilde{r}(t \mid x;\theta) = \sum_{d\in\mathcal{D}} \tilde{r}_d(t \mid x;\theta_d)$$

Daraus gewinnt man die Survivorfunktion

$$\tilde{G}(t \mid x;\theta) = \exp\left\{-\int_0^t \tilde{r}(\tau \mid x;\theta)\,\mathrm{d}\tau\right\}$$

Schließlich erhält man die zustandsspezifischen Dichtefunktionen

$$\tilde{f}(t,d \mid x;\theta) = \tilde{r}_d(t \mid x;\theta_d)\,\tilde{G}_d(t \mid x;\theta)$$

Um die Bedeutung dieser Dichtefunktionen zu erfassen, kann man sich vorstellen, daß jedes der potentiellen Modelle eine (durch X bedingte) Zufallsvariable $(\tilde{T}_\theta, \tilde{D}_\theta)$ beschreibt. Es gilt dann

$$\tilde{f}(t,d \mid x;\theta) = \lim_{\Delta\to 0} \frac{\mathrm{P}\,(t \le \tilde{T}_\theta < t+\Delta, \tilde{D}_\theta = d \mid X = x)}{\Delta}$$

Es ist also sinnvoll, für die Modellschätzung von folgendem Vergleich auszugehen:

$$\mathrm{P}\,(T=t, D=d \mid X=x) \;\approx\; \frac{\mathrm{P}\,(t \le \tilde{T}_\theta < t+\Delta, \tilde{D}_\theta = d \mid X = x)}{\Delta}$$

Wie in Abschnitt 3.5 besprochen worden ist, kann die diesem Vergleich zugrundeliegende Approximation unterschiedlich interpretiert werden. Wird die durch das Modell angenommene theoretische Verteilungsfunktion als Approximation der durch die Daten gegebenen empirischen Verteilungsfunktion aufgefaßt, kann das KL-Distanzmaß in der üblichen Form verwendet werden. Die Formulierung ist mit der für den diskreten Fall in (3.42) angegebenen identisch. Also kann auch bei stetigen Modellen die in (3.43) formulierte Log-Likelihood verwendet werden.

Berücksichtigung unvollständiger Beobachtungen

Wie unvollständige Beobachtungen bei der Modellschätzung berücksichtigt werden können, wenn es nur einen möglichen Zielzustand gibt, wurde in Abschnitt 3.4 diskutiert. Die dort entwickelten Grundgedanken können leicht auf eine Situation übertragen werden, in der es mehrere mögliche Zielzustände gibt. Wir betrachten kurz den Fall, daß einige Beobachtungen rechts zensiert und / oder links abgeschnitten sind. Die verfügbaren Beobachtungen sehen dann folgendermaßen aus. Es gibt für jedes der Individuen $i = 1, \ldots, N$ Angaben über (s_i, t_i, d_i, x_i). s_i $(s_i \ge 0)$ ist der Anfangszeitpunkt der Beobachtung, t_i $(t_i > s_i)$ ist der Endzeitpunkt der Beobachtung, d_i ist der zum Zeitpunkt t_i erreichte Endzustand der Episode, und x_i ist der Wert der Klassifizierungsvariablen X. Wenn $s_i > 0$, handelt es sich um eine links abgeschnittene Beobachtung; wenn $d_i = 0$, handelt es sich um eine rechts zensierte Beobachtung.[67]

Um eine geeignete Log-Likelihood für die Modellschätzung zu entwickeln, ist es zweckmäßig, Modelle für eine diskrete und für eine stetige Zeitachse zu unterscheiden. Außerdem vereinfacht es die Notation, wenn folgende Bezeichnungen verwendet werden: $\mathcal{N}$ sei die Menge aller Individuen $i = 1, \ldots, N$, $\mathcal{E}$ sei die Menge der Individuen, bei denen die Episode in einem beobachtbaren Ereignis endet, und $\mathcal{Z}$ sei die Menge der Individuen, bei denen die Beobachtung rechts zensiert ist.

a) Der Grundgedanke besteht darin, rechts zensierte Beobachtungen mit dem Wert der durch das Modell unterstellten Survivorfunktion zu vergleichen und bei links abgeschnittenen Beobachtungen eine Konditionierung auf den Anfangszeitpunkt der Prozeßbeobachtung vorzunehmen. Im diskreten Fall kann dementsprechend folgende Log-Likelihood verwendet

[67] Hierbei wird angenommen, daß der Wert 0 nicht zur Kodierung der Zustände in der Menge $\mathcal{D}$ verwendet wird.

werden:

$$\ell(\theta) = \sum_{i \in \mathcal{E}} \log \left\{ \frac{\tilde{f}(t_i, d_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} \right\} + \sum_{i \in \mathcal{Z}} \log \left\{ \frac{\tilde{G}(t_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} \right\} \quad (3.44)$$

Dies kann durch Verwendung der in (3.41) angegebenen Zerlegung vereinfacht werden zu

$$\ell(\theta) = \sum_{i \in \mathcal{E}} \log \left\{ \frac{\tilde{r}_{d_i}(t_i \mid x_i; \theta_{d_i})}{1 - \tilde{r}(t_i \mid x_i; \theta)} \right\} + \sum_{i \in \mathcal{N}} \log \left\{ \frac{\tilde{G}(t_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} \right\} \quad (3.45)$$

Schließlich kann auch noch die rechte Seite durch Übergangsraten formuliert werden, denn

$$\frac{\tilde{G}(t_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} = \prod_{l=s_i+1}^{t_i} (1 - \tilde{r}(l \mid x_i; \theta)) \quad (3.46)$$

Also erhält man aus (3.45) die Formulierung

$$\ell(\theta) = \sum_{i \in \mathcal{E}} \log \left\{ \frac{\tilde{r}_{d_i}(t_i \mid x_i; \theta_{d_i})}{1 - \tilde{r}(t_i \mid x_i; \theta)} \right\} + \sum_{i \in \mathcal{N}} \prod_{l=s_i+1}^{t_i} \log \{1 - \tilde{r}(l \mid x_i; \theta)\} \quad (3.47)$$

Dies zeigt, daß nur eine Spezifikation der zustandsspezifischen Übergangsraten erforderlich ist, um eine für die Modellschätzung geeignete Log-Likelihood bilden zu können.

b) Wird das Modell für eine stetige Zeitachse formuliert, kann grundsätzlich analog vorgegangen werden. Die Formulierung ändert sich nur dadurch etwas, daß die Zerlegung (3.41) im stetigen Fall die Form

$$\tilde{f}(t, d \mid x; \theta) = \tilde{r}_d(t \mid x; \theta_d)\, \tilde{G}(t \mid x; \theta)$$

annimmt, wobei jetzt $\tilde{f}(t, d \,|\, x; \theta)$ eine zweidimensionale Dichtefunktion ist. Wird diese Zerlegung verwendet, erhält man analog zu (3.45) die Log-Likelihood

$$\ell(\theta) = \sum_{i \in \mathcal{E}} \log \{\tilde{r}_{d_i}(t_i \mid x_i; \theta)\} + \sum_{i \in \mathcal{N}} \log \left\{ \frac{\tilde{G}(t_i \mid x_i; \theta)}{\tilde{G}(s_i \mid x_i; \theta)} \right\} \quad (3.48)$$

3.7 Modelle als vereinfachende Beschreibungen

Die in diesem Kapitel diskutierten statistischen Modelle sollen dem Zweck dienen, vereinfachende Beschreibungen von Prozessen, d.h. hier Lebensverläufen in Gesamtheiten von Individuen, zu ermöglichen. Bei dieser Zielsetzung gibt es zwei Aspekte, die relativ unbestimmt sind. Erstens bereits das Ziel, einen Sachverhalt oder Vorgang beschreiben zu wollen. Beschreibungen können stets im Hinblick auf unterschiedliche Aspekte vorgenommen werden. Die Auswahl von Aspekten beginnt bereits mit der Fixierung eines Zustandsraums und einer Zeitachse. Aber auch wenn nach dieser Fixierung ein Prozeß formal definiert werden kann, gibt es noch unterschiedliche Mögichkeiten, ihn zu beschreiben. Man kann sich auf gewisse „summarische“ Aspekte beschränken, zum Beispiel auf eine Berechnung mittlerer Verweildauern; oder man kann versuchen, den Prozeß*verlauf* zu beschreiben. Bei den hier behandelten Modellen steht dieser Prozeßverlauf im Mittelpunkt, und als zentrales Konzept, um diesen Verlauf zu beschreiben, diente uns der Begriff der Übergangsrate.

Zweitens gibt es einen Unbestimmtheitsaspekt, der die Genauigkeit der Beschreibung betrifft. Folgende, bereits zu Beginn dieses Kapitels angeführte Bemerkung von Cox und Snell [1981, S. 28] kann zur Verdeutlichung dienen: „The objective of statistical analysis is to discover what conclusions can be drawn from data and to present these conclusions in as simple and lucid a form as is consistent with accuracy.“ Ein Problem besteht offenbar darin, was „consistent with accuracy“ bedeuten soll. Jeder Sachverhalt kann mehr oder weniger genau beschrieben werden; ein absolutes Kriterium für Genauigkeit gibt es nicht. Noch wichtiger ist, daß der Modellbildung zwei sich tendenziell widersprechende Zielsetzungen zugrundeliegen. Einerseits soll eine *vereinfachende* Beschreibung eines Prozesses gegeben werden, andererseits soll er „möglichst genau“ beschrieben werden. Die Lösung dieses Dilemmas besteht darin, nach einer *angemessenen* Beschreibung zu suchen. Dies ist zwar eine sehr unbestimmte Formulierung, sie macht jedoch deutlich, daß die Frage, ob eine Beschreibung angemessen ist, hauptsächlich davon abhängt, welcher Zweck mit ihr verfolgt wird.

3.7.1 Übergangsraten- vs. Regressionsmodelle

Die in der Statistik übliche Vorgehensweise, um das Problem der Angemessenheit von Modellen zu reflektieren, orientiert sich am ND-Modell, wie es in Abschnitt 3.1.2 skizziert worden ist. Als Prototyp dient das klassische Regressionsmodell: $y_i = g(x_i, \epsilon_i)$. Für jedes Individuum $i = 1, \ldots, N$ wird der Sachverhalt y_i als eine Funktion des Sachverhalts x_i und der „Zufallseinflüsse“ ϵ_i angesehen. Das Ziel besteht darin, im Hinblick auf eine Grundgesamtheit von Individuen mithilfe einer Kenntnis der Sachverhalte x_i möglichst gute Prognosen der Sachverhalte y_i zu erreichen. In diesem

Kontext können die „Zufallseinflüsse“ ϵ_i als *Residuen* der Modellbildung – als „unerklärte Varianz“ – interpretiert werden, und aus der Zwecksetzung der Modellbildung ergibt sich zugleich ein gewisses Kriterium, um zwischen mehr oder weniger angemessenen Modellen unterscheiden zu können: Ein Modell ist dann angemessen (im statistischen Sinne optimal), wenn sich aus der Verteilung der Residuen keine Informationen gewinnen lassen, um die Prognosegüte des Modells zu verbessern. Implizit liefert dies zugleich eine Definition dafür, was es heißen soll, daß die Residuen „zufällig“ verteilt sind.

Versucht man, diesen Gedankengang zur Beurteilung von Übergangsratenmodellen zu nutzen, stellen sich jedoch mehrere Probleme.

a) Das erste Problem betrifft die Frage, wie die mehr oder weniger vage Vorstellung, daß ein Modell „angemessen“ ist, präzisiert werden kann. In der statistischen Literatur wird häufig von folgender Überlegung ausgegangen. Jedes Modell kann als Beschreibung eines Zufallsgenerators zur Erzeugung von Daten aufgefaßt werden, es beschreibt in diesem Sinne einen „datenerzeugenden Prozeß“. Also kann man versuchen, die Frage, ob ein Modell für eine gegebene Menge an Daten angemessen ist, folgendermaßen zu reformulieren: Ist es vorstellbar, daß die vorhandenen Daten mit dem durch das Modell beschriebenen Zufallsgenerator hätten erzeugt werden können? Diese Formulierung ist jedoch offensichtlich nicht sehr hilfreich, denn *vorstellbar* bzw. *möglich* ist dies in jedem Fall. Um diese Sackgasse zu vermeiden, wird deshalb die Frage so formuliert: Ist es *wahrscheinlich*, daß die vorliegenden Daten ein Ergebnis des durch das Modell beschriebenen Zufallsgenerators sind?

Alle im engeren Sinne statistischen, d.h. nicht nur deskriptiven Verfahren, um die Angemessenheit eines Modells zu prüfen, gehen von dieser Fragestellung aus. Es ist jedoch schwer, die in diesem Kontext erforderlichen Wahrscheinlichkeitsaussagen sinnvoll zu interpretieren.

b) Das Problem wird noch etwas komplizierter, wenn man berücksichtigt, daß mit der Modellbildung eine *vereinfachende* Beschreibung empirischer Daten intendiert wird. D.h. bei der Modellbildung wird von der empirischen Verteilung der Daten bewußt abgewichen, um sich ein einfacheres Bild des Sachverhalts machen zu können. Infolgedessen kann nicht erwartet werden, daß der durch das Modell beschriebene abstrakte Zufallsgenerator genau denjenigen Prozeß repräsentiert, durch den die Daten tatsächlich zustandegekommen sind. Damit wird jedoch eine entscheidende Voraussetzung der üblichen statistischen Verfahren, um die Angemessenheit von Modellen zu „testen“, fragwürdig. Im Kontext einer Diskussion von „Spezifikationstests“ für Übergangsratenmodelle hat zum Beispiel Arminger [1990, S. 253] diese Voraussetzung folgendermaßen formuliert:

„A crucial issue in the interpretation of results from fitting statistical models to empirical data is the underlying but rarely voiced assumption that the class of models fitted to the data contains the true model of the random mechanism generating the data. If this is not the case, the parameters estimated from a wrong model can easily be heavily biased and lead to wrong conclusions about the nature of relationships and causal mechanism.“

c) Schließlich sollte noch einmal die Frage nach dem Zweck der Modellbildung gestellt werden. Denn der primäre soziologische Sinn der Modellbildung liegt nicht darin, Prognosen für individuelles Verhalten zu ermöglichen, sondern die Modellbildung soll Einsichten in die Regeln vermitteln, denen die Individuen in einer Gesellschaft folgen. D.h. das soziologische Interesse richtet sich nicht in erster Linie auf ein im Einzelfall zu erwartendes Verhalten, sondern auf die Darstellung eines Spektrums unterschiedlicher Verhaltensweisen. Dazu dient insbesondere das Konzept der Übergangsrate, mit dem nicht Individuen, sondern ein Prozeßverlauf in einer Gesamtheit von Individuen beschrieben wird. Geht man von einer Situation aus, in der sich der interessierende Prozeß durch eine zweidimensionale Zufallsvariable (T, D) repräsentieren läßt (T = Verweildauer, D = Zielzustand), könnte man sagen, daß sich im Kontext des klassischen Regressionsmodells das Interesse auf den Erwartungswert dieser Zufallsvariablen, das soziologische Interesse dagegen auf die Verteilung dieser Zufallsvariablen richtet.

Darin kommt ein unterschiedliches Erkenntnisinteresse zum Ausdruck. Tatsächlich ist es durchaus möglich, Übergangsratenmodelle als Regressionsmodelle zu interpretieren. Man sieht dies besonders einfach in einer Situation, in der es nur einen möglichen Zielzustand gibt. Man braucht dann nur eine Zufallsvariable T, die Verweildauer bis zu einem Zustandswechsel, zu betrachten. Also kann man ein Regressionsmodell $T = g(X, \epsilon)$ formulieren: die Verweildauer ist eine Funktion von gewissen bebachtbaren Sachverhalten X und „Zufallseinflüssen“ ϵ. Bei dieser Betrachtungsweise gibt es keinen wesentlichen Unterschied zum klassischen Regressionsmodell. Gewisse Besonderheiten entstehen nur daraus, daß es der Beschreibung eines zeitlichen Prozesses dienen soll und infolgedessen in der Regel einige Beobachtungen rechts zensiert sind. Dies kann jedoch als ein bloß technisches Problem angesehen werden, das durch geeignete Schätzmethoden (Maximum-Likelihood-Verfahren) gelöst werden kann.

Das wesentliche Problem liegt jedoch in der Frage, welchem soziologischen Zweck die Modellbildung dienen soll. Ein Aspekt dieser Frage erscheint in der gelegentlich geführten Diskussion, ob bei Übergangsratenmodellen die „abhängige Variable“ beobachtbar ist. Einige Autoren verneinen dies; zum Beispiel sagen Wu und Tuma [1991, S. 356], “that the outcomes in these models are unobserved since they are defined in terms of an individual's probability of experiencing the event of interest“. Andere Autoren, zum Beispiel Petersen [1990, 1991], betonen demgegenüber, daß auch in diesem Fall der zu erklärende Sachverhalt beobachtbar ist, nämlich eine Verweildauer oder eine Sequenz von Zuständen. Offensichtlich sind beide Auffassungen möglich; die Frage ist, welchem Zweck die Modellbildung dienen soll. Aber auch dann, wenn man den Zweck der Modellbildung darin sieht, Einsichten in Übergangsraten (und deren Bedingungen) zu gewinnen, ist die Frage ihrer Beobachtbarkeit und soziologischen Interpretation

noch offen. Das angeführte Zitat von Wu und Tuma plädiert dafür, Übergangsraten als *den Individuen zurechenbare* Wahrscheinlichkeiten (bzw. Wahrscheinlichkeitsdichten) zu interpretieren. Daß es sich um einen nicht beobachtbaren Sachverhalt handelt, folgt aus dieser Interpretation, in der Wahrscheinlichkeiten als Eigenschaften von Individuen (und ihrer jeweiligen sozialen Umstände) angesehen werden. Diese Interpretation ist jedoch problematisch. Einerseits grundsätzlich im Hinblick auf die Frage, wie eine empirisch sinnvolle Definition des Wahrscheinlichkeitsbegriffs erreicht werden kann. Andererseits im Hinblick darauf, welchem Erkenntnisinteresse eine soziologische Lebensverlaufsforschung dienen sollte. Dies hat unmittelbare Folgen für die Frage der Beobachtbarkeit. Wenn man Wahrscheinlichkeitsaussagen als Aussagen über Gesamtheiten von Individuen (oder über Zufallsgeneratoren, mit denen Gesamtheiten individueller Ereignisse erzeugt werden können) ansieht, beziehen sich auch Übergangsraten nicht auf Individuen; sie charakterisieren nicht (unbeobachtbare) Eigenschaften von (jeweils bestimmten) Individuen, sondern – wie man im Hinblick auf eine soziologische Interpretation sagen könnte – Eigenschaften ihrer gesellschaftlichen Verhältnisse. Bei dieser Interpretation sind Übergangsraten durchaus beobachtbare Sachverhalte. Es handelt sich um empirisch erfaßbare, zunächst deskriptiv zu deutende Übergangswahrscheinlichkeiten. Dies gilt unabhängig davon, daß es für die Modellbildung sinnvoll sein kann, sie durch stetige Übergangsraten zu approximieren.

Um die unterschiedlichen Interpretationen zu illustrieren, beziehe ich mich noch einmal auf unsere SOEP-Daten über die Zeitdauer bis zur ersten Heirat, gemessen auf einer Prozeßzeitachse, die mit dem 15. Lebensjahr beginnt. Das üblicherweise mit einem Regressionsmodell verfolgte Ziel würde in diesem Fall darin bestehen, für jedes Individuum der Grundgesamtheit eine bedingte Prognose darüber abgeben zu können, in welchem Alter eine Heirat stattfinden wird. Um die Illustration einfach zu halten, sehe ich davon ab, daß einige Personen vermutlich überhaupt nicht heiraten werden. Das Regressionsmodell hat dann die Form $T = g(X, \epsilon)$. T ist die Zeitdauer bis zur Heirat, X repräsentiert Sachverhalte, von denen bei der Modellbildung angenommen wird, daß sie als Bedingungen der Verteilung von T gedeutet werden können, ϵ repräsentiert die nicht erfaßten „Zufallseinflüsse“. Im folgenden beschränke ich mich darauf, die Individuen nur nach ihren Geburtskohorten zu unterscheiden.

Abbildung 3.7.1 zeigt die verfügbaren Daten. Orientiert man sich an einem Regressionsmodells (als Prototyp eines ND-Modells), konzentriert sich das Interesse darauf, für jede Geburtskohorte ein durchschnittliches Heiratsalter voraussagen zu können. Da die Daten rechts zensiert und die Verteilungen für das kohortenspezifische Heiratsalter sehr unsymmetrisch sind, ist es zweckmäßig, den Median zu verwenden, d.h. dasjenige Alter, bei dem 50 % einer Geburtskohorte geheiratet haben. Abbildung 3.7.2 zeigt, wie sich dieses mittlere Heiratsalter in der Abfolge der Geburtskohorten entwickelt hat. Die Berechnung erfolgte mit dem Kaplan-Meier-Verfahren,

um rechts zensierte Beobachtungen berücksichtigen zu können.[68]

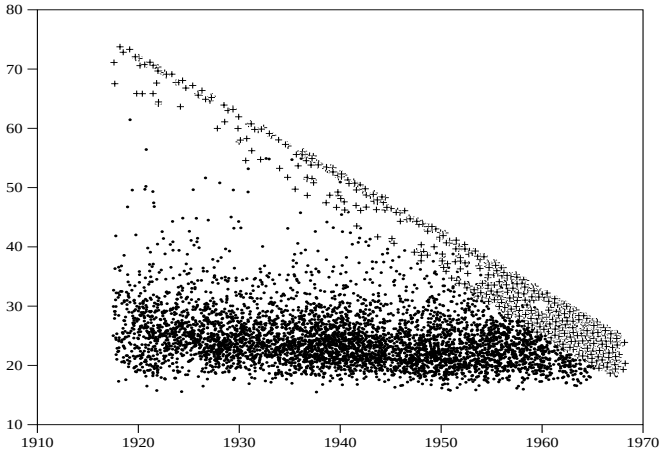


Abbildung 3.7.1. Alter bei der ersten Heirat (Ordinate) in Abhängigkeit vom Geburtsjahr (Abszisse) bei 6416 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben. Rechts zensierte Beobachtungen sind durch ein + markiert.

Zweifellos vermittelt Abbildung 3.7.2 eine Einsicht in das Heiratsverhalten. Sie zeigt, daß das mittlere Heiratsalter bis zu den etwa 1950 geborenen Kohorten gesunken, dann wieder angestiegen ist. Die Abbildung zeigt jedoch nicht, wie der Heiratsprozeß in den einzelnen Geburtskohorten verläuft. Dies entspricht der Zielsetzung der üblichen Regressionsmodelle. Sie sollen den Erwartungswert der abhängigen Variablen prognostizierbar machen, die Residuen sind im Hinblick auf diese Aufgabenstellung bloß ein unvermeidbarer Mangel des Modells, insofern es immer nur einige wenige Faktoren in Betracht ziehen kann.[69]

[68] Jeweils für die Personen *eines* Geburtsjahres wurde eine separate Schätzung durchgeführt. Ein Median kann bis zum Geburtsjahrgang 1963 berechnet werden. Bei den jüngeren Geburtskohorten kann infolge des kurzen Beobachtungszeitraums der Median nicht berechnet werden.

[69] Natürlich würde man auch in unserem Beispiel zahlreiche weitere Faktoren in Betracht ziehen müssen, insbesondere das Geschlecht und das (jeweils erreichte) Bildungsniveau. An dieser Stelle geht es jedoch nur um ein möglichst einfaches Beispiel, um den Unterschied zwischen Regressions- und Übergangsratenmodellen illustrieren zu können.

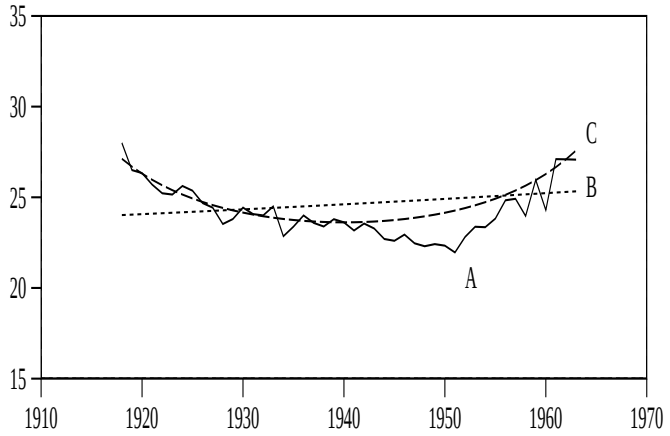


Abbildung 3.7.2. Mittleres Alter (Median) bei der ersten Heirat (Ordinate) in Abhängigkeit vom Geburtsjahr (Abszisse) bei 6416 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben. Durchgezogene Linie (A): Kaplan-Meier-Schätzungen, gestrichtelte Linien (B und C): Schätzungen mit einem erweiterten log-logistischen Modell.

Für die soziologische Interpretation kann jedoch den Residuen eine darüber hinausgehende Bedeutung gegeben werden. Sie sind nicht nur Ausdruck einer Grenze für die Prognostizierbarkeit des jeweils individuellen Verhaltens, sondern liefern eine substantielle Einsicht in die Vielfalt der individuellen Lebensverläufe und mithin einen Ausgangspunkt für die soziologische Fragestellung, ob in dieser Vielfalt gleichwohl gewisse Regeln bzw. Regelmäßigkeiten erkennbar sind. Für die soziologische Interpretation ist infolgedessen die Abbildung 3.7.1 wesentlich aufschlußreicher als Abbildung 3.7.2, in der nur das mittlere Heiratsalter dargestellt worden ist. Dieser Perspektivenwechsel kommt darin zum Ausdruck, daß sich die Modellbildung auf den Verlauf der Übergangsrate konzentriert. Rein formal betrachtet handelt es sich nur um eine austauschbare Darstellung für die Verteilung der Residuen. Wichtig ist demgegenüber, daß der Verlauf der Übergangsrate, als eine Funktion der Zeit, den Ablauf eines Prozesses beschreibt.

In unserem Beispiel geht es um den kohortenspezifischen Verlauf des Heiratsverhaltens. Die Basisinformation liefert Abbildung 3.7.1. Diese Information kann durch ein Übergangsratenmodell vereinfachend dargestellt werden. Verwendet man ein erweitertes log-logistisches Modell, erhält man die Abbildung 3.7.3.[70]

[70] Die Anregung zu dieser Darstellung verdanke ich Mayer und Huinink [1990a].

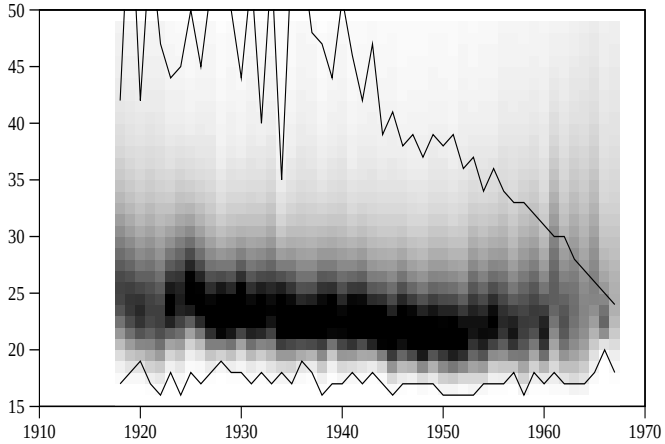


Abbildung 3.7.3. Geschätzte Wahrscheinlichkeitsdichten für das Alter bei der ersten Heirat (Ordinate) in Abhängigkeit vom Geburtsjahr (Abszisse) bei 6416 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben. Schätzung durch kohortenspezifische erweiterte log-logistische Modelle. Außerdem eingezeichnet sind die Altersintervalle, innerhalb derer Heiraten beobachtet werden können.

Um dieses Bild zu erzeugen, wurde für jeden Geburtsjahrgang ein separates erweitertes log-logistisches Modell geschätzt. Dann wurde die aus der Modellschätzung resultierende Wahrscheinlichkeitsdichte für Heiraten separat für jeden Geburtsjahrgang und jedes Lebensalter durch einen der Dichte proportionalen Grauwert eingezeichnet. Im Unterschied zu Abbildung 3.7.2 zeigt dieses Bild nicht nur Veränderungen in den kohortenspezifischen Verteilungen des Heiratsalters, es zeigt auch, daß der Anstieg des Heiratsalters bei den jüngeren Geburtskohorten damit verbunden ist, daß die kohortenspezifischen Heiratsverläufe gewissermaßen „diffuser“ werden.[71] Hier könnten dann weitere soziologische Fragestellungen ansetzen, zum Beispiel ausgehend von der Vermutung, daß sich durch die Ausbreitung nicht-ehelicher Lebensgemeinschaften die Lokalisierung von Heiraten innerhalb individueller Lebensverläufe verändert hat und möglicherweise eine engere Anbindung von Heiraten an die Geburt von Kindern erfolgt.[72]

[71] Noch deutlicher kommen die Veränderungen im Heiratsverhalten zum Ausdruck, wenn man nicht die Wahrscheinlichkeitsdichten, wie in Abbildung 3.7.3, sondern die Abhängigkeit der Übergangsraten von der Geburtskohorte und der Prozeßzeit betrachtet, vgl. Abbildung 3.7.4.

[72] Als eine interessante empirische Analyse dieser Fragen vgl. Manting [1994].

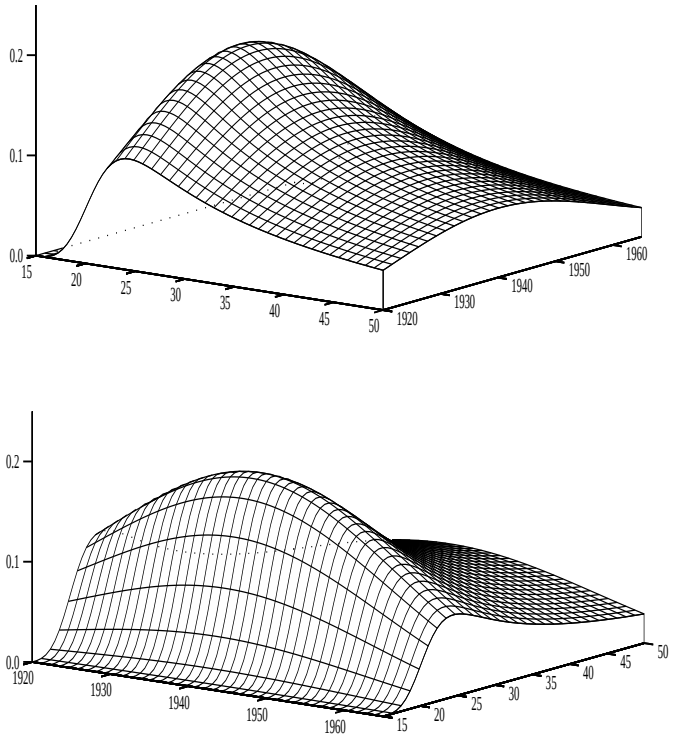


Abbildung 3.7.4. Mit einem erweiterten log-logistischen Modell geschätzte Übergangsraten für das Ereignis *erste Heirat*; Geburtsjahr als lineare und quadratische Kovariable. Stichprobe: 6416 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben.

3.7.2 Die Anpassung des Modells an die Daten

Wie kann geprüft werden, ob ein Übergangsratenmodell eine angemessene Repräsentation der jeweils verfügbaren Daten liefert? Wie bereits angedeutet worden ist, sind die bei Regressionsmodellen üblichen Methoden nicht ohne weiteres anwendbar. Nicht nur deshalb, weil die – aus der Regressionsperspektive – „unerklärte Varianz“ typischerweise sehr groß ist, so daß es wenig sinnvoll erscheint, die individuell beobachteten Verweildauern mit ihren durch das Modell berechenbaren Erwartungswerten zu

vergleichen, sondern auch weil ein solcher Vergleich nur mit nicht rechts zensierten Beobachtungen durchgeführt werden könnte. In der Literatur sind jedoch einige andere Verfahren diskutiert worden, um Übergangsratenmodelle mit den jeweils verfügbaren Daten zu vergleichen. Einige dieser Methoden sollen im folgenden dargestellt werden.

a) Eine naheliegende Möglichkeit besteht darin, Survivorfunktionen zu vergleichen: einerseits eine mit dem Kaplan-Meier-Verfahren gewonnene Schätzung, die ein unverzerrtes Bild der Daten liefert, und andererseits die durch ein Modell geschätzte Survivorfunktion.[73] Abbildung 3.7.4 illustriert diese Methode anhand unserer Beispieldaten für das Alter bei der ersten Heirat, wobei nur diejenigen Personen betrachtet werden, die im Zeitraum 1918 – 1927 geboren wurden (Geburtskohorte 1 in Tabelle 3.12).

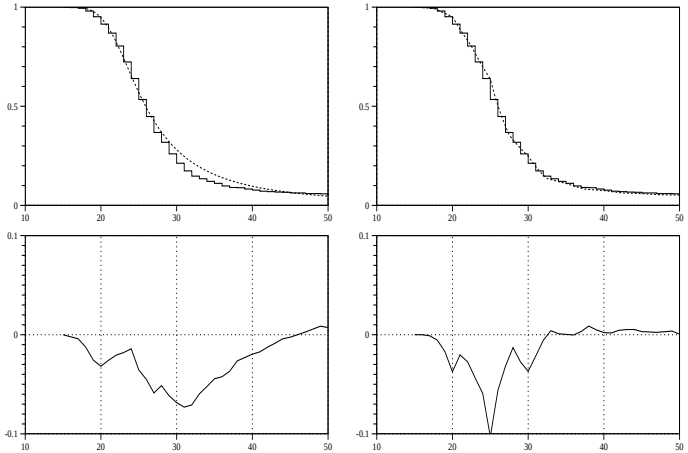


Abbildung 3.7.4. Survivorfunktionen für die Zeitdauer vom 15. Lebensjahr bis zur ersten Heirat bei 908 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1927 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben. Durchgezogene Linie: Schätzungen mit dem Kaplan-Meier-Verfahren. Gestrichelt: Schätzungen mit einem erweiterten log-logistischen Modell (linke Hälfte) und mit einem Exponentialmodell mit Zeitperioden der Länge 2.5 Jahre (rechte Hälfte). Untere Abbildungen: Differenzen zwischen den Survivorfunktionen.

Für die Modellbildung werden ein erweitertes log-logistisches Modell und ein Exponentialmodell mit Zeitperioden der Länge 2.5 Jahre verwendet. Scheinbar liefert das Exponentialmodell (rechte Seite) eine bes-

[73] Dieses Verfahren, um das Verhältnis von Modellen und Daten zu untersuchen, wird in der Literatur sehr häufig verwendet, vgl. zum Beispiel Wu und Tuma [1991]. Anstelle des Kaplan-Meier-Verfahrens kann auch die Sterbetafelmethode verwendet werden, wie zum Beispiel bei Diekmann [1990, 1991] illustriert wird.

sere Anpassung an die Daten; wenn man jedoch die Differenzen zwischen den Kaplan-Meier-Schätzungen und den Modellschätzungen für die Survivorfunktion betrachtet, gibt es bei beiden Modellen Schwächen in ihrer Fähigkeit, den empirischen Verlauf der Survivorfunktion angemessen zu repräsentieren.

Bei der Interpretation der Abbildungen muß allerdings berücksichtigt werden, daß die verfügbaren Daten nur eine begrenzte Genauigkeit aufweisen. Bei der für dieses Beispiel ausgewählten Geburtskohorte der zwischen 1918 und 1927 geborenen Personen liefert das SOEP fast ausschließlich nur Jahresangaben für den Heiratszeitpunkt. Infolgedessen gibt es ein verhältnismäßig breites Unbestimmtheitsband, innerhalb dessen die empirische Survivorfunktion nicht genau lokalisiert werden kann, vgl. Abbildung 3.7.5.

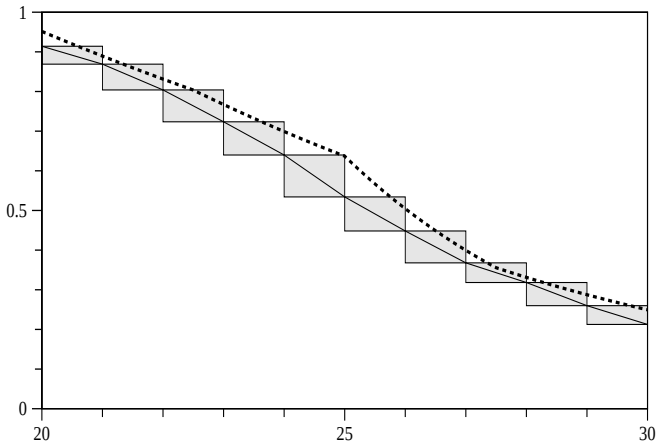


Abbildung 3.7.5 Vergrößerter Ausschnitt aus Abbildung 3.7.4, oben rechts.

Tatsächlich würde das Exponentialmodell mit Zeitperioden eine wesentlich bessere Anpassung an die Daten liefern, wenn genauere Zeitangaben verfügbar wären. Um dies zu illustrieren, simulieren wir eine größere Genauigkeit, indem jede Zeitangabe mit einer in einem Jahresintervall gleichverteilten Zufallszahl überlagert wird. Dann wird mit diesen Pseudo-Daten eine Kaplan-Meier-Schätzung der Survivorfunktion und eine Modellschätzung durchgeführt. Abbildung 3.7.6 zeigt, analog zu Abbildung 3.7.4, das Ergebnis. Offensichtlich gibt es eine fast perfekte Übereinstimmung.

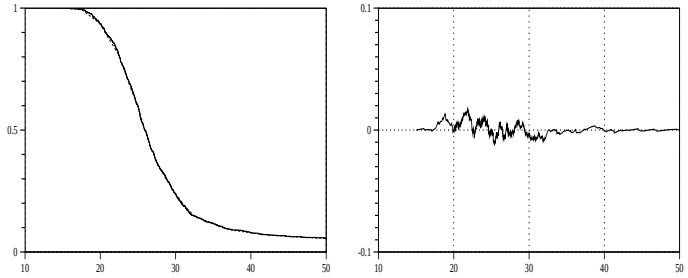


Abbildung 3.7.6 Vergleich von Survivorfunktionen unter Verwendung von Pseudo-Daten für das Alter bei der ersten Heirat.

b) Der Vergleich der durch ein Modell implizierten Survivorfunktion mit ihrer Kaplan-Meier-Schätzung setzt voraus, daß zunächst ein Modell geschätzt werden muß. Ein häufig einfacheres Verfahren besteht darin, geeignete Transformationen der mit dem Kaplan-Meier-Verfahren geschätzten Survivorfunktionen zu betrachten. Der Grundgedanke ist einfach.[74] Es wird nach einer geeigneten Transformation für die Zeitachse und für die mit dem Kaplan-Meier-Verfahren geschätzten Survivorfunktion gesucht, so daß eine graphische Darstellung des Zusammenhangs zwischen diesen beiden Transformationen näherungsweise eine grade Linie ergeben müßte. Besonders einfach ist dieses Verfahren, wenn geprüft werden soll, ob ein einfaches Exponentialmodell angemessen ist. Bei der Annahme eines Exponentialmodells erhält man für die Survivorfunktion die Formulierung $G(t) = \exp(-r\,t)$, wobei r eine konstante Übergangsrate ist. Also sollte der Graph der Funktion $-\log(\hat{G}(t))$ vs. t, wobei $\hat{G}(t)$ eine Kaplan-Meier-Schätzung der Survivorfunktion ist, näherungsweise eine grade Linie mit der Steigung r sein. Abbildung 3.7.7 gibt eine Illustration für die Zeitdauer bis zur ersten Heirat bei den SOEP-Personen aus der Teilstichprobe A, die im Zeitraum 1918 – 1927 geboren wurden. Wie zu vermuten war, liefert das einfache Exponentialmodell eine sehr schlechte Repäsentation der Daten.[75]

Ein Nachteil des Verfahrens liegt darin, daß es nur bei einfachen parametrischen Übergangsratenmodellen anwendbar ist. Bereits für das erweiterte log-logistische Modell stehen keine einfachen Transformationen zur Verfügung. Auch ist es nicht sinnvoll anwendbar für das für praktische Anwendungen besonders wichtige Exponentialmodell mit Zeitperioden. Allerdings stellt sich bei diesem Modell die Frage etwas anders. Die Frage

[74] Eine ausführliche Diskussion dieses Verfahrens findet sich bei Wu [1990]. Vgl. auch Blossfeld et al. [1989, S. 176ff].

[75] Gelegentlich wird einfache OLS-Regression verwendet, um eine lineare Approximation zu ermitteln. Besser ist es jedoch, die Rate aus der ML-Schätzung eines Übergangsratenmodells zu ermitteln, bei dem rechts zensierte Beobachtungen berücksichtigt of y axis, inen. Abbildung 3.7.7 zeigt, daß man wesentlich unterschiedliche Ergebnisse erhält.

ist nicht *ob*, sondern *wie* mit diesem Modell eine angemessene Repräsentation der Daten erreicht werden kann. Denn grundsätzlich kann durch geeignete Wahl der Zeitperioden stets eine hinreichende Approximation an den empirischen Ratenverlauf erreicht werden. Die Frage ist, wie eine möglichst zweckmäßige Definition von Zeitperioden vorgenommen werden sollte. Dafür kann bereits eine graphische Darstellung wie in Abbildung 3.7.7 Hinweise liefern, indem man sich überlegt, wie lineare Teilstücke gebildet werden können.

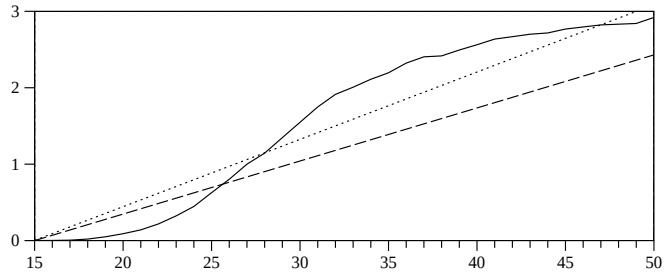


Abbildung 3.7.7 Darstellung von $-\log(\hat{G}(t))$ der Kaplan-Meier-Schätzung $\hat{G}(t)$ der Survivorfunktion für die Dauer bis zur ersten Heirat. Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1927 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben. Gestrichelt: lineare Funktion der Steigung $r = 0.0694$, d.i. die ML-Schätzung für eine konstante Übergangsrate für die gleichen Daten. Gepunktet: OLS-Regression.

c) Die bisher beschriebenen einfachen Verfahren werden problematisch, wenn das zu schätzende Modell zahlreiche Kovariablen enthält. Bei der Kaplan-Meier-Schätzung einer Survivorfunktion wird von der durch Kovariablen zu erfassenden Heterogenität der Personen in der zugrundeliegenden Gesamtheit abstrahiert. Gibt es zum Beispiel eine Klassifizierungsvariable X, mit der die Grundgesamtheit in m Teilgesamtheiten klassifiziert werden kann, liefert das Kaplan-Meier-Verfahren eine Mischung aus den in diesen Teilgesamtheiten möglicherweise unterschiedlichen Survivorfunktionen. Auch wenn das Modell für den Verlauf der Übergangsraten in den einzelnen Teilgesamtheiten angemessen ist, können infolgedessen bei einem Vergleich der mit dem Kaplan-Meier-Verfahren und mit der Modellschätzung gewonnenen Survivorfunktionen wesentliche Unterschiede auftreten, wenn diese Heterogenität nicht erfaßt wird.

Wenn die Grundgesamtheit nur in wenige Teilgesamtheiten klassifiziert wird, erscheint es möglich, die oben beschriebenen einfachen Vergleichsverfahren für jede Teilgesamtheit gesondert durchzuführen. Damit stößt man jedoch bei zunehmender Anzahl von Kovariablen sehr schnell an praktische Grenzen, bereits wegen der immer geringer werdenden Fallzahlen in den Teilgesamtheiten. Eine partielle Lösung dieses Problems liegt darin,

daß man die Modellschätzung unter Berücksichtigung der Kovariablen vornimmt und dann mithilfe des Modells eine Approximation an die mit dem Kaplan-Meier-Verfahren ermittelte Mischverteilung berechnet. Der Grundgedanke ist einfach. Das Modell soll eine approximative Darstellung der durch die Kovariablen bedingten Verteilung für die Verweildauer liefern, also

$$\mathrm{P}\,(T = t \mid X = x) \;\approx\; \tilde{f}(t \mid x; \hat{\theta})$$

Also kann man unter Verwendung der durch die Daten gegebenen Verteilung der Kovariablen folgende Näherung für die „unbedingte“ Survivorfunktion ableiten:

$$\begin{aligned} \mathrm{P}\,(T > t) &= \sum_{x \in \mathcal{X}} \mathrm{P}\,(T > t \mid X = x)\,\mathrm{P}\,(X = x) \\ &\approx \sum_{x \in \mathcal{X}} \tilde{G}(t \mid x; \hat{\theta})\,\mathrm{P}\,(X = x) \qquad (3.49) \\ &= \frac{1}{N} \sum_{i=1}^{N} \tilde{G}(t \mid x_i; \hat{\theta}) \end{aligned}$$

Der Ausdruck unten rechts kann mithilfe des geschätzten Modells berechnet werden. Andererseits liefert das Kaplan-Meier-Verfahren eine Schätzung für $\mathrm{P}(T > t)$. Beide Schätzungen können schließlich miteinander verglichen werden.

Zur Illustration betrachte ich wieder das Alter bei der ersten Heirat, gemessen auf einer Prozeßzeitachse, die mit dem 15. Lebensjahr beginnt. Es werden ein erweiteres log-logistisches Modell und ein Exponentialmodell mit Zeitperioden geschätzt.[76] Als Klassifizierungsvariable dient die Einteilung in Geburtskohorten entsprechend (3.12); die älteste Geburtskohorte wird als Referenzkategorie verwendet. Abbildung 3.7.8 vergleicht, analog zu Abbildung 3.7.4, die durch die Modelle geschätzten Survivorfunktionen mit ihrer Kaplan-Meier-Schätzung. Für die Berechnung der Survivorfunktionen wurde das in (3.49) beschriebene Verfahren verwendet. Die Abweichungen, jedenfalls beim Exponentialmodell mit Zeitperioden, sind im wesentlichen eine Folge der ungenau erfaßten Heiratszeitpunkte. Die Fluktuation der Differenzen liegt daran, daß bei diesen Zeitangaben Jahres- und Monatsangaben gemischt auftreten.

[76] Die Zeitperioden sind durch folgende Zeitpunkte auf der Prozeßzeitachse definiert: 0, 5, 7, 9, 11, 13, 15, 17, 20, 25 und 30 Jahre.

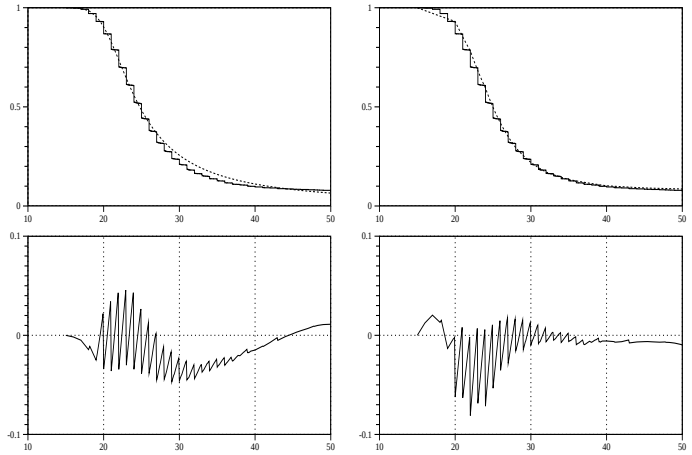


Abbildung 3.7.8. Survivorfunktionen für die Zeitdauer vom 15. Lebensjahr bis zur ersten Heirat für alle Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben. Durchgezogene Linie: Schätzungen mit dem Kaplan-Meier-Verfahren. Gestrichelt: Schätzung mit einem erweiterten log-logistischen Modell (linke Hälfte) und mit einem Exponentialmodell mit Zeitperioden (rechte Hälfte). Untere Abbildungen: Differenzen zwischen den Survivorfunktionen.

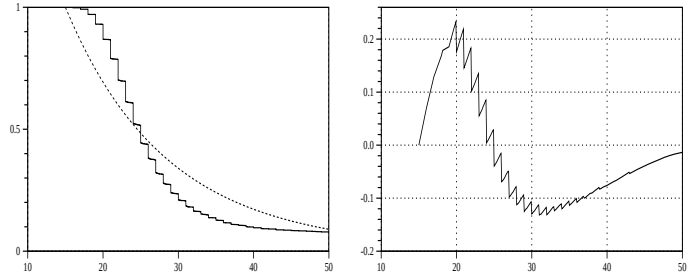
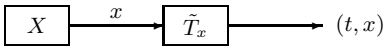


Abbildung 3.7.9. Analog zu Abbildung 3.7.8, jedoch unter Verwendung eines einfachen Exponentialmodells ohne Zeitperioden.

Dieses Verfahren, mithilfe eines Modells eine für die jeweiligen Kovariablen gemischte Survivorfunktion zu berechnen und dann mit ihrer Kaplan-Meier-Schätzung zu vergleichen, ist einfach anzuwenden, hat jedoch nur eine begrenzte Aussagekraft. Es zeigt nicht, ob das gewählte Modell *innerhalb* der durch die Kovariablen gebildeten Teilgesamtheiten angemessen ist, sondern nur, ob die schließlich resultierende Mischverteilung mit einer Kaplan-Meier-Schätzung verträglich ist. Immerhin kann das Verfahren gelegentlich darauf aufmerksam machen, daß ein Modell nicht angemessen ist. Abbildung 3.7.9 illustriert dies für ein einfaches Exponentialmodell ohne Zeitperioden.

d) Ein weiteres Verfahren, um die Angemessenheit von Übergangsratenmodellen zu prüfen, beruht auf der Idee, „verallgemeinerte Residuen“ zu definieren und ihre Verteilung zu untersuchen.[77] Um den Gedankengang zu verdeutlichen, nehmen wir an, daß die Daten in der Form (t_i, x_i, δ_i) vorliegen, für $i = 1, \ldots, N$ Individuen. t_i ist die Verweildauer, x_i der Wert des Kovariablenvektors, und δ_i ein Indikator für rechts zensierte Beobachtungen. Für diese Daten sei ein stetiges Übergangsratenmodell geschätzt worden, das durch eine Übergangsrate $\tilde{r}(t \mid x; \hat{\theta})$, eine Dichtefunktion $\tilde{f}(t \mid x; \hat{\theta})$ und eine Survivorfunktion $\tilde{G}(t \mid x; \hat{\theta})$ repräsentiert wird.

Dieses Modell kann nun als Beschreibung eines zweistufigen Zufallsgenerators betrachtet werden, der folgendermaßen aussieht:



Der erste Zufallsgenerator liefert einen Wert x für den Kovariablenvektor entsprechend der in den Daten gegebenen Verteilung von X. Der zweite Zufallsgenerator liefert eine Verweildauer t, wobei vom Modell $\tilde{f}(t \mid x; \hat{\theta})$ ausgegangen wird, also konditional auf den im ersten Schritt realisierten Wert des Kovariablenvektors. Auf diese Weise kann eine beliebige Menge von Pseudo-Beobachtungen (t_j^*, x_j^*) erzeugt werden $(j = 1, 2, 3, \ldots)$, so daß die Verteilung der x_j^* mit der durch die Daten gegebenen Verteilung von X und die konditionale Verteilung von t_j^* mit dem Modell übereinstimmt. Für jeden möglichen Wert $x \in \mathcal{X}$ kann man die auf diese Weise erzeugten Werte als Realisierungen einer Zufallsvariablen $\tilde{T}_x$ ansehen, deren Verteilung durch das Modell, d.h. durch $\tilde{f}(t \mid x; \hat{\theta})$ definiert ist.

Im nächsten Schritt wird eine Transformation der Zufallsvariablen $\tilde{T}_x$ betrachtet, so daß die Abhängigkeit vom jeweils realisierten Wert des Kovariablenvektors verschwindet. Als Transformation wird

$$\tilde{T}_x \longrightarrow J(\tilde{T}_x)\,, \text{ definiert durch } \; t \longrightarrow J(t) = \int_0^t \tilde{r}(\tau \mid x; \hat{\theta})\, \mathrm{d}\tau$$

verwendet, denn die Verteilung von $J(\tilde{T}_x)$ ist dann eine Standard-Exponentialverteilung, d.h. eine Exponentialverteilung mit der konstanten Rate 1. Man sieht das folgendermaßen, wobei ausgenutzt wird, daß es sich um eine monotone Transformation handelt:

$$\begin{aligned} \mathrm{P}\left(J(\tilde{T}_x) > J(t)\right) &= \mathrm{P}\left(\tilde{T}_x > t\right) = \tilde{G}(t \mid x; \hat{\theta}) \\ &= \exp\left\{-\int_0^t \tilde{r}(\tau \mid x; \hat{\theta})\, \mathrm{d}\tau\right\} = \exp\left\{-J(t)\right\} \end{aligned}$$

[77] Die Grundidee wurde von Cox und Snell [1968] vorgeschlagen. Zur Anwendung auf Übergangsratenmodelle vgl. insbesondere Lancaster [1985], Lancaster und Chesher [1987], Blossfeld et al. [1989].

Die Survivorfunktion für die transformierte Zufallsvariable $J(\tilde{T}_x)$ ist also die Survivorfunktion einer Standard-Exponentialverteilung und mithin unabhängig von x.

Diese Tatsache kann ausgenutzt werden, um zu prüfen, ob es plausibel erscheint, daß die als Stichprobe vorliegenden Daten aus dem durch das Modell beschriebenen Zufallsgenerator stammen könnten. Wenn dies der Fall ist, müßte die Anwendung der oben beschriebenen Transformation auf die vorliegenden Daten zu einer Menge transformierter Größen $e_i = J(t_i)$ führen, die näherungsweise einer Standard-Exponentialverteilung folgen.

Die Größen e_i werden als *verallgemeinerte Residuen* bezeichnet. Um zu prüfen, ob sie näherungsweise einer Standard-Exponentialverteilung folgen, kann ihre Survivorfunktion berechnet werden. Dafür sollte das Kaplan-Meier-Verfahren verwendet werden, um berücksichtigen zu können, daß einige Beobachtungen und mithin die ihnen korrespondierenden Residuen rechts zensiert sein können. Wenn $\tilde{G}_r(t)$ die auf diese Weise berechnete Survivorfunktion bezeichnet, kann man schließlich die Abbildung

$$t \longrightarrow -\log\left\{\tilde{G}_r(t)\right\} \tag{3.50}$$

betrachten. Wenn die Residuen näherungsweise einer Standard-Exponentialverteilung folgen, sollte diese Abbildung näherungsweise einer 45°-Linie entsprechen.

Zur Illustration berechne ich verallgemeinerte Residuen für einige typische Übergangsratenmodelle für das Alter bei der ersten Heirat, gemessen auf einer Prozeßzeitachse, die mit dem 15. Lebensjahr beginnt. Die Stichprobenabgrenzung entspricht den unter (c) angeführten Beispielen.

Abbildung 3.7.10 zeigt die in (3.50) definierte Darstellung. Sie zeigt zunächst, daß das Exponentialmodell mit Zeitperioden eine vergleichsweise beste Anpassung an die Daten ermöglicht. Alle anderen im engeren Sinne parametrischen Modelle, auch das erweiterte log-logistische Modell, erscheinen nicht flexibel genug, um die Daten angemessen repräsentieren zu können.

Ein gewisser Mangel der in Abbildung 3.7.10 verwendeten Darstellungsmethode liegt darin, daß sie keine Hinweise auf mögliche Gründe für eine mangelhafte Anpassung des Modells an die Daten gibt. Dies gilt auch für das Exponentialmodell mit Zeitperioden, das ebenfalls bei Residuen, deren Wert größer als 2 ist, einen zunehmend schlechteren Modellfit liefert. Da es keinen direkten Zusammenhang zwischen der Größe der Residuen und der in den Daten realisierten Verweildauern gibt, ist dies schwer zu interpretieren.

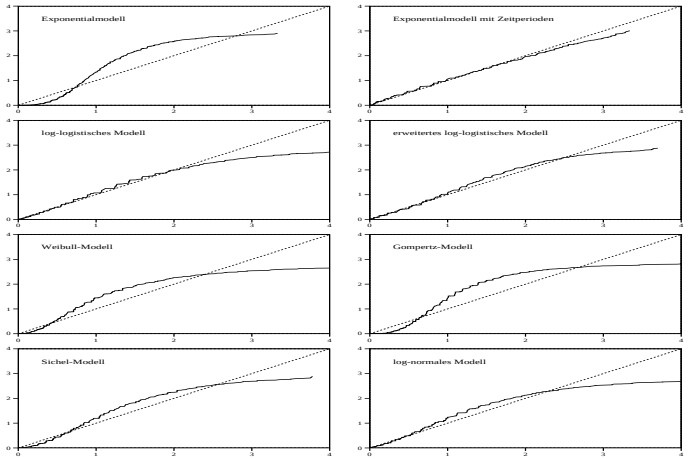


Abbildung 3.7.10. Darstellung verallgemeinerter Residuen für einige Standardmodelle zur Beschreibung der Zeitdauer vom 15. Lebensjahr bis zur ersten Heirat für alle Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1918 – 1968 geboren wurden und mindestens an den ersten drei Wellen teilgenommen haben. Abszisse: Residuen, Ordinate: Minus Logarithmus der Survivorfunktion der Residuen. Schätzung mit dem Kaplan-Meier-Verfahren.

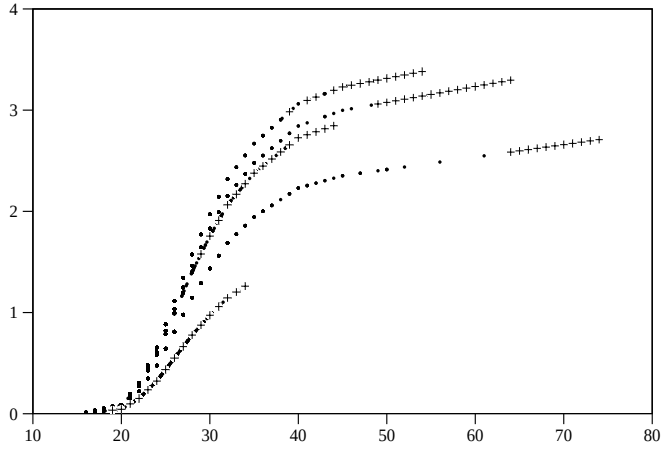


Abbildung 3.7.11. Streuungsdiagramm der mit einem Exponentialmodell mit Zeitperioden (entsprechend Abbildung 3.7.11) berechneten Residuen (Ordinate) vs. Alter bei der ersten Heirat (Abszisse). Rechts zensierte Beobachtungen sind durch ein + markiert.

Einige Hinweise erhält man jedoch, wenn man sich die Verteilung des Zusammenhangs zwischen Residuen und Verweildauer ansieht. Abbildung 3.7.11 zeigt dies in der Form eines Streuungsdiagramms für das Exponentialmodell mit Zeitperioden. Die Abszisse repräsentiert das Heiratsalter, die Ordinate die Residuen. Nicht zensierte Residuen sind durch einen Punkt, zensierte Residuen durch ein Kreuz markiert. Man erkennt, daß hauptsächlich die zensierten Beobachtungen zu vergleichsweise großen Residuen führen. Außerdem stützt diese Abbildung die bereits weiter oben erwähnte Vermutung, daß die nach Geburtskohorten differenzierten Verläufe der Übergangsraten nicht proportional sind.

Kapitel 4

Modelle für Bedingungen von Lebensverläufen

In den beiden vorangegangenen Kapiteln wurde diskutiert, wie mit statistischen Begriffen und Modellen Lebensverläufe *beschrieben* werden können. Der wesentliche Sinn statistischer Verfahren – im Kontext der Lebensverlaufsforschung – wurde darin gesehen, daß mit ihrer Hilfe Gesamtheiten vergleichbarer, d.h. durch ein Biographieschema vergleichbar gemachter Lebensverläufe als ein zeitlicher Prozeß beschrieben werden können. Zwar wurde bereits berücksichtigt, daß Individuen klassifiziert und dadurch die Lebensverläufe in Teilgesamtheiten verglichen werden können; aber auch dies bewegte sich noch im Rahmen einer deskriptiven Interpretation.

Die soziologische Lebensverlaufsforschung ist jedoch nicht nur an einer Beschreibung von Lebensverläufen interessiert ist, sondern möchte darüber hinaus Einsichten in ihre sozialen Bedingungen gewinnen. Diese Idee soll im folgenden etwas näher diskutiert werden. Zunächst wird in Abschnitt 4.1 erörtert, wie von Bedingungen von Lebensverläufen gesprochen werden kann. Dann wird in Abschnitt 4.2 dargestellt, wie die Modellbildung bei parallelen Prozessen vorgenommen werden kann.[1]

4.1 Bedingungen von Lebensverläufen

In der alltäglichen Reflexion von Lebensverläufen erscheint es unproblematisch, von Bedingungen zu sprechen, von denen ihre Entwicklung abhängig ist. Zweierlei erleichtert dabei die umgangssprachliche Kommunikation. Erstens kann auf differenzierte Weise auf unterschiedliche Arten von Bedingungen Bezug genommen werden; zweitens stellt die Umgangssprache eine Vielzahl von Formulierungen bereit, um an die Kontingenz von Lebensverläufen zu erinnern, d.h. daran, daß sich die Entwicklung individueller Lebensverläufe einer einfachen deterministischen Deutung entzieht. Demgegenüber steht die soziologische Lebensverlaufsforschung vor der Schwierigkeit, wie die vielfältige und schillernde Vorstellung einer Bedingtheit von Lebensverläufen in generalisierbarer Weise faßbar gemacht werden kann.

[1] Die Ausführungen dieses Kapitels verdanken sich in wesentlichen Punkten Diskussionen mit Hans-Peter Blossfeld. Wichtige Anregungen entstammen insbesondere seinen Überlegungen zur Frage, wie das Kausalitätsproblem im Hinblick auf ein zeitlich strukturiertes Bedingungsgefüge von Ereignissen reformuliert werden kann, vgl. Blossfeld [1994] sowie Blossfeld und Rohwer [1994].

Ob bzw. wie dies erreicht werden kann, hängt in gewisser Weise davon ab, worin das Ziel der Lebensverlaufsforschung gesehen wird. Wie in Abschnitt 3.1.2 überlegt wurde, können zwei Auffassungen unterschieden werden. Einerseits die Auffassung, daß das Ziel im Gewinnen von Einsichten liegt, mit denen *individuelle* Lebensverläufe erklärt werden können; andererseits die Auffassung, daß Einsichten in gesellschaftliche Verhältnisse als Bedingungen individueller Lebensverläufe gewonnen werden sollen. Je nachdem, von welcher Auffassung ausgegangen wird, stellt sich das Problem der Kontingenz individueller Lebensverläufe auf unterschiedliche Weise. Geht man von der ersten Auffassung aus, erscheint die Kontingenz individueller Lebensverläufe als eine Grenze für die Gewinnung verläßlichen Kausalwissens. Geht man von der zweiten Auffassung aus, besteht die Möglichkeit, von der Kontingenz individueller Lebensverläufe zu abstrahieren. Der empirische Gegenstand der Lebensverlaufsforschung besteht dann nicht in individuellen Lebensverläufen, sondern in Gesamtheiten von Lebensverläufen; die mit der statistischen Modellbildung intendierten Aussagen beziehen sich auf Eigenschaften von Gesamtheiten, nicht auf Eigenschaften von Individuen. Individuelle Lebensverläufe sind nur von Interesse, insoweit sie zur statistisch beschreibbaren Vielfalt einer Gesamtheit von Lebensverläufen beitragen.

Auch wenn man, wie in dieser Arbeit, der zweiten Auffassung zu folgen versucht, stellt sich allerdings das Problem, wie von *Bedingungen* von Lebensverläufen gesprochen werden kann. Dies ist deshalb ein Problem, weil nur im Hinblick auf Individuen von Bedingungen ihrer Lebensverläufe gesprochen werden kann. Gleichwohl ermöglicht die Betrachtung von Gesamtheiten von Lebensverläufen eine wesentliche Abstraktion. Sie erlaubt es, davon zu abstrahieren, wie Bedingungen im Einzelfall wirken. Der Nachweis, daß eine Bedingung wirksam ist, muß nicht im Einzelfall erbracht werden – was in der Regel kaum möglich ist –, sondern es genügt zu zeigen, daß sich in einer Gesamtheit von Lebensverläufen die Verteilung gewisser Merkmale auf „signifikante“ Weise verändert. Die Abstraktion von den konkreten Formen, in denen sich die Kontingenz der individuellen Lebensverläufe vollzieht, schließt es zwar nicht aus, darüber generalisierbare Vorstellungen zu entwickeln, d.h. hermeneutisch nachvollziehbare Geschichten zu erzählen, die verständlich erscheinen lassen, wie die Lebensverläufe einer Gesamtheit von Individuen durch historische Ereignisse und soziale Bedingungen geprägt werden. In gewisser Weise liegt hierin die Hauptaufgabe der Theoriebildung. Aber es wäre, wie ich glaube, ein Mißverständnis, den Sinn von Geschichten dieser Art darin sehen zu wollen, daß durch sie individuelle Lebensverläufe verständlich werden könnten. Es erscheint angemessener, ihren theoretischen Sinn darin zu sehen, daß sie Reflexionsmöglichkeiten, Begriffe und Modelle, anbieten, mit denen von individuellen Lebensgeschichten abstrahiert werden kann.

Selbst wenn man bereit ist, ein sehr weitgehendes Abstraktionsniveau zu akzeptieren, gibt es vermutlich kein einfaches Schema, in dem auf einheitliche Weise von Bedingungen von Lebensverläufen gesprochen werden kann. Zumindest zwei Betrachtungsweisen sollten unterschieden werden. (a) Zunächst die für die meisten soziologischen Theorieansätze zentrale Vorstellung, daß gesellschaftliche Verhältnisse Handlungsmöglichkeiten der Individuen strukturieren und dadurch Bedingungen für Lebensverläufe sind. Dies entspricht im wesentlichen der in der Einleitung skizzierten Auffassung, daß gesellschaftliche Verhältnisse durch soziale Regeln beschrieben werden können. (b) Andererseits die Vorstellung, daß die Entwicklung von Lebensverläufen von *Ereignissen* abhängig ist; zum Beispiel der Ausbruch einer Krankheit, ein Verkehrsunfall, eine Heirat, der Verlust eines Arbeitsplatzes, usw.

Diese Möglichkeiten, von Bedingungen von Lebensverläufen zu sprechen, schließen sich nicht aus, sondern können miteinander verknüpft werden. Denn bei allen Ereignissen, die nicht unmittelbar zum Tod führen, kann die Frage gestellt werden, wie die betroffenen Personen mit dem Ereignis umgehen. Diese Frage bezieht sich jedoch auf Handlungen, die wiederum sozialen Bedingungen unterliegen, d.h. sozialen Regeln folgen oder auch nicht. Insofern glaube ich, daß die grundlegende Aufgabe darin besteht, die sozialen Regeln zu beschreiben, an denen sich die Menschen in ihren Lebensverläufen orientieren. Die Bezugnahme auf Ereignisse dient in diesem Rahmen im wesentlichen nur dem Zweck, Regeln spezifischer, d.h. ereignisbezogen formulieren zu können.

4.1.1 Mehrdimensionale Zustandsräume

Um diese Überlegung formal zu präzisieren, kann an die bisherige Darstellung von Prozessen angeknüpft werden. Denn die in Kapitel 3 behandelten eindimensionalen Modelle liefern bereits eine einfache statistische Form, um von Ereignissen und Regeln als Bedingungen von Lebensverläufen sprechen zu können. Das Ereignis, das den weiteren Lebensverlauf bedingt, ist der Übergang in den Anfangszustand einer Episode; das den dann einsetzenden Episodenverlauf beschreibende System zustandsspezifischer Übergangsraten charakterisiert die Regeln, denen die Individuen – gewissermaßen als Reaktion auf dieses Ereignis – folgen.

Die weitere Überlegung besteht einfach darin, diesen Gedankengang zu verallgemeinern, d.h. eine Vielzahl von sich wechselseitig bedingenden Ereignissen zu betrachten. Anstelle eines eindimensionalen Zustandsraum werden jetzt mehrere Zustandsräume betrachtet.[2] Wie bisher gehen wir von einer Längsschnittgesamtheit Ω aus. Zunächst wird außerdem eine dis-

[2] Ereignisanalytische Modelle für parallele Prozesse sind in der Literatur bereits mehrfach diskutiert worden. Der im folgenden verfolgte Ansatz folgt Überlegungen, die u.a. von Gardner und Griffin [1986], Aalen [1987], Huinink [1992] und Pötter [1993] dargestellt worden sind. Ein Hinweis darauf, daß zeitabhängige Kovariablen am besten im Kontext paralleler Prozesse zu betrachten sind, wurde auch bereits von Blossfeld et al. [1986, S. 155 und 193] gegeben.

krete Prozeßzeitachse $\mathcal{T}$ vorausgesetzt. Ein parallel in m Zustandsräumen ablaufender Prozeß kann dann durch folgendes System von Zufallsvariablen repräsentiert werden:

$$(Y_t^{(1)}, \ldots, Y_t^{(m)}) : \Omega \longrightarrow \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_m \qquad t \in \mathcal{T} \tag{4.1}$$

$Y_t^{(j)}(\omega)$ ist der Zustand, den das Individuum $\omega \in \Omega$ zum Zeitpunkt t im j.ten Zustandsraum einnimmt. Während der Prozeß abläuft, können in allen Zustandsräumen simultan Ereignisse, d.h. Zustandsänderungen eintreten. Zur Veranschaulichung kann folgende Abbildung dienen.

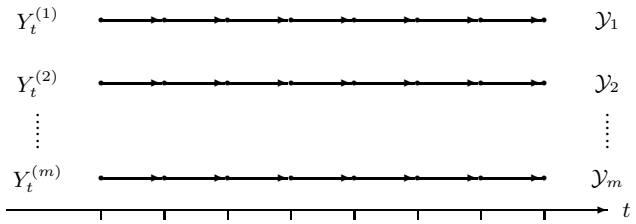


Abbildung 4.1.1 Parallele Prozesse auf einer diskreten Zeitachse.

Die Grundidee besteht darin, daß die Ereignisse, die in den verschiedenen Zustandsräumen stattfinden, sich wechselseitig bedingen können. Zum Beispiel kann im Zustandsraum $\mathcal{Y}_1$ dargestellt werden, wie die Lebensgemeinschaften der Individuen verlaufen, im Zustandsraum $\mathcal{Y}_2$ kann erfaßt werden, ob Kinder geboren werden, und der Zustandsraum $\mathcal{Y}_3$ kann dazu dienen, die Erwerbsbeteiligung der Individuen zu erfassen. Die simultane Betrachtung dieser Zustandsräume erlaubt es dann, die Frage zu stellen, wie sich die Bildung und Auflösung von Lebensgemeinschaften, die Geburt von Kindern und Ereignisse, die mit der Erwerbstätigkeit verbunden sind, wechselseitig beeinflussen.

Die in (4.1) formal definierte und in Abbildung 4.1.1 veranschaulichte Vorstellung paralleler Prozesse ist sehr allgemein. Es kann sich um Prozesse handeln, die unverbunden nebeneinander ablaufen oder die sich einseitig oder wechselseitig beeinflussen. Außerdem ist die Bezugnahme auf Individuen sehr allgemein. Die formale Zurechenbarkeit eines Prozesses zu den Individuen einer Grundgesamtheit impliziert nicht, daß es sich um Zustände handeln muß, die als Eigenschaften der Individuen definiert werden können. Zum Beispiel können Prozesse definiert werden, um gewisse Merkmale der sozialen Umgebung der Individuen zu repräsentieren, etwa Indikatoren für die Beschaffenheit lokaler Arbeitsmärkte. Dies sind offensichtlich keine Eigenschaften der Individuen, sie werden jedoch den Individuen formal zugerechnet, um damit zum Ausdruck zu bringen, daß sich deren Lebensverläufe innerhalb dieser lokalen Arbeitsmärkte bewegen. Diese formale Zurechnung erlaubt auch die Berücksichtigung räumlicher Mobilität; in diesem Beispiel wird stets auf Merkmale desjenigen Arbeitsmarkts

Bezug genommen, in dem sich ein Individuum vorübergehend aufhält.

Das in (4.1) definierte Schema paralleler Prozesse kann auch verwendet werden, um gewisse Aspekte sozialer Interaktion zu beschreiben. Dies erscheint jedenfalls dann möglich, wenn es sich um Personen handelt, die durch längerfristige Lebensgemeinschaften miteinander verbunden sind. Formal geschieht dies wiederum dadurch, daß jedem Individuum ein Prozeß (oder eine Menge von Prozessen) zugerechnet wird, der gewisse Aspekte des Lebensverlaufs seines jeweiligen Lebenspartners repräsentiert; der oder die Partner werden als Aspekte der sozialen Umgebung der für die Grundgesamtheit ausgewählten Individuen betrachtet.

In den restlichen Abschnitten dieses Kapitels soll überlegt werden, wie eine statistische Beschreibung und Analyse paralleler Prozesse so vorgenommen werden kann, daß dadurch Einsichten in Bedingungen von Lebensverläufen gewonnen werden können. Im Mittelpunkt steht die Frage, in welcher Weise davon gesprochen werden kann, daß sich Ereignisse wechselseitig bedingen.

4.1.2 Soziale Regeln und (probabilistische) Kausalität

Um einen Zugang zur Frage zu finden, wie *interdependente* Prozesse verstanden werden können, muß zunächst überlegt werden, was es heißen soll, daß sich Ereignisse bedingen. Das traditionelle Denkschema kreist um den Begriff der Kausalität. Die Grundvorstellung ist, daß zumindest ein Teil der in der Welt stattfindenden Ereignisse durch kausale Gesetzmäßigkeiten miteinander verknüpft ist. Wenn man versucht, diese Vorstellung zu verwenden, um den Zusammenhang zwischen Ereignissen in menschlichen Lebensverläufen zu verstehen, erscheint es vor allem wichtig, sich darüber klar zu werden, wodurch solche Ereignisse miteinander verbunden sind. Dies hängt sicherlich auch von der Art der Ereignisse ab. Handelt es sich zum Beispiel um einen Flugzeugabsturz, hat dieses Ereignis auf die Lebensverläufe der Insassen in den meisten Fällen eine Wirkung, die von ihrer subjektiven Verfassung, ihren Wünschen, Hoffnungen, Absichten und Präferenzen unabhängig ist. Jeder versteht den Satz, daß der Flugzeugabsturz die Ursache für den Tod der Insassen gewesen ist. Anders verhält es sich bei Ereignissen, die nicht unmittelbar zum Tod führen. Bei Ereignissen dieser Art hängt der weitere Lebensverlauf nicht nur von dem Ereignis, sondern auch davon ab, wie sich das Individuum, um dessen Lebensverlauf es sich handelt, verhält. Erst als Folge dieses Verhaltens kommt es zu einer neuen Situation.

Als Prototyp für diese Situation kann man eine einfache Episode mit konkurrierenden Risiken betrachten. Zunächst findet ein Ereignis statt, das die Episode, d.h. den Übergang in den Ausgangszustand der Episode einleitet. Dann kann, nach mehr oder weniger langer Zeit, eines von mehreren möglichen Folgeereignissen stattfinden. Nimmt man auf jeweils bestimmte Individuen Bezug, ist der Zusammenhang kontingent. Betrachtet man

jedoch Gesamtheiten von Individuen, kann der Zusammenhang statistisch beschrieben werden, durch ein System zustandsspezifischer Übergangsraten, und man kann dies als eine Beschreibung der sozialen Regeln ansehen, denen „die Individuen“ in der durch das Anfangsereignis hervorgebrachten Situation folgen. Folgende Abbildung veranschaulicht den Zusammenhang zwischen zwei Ereignissen.

$$E_1 \xrightarrow[\text{soziale Regeln}]{\text{individuelles Verhalten}} E_2$$

Das Ereignis E_2 wird „als Folge“ des Ereignisses E_1 durch individuelles Verhalten hervorgebracht. Betrachtet man die Situation im Hinblick auf ein jeweils bestimmtes Individuum, gibt es keinen gesetzmäßigen Zusammenhang. Wenn man jedoch Gesamtheiten von Individuen und, in diesem Zusammenhang, Gesamtheiten von Ereignissen des gleichen Typs betrachtet, läßt sich ein statistischer Zusammenhang zwischen den beiden Ereignissen ermitteln, der sich – nicht immer, aber in vielen Fällen – durch die Vorstellung interpretieren läßt, daß die beteiligten Individuen sozialen Regeln folgen.

a) Eine wesentliche Sinnvoraussetzung dieser Interpretation liegt darin, daß nicht einzelne Individuen, sondern Gesamtheiten von Individuen betrachtet werden. Denn nur durch eine Bezugnahme auf Gesamtheiten von Individuen kann der Begriff einer sozialen Regel empirisch erklärt werden.

b) Ob man den für eine Gesamtheit von Individuen ermittelbaren statistischen Zusammenhang zwischen Ereignissen als eine Gesetzmäßigkeit bezeichnen will, ist – wie ich in der Einleitung deutlich zu machen versucht habe – weitgehend nur eine Frage der Rhetorik. In dieser Arbeit spreche ich stattdessen von sozialen Regeln oder Regelmäßigkeiten, um damit zum Ausdruck zu bringen, daß es sich um Beschreibungen von Verhaltensweisen in Gesamtheiten von Individuen handelt und mithin um Sachverhalte, die sich im Zeitablauf verändern können.

c) Den wesentlichen Sinn der Bezugnahme auf Gesamtheiten von Individuen sehe ich darin, daß dadurch von der Kontingenz der individuellen Lebensverläufe und dem damit verbundenen Problem, ob sich die Lebensverläufe jeweils konkreter, bestimmter Individuen kausal erklären lassen, abstrahiert werden kann. Denn bei den in der Lebensverlaufsforschung typischerweise betrachteten Ereignissen erscheint es unsinnig, auf der Ebene der Individuen einen deterministischen Kausalzusammenhang anzunehmen. Das Ereignis, daß ein Kind ein gewisses Alter erreicht, ist nicht die Ursache dafür, daß es in die Schule kommt. Das Ereignis, daß in einer nicht-ehelichen Lebensgemeinschaft ein Kind geboren wird, ist nicht die Ursache dafür, daß dann geheiratet (oder nicht geheiratet) wird. Angemessener erscheint etwa folgende Formulierung: Durch Ereignisse dieser Art entstehen soziale Situationen, in denen ein sich an Regeln orientierendes Verhalten darüber entscheidet, was schließlich geschieht.

d) Gleichwohl erscheint es sinnvoll, an der Vorstellung festzuhalten, daß sich Ereignisse *bedingen* können. Die Schwangerschaft bzw. Geburt eines Kindes kann zwar im allgemeinen nicht sinnvoll als Ursache einer nachfolgenden Heirat angesehen werden, wohl aber – in vielen Fällen – als eine Bedingung. Der wesentliche Unterschied liegt darin, daß sinnvoll von Bedingungen gesprochen werden kann, ohne damit sogleich notwendige oder hinreichende Bedingungen (im logischen Sinne) zu meinen.[3] Die Schwangerschaft bzw. Geburt eines Kindes ist zunächst insofern eine Bedingung, als sie sicherlich bei der Entscheidung, ob daraufhin geheiratet werden soll, eine Rolle gespielt hat. Darüberhinaus kann insoweit von einer effektiven Bedingung gesprochen werden, als es möglich ist, eine soziale Regel bzw. Regelmäßigkeit zu ermitteln, die die beiden mögichen Ereignisse miteinander verknüpft.

e) Man könnte einwenden, daß diese Beispiele nur zeigen, daß der „kausale Mechanismus“, der die Abfolge von Ereignissen in Lebensverläufen vermittelt, noch nicht genügend detailliert spezifiziert und erfaßt worden ist, und daß insofern mit Beispielen dieser Art nicht gezeigt werden könne, daß eine kausale Erklärung des Ablaufs *individueller* Lebensverläufe unmöglich sei. Dieser Einwand verfehlt jedoch, wie ich glaube, das Problem. Unser Problem ist nicht, ob es „im Prinzip“ möglich ist, menschliche Handlungen (und infolgedessen Lebensverläufe) kausal zu erklären; das ist eine offene Frage.[4] Das Problem ist vielmehr, wie der Zusammenhang *derjenigen* Ereignisse, die typischerweise im Kontext der soziologischen Lebensverlaufsforschung betrachtet werden, beschrieben und interpretiert werden kann.[5] Das Argument ist, daß die Erfahrung zeigt, daß es zwischen *diesen* Ereignissen keinen deterministischen Zusammenhang gibt und daß diese Tatsache in einer angemessenen Beschreibung zum Ausdruck kommen sollte. Dieses Argument gilt, wie ich glaube, unabhängig davon, ob man an die Existenz eines „kausalen Mechanismus“ glaubt, der den Zusammenhang der beoachtbaren Ereignisse hervorgebracht haben könnte.

---

[3] Vgl. Scriven [1966], der auf sehr interessante Weise den Versuch unternommen hat, das Reden von Kausalzusammenhängen aus seiner durch die traditionelle Philosophie vorgenommenen Fixierung auf allgemeine Gesetzmäßigkeiten zu befreien, um den vielfältigen umgangssprachlichen und realwissenschaftlichen Vorstellungen über Bedingungszusammenhänge Raum zu verschaffen.

[4] Diese Frage wird üblicherweise im Hinblick auf menschliche Handlungen gestellt und diskutiert, ein ähnliches Problem kausaler Erklärbarkeit kann jedoch auch anhand von Situationen reflektiert werden, in denen ein Ereigniszusammenhang explizit durch Zufallsgeneratoren vermittelt ist; ein interessantes Beispiel wurde von Dretske und Snyder [1972] diskutiert.

[5] Die Bezugnahme auf Ereignisse eines bestimmten Typs ist hier offenbar wichtig. Wie bereits zu Beginn dieses Abschnitts gesagt worden ist, gibt es sicherlich Ereignisse, deren Folgen für menschliche Lebensverläufe sich am angemessensten als einfache kausale Folgen darstellen lassen. Im Mittelpunkt der soziologischen Lebensverlaufsforschung stehen jedoch Ereignisse, deren Folgen auch davon abhängen, wie sich die Individuen verhalten.

f) Unsere Betrachtungsweise des Zusammenhangs von Ereignissen im Kontext von Lebensverläufen schließt es nicht aus, nach vermittelnden Ereignissen Ausschau zu halten. Für die Frage, wie der Zusammenhang solcher Ereignisse angemessen beschrieben und theoretisch gedeutet werden kann, ergeben sich daraus jedoch keine neuen Gesichtspunkte. Der Gewinn liegt darin, daß die sozialen Regeln, denen Menschen in der Entwicklung ihrer Lebensverläufe folgen, detaillierter beschrieben werden können. Tatsächlich stößt jeder Versuch, sehr detaillierte Ereignisketten in Erfahrung zu bringen, rasch an Grenzen. Das Hauptproblem liegt nicht in der Verfügbarkeit von Daten, sondern darin, daß bei einer immer detaillierter werdenden Beschreibung schließlich keine sinnvollen Gesamtheiten von Individuen mehr gebildet werden können und infolgedessen eine wesentliche Sinnvoraussetzung für statistische Analysen des Zusammenhangs von Ereignissen verlorengeht.[6]

g) Eine Alternative zu traditionellen Kausalvorstellungen scheint der Begriff einer „probabilistischen Kausalität“ zu liefern. Damit kann, wie es scheint, auf zumindest rhetorisch einfache Weise der Tatsache Rechnung getragen werden, daß es im allgemeinen keinen deterministischen Zusammenhang zwischen Ereignissen im Kontext menschlicher Lebensverläufe gibt. Abgesehen von dieser rein negativen Konnotation liefert der Begriff jedoch keine klare Vorstellung. Dies zeigt bereits die inzwischen breit verzeigte und kontroverse Diskussion.[7] Dies ist verständlich, weil bei der Definition des Begriffs üblicherweise auf das Konzept einer bedingten Wahrscheinlichkeit zurückgegriffen wird. Infolgedessen treten alle Probleme auf, die mit der Frage verbunden sind, wie eine empirisch angemessene Interpretation von Wahrscheinlichkeitsaussagen erreicht werden kann. Insbesondere entsteht die Frage, ob bei der Formulierung von Wahrscheinlichkeitsaussagen auf Gesamtheiten Bezug genommen werden muß, so daß von relativen Häufigkeiten gesprochen werden kann, oder ob auch über singuläre Ereignisse bzw. Zusammenhänge von Ereignissen sinnvolle Wahrscheinlichkeitsaussagen getroffen werden können.

Betrachten wir als Beispiel noch einmal nicht-eheliche Lebensgemeinschaften. Man kann feststellen, daß die Wahrscheinlichkeit, daß eine Heirat stattfindet, wesentlich größer wird, wenn eine Schwangerschaft eintritt. Man könnte infolgedessen sagen, daß die Schwangerschaft bzw. die Geburt eines Kindes eine „probabilistische Ursache“ für die Heirat ist. Aber welchen Erkenntnisgewinn liefert diese Formulierung? Tatsächlich liefert sie nur eine Reformulierung der empirisch ermittelbaren Tatsache, daß die relative Häufigkeit einer Heirat in nicht-ehelichen Lebensgemeinschaften größer ist, nachdem die Frau schwanger geworden bzw. ein Kind geboren worden ist. Man könnte sagen, daß wir nicht nur an einer Beschreibung dieser Tatsache interessiert sind, sondern wissen möchten, welcher Art der Zusammenhang zwischen den Ereignissen „Schwangerschaft“ und „Heirat“ in nicht-ehelichen Lebensgemeinschaften ist. Um dies herauszufinden, hilft jedoch die Formulierung, daß sie durch eine „probabilistische Gesetzmäßigkeit“ verbunden sind, nicht weiter. Es erscheint sinnvoller, sich zu überlegen, wodurch der Zusammenhang der Ereignisse tatsächlich vermittelt wird – nämlich durch Überlegungen und Entscheidungen der beteiligten Personen –, und die statistisch ermittelbare Regelmäßigkeit als eine abstrakte – nämlich von den Details der Vermittlung abstrahierende – Beschreibung des Ereigniszusammenhangs aufzufassen. Da hierbei darauf Bezug genommen wird, daß der Ereigniszusammenhang durch Individuen hergestellt wird, kann sich die Frage anschließen, ob der statistisch ermittelte und insoweit rein deskriptive Zusammenhang auf eine soziologisch sinnvolle Weise durch die Vorstellung gedeutet werden kann, daß die Individuen, die den Ereigniszusammenhang vermitteln, dabei gewissen sozialen Regeln folgen.

h) Wenn *demgegenüber* die Vorstellung angeführt wird, daß der Zusammenhang von Ereignissen im Kontext von Lebensverläufen am besten durch „probabilistische Gesetzmäßigkeiten“ charakterisiert werden kann, scheint die Intention hauptsächlich darin zu liegen, Fragen nach der Beschaffenheit des die Ereignisse vermittelnden Zusammenhangs gegenstandslos zu machen.[8] Diese Haltung ist verständlich, wenn sich das Erkenntnisinteresse nur auf die konditionale Prognostizierbarkeit von Ereignissen richtet; dann erscheint es sogar sinnvoll, (probabilistische) Kausalität durch die Möglichkeit *zu definieren*, daß Vorhersagen gemacht werden können.[9] Wenn man demgegenüber das primäre Erkenntnisinteresse der soziologischen Lebensverlaufsforschung darin sieht, Einsichten in gesellschaftliche Verhältnisse (als Bedingungen von Lebensverläufen) zu gewinnen, erscheint es sinnvoller, der Frage nach der Beschaffenheit von Ereigniszusammenhängen (im Kontext von Lebensverläufen) nicht auszuweichen, sondern in ihr eine zulässige Frage soziologischer Theoriebildung zu sehen.[10]

[6] Detaillierte Beschreibungen einzelner Lebensverläufe können natürlich interessante Einsichten vermitteln, insbesondere unserem Verständnis für deren Kontingenz dienen; sie liefern jedoch keine Einsichten in die sozialen Regeln, denen „die Individuen“ folgen.

[7] Vgl. die ausführliche Diskussion bei Eells [1991].

[8] Vgl. dazu die interessanten Ausführungen von Rosen [1983], die zu zeigen versucht, daß probabilistische Vorstellungen über Kausalität in vielen Fällen hauptsächlich dazu herangezogen werden, um an der Vorstellung festhalten zu können, daß Ereigniszusammenhänge durch Gesetzmäßigkeiten beherrscht werden.

[9] Eine in der Ökonometrie weitgehend akzeptierte Variante einer solchen pragmatischen Definition von Kausalität wurde von Granger gegeben. Ihr Grundgedanke besteht darin, daß ein Prozeß A durch einen Prozeß B kausal beeinflußt wird, wenn die Kenntnis von B zusätzliche Informationen liefert, um die Entwicklung von A vorauszusagen. Vgl. Granger [1982].

[10] Der Versuch, den Zusammenhang von Ereignissen im Kontext von Lebensverläufen durch Rückgriff auf eine Vorstellung sozialer Regeln zu deuten, schließt natürlich die Möglichkeit, Ereignisse vorauszusagen, nicht aus. Allerdings zeigt diese Deutung auch, warum solche Voraussagen problematisch und unsicher sind: weil sich soziale Regeln

4.1.3 Interdependente Prozesse

Es bleibt zu überlegen, wie interdependente Prozesse sinnvoll beschrieben werden können. Ausgangspunkt ist die Vorstellung, daß sich Ereignisse – im Kontext von Lebensverläufen – bedingen, so daß davon gesprochen werden kann, daß ein Ereignis „eine Folge“ eines oder mehrerer anderer Ereignisse ist. (Man kann natürlich Aussagen dieser Art auch dann für sinnvoll halten, wenn man ihre im vorangegangenen Abschnitt versuchte Interpretation ablehnt.) Prototyp ist ein Zusammenhang zwischen zwei Ereignissen:

$$E_1 \xrightarrow{\quad r(t) \quad} E_2$$

Unter Bezugnahme auf eine Gesamtheit von Lebensverläufen, bei denen die Ereignisfolge $E_1 \rightarrow E_2$ auftritt, kann man ihren Zusammenhang durch eine Übergangsrate $r(t)$ beschreiben. Durch sie kann beschrieben werden, wie das Ereignis E_1 das Ereignis E_2 bedingt. Sie berücksichtigt nicht nur die zeitliche Folge der Ereignisse, sondern beschreibt darüber hinaus die „Zeitstruktur“, in der die Ereignisse aufeinander folgen. Es erscheint deshalb sinnvoll, den Begriff der Übergangsrate als einen Elementarbegriff anzusehen, mit dem der Zusammenhang von Ereignissen im Kontext von Lebensverläufen beschrieben werden kann (unbeschadet der Frage, wie eine soziologisch angemessene Deutung solcher Ereigniszusammenhänge erreicht werden kann).

Diese Basisvorstellung kann verwendet werden, um auch parallele Prozesse so zu beschreiben, daß ihr Ablauf als eine Folge von Ereignissen sichtbar wird. Um dies zu erreichen, ist es zunächst nur erforderlich, aus den einzelnen Zustandsräumen der parallel ablaufenden Prozesse einen gemeinsamen Zustandsraum zu bilden. Geht man von den in (4.1) eingeführten Bezeichnungen aus, wird dies durch die Definition einer Folge von Zufallsvariablen

$$Y_t = (Y_t^{(1)}, \ldots, Y_t^{(m)}) : \Omega \longrightarrow \mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_m \qquad t \in \mathcal{T} \tag{4.2}$$

erreicht. Diese Folge von Zufallsvariablen repräsentiert einen mehrdimensionalen Prozeß. Innerhalb dieses Prozesses findet immer dann ein Ereignis statt, wenn sich in mindestens einem der beteiligten Zustandsräume der Zustand eines Individuums verändert. Für jedes Individuum kann also dieser mehrdimensionale Prozeß als eine Folge von Ereignissen beschrieben werden.

Allerdings zielt die statistische Beschreibung nicht darauf ab, wie der Prozeß bei jedem einzelnen Individuum verläuft, die Beschreibung soll vielmehr zeigen, wie der Zusammenhang der Ereignisse in der Gesamtheit der individuellen Lebensverläufe beschaffen ist und wie sich die parallel ablaufenden Prozesse wechselseitig bedingen. Dafür gibt es unterschiedliche

im Zeitablauf verändern. Insoweit hängt die Möglichkeit von Voraussagen vom Ausmaß und von der Geschwindigkeit sozialen Wandels ab.

Möglichkeiten.

a) Die einfachste Möglichkeit besteht darin, für jeden Zustand, der im gemeinsamen Zustandsraum $\mathcal{Y}$ eingenommen werden kann, eine Situation konkurrierender Risiken zu formulieren. D.h. für jeden Zustand $y \in \mathcal{Y}$ wird die Gesamtheit der möglichen Folgezustände betrachtet. Der Episodenverlauf, der mit einem Übergang in den Zustand y beginnt, kann dann durch ein System zustandsspezifischer Übergangsraten beschrieben werden. Diese Vorgehensweise knüpft auf einfache Weise an die in Kapitel 3 diskutierten Modelle an, sie hat aber offensichtlich den Mangel, daß sie keine Einsichten darüber liefert, wie sich die parallel ablaufenden Prozesse wechselseitig bedingen.

b) Um Einsichten in die Interdependenz der parallel ablaufenden Prozesse zu gewinnen, muß eine Modellformulierung erreicht werden, die es erlaubt, Aussagen von der Art zu formulieren, daß ein Ereignis im Zustandsraum $\mathcal{Y}_j$ ein Ereignis im Zustandsraum $\mathcal{Y}_k$ „zur Folge hat“. Um der Vorstellung zu genügen, daß es sich um interdependente Prozesse handelt, muß die Möglichkeit zugelassen werden, daß jedes Ereignis, das sich in einem der beteiligten Zustandsräume ereignen kann, eine Bedingung für jedes andere Ereignis sein kann. Allerdings sollte dabei die zeitliche Ordnung der Ereignisse berücksichtigt werden. Damit sinnvoll davon gesprochen werden kann, daß ein Ereignis E_2 durch ein Ereignis E_1 bedingt wird, muß das Ereignis E_1 dem Ereignis E_2 zeitlich vorausgehen.

Daß nur die jeweilige Vergangenheit sinnvoll als Bedingung der jeweiligen Zukunft eines Prozesses angesehen werden sollte, wurde bereits in Abschnitt 2.4.3 (S. 82) als ein grundlegendes Prinzip zur Beschreibung zeitlicher Prozesse formuliert. Im Kontext einer statistischen Beschreibung von Prozessen handelt es sich um ein Prinzip für die Bildung bedingter Wahrscheinlichkeiten, im Hinblick auf den in (4.2) definierten parallelen Prozeß:

$$\mathrm{P}\left(Y_t = y_t \mid Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \ldots\right) \tag{4.3}$$

Folgt man dem Prinzip, daß ein Prozeß so beschrieben werden sollte, daß sichtbar wird, wie die jeweilige Vergangenheit die jeweilige Gegenwart (und mithin Zukunft) bedingt, kann man noch einen Schritt weiter gehen und das Prinzip – im Hinblick auf parallel ablaufende Prozesse – folgendermaßen formulieren: Der Zustand, der in jedem der Teilprozesse zum Zeitpunkt t eingenommen wird, hängt nur von der *gemeinsamen* Vorgeschichte bis zum Zeitpunkt $t-1$ ab, nicht jedoch davon, welche Zustände zum Zeitpunkt t in den jeweils anderen Zustandsräumen eingenommen werden. In statistischer Formulierung entspricht dies folgendem Prinzip einer konditionalen Unabhängigkeit:

$$\begin{aligned}\mathrm{P}\left(Y_t = y_t \mid Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \ldots\right) = \\ \prod_{j=1}^{m} \mathrm{P}\left(Y_t^{(j)} = y_t^{(j)} \mid Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \ldots\right)\end{aligned} \tag{4.4}$$

Dieses Prinzip folgt unmittelbar aus der Annahme, daß nur die jeweilige Vergangenheit die jeweilige Gegenwart (und mithin die Zukunft) beeinflussen kann. Es impliziert, daß sich simultan stattfindende Ereignisse nicht bedingen, weder wechselseitig noch einseitig.

Ich möchte betonen, daß es sich hierbei nicht um eine Annahme im gewöhnlichen Sinne des Wortes handelt, also um eine Hypothese, die empirisch infrage gestellt werden könnte. Zweifellos ist es möglich, einen Prozeß, insbesondere eine Menge parallel ablaufender Prozesse, auf viele unterschiedliche Weisen zu beschreiben. Es gibt keinen logischen Einwand dagegen, jedes Ereignis als Bedingung jedes anderen Ereignisses zu betrachten.[11] Da der Begriff einer bedingten Wahrscheinlichkeit selbst zeitlos ist, können mit seiner Hilfe Bedingungsverhältnisse für beliebige Ereignisse angenommen werden. Zum Beispiel liefert der Ausdruck $\mathrm{P}(Y_{t-1} \mid Y_t)$ eine Formulierung, die innerhalb des für die Definition von (4.1) vorausgesetzten Wahrscheinlichkeitsraums sinnvoll gebildet werden kann. Daß diese Möglichkeit besteht, liefert natürlich keinen Beweis dafür, daß die Zukunft die Vergangenheit beeinflussen kann.

Ich betrachte die in (4.4) formulierte Annahme als ein Prinzip, das angibt, wie parallel ablaufende Prozesse sinnvoll beschrieben werden können. Folgt man diesem Prinzip, gelangt man zu Beschreibungen, die mit unseren traditionellen Zeitvorstellungen vereinbar sind. Andere Formen, parallel ablaufende Prozesse zu beschreiben, sind jedoch vorstellbar und auch gelegentlich vorgeschlagen worden. Zum Beispiel gehen Courgeau und Lelièvre [1992, S. 86f] ausdrücklich von der Vorstellung aus, daß sich simultan stattfindende Ereignisse wechselseitig bedingen können. Konsequenterweise bemerken sie: „We shall therefore consider our model as beeing outside the framework of causal explanation, and shall consider it rather as an analysis of phenomena and their interactions.“[12]

4.1.4 Exogene und endogene Bedingungen

Der hier verfolgte Versuch, einen Zugang zur Beschreibung parallel ablaufender Prozesse zu finden, führt dazu, daß auch das Problem „exogener“ und „endogener“ Bedingungen von Lebensverläufen in einer Form erscheint, die etwas von seiner üblichen Behandlung abweicht. Einerseits erfolgt, wie schon ausgeführt worden ist, eine Abkehr von der Vorstellung eines statischen Systems interdependenter Variablen. Das in diesem Kontext unlösbar erscheinende Problem, kausale Abhängigkeiten zu identifizieren, wird in gewisser Weise gegenstandslos, wenn man stattdessen

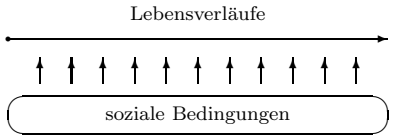
[11] Man kann dies als eine Folge dessen ansehen, daß die in unserer bisherigen philosophischen Tradition ausgearbeitete Logik im wesentlichen zeitlos ist. Inzwischen gibt es jedoch eine breite philosophische Diskussion über die daraus für die Beschreibung zeitlicher Prozesse resultierenden Probleme; vgl. als Einführung in diese Diskussion den Sammelband von Kienzle [1994].

[12] Vgl. auch Corgeau und Lelièvre [1988].

von der Vorstellung paralleler Prozesse ausgeht und das in (4.4) formulierte Beschreibungsprinzip zugrunde legt. Es gibt dann für jeden Teilprozeß genau eine Möglichkeit, die Frage nach möglichen Bedingungen zu stellen, nämlich als Vermutung, daß ein Ereignis, das in diesem Teilprozeß zu einem gewissen Zeitpunkt eintritt, von der *gemeinsamen* Vorgeschichte aller Prozesse bis zu diesem Zeitpunkt abhängig sein kann. Diese Vermutung liefert gewissermaßen den theoretischen Rahmen, um aus empirisch verfügbaren Lebensverlaufsdaten Einsichten in mögliche Abhängigkeiten zu gewinnen.

Zweitens zeigt sich, daß die übliche Formulierung, daß individuelle Lebensverläufe von „äußeren Bedingungen“ abhängig sind, in gewisser Weise irreführend ist. An die Stelle dessen tritt die Vorstellung, daß die Entwicklung eines Prozesses – bzw. einer Menge paralleler Prozesse – zum Zeitpunkt t *nur* von seiner Vorgeschichte, die bis zu diesem Zeitpunkt abgelaufen ist, abhängig ist.

Zumindest ein Aspekt der üblichen Vorstellung von der sozialen Bedingtheit von Lebensverläufen kann durch folgendes Bild veranschaulicht werden:



Bei näherem Hinsehen erweist sich die in diesem Bild zum Ausdruck kommende Betrachtungsweise jedoch als problematisch. Das Problem hat zwei Aspekte. Erstens stellt sich die Frage, ob soziale Bedingungen sinnvoll als exogene Bedingungen von Lebensverläufen verstanden werden können. Denn wenn man sagt, daß A durch B beeinflußt wird, meint man üblicherweise, daß A davon abhängt, wie B beschaffen ist, aber nicht umgekehrt. Es ist jedoch unklar, ob bzw. in welcher Weise dieses übliche Verständnis der Bedeutung von *beeinflussen* auf das Verhältnis von Lebensverläufen und ihren sozialen Bedingungen angewendet werden kann. Zweitens stellt sich die Frage, wie überhaupt individuelle Lebensverläufe von ihren sozialen Bedingungen abgegrenzt werden können.

a) Man könnte versuchen, soziale Bedingungen dadurch zu charakterisieren, daß sie sich einer Kontrolle durch die Individuen entziehen. Diese Charakterisierung, obwohl sie zutreffend erscheint,[13] liefert jedoch keine befriedigende Lösung des Abgrenzungsproblems. Denn einerseits können die Individuen, zumindest in gewissen Grenzen, wählen, unter welchen

[13] Formulierungen, in denen das Wort *Individuum* im Plural verwendet wird, sind meistens, auch in diesem Fall, zweideutig. Gemeint ist hier, daß jedes einzelne Individuum nicht in der Lage ist, gesellschaftliche Verhältnisse zu verändern. Dies schließt nicht aus, daß „die Individuen“ in der Form kollektiven Handelns auf die Verfassung ihrer gesellschaftlichen Verhältnisse Einfluß nehmen können.

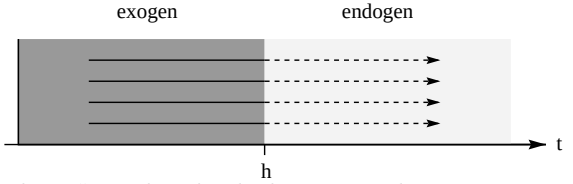
sozialen Bedingungen sie leben möchten. Zum Beispiel durch räumliche Mobilität, oder durch die Auswahl von Arbeitsplätzen, Lebenspartnern und Freunden. Andererseits trifft die Bemerkung, daß soziale Bedingungen nicht individuell *kontrolliert* werden können, in gewisser Weise auch auf die individuellen Lebensverläufe zu. Zu jedem Zeitpunkt in der Entwicklung eines Lebensverlaufs können gewisse Entscheidungen getroffen werden, aber der resultierende Lebensverlauf, das realisierte Ergebnis der vielen einzelnen Entscheidungen, entzieht sich dieser Betrachtungsweise. Eine bereits in der Einleitung zitierte Bemerkung Dantos [1965, S. 465] charakterisiert den Sachverhalt sehr treffend so: „Da wir nicht wissen, wie unsere Handlungen aus dem Gesichtswinkel der Historie gesehen werden, ermangeln wir dementsprechend auch der Kontrolle über die Gegenwart."

b) Ein zweiter problematischer Aspekt der Abgrenzung von A (sozialen Bedingungen) und B (Lebensverläufen) wird sichtbar, wenn man sich noch einmal überlegt, in welcher Weise von einer Beeinflussung gesprochen werden kann. Das übliche Verständnis der Formulierung *A beeinflußt B* impliziert, daß B keinen Einfluß darauf nehmen kann, wie es durch A beeinflußt wird. Für das Verhältnis zwischen sozialen Bedingungen und Lebensverläufen gilt dies jedoch nicht. Die Individuen können nicht nur, in gewissen Grenzen, ihre sozialen Lebensbedingungen wählen, es hängt auch mehr oder weniger von ihnen selbst ab, in welcher Weise sie durch ihre sozialen Bedingungen beeinflußt werden. Dies ist unmittelbar sichtbar, wenn man soziale Verhältnisse durch „geltende" Regeln charakterisiert. Die Individuen können diesen Regeln folgen, aber sie sind nicht dazu gezwungen. Genauso verhält es sich, wenn man soziale Bedingungen dadurch charakterisiert, daß sie die Handlungsmöglichkeiten der Individuen strukturieren. Das einfache Schema, daß die Gesellschaft die Handlungsmöglichkeiten vorgibt, innerhalb derer die Individuen ihre Wahl treffen können, kann zwar in vielen Fällen zu einer in erster Näherung sinnvollen Beschreibung führen. Dieses Schema erfaßt jedoch nicht, daß bereits die Wahrnehmung von Handlungsmöglichkeiten von den Individuen abhängt, von ihrer Phantasie, ihrer Risikobereitschaft und ihrer moralischen Verfassung.

c) Schließlich kann man sich die Frage stellen, wie überhaupt Individuen von ihrer sozialen Umgebung unterschieden werden können. Die Frage ist natürlich nicht phänomenologisch gemeint, sondern: Was kann dem Individuum als Individuum zugerechnet werden, was sollte eher als eine Eigenschaft seiner sozialen Umgebung betrachtet werden? In dieser Form stellt sich die Frage bereits bei scheinbar elementaren Eigenschaften wie „Geschlecht" und „Bildung", die üblicherweise den Individuen zugerechnet werden. Denn man sieht sofort, daß diese Zurechnung problematisch ist, daß diese Eigenschaften nur definiert werden können, wenn darauf Bezug genommen wird, welche Bedeutung ihnen in der jeweiligen sozialen Umgebung gegeben wird.[14]

[14] Eine interessante Analyse dieses Sachverhalts stammt von Mandelbaum [1955]. Ich stimme jedoch weitgehend Danto's [1965, insb. S. 445ff] kritischer Diskussion des Standpunkts von Mandelbaum zu, insbesondere seiner Auffassung, daß die Tatsache, daß Individuen zugeschriebene Eigenschaften eine „sozial definierte" Bedeutung haben, am angemessensten durch Bezugnahme auf soziale Regeln erklärt werden kann und nicht im Widerspruch zu einer an individuellen Akteuren orientierten Betrachtungsweise gesellschaftlicher Verhältnisse steht.

Diese Probleme stellen sich, wenn man versucht, *individuelle* Lebensverläufe zu beschreiben und Einsichten in ihre „Bedingtheit" zu gewinnen. Geht man jedoch davon aus, daß die soziologische Lebensverlaufsforschung nicht unmittelbar an Individuen interessiert ist, *sondern* an einer Beschreibung gesellschaftlicher Verhältnisse, erscheint es möglich, von diesen Problemen zu abstrahieren. Das formale Hilfsmittel dafür liefert die in (4.1) definierte und in Abbildung 4.1.1 veranschaulichte Vorstellung paralleler Prozesse. Bei dieser Definition wird zwar auf Individuen Bezug genommen, aber die Frage, ob die innerhalb eines Teilprozesses definierten Zustände als Eigenschaften der Individuen oder als Eigenschaften ihrer sozialen Umgebung anzusehen sind, kann unbeantwortet bleiben; sie ist für die statistische Beschreibung des parallelen Ablaufs der Prozesse nicht von Bedeutung. Gegenstand der Beschreibung ist der gemeinsame Prozeß. Die Frage, wie sich die Teilprozesse wechselseitig beeinflussen, stellt sich nicht in der Form, wie ein Prozeß einen anderen Prozeß beeinflußt, vielmehr in der Form der Frage, wie die Vorgeschichte des gemeinsamen Prozesses die Gegenwart jedes Teilprozesses – und dadurch die Zukunft des gemeinsamen Prozesses – beeinflußt. An die Stelle des oben angeführten Bildes, in dem individuelle Lebensverläufe durch (exogene) soziale Bedingungen beeinflußt werden, tritt also ein Bild der folgenden Art:



„Exogen" und „endogen" werden durch diese Betrachtungsweise zu relativen Begriffen. Es erscheint sinnlos, sich darunter Eigenschaften von Sachverhalten vorzustellen, die Begriffe verweisen vielmehr auf die relative Stellung von Sachverhalten im zeitlichen Ablauf eines Prozesses. Nur Ereignisse, die in der Vergangenheit stattgefunden haben, können als mögliche exogene Bedingungen des Prozeßverlaufs in der Gegenwart in Betracht gezogen werden.

4.1.5 Zeitabhängige Kovariablen

Folgt man dieser Betrachtungsweise paralleler Prozesse, erhält man auch eine neue Sichtweise des Problems zeitabhängiger Kovariablen in Übergangsratenmodellen. Üblicherweise wird von der Fragestellung ausgegan-

gen, wie ein Prozeß Z_t einen Prozeß Y_t beeinflußt. Diese Fragestellung erzeugt jedoch unmittelbar das Problem, daß auch umgekehrt der Prozeß Y_t den Prozeß Z_t beeinflussen kann und daß dann unklar wird, wie diese Wechselwirkungen bei der Modellbildung angemessen berücksichtigt werden kann. Üblicherweise wird deshalb betont, daß die Berücksichtigung von Z_t in der Form zeitabhängiger Kovariablen in einem Übergangsratenmodell für den Y_t-Prozeß vorausetzt, daß der Z_t-Prozeß eine „exogene" Bedingung für den Y_t-Prozeß bildet.

Wie ich zu zeigen versucht habe, resultiert diese Problemformulierung daraus, daß die Fragestellung, wie ein Prozeß Z_t einen anderen Prozeß Y_t beeinflußt, zum Verständnis der parallel ablaufenden Prozesse ungeeignet ist; daß eine angemessenere Fragestellung darin liegt, wie die gemeinsame Vorgeschichte der Prozesse Y_t und Z_t die jeweils gegenwärtige Entwicklung in den beiden Teilprozessen – und dadurch die Zukunft des gemeinsamen Prozesses – beeinflußt. Bevor ich im nächsten Abschnitt versuche, die Implikationen dieser Überlegung für die statistische Modellbildung auszuführen, soll zunächst dargestellt werden, wie es zu der üblichen Betrachtungsweise kommt, daß zeitabhängige Kovariablen einer gewissen Exogenitätsbedingung genügen müssen. Für die formale Darstellung stütze ich mich hauptsächlich auf Lancaster [1990, S. 24ff], allerdings gehe ich im folgenden, wie bisher, von einem deskriptiven Wahrscheinlichkeitsraum aus.

Die beiden Prozesse, Y_t und Z_t, können dann durch eine Folge von Zufallsvariablen

$$(Y_t, Z_t) : \Omega \longrightarrow \mathcal{Y} \times \mathcal{Z} \qquad t \in \mathcal{T} \tag{4.5}$$

repräsentiert werden. Ω ist die Längsschnittsgesamtheit der Individuen, auf deren Lebensverläufe Bezug genommen wird. Y_t und Z_t sind jeweils Folgen von Zufallsvariablen, durch die gewisse Aspekte der Lebensverläufe der Individuen aus Ω bzw. ihrer sozialen Umgebung erfaßt werden. Der Zeitindex t bezieht sich auf eine diskrete Prozeßzeitachse $\mathcal{T}$.

Die übliche Fragestellung ist, wie mithilfe eines Modells für den Prozeß Y_t erfaßt werden kann, daß seine Entwicklung durch den Ablauf des Z_t-Prozesses beeinflußt wird. Der Asymmetrie der Fragestellung folgend, können zunächst folgende neue Zufallsvariablen gebildet werden:

$$(T, D, Z_0, Z_1, Z_2, \ldots) : \Omega \longrightarrow \mathcal{T} \times \mathcal{D} \times \mathcal{Z} \times \mathcal{Z} \times \mathcal{Z} \ldots \tag{4.6}$$

T ist die Verweildauer im Ausgangszustand einer Episode des Y_t-Prozesses. $\mathcal{D}$ ist die Menge der möglichen Endzustände dieser Episode, wobei angenommen wird, daß jede Episode nach einer endlichen Verweildauer in einem der Zustände aus $\mathcal{D}$ enden muß. Die Variablen $Z_0, Z_1, Z_2, \ldots$ beschreiben den parallel zum Episodenverlauf sich entwickelnden Z_t-Prozeß. Um auf einfache Weise auf diesen Prozeß verweisen zu können, wird die

Bezeichung

$$\bar{Z}_t = (Z_0, Z_1, Z_2, \ldots, Z_t)$$

verwendet. Mögliche Realisierungen für solche Folgen von Zufallsvariablen werden mit $\bar{z}_t$ bezeichnet. Der empirische Ausgangspunkt besteht dann in den in der Grundgesamtheit gegebenen Wahrscheinlichkeiten

$$\mathrm{P}\,(T = t, D = d, \bar{Z}_t = \bar{z}_t) \tag{4.7}$$

Es sind dies die Wahrscheinlichkeiten dafür, daß in der Gesamtheit der betrachteten Episoden zum Zeitpunkt t ein Übergang in den Zustand d stattfindet *und* daß der Z_t-Prozeß bis zum Zeitpunkt t den Verlauf $\bar{z}_t$ aufweist.

Um eine statistische Formulierung für die Idee zu gewinnen, daß der Episodenverlauf vom Verlauf des Z_t-Prozesses abhängt, müssen bedingte Wahrscheinlichkeiten gebildet werden. Zu überlegen ist, wie dies auf sinnvolle Weise erreicht werden kann. Zunächst können die empirischen Wahrscheinlichkeiten (4.7) durch zustandsspezifische Übergangsraten und eine Survivorfunktion dargestellt werden:

$$\mathrm{P}\,(T = t, D = d, \bar{Z}_t = \bar{z}_t) = r_d(t \mid \bar{Z}_t = \bar{z}_t)\,\mathrm{P}\,(T \geq t, \bar{Z}_t = \bar{z}_t) \tag{4.8}$$

Dies liefert den ersten wesentlichen Schritt, um den Prozeß in seiner Dynamik beschreiben zu können. Auf der rechten Seite steht zunächst eine zustandsspezifische Übergangsrate:

$$r_d(t \mid \bar{Z}_t = \bar{z}_t) \;=\; \mathrm{P}\,(T = t, D = d \mid T \geq t, \bar{Z}_t = \bar{z}_t)$$

Es ist die Wahrscheinlichkeit, daß zum Zeitpunkt t ein Übergang in den Zustand d stattfindet, *unter der Bedingung*, daß sich der Z_t-Prozeß *bis zum Zeitpunkt* t entsprechend $\bar{z}_t$ entwickelt hat. Diese Übergangsrate liefert eine sinnvoll kausal interpretierbare Darstellung des Episodenverlaufs in seiner Abhängigkeit von Z_t. Dann folgt jedoch, auf der rechten Seite von (4.8), noch ein zweiter Ausdruck, in dem sich zwei Aspekte des Prozesses gewissermaßen vermischen: erstens eine Aussage über den Episodenverlauf, daß bis zum Zeitpunkt t noch kein Zustandswechsel eingetreten ist; zweitens eine Aussage über den Verlauf des Z_t-Prozesses. Es muß also überlegt werden, wie sich diese „Vermischung", d.h. die Möglichkeit einer wechselseitigen Beeinflussung, sinnvoll auflösen läßt.

Das übliche Verfahren, um diese Vermischung aufzulösen, besteht darin, nach einer geeigneten statistischen Formulierung für die Idee zu suchen, daß der Verlauf des Z_t-Prozesses eine *exogene* Bedingung für den Episodenverlauf ist, d.h. daß der Z_t-Prozeß zwar den Episodenverlauf beeinflussen kann, jedoch nicht umgekehrt. Die Ableitung einer solchen Exogenitätsbedingung kann folgendermaßen erreicht werden. Zunächst wird der zweite

Ausdruck auf der rechten Seite von (4.8) gesondert betrachtet. Wenn $t = 1$ ist, erhält man

$$\mathrm{P}\,(T \geq 1, \bar{Z}_1 = \bar{z}_1) \;=\; \mathrm{P}\,(\bar{Z}_1 = \bar{z}_1) \tag{4.9}$$

da der Anfangszustand frühestens zum Zeitpunkt $t = 1$ verlassen werden kann. In diesem Fall gibt es also keine Abhängigkeit. Wenn $t > 1$ ist, erhält man

$$\begin{aligned}&\mathrm{P}\,(T \geq t, \bar{Z}_t = \bar{z}_t) \;= \\ &\quad \mathrm{P}\,(Z_t = z_t \mid T \geq t, \bar{Z}_{t-1} = \bar{z}_{t-1})\,\mathrm{P}\,(T \geq t, \bar{Z}_{t-1} = \bar{z}_{t-1})\end{aligned} \tag{4.10}$$

Der zweite Ausdruck auf der rechten Seite kann durch sukzessives Konditionieren als ein Produkt bedingter Wahrscheinlichkeiten geschrieben werden:

$$\begin{aligned}&\mathrm{P}\,(T \geq t, \bar{Z}_{t-1} = \bar{z}_{t-1}) \;= \\ &\prod_{\tau=1}^{t-1} \mathrm{P}\,(T > \tau, Z_\tau = z_\tau \mid T > \tau - 1, \bar{Z}_{\tau-1} = \bar{z}_{\tau-1}) \;= \\ &\prod_{\tau=1}^{t-1} \mathrm{P}\,(T > \tau \mid T \geq \tau, \bar{Z}_\tau = \bar{z}_\tau)\,\mathrm{P}\,(Z_\tau = z_\tau \mid T > \tau - 1, \bar{Z}_{\tau-1} = \bar{z}_{\tau-1}) \;= \\ &\prod_{\tau=1}^{t-1} \big(1 - \mathrm{P}\,(T \leq \tau \mid T \geq \tau, \bar{Z}_\tau = \bar{z}_\tau)\big)\,\mathrm{P}\,(Z_\tau = z_\tau \mid T \geq \tau, \bar{Z}_{\tau-1} = \bar{z}_{\tau-1}) \;= \\ &\prod_{\tau=1}^{t-1} \big(1 - r(\tau \mid \bar{Z}_\tau = \bar{z}_\tau)\big)\,\mathrm{P}\,(Z_\tau = z_\tau \mid T \geq \tau, \bar{Z}_{\tau-1} = \bar{z}_{\tau-1})\end{aligned} \tag{4.11}$$

Dabei ist $r(\tau \mid \bar{Z}_\tau = \bar{z}_\tau)$ die unspezifische Abgangsrate aus dem Ausgangszustand. Faßt man (4.10) und (4.11) zusammen, erhält man

$$\begin{aligned}&\mathrm{P}\,(T \geq t, \bar{Z}_t = \bar{z}_t) \;= \\ &\quad \prod_{\tau=1}^{t-1} \big(1 - r(\tau \mid \bar{Z}_\tau = \bar{z}_\tau)\big) \prod_{\tau=1}^{t} \mathrm{P}\,(Z_\tau = z_\tau \mid T \geq \tau, \bar{Z}_{\tau-1} = \bar{z}_{\tau-1})\end{aligned}$$

Schließlich kann man diesen Ausdruck in (4.8) einsetzen und erhält

$$\begin{aligned}&\mathrm{P}\,(T = t, D = d, \bar{Z}_t = \bar{z}_t) = \\ &r_d(t \mid \bar{Z}_t = \bar{z}_t) \prod_{\tau=1}^{t-1} \big(1 - r(\tau \mid \bar{Z}_\tau = \bar{z}_\tau)\big) \prod_{\tau=1}^{t} \mathrm{P}\,(Z_\tau = z_\tau \mid T \geq \tau, \bar{Z}_{\tau-1} = \bar{z}_{\tau-1})\end{aligned} \tag{4.12}$$

Der in (4.9) beschriebene Fall, daß $t = 1$ ist, kann als ein Spezialfall dieser Gleichung angesehen werden.

Die in (4.12) angegebene Formulierung ist offenbar analog zu der in Abschnitt 3.6 gegebenen Formulierung für eine Situation, in der es nur eine zeitunabhängige Klassifizierungsvariable gibt. Der Unterschied besteht darin, daß in (4.12) eine Wahrscheinlichkeit für den Z_t-Prozeß hinzukommt, die von der Bedingung abhängt, daß der Ausgangszustand bis zum jeweiligen Zeitpunkt noch nicht verlassen worden ist.

Folgt man der Vorstellung, daß der Z_t-Prozeß eine exogene Bedingung des Y_t-Prozesses ist, erscheint es sinnvoll, folgende Exogenitätsbedingung zu formulieren:[15]

$$\mathrm{P}\,(Z_t = z_t \mid T \geq t, \bar{Z}_{t-1} = \bar{z}_{t-1}) = \mathrm{P}\,(Z_t = z_t \mid \bar{Z}_{t-1} = \bar{z}_{t-1}) \tag{4.13}$$

Sie bedeutet, daß der Verlauf des Z_t-Prozesses nur von seiner eigenen Vorgeschichte abhängt, nicht jedoch davon, wann und wie ein Zustandswechsel im Y_t-Prozeß erfolgt. Wenn diese Exogenitätsbedingung erfüllt ist, folgt

$$\prod_{\tau=1}^{t} \mathrm{P}\,(Z_\tau = z_\tau \mid T \geq \tau, \bar{Z}_{\tau-1} = \bar{z}_{\tau-1}) = \mathrm{P}\,(\bar{Z}_t = \bar{z}_t)$$

und (4.12) kann folgendermaßen geschrieben werden:

$$\begin{aligned}&\mathrm{P}\,(T = t, D = d, \bar{Z}_t = \bar{z}_t) = \\ &\quad r_d(t \mid \bar{Z}_t = \bar{z}_t) \prod_{\tau=1}^{t-1} \big(1 - r(\tau \mid \bar{Z}_\tau = \bar{z}_\tau)\big)\,\mathrm{P}\,(\bar{Z}_t = \bar{z}_t)\end{aligned} \tag{4.14}$$

Diese Gleichung kann als Ausgangspunkt für die Konstruktion eines Modells verwendet werden, um zu erfassen, wie der Verlauf des Y_t-Prozesses durch den Z_t-Prozeß bedingt wird. Im wesentlichen kann genau so vorgegangen werden, wie es in Abschnitt 3.6 im Hinblick auf einfache Klassifizierungsvariablen beschrieben worden ist. Der einzige Unterschied besteht darin, daß sich die Werte der Z_t-Variablen während des (im Rahmen des Y_t-Prozesses definierten) Episodenverlaufs verändern können und daß dies in der Spezifikation von Modellen für die zustandsspezifischen Übergangsraten berücksichtigt werden muß.

Allerdings scheint es so, daß Gleichung (4.14) nur verwendet werden kann, wenn die in (4.13) formulierte Exogenitätsbedingung vorausgesetzt werden kann, und diese Annahme erscheint in den meisten Fällen nicht plausibel. Tatsächlich ist diese Annahme jedoch nicht erforderlich, um (4.14) als einen sinnvollen Ausgangspunkt für die Modellbildung zu begründen. Dieser Eindruck entsteht nur, weil hier von der Fragestellung ausgegangen wurde, wie „der“ Z_t-Prozeß „den“ Y_t-Prozeß bedingt. Geht man stattdessen davon aus, daß sich beide Prozesse parallel entwickeln und wechselseitig bedingen können, erhält man nicht nur eine andere Sichtweise des Exogenitätsproblems, sondern auch einen alternativen Zugang zur Modellbildung. Das wird in Abschnitt 4.2 näher ausgeführt.

[15] Vgl. Lancaster [1990, S. 28].

4.1.6 Ereignisse, Zustände und Eigenschaften von Individuen

Im Mittelpunkt unseres Versuchs, Lebensverläufe als „bedingt“ zu verstehen, steht die Vorstellung einer zeitlich geordneten Folge von Ereignissen, die sich wechselseitig bedingen. Dabei wird vorausgesetzt, daß nur bereits vergangene Ereignisse die jeweils in der Gegenwart stattfindenden Ereignisse bedingen können. Ich nenne dies eine *ereignisanalytische* Betrachtungsweise. Es sollte überlegt werden, ob sie einen hinreichend allgemeinen Rahmen zur Verfügung stellt, in dem auch die üblichen Vorstellungen über die Bedingtheit von Lebensverläufen formuliert werden können.

a) Eine auch in dieser Arbeit häufig verwendete Formulierung spricht von „sozialen Bedingungen“ individueller Lebensverläufe. In einer ereignisanalytischen Betrachtungsweise kann dies nicht so verstanden werden, daß gesellschaftliche Verhältnisse gewissermaßen exogene Faktoren sind, die – so wie das Wetter – individuelle Lebensverläufe beeinflussen. Um den Gegensatz etwas überspitzt zu formulieren, könnte man vielmehr sagen, daß mit dem Wort „gesellschaftliche Verhältnisse“ auf die Form verwiesen wird, in der sich Gesamtheiten von individuellen Lebensverläufen entwickeln. Obwohl es vielleicht möglich ist, Eigenschaften von Individuen und Eigenschaften ihrer gesellschaftlichen Verhältnisse zu unterscheiden, ist diese Unterscheidung für eine ereignisanalytische Betrachtung nicht wesentlich. Die Bezugnahme auf gesellschaftliche Verhältnisse kommt vielmehr in zwei anderen Aspekten der Modellbildung zum Ausdruck. Erstens in der Konstruktion eines Biographieschemas, denn es ist evident, daß dadurch jeweils bestimmte gesellschaftliche Verhältnisse vorausgesetzt werden. Der zweite Aspekt betrifft die Frage, wie der Zusammenhang der Ereignisse, die in dem jeweils vorausgesetzten Biographieschema möglich sind, beschaffen ist. Dieser Ereigniszusammenhang kann statistisch beschrieben und durch eine soziologische Interpretation als Ausdruck sozialer Regeln verstanden werden.

b) Es erscheint sinnvoll, Ereignisse und Zustände zu unterscheiden. Ein Zustand kann als eine Situation definiert werden, die „für eine gewisse Zeit“ besteht. Im üblichen Sprachgebrauch ist dies die grundlegende Form für Beschreibungen. Der Begriff „Ereignis“ ist demgegenüber abstrakt. Er dient dazu, um auf allgemeine Weise die Erfahrung zum Ausdruck bringen zu können, daß sich Zustände verändern. Eine abstrakte Definition kann insofern dadurch vorgenommen werden, daß man Ereignisse als Zustandswechsel bezeichnet. Zwar klärt diese Definition nicht, was Ereignisse eigentlich sind; sie macht jedoch deutlich, daß mit dem Begriff „Ereignis“ auf Veränderungen von Zuständen verwiesen wird und daß infolgedessen die Art und Weise, wie sinnvoll von Ereignissen gesprochen werden kann, davon abhängt, auf welche Art von Zuständen bezug genommen wird. Dies hat insbesondere Konsequenzen für die Frage, ob man sich Ereignisse so vorstellen kann, daß sie zu einem Zeitpunkt stattfinden. Die Vorstellung einer stetigen Zeitachse erscheint (als eine Idealisierung) sinnvoll, wenn sich Zustände kontinuierlich verändern können. Dann wird jedoch zugleich der Ereignisbegriff fragwürdig, denn es erscheint nicht angemessen, sich eine kontinuierliche Bewegung als eine Folge von Ereignissen vorzustellen.

Geht man von einer diskreten Zeitachse aus, kann in gewisser Weise von diesen Schwierigkeiten abstrahiert werden. Man kann dann sagen, daß ein Ereignis stattgefunden hat, wenn zum Zeitpunkt t ein anderer Zustand besteht als zum Zeitpunkt $t-1$. Das Wort „Ereignis“ kann dann diesen Zustandswechsel bezeichnen, und es erscheint nicht erforderlich, darüber nachzudenken, ob bzw. wie das Ereignis zwischen den beiden Zeitpunkten stattgefunden hat, ob von einem Ereignis oder von einem kontinulierlichen Zustandswechsel gesprochen werden sollte. In dieser Arbeit gehe ich – auch deshalb – von einer diskreten Zeitachse aus (was selbstverständlich nicht die Möglichkeit ausschließt, stetige Modelle zu bilden).

Akzeptiert man diese Überlegung, kann etwas genauer gesagt werden, inwiefern die Beschreibung eines Episodenverlaufs durch eine Übergangsrate eine Einsicht in Bedingungen dieses Ablaufs vermitteln kann. Gelegentlich wird dies so formuliert, daß der Ablauf sowohl vom Ausgangszustand als auch davon abhängt, wieviel Zeit seit dem Episodenbeginn verstrichen ist. In einer ereignisanalytischen Betrachtungsweise müßte zunächst genauer gesagt werden, daß der Episodenverlauf durch das Ereignis bedingt wird, dessen Stattfinden die Episode eingeleitet hat. Dann bleibt allerdings noch die Frage, was es heißen soll, daß der Episoden*verlauf* durch dieses Ereignis bedingt wird. Denn es ist üblich, von der Vorstellung auszugehen, daß sich nur Ereignisse bedingen können; der Episodenverlauf ist jedoch kein Ereignis.

Der Episodenverlauf ist zwar kein Ereignis, er besteht jedoch in gewisser Weise aus einer Folge von Ereignissen, nämlich aus der Folge derjenigen Ereignisse, durch die die Individuen den Anfangszustand der Episode wieder verlassen. Insofern kann die abstrakte Vorstellung, daß sich Ereignisse bedingen, auch in diesem Fall verwendet werden: das Ereignis, mit dem eine Episode beginnt, wird als Bedingung des Ereignisses betrachtet, mit dem die Episode endet. Übergangsraten können in dieser Betrachtungsweise als Hilfsmittel für eine statistische Beschreibung eines Bedingungszusammenhangs zwischen Ereignissen interpretiert werden.

Die Formulierung, daß die Übergangsrate von der Verweildauer im Ausgangszustand einer Episode „abhängt“, wird bei dieser Interpretation allerdings fragwürdig. Die Verweildauer ist kein Ereignis, insofern kann sie den Episodenverlauf nicht „bedingen“. Es erscheint angemessener, sich die Verweildauerabhängigkeit als eine Eigenschaft der Übergangsrate vorzustellen, d.h. sie als eine mathematische Funktion der Zeit zu betrachten. Die Verweildauerabhängigkeit liefert keine Erklärung, sondern charakterisiert die Übergangsrate, die ihrerseit beschreibt, wie das Anfangsereignis einer Episode diejenigen Ereignisse bedingt, durch die die Episode beendet werden kann. Im Hinblick auf diesen Bedingungszusammenhang ist die

Zeitabhängigkeit der Übergangsrate eine rein deskriptive Eigenschaft.

In gewisser Weise bedeutet dies, daß man den Ablauf einer Episode nur beschreiben kann, d.h. man kann beschreiben, wie das Ereignis, mit dem die Episode beginnt, das Ereignis, durch das sie endet, bedingt. Die Absicht, auch noch diesen Zusammenhang zu erklären, erscheint nicht sinnvoll. Diese Aussage muß jedoch in zweierlei Hinsicht präzisiert werden. Erstens wird keineswegs ausgeschlossen, nach jeweils vermittelnden Ereignissen zu suchen. Gelingt dies, kann man insoweit von einer (partiellen) Erklärung eines Episodenverlaufs sprechen. Man erklärt einen Episodenverlauf, der mit einem Ereignis E_1 beginnt und mit einem Ereignis E_2 endet, dadurch, daß man zeigt, wie sein Ablauf durch andere Ereignisse, die zwischen E_1 und E_2 stattfinden, bedingt wird. Die Erklärung besteht darin, daß man den Episodenverlauf als eine Folge von sich bedingenden Ereignissen *beschreibt*.

Zweitens erscheint wichtig, um was für Ereignisse es sich handelt. Bei der Beschreibung von Lebensverläufen handelt es sich nicht um spontan stattfindende Elementarereignisse, sondern um Ereignisse, die durch menschliche Akteure hervorgebracht werden. Der Bedingungszusammenhang von Ereignissen kann deshalb – zwar nicht immer, aber in vielen Fällen – durch eine Bezugnahme auf Verhaltensweisen von Individuen interpretiert werden. Dies unterscheidet Lebensverläufe von Ereigniszusammenhängen, die jenseits der Reichweite menschlicher Akteure ablaufen. Dadurch entsteht zugleich eine spezifische Aufgabe für die soziologische Theoriebildung. Sie braucht sich nicht darauf zu beschränken, Ereigniszusammenhänge zu beschreiben, sondern kann darüberhinaus versuchen, den Zusammenhang von Ereignissen auch dadurch zu erklären, daß die Menschen, die diese Ereignisse hervorbringen, dabei sozialen Regeln folgen.

c) Schließlich bleibt noch zu überlegen, wie die üblichen Formulierungen verstanden werden können, in denen auf Eigenschaften von Individuen verwiesen wird, um Lebensverläufe zu erklären. Zum Beispiel wird oft gesagt, daß das Heiratsverhalten vom Geschlecht, von der Zugehörigkeit zu einer Geburtskohorte, vom Bildungsgrad usw. „abhängt“. Was könnte damit gemeint sein?

Zunächst eine rein deskriptive Feststellung, daß sich in einer Gesamtheit von Individuen Teilgesamtheiten unterscheiden lassen, bei denen ein Prozeß unterschiedlich abläuft. Wenn nur dies gemeint ist, erscheint die Formulierung, daß der Prozeßverlauf davon „abhängt“, in welcher Teilgesamtheit er stattfindet, unproblematisch. Denn es ist dann klar, daß eigentlich kein Bedingungsverhältnis – so wie wir es uns zwischen Ereignissen vorstellen können – gemeint ist. Unproblematisch erscheint dies insbesondere aus einer soziologischen Perspektive, denn man könnte dann einfach darauf hinweisen, daß die Individuen in den verschiedenen Teilgesamtheiten unterschiedlichen sozialen Regeln folgen. Frauen und Männer verhalten sich unterschiedlich, d.h. sie folgen unterschiedlichen sozialen Regeln, usw.

Dies erscheint unproblematisch, bleibt aber offensichtlich unbefriedigend. Ich glaube jedoch, daß man mit einer ereignisanalytisch konzipierten Lebensverlaufsforschung einen Schritt weiter gehen kann. Der entscheidende Schritt kann darin gesehen werden, daß durch diese Betrachtungsweise das traditionelle statische Verständnis von Individuen – und infolgedessen die Methode, sie statisch zu klassifizieren – infrage gestellt werden kann. Aus der Perspektive einer ereignisanalytischen Lebensverlaufsforschung kann der Begriff „Individuum“ von der Vorstellung eines Lebensverlaufs nicht getrennt werden. Das Verhältnis der beiden Begriffe ist kompliziert. Aber zumindest zwei Aspekte bzw. Verwendungsweisen des Begriffs „Individuum“ können in diesem Zusammenhang unterschieden werden. Einerseits kann man das Wort verwenden, um auf das Subjekt eines Lebensverlaufs zu verweisen und dadurch eine zeitliche Identität von Lebensverläufen zu begründen. Zweitens kann man sich ein Individuum als ein transitorisches Ergebnis der Entwicklung seines Lebensverlaufs vorstellen; Individuen sind dann jeweils das, was sie durch ihre bisherigen Lebensverläufe geworden sind.

Wichtig erscheint, daß sich infolgedessen alle üblicherweise den Individuen zugeschriebenen Eigenschaften als Eigenschaften ihrer Lebensverläufe ansehen lassen.[16] Daran lassen sich zwei für die Theoriebildung wichtige Überlegungen anschließen. Erstens kann man die Bezugnahme auf Eigenschaften von Individuen als eine Bezugnahme auf Aspekte der jeweiligen Vorgeschichte ihres Lebensverlaufs verstehen. Der Versuch, Lebensverläufe durch Eigenschaften ihrer Subjekte zu erklären, wird dann mit einer ereignisanalytischen Betrachtungsweise vereinbar, d.h. mit dem Basisprinzip, den Ablauf eines kontingenten Prozesses nur durch Rückgriff auf seine jeweilige Vorgeschichte zu erklären.

Zweitens können viele der den Individuen üblicherweise statisch zugerechneten Eigenschaften durch eine Bezugnahme auf Ereignisse, die sich im jeweils vergangenen Lebensverlauf ereignet haben, präziser beschrieben und als transitorische Eigenschaften von Lebensverläufen erfaßt werden. Zum Beispiel: statt eine Person durch ein gewisses Bildungsniveau zu charakterisieren, kann explizit auf die Ereignisse Bezug genommen werden, durch die im bisherigen Lebensverlauf dieses Bildungsniveau zustande gekommen ist; oder: statt zu sagen, daß eine Person verheiratet ist, kann darauf Bezug genommen werden, in welchem Alter sie geheiratet hat und wie lange sie bereits verheiratet ist. Der theoretische Gewinn liegt darin, daß eine bloß statische Klassifikation von Individuen vermieden und daß die Bezugnahme auf Eigenschaften von Individuen insoweit in das allgemeine Schema einer ereignisanalytischen Erklärung von Lebensverläufen eingeordnet werden kann. Daß das Heiratsverhalten vom Bildungsniveau

[16] Wie in Abschnitt 4.1.4 bemerkt worden ist, sollte auch nicht vergessen werden, daß es sich um sozial definierte Eigenschaften handelt. Aber von diesem Aspekt der Zurechnung von Eigenschaften zu Individuen kann im vorliegenden Diskussionszusammenhang abstrahiert werden.

abhängt, läßt sich dann zum Beispiel dadurch erklären, daß man eine Folge sich bedingender Ereignisse beschreibt.[17]

4.1.7 Unbeobachtete Heterogenität

Die statistische Beschreibung von Lebensverläufen setzt ein Biographieschema voraus, das die jeweils möglichen Lebensverläufe fixiert. Ist ein solches Biographieschema gegeben, können Episoden definiert und kann ihr Ablauf durch Übergangsraten beschrieben werden. Jede Beschreibung dieser Art liefert zugleich eine gewisse Erklärung: sie zeigt, wie es von einem Ausgangszustand zu einem Endzustand kommt.

Gibt es ein Kriterium für die Güte solcher Erklärungen? Eine naheliegende Antwort könnte folgendermaßen formuliert werden: je mehr vermittelnde Ereignisse innerhalb eines Biographieschemas berücksichtigt werden können, desto bessere Erklärungen, d.h. Einsichten in den Bedingungszusammenhang von Ereignissen, sind möglich. Diese Antwort ist jedoch fragwürdig, weil sie offen läßt, wie mit den Grenzen eines solchen Erklärungsziels umgegangen werden sollte. Diese Grenzen ergeben sich daraus, daß individuelle Lebensverläufe kontingent sind, so daß die Vorstellung, individuelle Lebensverläufe als Sequenzen von sich kausal bedingenden Ereignissen erklären zu können, sinnlos wird. Statistische Modelle zur Beschreibung von Lebensverläufen können so verstanden werden, daß mit ihrer Hilfe von dieser Kontingenz abstrahiert werden kann. Dies kann jedoch nur durch einen Perspektivenwechsel erreicht werden, indem sich die statistische Beschreibung nicht auf individuelle Lebensverläufe, sondern auf Gesamtheiten von Lebensverläufen richtet. Daraus folgt jedoch, daß es (bei dieser Interpretation) nicht möglich ist, mithilfe statistischer Modelle eine vollständige Erklärung des Ablaufs individueller Lebensverläufe zu erreichen.

Daraus ergeben sich, wie ich glaube, zwei Konsequenzen für die Art der Erklärungen, die in der soziologischen Lebensverlaufsforschung angestrebt werden können. Erstens kann die Güte solcher Erklärungen nicht darin gesehen werden, wie weit man sich mit einer ereignisanalytischen Beschreibung von Lebensverläufen dem Gedanken an ihre kausale Erklärbarkeit nähern kann. Die Frage ist nicht, wie mithilfe dieses Verfahrens „immer mehr“ vermittelnde Ereignisse erfaßt werden können. In der Regel ist es schon schwer, auch nur einen kleinen Teil derjenigen Ereignisse zu berücksichtigen, die in der alltäglichen Reflexion von Lebensverläufen immer schon zu ihrer Erklärung herangezogen werden. Es erscheint aussichtlos, mit statistischen Modellen gegen die Differenziertheit dieser Erklärungen zu konkurrieren. Es gibt jedoch einen anderen – soziologischen – Gesichtspunkt, mit dem die Verwendung statistischer Modelle zur Beschreibung

[17] Als Beispiele für die Anwendungen dieser Überlegungen in der Lebensverlaufsforschung vgl. exemplarisch Blossfeld und Huinink [1989], Blossfeld und Jaenichen [1990].

von Lebensverläufen gerechtfertigt werden kann. Denn die alltägliche Reflexion von Lebensverläufen ist zwar differenziert, sie basiert jedoch auf der Annahme sozialer Regeln, deren Geltung ungewiß ist. Es ist üblich, Lebensverläufe durch soziale Regeln zu erklären, aber die im Alltag verfügbare Empirie ist typischerweise sehr beschränkt und von subjektiven Erfahrungen abhängig. Hierauf bezogen kann einer systematischen Sammlung und statistischen Analyse von Lebensverlaufsdaten dadurch eine Bedeutung gegeben werden, daß durch sie eine objektivere Einsicht in soziale Regeln möglich wird. Wichtig erscheint mir, daß dadurch dem hier betonten Perspektivenwechsel ein soziologischer Sinn gegeben werden kann. Statt sich an der aussichtslosen Vorstellung zu orientieren, daß es gelingen könnte, individuelle Lebensverläufe zu erklären, kann man sich die Aufgabe vornehmen, objektivierbare Einsichten in soziale Regeln bzw. Regemäßigkeiten zu gewinnen.[18]

Eine zweite Konsequenz betrifft die Frage, wie mit der Tatsache umgegangen werden sollte, daß die Art der Erklärungen, die mithilfe statistischer Modelle erreicht werden kann, vor allem davon abhängt, was für ein Biographieschema vorausgesetzt wird. Es ist seit einiger Zeit Mode geworden, diese Frage unter dem Stichwort „unbeobachtete Heterogenität“ zu diskutieren. Als Leitfaden, um diese Diskussion zu verstehen, kann die Frage dienen, ob bzw. in welcher Weise Einsichten in zeitabhängige Übergangsraten gewonnen werden können.

Man kann versuchen, dies dadurch herauszufinden, daß man einen Modellansatz verwendet, der die Möglichkeit zeitabhängiger Übergangsraten zuläßt. Als Ergebnis der Schätzung eines solchen Modells findet man dann, ob bzw. wie die Übergangsrate von der Verweildauer im Ausgangszustand einer Episode abhängt. Dagegen kann jedoch eingewendet werden, daß die Verlaufsform der Übergangsrate insbesondere davon abhängt, wie das Biographieschema beschaffen ist bzw. welche Kovariablen in der Modellschätzung berücksichtigt werden. Wie bereits in Abschnitt 2.4.3 besprochen worden ist, kann man sich vorstellen, daß die Grundgesamtheit aus mehreren Teilgesamtheiten besteht, in denen es einen unterschiedlichen

[18] Dies beinhaltet nicht nur eine empirische Überprüfung der Frage, ob bzw. in welcher Weise die in der alltäglichen Reflexion von Lebensverläufen und gesellschaftlichen Verhältnissen unterstellten Regeln tatsächlich gelten, sondern verlangt häufig auch eine Reformulierung und Präzisierung der umgangssprachlichen Formulierung solcher Regeln. Dies betrifft zum Beispiel Regeln, in denen auf eine Unterscheidung von Jugend- und Erwachsenenphase in den individuellen Lebensverläufen Bezug genommen wird. Unterscheidungen dieser Art sind offenbar nicht selbstverständlich und bedürfen aus soziologischer Sicht einer Präzisierung, bevor empirische Einsichten gewonnen werden können. Vgl. exemplarisch im Hinblick auf die Unterscheidung von Jugend- und Erwachsenenphase die Arbeit von Blossfeld und Nuthmann [1990]. Ein weitere Problemstellung ergibt sich daraus, daß zwar einerseits zahlreiche Aspekte von Lebensverläufen durch Rechtsnormen geregelt werden, vgl. zum Beispiel die Betonung von an das Lebensalter geknüpfter Rechtsnormen bei Kohli [1985], daß jedoch andererseits daraus nicht ohne weiteres tatsächlich homogene Lebensverläufe bzw. Übergangsphasen in Lebensverläufen resultieren.

Verlauf der Übergangsrate – eine unterschiedliche Form der Vermittlung von Ereigniszusammenhängen – gibt. Wenn dies bei der Modellbildung nicht berücksichtigt wird, liefert die Modellschätzung schließlich nur eine Einsicht in eine Mischverteilung, die aus einer zeitlichen Mischung unterschiedlicher Übergangsraten entsteht. Die Frage ist, welche Bedeutung dieser Überlegung zukommt.

a) In gewisser Weise ist der Einwand trivial, denn er zeigt nur, daß man Sachverhalte unterschiedlich beschreiben kann. Bei jeder statistischen Beschreibung von Lebensverläufen muß ein Biographieschema vorausgesetzt werden; und infolgedessen sind alle daraufhin gewonnenen Beschreibungen und Modelle von dieser Voraussetzung abhängig. Die Fixierung eines Biographieschemas liegt im Ermessen des jeweiligen Modellkonstrukteurs und reflektiert die von ihm für die Beschreibung von Lebensverläufen für wichtig erachteten Aspekte (und natürlich auch die jeweils verfügbaren Daten). Wird das Biographieschema verändert, verändert sich auch die resultierende Beschreibung. Dies gilt vor allem für die durch das Biographieschema vorgenommene Definition von Episoden, insbesondere für die jeweils in Betracht gezogenen Zielzustände. Es ist klar, daß man bei der Beschreibung eines Episodenverlaufs unterschiedliche Ergebnisse erzielt, wenn nur ein möglicher Endzustand betrachtet oder wenn eine Mehrzahl konkurrierender Risiken unterschieden wird. Ebenso sind die Ergebnisse auch davon abhängig, welche Merkmale der individuellen Lebensverläufe in der Modellbildung berücksichtigt werden.[19]

Diese Abhängigkeit statistischer Beschreibungen vom jeweils vorausgesetzten Beschreibungsrahmen ist grundsätzlich nicht zu vermeiden. Sie schließt selbstverständlich eine konstruktive Kritik nicht aus. Liefert jemand eine statistische Beschreibung gewisser Aspekte von Lebensverläufen, kann die Abhängigkeit der Beschreibung vom vorausgesetzten Biographieschema zum Ausgangspunkt einer Kritik werden, indem man zeigt, daß eine Veränderung des Biographieschemas (darin eingeschlossen die Berücksichtigung weiterer Merkmale von Lebensverläufen) zu anderen – besseren – Einsichten in den zu beschreibenden Sachverhalt führt. Examplarisch kann man an den in Abschnitt 3.4.1 dargestellten Vergleich des Heiratsverhaltens in den beiden Teilstichproben A (überwiegend Deutsche) und B (Ausländer) des SOEP denken. Zunächst erscheint es so, daß in der Teilstichprobe A deutlich später geheiratet wird als in der Teilstichprobe B. Mithilfe einer einfachen Unterscheidung von Geburtskohorten kann jedoch eine wesentlich differenziertere Einsicht in das unterschiedliche Heiratsverhalten gewonnen werden.

[19] Schließlich sollte auch nicht vergessen werden, daß die jeweils verfügbaren Daten ungenau und unvollständig sind und daß bei der praktischen Datenanalyse meistens mehr oder weniger problematische Entscheidungen getroffen werden müssen, deren Einfluß auf die schließlich erzielten Ergebnisse kaum abschätzbar ist. Während die statistische Theorie weit fortgeschritten ist, wird diesen „datentechnischen“ Problemen oft keine hinreichende Aufmerksamkeit geschenkt. Vgl. als exemplarische Diskussion dieser Probleme Ludwig-Mayerhofer [1992].

b) Die unter dem Stichwort „unbeobachtete Heterogenität“ geführte Diskussion kreist um die Frage, ob und ggf. wie bei der Modellbildung berücksichtigt werden sollte, daß die Ergebnisse der Modellschätzung auch von Sachverhalten abhängig sein *könnten*, die bei der Modellspezifikation nicht explizit erfaßt worden sind. Insbesondere im Hinblick auf die Frage der Verweildauerabhängigkeit von Übergangsraten wird argumentiert, daß man sich nicht sicher sein kann, ob eine empirisch ermittelte Verweildauerabhängigkeit „tatsächlich“ der Fall ist, weil sie auch das Ergebnis „unbeobachteter Heterogenität“ (der individuellen Lebensverläufe) sein könnte.

Man kann dieser Vermutung einerseits, wie unter (a) angedeutet wurde, eine konstruktive Bedeutung geben: sie kann zum Anlaß genommen werden, nach möglicherweise angemesseneren Beschreibungen zu suchen. In der Literatur wird ihr indessen vielfach eine grundsätzliche Bedeutung gegeben, indem versucht wird, eine „wirkliche“ von einer „bloß scheinbaren“ Verweildauerabhängigkeit von Übergangsraten zu unterscheiden. Ich glaube jedoch, daß der Versuch, eine solche Unterscheidung zu treffen, irreführend ist. Die entscheidende Frage ist, worauf sich Aussagen über Übergangsraten beziehen. In der Literatur wird häufig angenommen, daß sich Übergangsraten als Eigenschaften von Individuen interpretieren lassen. Geht man von dieser Interpretation aus, wird es jedoch sinnlos bzw. zu einer rein spekulativen Frage, ob es verweildauerabhängige Übergangsraten überhaupt gibt. Nichts spricht dann gegen die alternative Spekulation, für jedes Individuum eine ihm eigene Übergangsrate anzunehmen und jede bei einer Gesamtheit von Individuen empirisch beobachtbare Verweildauerabhängigkeit als einen Mischungseffekt, d.h. als eine Folge (beobachteter und) unbeobachteter Heterogenität der Individuen zu interpretieren.

Ich gehe demgegenüber, wie in Abschnitt 2.3 dargestellt worden ist, davon aus, daß der Begriff der Übergangsrate (wie alle nicht subjektiven Wahrscheinlichkeitsbegriffe) nur im Hinblick auf Gesamtheiten sinnvoll verstanden werden können. Übergangsraten sind dann nicht Eigenschaften von Individuen, sondern Eigenschaften von Gesamtheiten von Individuen. Sie beschreiben den Zusammenhang von Ereignissen in einer Gesamtheit von Lebensverläufen. Stellt man zum Beispiel fest, daß es bei einer gewissen Gesamtheit von Episoden eine zeitabhängige Übergangsrate gibt, ist dies eine Aussage über diese Gesamtheit. Eine solche Aussage wird nicht falsch oder sinnlos dadurch, daß es möglich sein kann, die Gesamtheit in Teilgesamtheiten zu zerlegen, bei denen sich unterschiedliche Übergangsraten feststellen lassen. Es wäre auch irreführend, davon zu sprechen, daß es in der zunächst vorausgesetzten Gesamtheit von Episoden nur eine „scheinbare“ Verweildauerabhängigkeit gibt. Es erscheint angemessener, die Situation so darzustellen: daß die Einsicht, daß es in den Teilgesamtheiten unterschiedliche Verläufe der Übergangsraten gibt, eine Möglichkeit liefert, den in der Gesamtheit beobachtbaren Verlauf zu erklären. Das Wort „erklären“ hat hier die gleiche Bedeutung, wie oben in Abschnitt 4.1.6

erläutert wurde: ein Episodenverlauf, der von einem Ereignis E_1 zu einem Ereignis E_2 führt, wird (partiell) erklärt, indem beschrieben wird, wie sein Ablauf durch vermittelnde Ereignisse bedingt wird.

c) Welche Konsequenzen aus der Existenz von zweifellos immer vorhandener „unbeobachteter Heterogenität“ gezogen werden (sollten), hängt schließlich davon ab, welches Erkenntnisinteresse mit der Modellbildung verfolgt wird. Wie in Abschnitt 3.7.1 ausgeführt wurde, kann man Übergangsratenmodelle auch als Regressionsmodelle interpretieren. Bei dieser Interpretation liefert die Übergangsrate eine Charakterisierung der Residuen des Modells. Zielt das Erkenntnisinteresse darauf ab, Modelle zu finden, mit denen sich individuelle Lebensverläufe vorhersagen lassen, erscheinen diese Residuen als ein Ausdruck der bisher noch „unerklärten Varianz“. Die Idee, daß diese „unerklärte Varianz“ beliebig klein gemacht werden könnte, impliziert, daß auch das Konzept der Übergangsrate und ihre mögliche Verweildauerabhängigkeit nur als Ausdruck einer noch unzureichenden Modellbildung angesehen werden kann. Bezieht man die Modellbildung demgegenüber auf die soziologische Aufgabe, Einsichten in soziale Regeln zu gewinnen, denen die Individuen als Vermittler von Ereigniszusammenhängen folgen, kommt in der Verteilung der Residuen nicht eine Grenze der Erklärungsleistung des Modells zum Ausdruck, sondern ihre Beschreibung durch eine meistens zeitabhängige Übergangsrate liefert einen Teil des für die soziologische Theoriebildung intendierten Wissens.

d) Schließlich sollte auch nicht vergessen werden, daß die Interpretation von Modellen, die sich um eine explizite Berücksichtigung „unbeobachteter Heterogenität“ bemühen, typischerweise spekulativ und fragwürdig ist. Um dies zu verdeutlichen, gehe ich von einer einfachen Episode mit konkurrierenden Risiken aus. Der empirische Sachverhalt wird durch die Zufallsvariablen (T, D) beschrieben. T ist die auf einer Prozeßzeitachse gemessene Verweildauer im Ausgangszustand, D liefert eine Angabe über den Zielzustand, in den der Ausgangszustand am Ende der Episode verlassen wird.

Um in dieser Situation „unbeobachtete Heterogenität“ zu berücksichtigen, kann die Existenz einer weiteren, nicht beobachtbaren Zufallsvariable U angenommen werden, so daß der Ausgangspunkt für die Modellbildung in der dreidimensionalen Zufallsvariable (T, D, U) besteht. Die Frage ist, ob sich Einsichten in die durch U bedingte Verteilung von (T, D) gewinnen lassen.

Zunächst muß betont werden, daß die Zufallsvariable U per Definition nicht beobachtet werden kann. Eine Verknüpfung mit beobachtbaren Episodenverläufen ist deshalb nur möglich, wenn die Verteilung von U parametrisch spezifiziert wird. Dafür gibt es zahlreiche unterschiedliche Möglichkeiten. Im folgenden gehe ich exemplarisch von einer diskreten Mischverteilung aus, d.h. es wird angenommen, daß U nur eine gewisse Anzahl von Werten u_k $(k = 1, \ldots, K)$ annehmen kann, mit den korrespondierenden Wahrscheinlichkeiten π_k.[20]

Hat man diese (oder eine andere) Annahme über die Verteilung von U getroffen, kann man die (Rand-)Verteilung von (T, D) ableiten. Es ist eine Mischung aus den Verteilungen von (T, D, U) über alle möglichen Realisierungen der Zufallsvariablen U. Bei einer diskreten Mischverteilung erhält man

$$\mathrm{P}\,(T = t, D = d) = \sum_{k=1}^{K} \pi_k \,\mathrm{P}\,(T = t, D = d \mid U = u_k)$$

Ebenso erhält man eine Mischung für die Survivorfunktion

$$\mathrm{P}\,(T > t) = \sum_{k=1}^{K} \pi_k \,\mathrm{P}\,(T > t \mid U = u_k)$$

und schließlich auch für die zustandsspezifischen Übergangsraten. Man kann dies als eine Vorgehensweise betrachten, die zu einem komplexeren Modell für die Zufallsvariable (T, D) führt, deren Realisierungen beobachtbar sind. Aus der Perspektive der Frage, wie ein möglichst angemessenes Modell für (T, D) gefunden werden kann, kann man also den Sinn einer Berücksichtigung „unbeobachteter Heterogenität“ darin sehen, daß dadurch für die Modellschätzung komplexere, weniger restriktive Modellklassen verwendet werden können. Es stellt sich jedoch die Frage, ob eine darüber hinausgehende Interpretation möglich ist.

Es liegt nahe, ein Mischverteilungsmodell der oben beschriebenen Art so zu interpretieren, daß es die Existenz von K unterschiedlichen Gruppen in der Grundgesamtheit „zeigt“, bei denen (möglicherweise) die Zufallsvariablen (T, D) eine jeweils unterschiedliche Verteilung haben.[21] Zum Beispiel konstruiert Schneider [1991, S. 186f] ein Mischungverteilungsmodell mit zwei diskreten Komponenten für Arbeitslosigkeitsepisoden und interpretiert dann die beiden Komponenten dieser Mischung als unterschiedliche Typen von Episodenverläufen für zwei Subpopulationen. Diese Interpretation ist jedoch problematisch, denn die Existenz unterschiedlicher *Gruppen von Individuen* kann mit einem solchen Modell nicht bewiesen werden; die Annahme, daß es Gruppen *von Individuen* gibt, wird der Interpretation spekulativ vorausgesetzt. Die mathematisch gegebene Möglichkeit zur Konstruktion eines Mischverteilungsmodells liefert tatsächlich kei-

[20] Der Vorschlag, unbeobachtete Heterogenität in Übergangsratenmodellen durch diskrete Mischverteilungen zu berücksichtigen, stammt von Heckman und Singer [1982], [1984]. Diskussionen dieses Vorschlags finden sich u.a. bei Trussell und Richards [1985], Yamaguchi [1986] und Schneider [1991, S. 174ff].

[21] Es geht hier nur um die Frage, ob die Komponenten einer solchen Mischverteilung sinnvoll interpretiert werden können, nicht darum, ob und ggf. wie sie identifiziert und berechnet werden können. Zur Frage der Identifizierbarkeit vgl. Elbers und Ridder [1982], Manton et al. [1992].

nerlei Begründung für die Annahme, daß die Komponenten der Mischverteilung *individuell zurechenbar* sind.

Selbst bei einer strikt individuenbezogenen und probabilistischen Betrachtungsweise läßt ein Mischungverteilungsmodell zwei unterschiedliche Interpretationen zu, über die jedoch mit den verfügbaren Beoachtungen nicht entschieden werden kann. Die erste Interpretation beruht auf der Annahme, daß es zwei Zufallsgeneratoren gibt, die sequentiell wirksam werden. Jedes Individuum erhält zunächst (konditional auf seine beobachtbaren und im Modell berücksichtigten Merkmale) durch den Zufallsgenerator $\mathcal{G}_1$ eine Realisierung der Zufallsvariable U, dann durch den Zufallsgenerator $\mathcal{G}_2$ eine Realisierung der Zufallsvariable (T, D). Diese sequentielle Interpretation erlaubt es, von unbeobachteten Unterschieden der Individuen bzw. von Gruppen von Individuen zu sprechen, die durch den Zufallsgenerator $\mathcal{G}_1$ zustande kommen. Stattdessen kann man jedoch auch annehmen, daß es nur einen Zufallsgenerator $\mathcal{G}$ gibt, der aus der Mischung von $\mathcal{G}_1$ und $\mathcal{G}_2$ besteht. Die Individuen ziehen dann zum Episodenbeginn ihr Los nicht sequentiell aus $\mathcal{G}_1$ und $\mathcal{G}_2$, sondern unmittelbar aus $\mathcal{G}$. Bei dieser Interpretation gibt es offensichtlich nur noch „homogene“ Zufallseinflüsse, die – konditional auf die beobachtbaren Unterschiede der Individuen – ihren weiteren Episodenverlauf beeinflussen. Mithilfe der empirisch verfügbaren Beobachtungen kann zwischen diesen beiden Interpretationen nicht entschieden werden. Jeder Versuch, die Residuen eines Modells für beobachtete Sachverhalte in Komponenten zu zerlegen, ist unvermeidlich spekulativ und problematisch.

4.2 Modelle zur Analyse mehrdimensonaler Prozesse

In diesem Abschnitt soll dargestellt werden, wie Übergangsratenmodelle bei parallel ablaufenden Prozessen konstruiert werden können. Ausgangspunkt ist die in (4.1) definierte und in Abbildung 4.1.1 veranschaulichte Vorstellung paralleler Prozesse. Um die Darstellung zu vereinfachen, gehe ich im folgenden nur von zwei Prozessen aus, die ich als den A-Prozeß und als den B-Prozeß bezeichne. Das wesentliche Ergebnis der Überlegungen besteht darin, daß sich ein Modell für den Ereigniszusammenhang in den parallel ablaufenden Prozessen in zwei separate Modelle zerlegen läßt, ein Modell für den A-Prozeß, bei dem die Interaktion mit dem B-Prozeß dadurch berücksichtigt werden kann, daß er durch zeitabhängige Kovariablen repräsentiert wird, und ein entsprechendes Modell für den B-Prozeß und seine Interaktion mit dem A-Prozeß. Die Überlegungen erfolgen zunächst für eine diskrete, dann für eine stetige Zeitachse.

4.2.1 Diskrete Ereigniszeitpunkte

Wie bisher gehen wir von einem deskriptiven Wahrscheinlichkeitsraum für eine Grundgesamtheit Ω aus. In diesem Rahmen werden zwei Folgen von Zustandsvariblen definiert:

$$Y_t^A : \Omega \longrightarrow \mathcal{Y}^A \qquad (4.15)$$
$$Y_t^B : \Omega \longrightarrow \mathcal{Y}^B$$

$\mathcal{Y}^A$ ist ein endlicher Zustandsraum für den Prozeß A, $\mathcal{Y}^B$ ist der endliche Zustandsraum für den Prozeß B. Der Index t bezieht sich auf eine diskrete Prozeßzeitachse $t = 0, 1, 2, \ldots$ Beide Prozesse beginnen zum Zeitpunkt $t = 0$ in einem jeweils spezifischen Anfangszustand: $Y_0^A = y_a^A \in \mathcal{Y}^A$ und $Y_0^B = y_a^B \in \mathcal{Y}^B$. Inbezug auf das Ende des gemeinsamen Prozesses gibt es drei Möglichkeiten.

(a) Man kann annehmen, daß der Teilprozeß A irgendwann in einem absorbierenden Endzustand y_e^A enden muß und daß der Teilprozeß B mindestens bis zu diesem Ereignis definiert ist. (b) Man kann umgekehrt annehmen, daß der Teilprozeß B irgendwann in einem absorbierenden Endzustand y_e^B enden muß und daß der Teilprozeß A mindestens bis zu diesem Ereignis definiert ist. (c) Schließlich kann man annehmen, daß beide Teilprozesse in einem absorbierenden Endzustand enden können und daß der gemeinsame Prozeß dann beendet wird, wenn zum erstenmal in einem der beiden Teilprozesse ein absorbierender Endzustand erreicht wird. Ich gehe im folgenden von dieser dritten Möglichkeit aus.

Die Prozesse A und B laufen parallel auf der Ebene von Individuen. Zum Zeitpunkt t befindet sich das Individuum $\omega \in \Omega$ gleichzeitig in den Zuständen $Y_t^A(\omega)$ und $Y_t^B(\omega)$. Um die Darstellung übersichtlich zu halten, werden zunächst keine weiteren Kovariablen betrachtet. Die Aufgabe

besteht darin, geeignete Modelle für den gemeinsamen, parallelen Verlauf der beiden Prozesse zu finden. Wie bereits diskutiert worden ist, gibt es zwei Möglichkeiten.

a) Eine erste Möglichkeit besteht darin, einen aus den Prozessen A und B konstruierbaren gemeinsamen Prozeß zu betrachten und mögliche Interdependenzen implizit zu lassen. Ausgehend von (4.15) kann der gemeinsame Prozeß auf folgende Weise gebildet werden:

$$Y_t : \Omega \longrightarrow \mathcal{Y} \quad \text{mit} \quad Y_t(\omega) = \big(Y_t^A(\omega), Y_t^B(\omega)\big) \in \mathcal{Y} = \mathcal{Y}^A \times \mathcal{Y}^B \tag{4.16}$$

Der Zustandsraum $\mathcal{Y}$ für den gemeinsamen Prozeß wird als kartesisches Produkt der beiden Zustandsräume $\mathcal{Y}^A$ und $\mathcal{Y}^B$ definiert, besteht also aus allen geordneten Zustandskombinationen aus diesen beiden Zustandsräumen.[22] Der Anfangszustand ist $y_a = (y_a^A, y_a^B)$. Der Prozeß endet, wenn einer der möglichen Endzustände, d.h. ein Zustand aus der Menge

$$\mathcal{Y}_e = \big\{(y_e^A, y^B) \mid y^B \in \mathcal{Y}^B\big\} \cup \big\{(y^A, y_e^B) \mid y^A \in \mathcal{Y}^A\big\} \tag{4.17}$$

eintritt. Der gemeinsame Prozeß kann dann auf übliche Weise modelliert werden, ggf. in Abhängigkeit von weiteren exogenen Kovariablen.

Diese Herangehensweise an die Modellbildung bei interdependenten Prozessen hat aber offensichtlich zwei Nachteile. Erstens ist der resultierende gemeinsame Zustandsraum in der Regel sehr komplex, so daß für viele der möglichen Zustandsänderungen nur wenige Beobachtungen verfügbar sind. Zweitens erhält man keinerlei Aufschluß darüber, wie sich die beiden Prozesse wechselseitig beeinflussen.

b) Ich gehe deshalb im folgenden von dem in Abschnitt 4.1.3 diskutierten Darstellungsprinzip für parallele Prozesse aus. Die Grundidee ist, daß die Entwicklung in jedem Teilprozeß zu jedem Zeitpunkt nur von der bis zu diesem Zeitpunkt realisierten Vorgeschichte des *gemeinsamen* Prozesses abhängt. Zu überlegen ist, wie mithilfe dieses Prinzips eine geeignete Zerlegung des gemeinsamen Prozesses erreicht werden kann.

Hierfür gehe ich von einer Darstellung des gemeinsamen Prozesses (4.16) als eine Folge von Episoden aus. Eine Episode endet, und zugleich beginnt eine neue Episode, wenn im gemeinsamen Zustandsraum $\mathcal{Y}$ einer neuer Zustand erreicht wird. Zur Beschreibung kann der in Abschnitt 2.4.5 eingeführte begriffliche Rahmen verwendet werden. Die Zufallsvariable L gibt die Anzahl der Episoden an. T_l ist der Zeitpunkt, zu dem die l.te Episode endet, und D_l ist der Endzustand der l.ten Episode, zugleich der Anfangszustand der folgenden Episode. Mit H_l wird die Geschichte des gemeinsamen Prozesses bis zum Ende der l.ten Episode bezeichnet. Entsprechend (2.18) in Abschnitt 2.4.5 kann dann die Basisgleichung für eine

[22] Wenn einige Zustandskombinationen nicht vorkommen können, kann $\mathcal{Y}$ als eine geeignete Teilmenge des Produkts $\mathcal{Y}^A \times \mathcal{Y}^B$ definiert werden.

Episodendarstellung des gemeinsamen Prozesses folgendermaßen geschrieben werden:[23]

$$\mathrm{P}\,(H_l = h_l \mid L \geq l) = \prod_{k=1}^{l} \mathrm{P}\,(D_k = d_k, T_k = t_k \mid H_{k-1} = h_{k-1}, L \geq k)$$

Um die Zerlegbarkeit dieses Prozesses erörtern zu können, ist es zweckmäßig, neue Zustandsvariablen

$$S_{l,t} : \Omega \longrightarrow \mathcal{Y} \qquad l = 1, \ldots, L_{\max} = \max\{L(\omega) \mid \omega \in \Omega\}, \;\; t = 0, 1, 2, \ldots$$

zu definieren, die sich am Episodenverlauf (nicht unmittelbar an der Prozeßzeit wie die Variablen Y_t) orientieren, also

$$S_{l,t}(\omega) \;\; = \;\; \begin{cases} D_{l-1}(\omega) & \text{wenn } \; t < T_l(\omega) \;\; \text{und} \;\; L(\omega) \geq l \\ D_l(\omega) & \text{wenn } \; t \geq T_l(\omega) \;\; \text{und} \;\; L(\omega) \geq l \\ y_e^* & \text{wenn } \; L(\omega) < l \end{cases}$$

y_e^* ist hierbei ein beliebiger Zustand aus der Menge $\mathcal{Y}_e$, d.h. aus der Menge der möglichen Endzustände des gemeinsamen Prozesses. Als Abkürzung wird die Bezeichnung

$$\bar{S}_{l,t} = (S_{l,t_{l-1}}, \ldots, S_{l,t})$$

verwendet, um den Prozeßverlauf in der l.ten Episode des gemeinsamen Prozesses bis zum Zeitpunkt $t \geq t_{l-1}$ darzustellen; t_{l-1} ist der Anfangszeitpunkt der l.ten Episode des gemeinsamen Prozesses.

Sei jetzt $t \geq t_{l-1}$ ein beliebiger Zeitpunkt in der l.ten Episode des gemeinsamen Prozesses. Der Prozeßverlauf bis zu diesem Zeitpunkt kann dann durch

$$\mathrm{P}\,(\bar{S}_{l,t} = \bar{s}_{l,t}, H_{l-1} = h_{l-1}, L \geq l) \tag{4.18}$$

beschrieben werden. $\bar{s}_{l,t}$ ist eine der möglichen Realisationen des Prozeßverlaufs in der l.ten Episode bis zum Zeitpunkt t.

Diese Beschreibung des Prozeßverlaufs kann nun durch sukzessives Konditionieren auf die jeweilige Vorgeschichte folgendermaßen umgeschrieben werden:

$$\begin{aligned} &\mathrm{P}\,(\bar{S}_{l,t} = \bar{s}_{l,t}, H_{l-1} = h_{l-1}, L \geq l) \; = \\ &\qquad \prod_{\tau=1}^{t} \mathrm{P}\,(S_{l,\tau} = s_{l,\tau} \mid \bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l) \end{aligned} \tag{4.19}$$

[23] Hier und im folgenden wird angenommen, daß die in Abschnitt 2.4.5 diskutierte Bedingung (2.17) erfüllt ist, daß also der Episodenverlauf bis zu einer Episode l nicht davon abhängt, wie viele weitere Episoden noch folgen werden.

Anhand dieser Darstellung kann überlegt werden, wie sich die Interdependenz der beiden Teilprozesse so beschreiben läßt, daß sichtbar wird, wie sie sich wechselseitig bedingen. Im Anschluß an das in Abschnitt 4.1.3 diskutierte Basisprinzip (4.4) kann folgende Annahme formuliert werden:[24]

$$\begin{aligned} &\mathrm{P}\left(S_{l,\tau} = s_{l,\tau} \mid \bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l\right) = \\ &\quad \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid \bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l\right) \cdot \\ &\quad \mathrm{P}\left(S^b_{l,\tau} = s^b_{l,\tau} \mid \bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l\right) \end{aligned} \tag{4.20}$$

Die hochgestellten Indizes a und b sollen andeuten, daß die jeweilige Projektion auf den Zustandsraum $\mathcal{Y}^A$ bzw. $\mathcal{Y}^B$ gemeint ist;[25] also: $S_{l,\tau} = (S^a_{l,\tau}, S^b_{l,\tau})$ und $s_{l,\tau} = (s^a_{l,\tau}, s^b_{l,\tau})$. Bedingung (4.20) besagt, daß die Wahrscheinlichkeiten, zum Zeitpunkt τ (in der l.ten Episode) im Teilprozeß A den Zustand $s^a_{l,\tau}$ und im Teilprozeß B den Zustand $s^b_{l,\tau}$ einzunehmen, unabhängig voneinander sind, wenn dabei auf den *gemeinsamen* Prozeßverlauf bis zum Zeitpunkt $\tau - 1$ konditioniert wird.

Die in (4.20) formulierte Annahme ist mit dem in Abschnitt 4.1.3 diskutierten Basisprinzip (4.4) identisch. Wird sie vorausgesetzt, kann die in (4.19) gegebene Prozeßbeschreibung folgendermaßen zerlegt werden:

$$\begin{aligned} &\mathrm{P}\left(\bar{S}_{l,t} = \bar{s}_{l,t}, H_{l-1} = h_{l-1}, L \geq l\right) = \\ &\quad \prod_{\tau=1}^{t} \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid \bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l\right) \cdot \\ &\quad \prod_{\tau=1}^{t} \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid \bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l\right) \end{aligned} \tag{4.21}$$

Der erste Term auf der rechten Seite beschreibt die Entwicklung im Teilprozeß A unter der Bedingung der jeweils vorangegangenen gemeinsamen Prozeßgeschichte, der zweite Term auf der rechten Seite liefert eine analoge Beschreibung für die Entwicklung des Teilprozesses B.

Diese Darstellung kann schließlich so verändert werden, daß beide Teilprozesse eine ihnen eigene Episodendarstellung erhalten und die Abhängigkeit vom jeweils anderen Teilprozeß in der üblichen Form einer Abhängigkeit von zeitveränderlichen Kovariablen erscheint. Dies ist im wesentlichen nur ein terminologisches Problem. Da die erforderliche Überlegung für beide Teilprozesse vollständig analog verläuft, beschränke ich mich auf eine Darstellung für den Teilprozeß A.

Zunächst muß für diesen Teilprozeß eine Episodendarstellung definiert

[24] Vgl. Pötter [1993, S. 773], der dies als Annahme einer „local conditional independence“ bezeichnet.

[25] Die Indizes werden kleingeschrieben, um Projektionen des gemeinsamen Prozesses von separat für die beiden Teilprozesse definierten Symbolen zu unterscheiden.

werden. Dies geschieht in der üblichen Weise durch Einführung der Zufallsvariablen D^A_k, T^A_k und L^A. k ist hierbei die Ordnungsnummer für die Episoden im Teilprozeß A. T^A_k ist der Zeitpunkt auf der zugrunde liegenden Prozeßzeitachse, bei dem die k.te Episode des Teilprozesses A endet. D^A_k ist der Endzustand der k.ten Episode. Die Zufallsvariable L^A gibt die Anzahl der Episoden an, die ein Individuum im Teilprozeß A erreicht.

Außerdem ist eine Notation erforderlich, um den Verlauf des Teilprozesses B in der Episodenstruktur des Teilprozesses A ausdrücken zu können. Um deutlich zu machen, daß der Teilprozeß B als eine Folge zeitabhängiger Kovariablen für den Teilprozeß A betrachtet werden kann, werden die Zufallsvariablen

$$X^B_{k,t} : \Omega \longrightarrow \mathcal{Y}^B \quad \text{definiert durch} \quad X^B_{k,t}(\omega) = Y^B_t(\omega) \quad \text{wenn} \quad L^A(\omega) \geq k$$

eingeführt. Als Abkürzung für die Entwicklung des Teilprozesses B in der k.ten Episode des Teilprozesses A, bis zu einem Zeitpunkt t, wird die Formulierung

$$\bar{X}^B_{k,t} = \left(X^B_{k,t^A_{k-1}}, \ldots, X^B_{k,t}\right)$$

verwendet. t^A_{k-1} ist der Zeitpunkt, zu dem die k.te Episode des Teilprozesses A beginnt.

Mithilfe dieser Notation kann die Darstellung des Teilprozesses A in (4.21), also

$$\prod_{\tau=1}^{t} \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid \bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l\right) \tag{4.22}$$

umgeschrieben werden. Diese Darstellung geht davon aus, daß der gemeinsame Prozeß bis zu einem Zeitpunkt t in seiner l.ten Episode betrachtet wird. Es ist dies ein Produkt bedingter Wahrscheinlichkeiten, wobei die jeweilige Vorgeschichte als Bedingung auftritt. In der Bedingung $H_{l-1} = h_{l-1}$ ist insbesondere die Information enthalten, wann im Teilprozeß A zum letztenmal ein Zustandswechsel stattgefunden hat. Es sei dies der Zeitpunkt t^A_{k-1}, also der Anfangszeitpunkt einer Episode des Teilprozesses A mit der Ordnungsnummer k und dem Ausgangszustand d^A_{k-1}. Ggf. ist $t^A_{k-1} = 0$, also $k = 1$.

Die in den Bedingungen der in (4.22) vorkommenden Wahrscheinlichkeiten enthaltene Information kann folgendermaßen reformuliert werden:

$$\begin{aligned} &\left(\bar{S}_{l,\tau-1} = \bar{s}_{l,\tau-1}, H_{l-1} = h_{l-1}, L \geq l\right) \equiv \\ &\quad \left(T^A_k \geq \tau, \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l\right) \end{aligned}$$

$T^A \geq \tau$ bedeutet, daß bis zum Zeitpunkt $\tau - 1$ in der k.ten Episode des Teilprozesses A der Ausgangszustand d^A_{k-1} noch nicht verlassen wurde.

$\bar{X}^B_{k,\tau-1}$ beschreibt den Teilprozeß B in der k.ten Episode des Teilprozesses A bis zum Zeitpunkt $\tau-1$ (einschließlich); $\bar{x}^B_{k,\tau-1}$ ist die in der Vorgeschichte enthaltene Information über die Realisierung dieses Teilprozesses. Schließlich liefert $H^A_{k-1} = h^A_{k-1}$ eine Beschreibung der Vorgeschichte des *gemeinsamen* Prozesses bis zum Beginn der k.ten Episode im Teilprozeß A. Durch diese Reformulierung der Vorgeschichte kann (4.22) jetzt folgendermaßen geschrieben werden:

$$\prod_{\tau=1}^{t} \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid T^A_k \geq \tau, \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l\right) \tag{4.23}$$

Jetzt gibt es zwei Möglichkeiten. Erstens $S^a_{l,t} = d^A_{k-1}$, d.h. daß zum Zeitpunkt t der Anfangszustand der k.ten Episode des Teilprozesses A noch nicht verlassen worden ist. In diesem Fall kann (4.23) folgendermaßen umgeschrieben werden:

$$\prod_{\tau=1}^{t} \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid T^A_k \geq \tau, \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l\right) = \tag{4.24}$$

$$\prod_{\tau=1}^{t} \mathrm{P}\left(T^A_k > \tau \mid T^A_k \geq \tau, \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l\right) =$$

$$\prod_{\tau=1}^{t} \left(1 - \mathrm{P}\left(T^A_k = \tau \mid T^A_k \geq \tau, \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l\right)\right) =$$

$$\prod_{\tau=1}^{t} \left(1 - r^{A,k}(\tau \mid \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l)\right)$$

Hierbei bezeichnet $r^{A,k}(\cdot)$ die Abgangsrate aus dem Ausgangszustand der k.ten Episode im Teilprozeß A.[26]

Die zweite Möglichkeit besteht darin, daß zum Zeitpunkt t ein Übergang in einen neuen Zustand, in den Endzustand der k.ten Episode des Teilprozesses A stattfindet, also $S^a_{l,t} = d^A_k \neq d^A_{k-1}$. In diesem Fall liefert eine analoge Reformulierung von (4.23) die Darstellung

$$\prod_{\tau=1}^{t} \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid T^A_k \geq \tau, \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l\right) = \tag{4.25}$$

$$r^{A,k}_{d^A_k}(t \mid \bar{X}^B_{k,t-1} = \bar{x}^B_{k,t-1}, H^A_{k-1} = h^A_{k-1}, L \geq l) \cdot$$

$$\prod_{\tau=1}^{t-1} \left(1 - r^{A,k}(\tau \mid \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l)\right)$$

[26] Definiert als die Summe der zustandsspezifischen Übergangsraten für diese Episode.

Hierbei bezeichnet $r^{A,k}_{d^A_k}(\cdot)$ die zustandsspezifische Übergangsrate für einen Übergang in den Zielzustand d^A_k in der k.ten Episode des Teilprozesses A.

Um die beiden in (4.24) und (4.25) dargestellten Möglichkeiten in eine gemeinsame Formulierung bringen zu können, ist es zweckmäßig, Zensierungsindikatoren einzuführen:

$$\delta^A_t = \begin{cases} 1 & \text{wenn zum Zeitpunkt } t \text{ im Teilprozeß A ein} \\ & \text{Zustandswechsel auftritt} \\ 0 & \text{andernfalls} \end{cases}$$

(4.24) und (4.25) können dann folgendermaßen zusammengefaßt werden:

$$\prod_{\tau=1}^{t} \mathrm{P}\left(S^a_{l,\tau} = s^a_{l,\tau} \mid T^A_k \geq \tau, \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l\right) =$$

$$r^{A,k}_{d^A_k}(t \mid \bar{X}^B_{k,t-1} = \bar{x}^B_{k,t-1}, H^A_{k-1} = h^A_{k-1}, L \geq l)^{\delta^A_t} \cdot$$

$$\prod_{\tau=1}^{t-\delta^A_t} \left(1 - r^{A,k}(\tau \mid \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l)\right)$$

Ganz analog kann man, ausgehend vom zweiten Term auf der rechten Seite von (4.21), den Teilprozeß B als eine Episodenfolge darstellen, bei der die Zustände im Teilprozeß A als zeitabhängige Kovariablen auftreten. Dies liefert schließlich folgende Darstellung für den in (4.18) definierten gemeinsamen Prozeßverlauf bis zu einem Zeitpunkt t in der l.ten Episode des gemeinsamen Prozesses:

$$\mathrm{P}\left(\bar{S}_{l,t} = \bar{s}_{l,t}, H_{l-1} = h_{l-1}, L \geq l\right) =$$

$$r^{A,k}_{d^A_k}(t \mid \bar{X}^B_{k,t-1} = \bar{x}^B_{k,t-1}, H^A_{k-1} = h^A_{k-1}, L \geq l)^{\delta^A_t} \cdot$$

$$\prod_{\tau=1}^{t-\delta^A_t} \left(1 - r^{A,k}(\tau \mid \bar{X}^B_{k,\tau-1} = \bar{x}^B_{k,\tau-1}, H^A_{k-1} = h^A_{k-1}, L \geq l)\right) \cdot$$

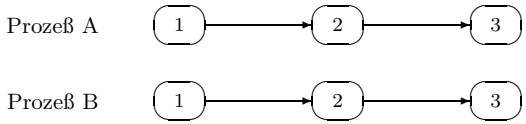
$$r^{B,k}_{d^B_k}(t \mid \bar{X}^A_{k,t-1} = \bar{x}^A_{k,t-1}, H^B_{k-1} = h^B_{k-1}, L \geq l)^{\delta^B_t} \cdot$$

$$\prod_{\tau=1}^{t-\delta^B_t} \left(1 - r^{B,k}(\tau \mid \bar{X}^A_{k,\tau-1} = \bar{x}^A_{k,\tau-1}, H^B_{k-1} = h^B_{k-1}, L \geq l)\right)$$

Bei der Ableitung dieser Darstellung sind wir von einem beliebigen Zeitpunkt in einer beliebigen Episode des gemeinsamen Prozesses ausgegangen; sie kann also für alle Zeitpunkte bis zum Ende des gemeinsamen Prozesses verwendet werden. Sie kann insbesondere verwendet werden, um den gemeinsamen Prozeß induktiv durch eine Folge von Episoden für den Teilprozeß A und eine Folge von Episoden für den Teilprozeß B zu beschreiben.

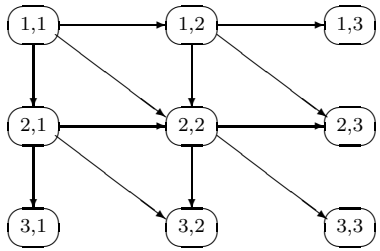
Beispiele mit simulierten Daten

Als Beispiel betrachten wir zwei Prozesse mit jeweils drei nichtwiederholbaren Zuständen. Die möglichen Übergänge sehen folgendermaßen aus:



Jeder Prozeß beginnt in einem Anfangszustand (1), dann erfolgt ein Übergang in den Zustand (2), schließlich wird in den absorbierenden Endzustand (3) gewechselt.

Beide Prozesse laufen parallel. Bildet man einen gemeinsamen Prozeß, sehen die Zustände und die möglichen Übergänge folgendermaßen aus:



Jede Zustandskombination mit dem Teilzustand (3) an erster oder an zweiter Stelle ist ein absorbierender Endzustand des gemeinsamen Prozesses.

Um einige Beispielrechnungen durchführen zu können, werden Zufallsdaten erzeugt. Für die Datenerzeugung wird ein logistisches Regressionsmodell für die Übergangsraten verwendet. Der Algorithmus zur Datenerzeugung für N Individuen sieht folgendermaßen aus:

(1) Erstes Individuum: $i = 1$

(2) Anfangszeit: $t = 0$

(3) Anfangszustand: $Y_t^A = 1, \;\; Y_t^B = 1$

(4) Nächster Zeitpunkt: $t = t + 1$

(5) Berechnung der Wahrscheinlichkeit, p_A, für eine Zustandsänderung im Prozeß A in Abhängigkeit von der Vorgeschichte.

(6) Ziehen einer gleichverteilten Zufallszahl r im Intervall $[0, 1]$. Wenn $r \leq p_A$, dann $Y_t^A = Y_{t-1}^A + 1$, andernfalls $Y_t^A = Y_{t-1}^A$.

(7) Berechnung der Wahrscheinlichkeit, p_B, für eine Zustandsänderung im Prozeß B in Abhängigkeit von der Vorgeschichte.

(8) Ziehen einer gleichverteilten Zufallszahl r im Intervall $[0, 1]$. Wenn $r \leq p_B$, dann $Y_t^B = Y_{t-1}^B + 1$, andernfalls $Y_t^B = Y_{t-1}^B$.

(9) Wenn $Y_t^A \neq 3$ und $Y_t^B \neq 3$, Fortsetzung bei (4).

(10) Nächstes Individuum: $i = i + 1$.

(11) Wenn $i \leq N$, Fortsetzung bei (2).

(12) Ende.

Im folgenden wird dieser Algorithmus verwendet, um für einige einfache Situationen Daten zu erzeugen und Modellschätzungen durchzuführen. Die Anzahl der Individuen ist stets $N = 1000$.

a) Als erstes Beispiel wird angenommen, daß sich die beiden Prozesse unabhängig voneinander entwickeln. Zur Datenerzeugung wird folgendes Modell angenommen:

$$r_2^{A,1} = \frac{\exp(\alpha_2^{A,1})}{1 + \exp(\alpha_2^{A,1})} \qquad r_3^{A,2} = \frac{\exp(\alpha_3^{A,2})}{1 + \exp(\alpha_3^{A,2})} \tag{4.26}$$

$$r_2^{B,1} = \frac{\exp(\alpha_2^{B,1})}{1 + \exp(\alpha_2^{B,1})} \qquad r_3^{B,2} = \frac{\exp(\alpha_3^{B,2})}{1 + \exp(\alpha_3^{B,2})}$$

Die Bezeichnungen entsprechen der bisher verwendeten Notation; zum Beispiel ist $r_d^{A,l}$ ist die Rate in der l.ten Episode des Prozesses A für einen Übergang in den Zustand d. Für die Modellparameter werden die Werte

$$\alpha_2^{A,1} = -2.2 \quad \alpha_3^{A,2} = -2.9 \quad \alpha_2^{B,1} = -2.9 \quad \alpha_3^{B,2} = -2.2$$

angenommen. Dies entspricht den Übergangsraten

$$r_2^{A,1} = 0.052 \quad r_3^{A,2} = 0.100 \quad r_2^{B,1} = 0.100 \quad r_3^{B,2} = 0.052$$

Die Datenerzeugung wird für jeden Prozeß getrennt vorgenommen, d.h. jeder Prozeß endet, wenn in ihm ein absorbierender Endzustand (3) erreicht wird, unabhängig davon, wie lange der jeweils andere Prozeß dauert. Eine Schätzung des Modells (4.26) ergibt:

$$\hat{\alpha}_2^{A,1} = -2.1805 \quad \hat{\alpha}_3^{A,2} = -2.9186 \quad \hat{\alpha}_2^{B,1} = -2.9146 \quad \hat{\alpha}_3^{B,2} = -2.1839$$

Die damit geschätzten Übergangsraten sind

$$\hat{r}_2^{A,1} = 0.102 \quad \hat{r}_3^{A,2} = 0.051 \quad \hat{r}_2^{B,1} = 0.051 \quad \hat{r}_3^{B,2} = 0.101$$

b) Wie in (a) wird angenommen, daß die Übergangsraten nicht vom Zustand im jeweils anderen Prozeß abhängen. Jeder Prozeß endet jedoch, wenn zum erstenmal in einem der beiden Prozesse ein absorbierender Endzustand erreicht wird. Im Unterschied zu (a) gibt es jetzt in beiden Episoden beider Prozesse rechts zensierte Verweildauern. Die Datenerzeugung mit dem oben angegebenen Algorithmus führt in diesem Fall zu folgender Situation: Im Prozeß A gibt es in der ersten Episode 15% und in der zweiten Episode 37% rechts zensierte Fälle. Im Prozeß B sind es in der ersten

Episode 32% und in der zweiten Episode 26% rechts zensierte Fälle.[27] Die Schätzung des Modells (4.26) ergibt jetzt:

$$\hat{\alpha}_2^{A,1} = -2.2218 \quad \hat{\alpha}_3^{A,2} = -2.8840 \quad \hat{\alpha}_2^{B,1} = -2.9171 \quad \hat{\alpha}_3^{B,2} = -2.2983$$

Dem entsprechen die geschätzten Übergangsraten

$$\hat{r}_2^{A,1} = 0.098 \quad \hat{r}_3^{A,2} = 0.053 \quad \hat{r}_2^{B,1} = 0.051 \quad \hat{r}_3^{B,2} = 0.091$$

Die Schätzwerte haben sich infolge der rechts zensierten Fälle (aber auch wegen eines etwas anderen Verlaufs der Zufallsdatenerzeugung) geringfügig verändert; die zur Datenerzeugung verwendeten Parameter können aber immer noch gut rekonstruiert werden.

c) Jetzt wird angenommen, daß es eine einseitige Beeinflussung des Prozesses A durch den Prozeß B gibt (jedoch nicht umgekehrt). Es wird eine Variable

$$X_t^B \quad = \quad \begin{cases} 1 & \text{wenn Prozeß B zum Zeitpunkt } t \text{ im Zustand 2} \\ 0 & \text{andernfalls} \end{cases}$$

definiert, um den Zustand des Prozesses B zu erfassen. Das Modell für die Datenerzeugung sieht folgendermaßen aus:

$$r_2^{A,1}(t) = \frac{\exp(\alpha_2^{A,1} + X_t^B\,\beta_2^{A,1})}{1+\exp(\alpha_2^{A,1}) + X_t^B\,\beta_2^{A,1})} \qquad r_2^{B,1} = \frac{\exp(\alpha_2^{B,1})}{1+\exp(\alpha_2^{B,1})} \tag{4.27}$$

$$r_3^{A,2}(t) = \frac{\exp(\alpha_3^{A,2} + X_t^B\,\beta_3^{A,2})}{1+\exp(\alpha_3^{A,2}) + X_t^B\,\beta_3^{A,2})} \qquad r_3^{B,2} = \frac{\exp(\alpha_3^{B,2})}{1+\exp(\alpha_3^{B,2})}$$

Für die Modellparameter werden die Werte

$$\alpha_2^{A,1} = -2.2 \quad \beta_2^{A,1} = -0.7 \quad \alpha_3^{A,2} = -2.9 \quad \beta_3^{A,2} = \quad 0.7$$
$$\alpha_2^{B,1} = -2.2 \quad \alpha_3^{B,2} = -2.9$$

angenommen. Eine Schätzung des Modells (4.27) ergibt

$$\hat{\alpha}_2^{A,1} = -2.2195 \quad \hat{\beta}_2^{A,1} = -0.6592 \quad \hat{\alpha}_3^{A,2} = -2.8751 \quad \hat{\beta}_3^{A,2} = \quad 0.6403$$
$$\hat{\alpha}_2^{B,1} = -2.1813 \quad \hat{\alpha}_3^{B,2} = -2.9066$$

Die Übergangsraten im Prozeß hängen jetzt davon ab, in welchem Zustand sich der Prozeß B befindet. Zum Beispiel ergibt sich für die erste Episode des Prozesses A: Wenn sich Prozeß B im Zustand 1 befindet, ist die Übergangsrate

$$r_2^{A,1}(t \mid X_t^B = 0) = 0.100 \qquad \text{geschätzt: } 0.098$$

[27] Da Zufallsdaten erzeugt werden, hängen die genauen Werte vom jeweils verwendeten Zufallsgenerator ab.

Wenn sich Prozeß B im Zustand 2 befindet, ist die Rate

$$r_2^{A,1}(t \mid X_t^B = 1) = 0.052 \qquad \text{geschätzt: } 0.053$$

d) Jetzt wird angenommen, daß sich die Prozesse A und B wechselseitig beeinflussen. Um das Modell formulieren zu können, wird zusätzlich zur Variable X_t^B die Variable

$$X_t^A \quad = \quad \begin{cases} 1 & \text{wenn Prozeß A zum Zeitpunkt } t \text{ im Zustand 2} \\ 0 & \text{andernfalls} \end{cases}$$

definiert, um den Zustand des Prozesses A zu erfassen. Das Modell für die Datenerzeugung sieht folgendermaßen aus:

$$r_2^{A,1}(t) = \frac{\exp(\alpha_2^{A,1} + X_t^B\,\beta_2^{A,1})}{1+\exp(\alpha_2^{A,1}) + X_t^B\,\beta_2^{A,1})} \qquad r_2^{B,1}(t) = \frac{\exp(\alpha_2^{B,1} + X_t^A\,\beta_2^{B,1})}{1+\exp(\alpha_2^{B,1} + X_t^A\,\beta_2^{B,1})} \tag{4.28}$$

$$r_3^{A,2}(t) = \frac{\exp(\alpha_3^{A,2} + X_t^B\,\beta_3^{A,2})}{1+\exp(\alpha_3^{A,2}) + X_t^B\,\beta_3^{A,2})} \qquad r_3^{B,2}(t) = \frac{\exp(\alpha_3^{B,2} + X_t^A\,\beta_3^{B,2})}{1+\exp(\alpha_3^{B,2} + X_t^A\,\beta_3^{B,2})}$$

Für die Datenerzeugung werden die Parameter

$$\alpha_2^{A,1} = -2.2 \quad \beta_2^{A,1} = -0.7 \quad \alpha_3^{A,2} = -2.9 \quad \beta_3^{A,2} = \quad 0.7$$
$$\alpha_2^{B,1} = -2.9 \quad \beta_2^{B,1} = \quad 0.7 \quad \alpha_3^{B,2} = -2.2 \quad \beta_3^{B,2} = -0.7$$

angenommen. Eine Schätzung des Modells (4.28) ergibt

$$\hat{\alpha}_2^{A,1} = -2.2231 \quad \hat{\beta}_2^{A,1} = -0.7437 \quad \hat{\alpha}_3^{A,2} = -2.8386 \quad \hat{\beta}_3^{A,2} = \quad 0.6519$$
$$\hat{\alpha}_2^{B,1} = -2.9553 \quad \hat{\beta}_2^{B,1} = \quad 0.7821 \quad \hat{\alpha}_3^{B,2} = -2.3026 \quad \hat{\beta}_3^{B,2} = -0.6430$$

Die folgende Tabelle zeigt die zur Datenerzeugung verwendeten und die geschätzten Übergangsraten.

Rate	Zustand	Datenerzeugung	Geschätzt
$r_2^{A,1}$	$X_t^B = 0$	0.100	0.098
$r_2^{A,1}$	$X_t^B = 1$	0.052	0.049
$r_3^{A,2}$	$X_t^B = 0$	0.052	0.055
$r_3^{A,2}$	$X_t^B = 1$	0.100	0.101
$r_2^{B,1}$	$X_t^A = 0$	0.052	0.049
$r_2^{B,1}$	$X_t^A = 1$	0.100	0.102
$r_3^{B,2}$	$X_t^A = 0$	0.100	0.091
$r_3^{B,2}$	$X_t^A = 1$	0.052	0.050

4.2.2 Stetig approximierte Ereigniszeitpunkte

Jetzt wird eine Situation betrachtet, in der die beiden Prozesse A und B auf einer stetigen Zeitachse definiert sind. Wiederum geht es um die Frage, ob bzw. wie der gemeinsame Prozeß in zwei Teilprozesse zerlegt werden kann, die sich wechselseitig bedingen.

Zunächst wird wieder eine Notation eingeführt, um die beiden Teilprozesse als Folgen von Episoden zu beschreiben. Beide Prozesse beginnen auf einer gemeinsamen Prozeßzeitachse zum Zeitpunkt $t = 0$. Die Zustandsräume sind $\mathcal{Y}^A$ bzw. $\mathcal{Y}^B$. Der Prozeß A beginnt mit dem Anfangszustand $y_a^A \in \mathcal{Y}^A$ und endet nach einer variablen Anzahl von Episoden im absorbierenden Endzustand $y_e^A \in \mathcal{Y}^A$. Die Anzahl der Episoden wird durch die Zufallsvariable L^A erfaßt. Der Endzeitpunkt der k.ten Episode wird durch die Zufallsvariable T_k^A, ihr Endzustand durch die Zufallsvariable D_k^A mit Werten in $\mathcal{Y}^A$ erfaßt. Entsprechende Bezeichnungen werden für den Teilprozeß B verwendet.

Diese Notation entspricht dem zuvor behandelten Fall, in dem von einer diskreten Zeitachse ausgegangen worden ist. Der wesentliche Unterschied besteht nur darin, daß T_l^A und T_l^B jetzt stetige Zufallsvariablen sind.

Wie im diskreten Fall kann jetzt ein gemeinsamer Prozeß definiert werden, mit möglichen Zuständen im Zustandsraum $\mathcal{Y} = \mathcal{Y}^A \times \mathcal{Y}^B$. Der Anfangszustand ist $y_a = (y_a^A, y_a^B)$; die Menge der Endzustände ist so definiert, wie in (4.17) für den diskreten Fall angegeben worden ist. Eine neue Episode beginnt stets dann, wenn sich der im gemeinsamen Zustandsraum $\mathcal{Y}$ definierte Zustand verändert. Die Zufallsvariable L erfaßt die Anzahl der Episoden im gemeinsamen Prozeß. T_l ist die Endzeit und D_l ist der Endzustand der l.ten Episode des gemeinsamen Prozesses.

Wir betrachten jetzt eine beliebige, die l.te Episode des gemeinsamen Prozesses. Die Vorgeschichte bis zum Beginn dieser Episode sei durch $H_{l-1} = h_{l-1}$ gegeben; sie enthält insbesondere den Anfangszustand d_{l-1} und den Anfangszeitpunkt t_{l-1} für die l.te Episode. Als Bedingung wird $L \geq l$ vorausgesetzt, d.h. es werden nur diejenigen Individuen betrachtet, die mindestens eine l.te Episode haben. Außerdem wird, wie im diskreten Fall, die in Abschnitt 2.4.5 diskutierte Bedingung (2.17) angenommen, d.h. daß der Verlauf einer Episode nicht davon abhängt, wieviele weitere Episoden ihr noch folgen. Um die Möglichkeiten einer Zerlegung des gemeinsamen Prozesses in sich wechselseitig bedingende Teilprozesse in Analogie zu den Aufführungen im diskreten Fall diskutieren zu können, wird zunächst von einer beliebigen Einteilung der bei t_{l-1} beginnenden Zeitachse in gleichlange zeitliche Teilintervalle ausgegangen. Die die Zeitintervalle abgrenzenden Zeitpunkte seien durch

$$\tau_0, \tau_1, \tau_2, \ldots, \tau_n, \ldots$$

gegeben. $[\tau_{j-1}, \tau_j)$ ist das j.te Zeitintervall. Das erste Zeitintervall beginnt bei $\tau_0 = t_{l-1}$.

Jetzt wird folgende Wahrscheinlichkeit betrachtet:

$$\mathrm{P}\left(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid H_{l-1} = h_{l-1}, L \geq l\right) \tag{4.29}$$

Es ist dies die Wahrscheinlichkeit, daß die l.te Episode des gemeinsamen Prozesses im Zeitintervall $[\tau_{n-1}, \tau_n)$ mit einem Übergang in den Zustand $d_l \in \mathcal{Y}$ endet, wobei als Bedingung die Vorgeschichte des Prozesses sowie eine Teilnahme an der l.ten Episode vorausgesetzt wird. Zu überlegen ist, unter welchen Bedingungen diese Wahrscheinlichkeit geeignet zerlegt werden kann.

Zunächst kann sicherlich folgende Zerlegung in zwei bedingte Wahrscheinlicheiten vorgenommen werden:

$$\begin{aligned}\mathrm{P}\left(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid H_{l-1} = h_{l-1}, L \geq l\right) = \\ \mathrm{P}\left(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid T_l \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l\right) \cdot \\ \mathrm{P}\left(T_l \geq \tau_{n-1} \mid H_{l-1} = h_{l-1}, L \geq l\right)\end{aligned} \tag{4.30}$$

Ich betrachte zuerst den zweiten Term auf der rechten Seite, d.h. die Wahrscheinlichkeit, daß der Ausgangszustand der l.ten Episode bis zum Zeitpunkt τ_{n-1} nicht verlassen wird. Dieser Ausdruck kann folgendermaßen als ein Produkt bedingter Wahrscheinlichkeiten geschrieben werden:

$$\begin{aligned}\mathrm{P}\left(T_l \geq \tau_{n-1} \mid H_{l-1} = h_{l-1}, L \geq l\right) = \\ \prod_{j=1}^{n-1} \mathrm{P}\left(T_l \geq \tau_j \mid T_l \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l\right)\end{aligned} \tag{4.31}$$

Jetzt kann wieder auf das in Abschnitt 4.1.3 diskutierte Basisprinzip einer konditionalen Unabhängigkeit der beiden Teilprozesse zurückgegriffen werden. Im Hinblick auf eine stetige Zeitachse kann es folgendermaßen formuliert werden: Wenn die Zeitintervalle $[\tau_{j-1}, \tau_j)$ sehr kurz sind, verlaufen die Teilprozesse A und B während dieser Intervalle unabhängig voneinander, vorausgesetzt daß die gemeinsame Prozeßentwicklung jeweils bis zum Anfangszeitpunkt des Intervalls gegeben ist.[28] Wird dies vorausgesetzt,

[28] Diese Formulierung ist insofern unpräzise, als nicht genau klar wird, was mit einem „sehr kurzen“ Zeitintervall gemeint ist. Es gibt zwei Möglichkeiten, um mit dieser Unschärfe umzugehen. Man kann gewisse Stetigkeitsannahmen voraussetzen, um in der mathematischen Ableitung einen Grenzübergang zu „infinitesimal kleinen“ Zeitintervallen vornehmen zu können; diesen Weg verfolgen Gardner und Griffin [1986]. Man kann andererseits die Annahme einer stetigen Zeitachse als ein Hilfsmittel für die Konstruktion eines stetigen Modells betrachten, mit dem ein empirischer Prozeß approximativ beschrieben werden soll. Für den empirischen Prozeß in einer endlichen Grundgesamtheit von Individuen gibt es dann nur endlich viele Ereigniszeitpunkte, insbesondere eine minimale Zeitdauer für je zwei nicht gleichzeitig stattfindende Ereignisse. Für die folgenden Ausführungen genügt es, sich unter einem „sehr kurzen“ Zeitintervall ein Zeitintervall vorzustellen, das kürzer ist als diese minimale Zeitdauer. Dadurch wird auch die problematische Annahme überflüssig, daß es auf einer stetigen Zeitachse keine (bzw. nur mit Wahrscheinlichkeit Null stattfindenden) simultanen Ereignisse geben kann.

kann jeder der Faktoren auf der rechten Seite von (4.31) folgendermaßen zerlegt werden:

$$\begin{aligned}&\mathrm{P}\,(T_l \geq \tau_j \mid T_l \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \mathrm{P}\,(T^A_{l_A} \geq \tau_j \mid T^A_{l_A} \geq \tau_{j-1}, T^B_{l_B} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l) \cdot \\ &\quad \mathrm{P}\,(T^B_{l_B} \geq \tau_j \mid T^B_{l_B} \geq \tau_{j-1}, T^A_{l_A} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l)\end{aligned} \tag{4.32}$$

l_A ist die Nummer derjenigen Episode des Teilprozesses A, die mit dem Eintritt in den Zustand $d^A_{l_A-1} = d^a_{l-1}$ begonnen hat, d.h. mit dem Anfangszustand der l.ten Episode des gemeinsamen Prozesses im Zustandsraum $\mathcal{Y}^A$. Entsprechend ist l_B für den Teilprozeß B definiert. Diese Definitionen sind möglich, weil die erforderliche Information durch die Vorgeschichte h_{l-1} gegeben ist.

Die Zerlegung (4.32) kann auch durch Übergangsraten ausgedrückt werden. Sei nämlich

$$r^{A,l_A}(t \mid \ldots) = \sum_{d \in \mathcal{Y}^A} r^{A,l_A}_d(t \mid \ldots)$$

die Abgangsrate aus dem Anfangszustand der l_A.ten Episode des Teilprozesses A, erhält man für den ersten Term auf der rechten Seite von (4.32) die Darstellung

$$\begin{aligned}&\mathrm{P}\,(T^A_{l_A} \geq \tau_j \mid T^A_{l_A} \geq \tau_{j-1}, T^B_{l_B} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \exp\left\{-\int_{\tau_{j-1}}^{\tau_j} r^{A,l_A}(\tau \mid T^B_{l_B} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l)\, d\tau\right\}\end{aligned} \tag{4.33}$$

Ganz analog gewinnt man eine Darstellung für den zweiten Term auf der rechten Seite von (4.32). Setzt man nun, analog für die Teilprozesse A und B, (4.33) in (4.32) ein, und dann (4.32) in (4.31), erhält man

$$\begin{aligned}&\mathrm{P}\,(T_l \geq \tau_{n-1} \mid H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \prod_{j=1}^{n-1} \exp\left\{-\int_{\tau_{j-1}}^{\tau_j} r^{A,l_A}(\tau \mid T^B_{l_B} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l)\, d\tau\right\} \cdot \\ &\quad \prod_{j=1}^{n-1} \exp\left\{-\int_{\tau_{j-1}}^{\tau_j} r^{B,l_B}(\tau \mid T^A_{l_A} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l)\, d\tau\right\}\end{aligned} \tag{4.34}$$

Jetzt wird der erste Term auf der rechten Seite von (4.30) betrachtet, also

$$\mathrm{P}\,(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid T_l \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) \tag{4.35}$$

Es ist die Wahrscheinlichkeit, daß die l.te Episode des gemeinsamen Prozesses im Zeitintervall $[\tau_{n-1}, \tau_n)$ mit einem Übergang in den Zielzustand d_l endet. Man kann drei Fälle unterscheiden.

a) Es ändert sich nur der Zustand im Teilprozeß A. Die oben formulierte Annahme bedingter Unabhängigkeit vorausgesetzt, kann mithilfe der bereits eingeführten Indizes l_A und l_B für die korrespondierenden Episoden der beiden Teilprozesse die Gleichung (4.35) folgendermaßen umgeschrieben werden:

$$\begin{aligned}&\mathrm{P}\,(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid T_l \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \mathrm{P}\,(D^A_{l_A} = d^A_{l_A}, \tau_{n-1} \leq T^A_{l_A} < \tau_n, T^B_{l_B} \geq \tau_n \mid T_l \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \mathrm{P}\,(D^A_{l_A} = d^A_{l_A}, \tau_{n-1} \leq T^A_{l_A} < \tau_n \mid T^A_{l_A} \geq \tau_{n-1}, T^B_{l_B} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) \cdot \\ &\quad \mathrm{P}\,(T^B_{l_B} \geq \tau_n \mid T^B_{l_B} \geq \tau_{n-1}, T^A_{l_A} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) \cdot\end{aligned} \tag{4.36}$$

b) Im Intervall $[\tau_{n-1}, \tau_n)$ ändert sich nur der Zustand im Teilprozeß B. Durch Vertauschen der Indizes A und B erhält man dann ebenfalls die Formulierung (4.36).

c) Im Intervall $[\tau_{n-1}, \tau_n)$ ändern sich die Zustände in beiden Teilprozessen. In diesem Fall kann (4.35) folgendermaßen umgeschrieben werden:

$$\begin{aligned}&\mathrm{P}\,(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid T_l \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \mathrm{P}\,(D^A_{l_A} = d^A_{l_A}, \tau_{n-1} \leq T^A_{l_A} < \tau_n, D^B_{l_B} = d^B_{l_B}, \tau_{n-1} \leq T^B_{l_B} < \tau_n \mid \\ &\qquad\qquad T_l \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \mathrm{P}\,(D^A_{l_A} = d^A_{l_A}, \tau_{n-1} \leq T^A_{l_A} < \tau_n \mid T^A_{l_A} \geq \tau_{n-1}, T^B_{l_B} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) \cdot \\ &\quad \mathrm{P}\,(D^B_{l_B} = d^B_{l_B}, \tau_{n-1} \leq T^B_{l_B} < \tau_n \mid T^B_{l_B} \geq \tau_{n-1}, T^A_{l_A} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l)\end{aligned} \tag{4.37}$$

Hierbei wird wiederum die Annahme verwendet, daß die Wahrscheinlichkeiten für das Auftreten von Zustandsänderungen in den Teilprozessen A und B im Zeitintervall $[\tau_{n-1}, \tau_n)$ unabhängig sind, wenn der gemeinsame Prozeß bis zum Zeitpunkt τ_{n-1} gegeben ist.

Die drei Fälle können mithilfe von Zensierungsindikatoren zusammengefaßt werden. Definiert man

$$\delta^A_{l,j} = \begin{cases} 1 & \text{wenn sich } d^a_l\text{, der Zustand im Teilprozeß A,} \\ & \text{im Zeitintervall } [\tau_{j-1}, \tau_j) \text{ der } l\text{.ten Episode des} \\ & \text{gemeinsamen Prozesses ändert} \\ 0 & \text{andernfalls} \end{cases}$$

und entsprechend $\delta^B_{l,j}$ für den Teilprozeß B, gewinnt man aus (4.36) und (4.37) die Darstellung

$$\begin{aligned}&\mathrm{P}\,(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid T_l \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l) = \\ &\quad \mathrm{P}\,(D^A_{l_A} = d^A_{l_A}, \tau_{n-1} \leq T^A_{l_A} < \tau_n \mid T^A_{l_A} \geq \tau_{n-1}, T^B_{l_B} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l)^{\delta^A_{l,n}} \cdot \\ &\quad \mathrm{P}\,(T^A_{l_A} \geq \tau_n \mid T^A_{l_A} \geq \tau_{n-1}, T^B_{l_B} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l)^{1-\delta^A_{l,j}} \cdot \\ &\quad \mathrm{P}\,(D^B_{l_B} = d^B_{l_B}, \tau_{n-1} \leq T^B_{l_B} < \tau_n \mid T^B_{l_B} \geq \tau_{n-1}, T^A_{l_A} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l)^{\delta^B_{l,n}} \cdot \\ &\quad \mathrm{P}\,(T^B_{l_B} \geq \tau_n \mid T^B_{l_B} \geq \tau_{n-1}, T^A_{l_A} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l)^{1-\delta^B_{l,j}}\end{aligned} \tag{4.38}$$

Unter der Annahme einer bedingten Unabhängigkeit der beiden Teilprozesse in kleinen Zeitintervallen läßt sich also eine Zerlegung für die beiden Terme auf der rechten Seite von (4.30) gewinnen. Setzt man (4.34) und (4.38) in (4.30) ein, erhält man schließlich folgende Darstellung für die in (4.29) angegebene Wahrscheinlichkeit, von der wir ausgegangen waren:[29]

$$
\begin{aligned}
&\mathrm{P}\left(D_l = d_l, \tau_{n-1} \leq T_l < \tau_n \mid H_{l-1} = h_{l-1}, L \geq l\right) = \\
&\quad \mathrm{P}\left(D^A_{l_A} = d^A_{l_A}, \tau_{n-1} \leq T^A_{l_A} < \tau_n \mid T^A_{l_A} \geq \tau_{n-1}, T^B_{l_B} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l\right)^{\delta^A_{l,n}} \cdot \\
&\quad \prod_{j=1}^{n-\delta^A_{l,n}} \exp\left\{ - \int_{\tau_{j-1}}^{\tau_j} r^{A,l_A}(\tau \mid T^B_{l_B} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l)\, d\tau \right\} \cdot \\
&\quad \mathrm{P}\left(D^B_{l_B} = d^B_{l_B}, \tau_{n-1} \leq T^B_{l_B} < \tau_n \mid T^B_{l_B} \geq \tau_{n-1}, T^A_{l_A} \geq \tau_{n-1}, H_{l-1} = h_{l-1}, L \geq l\right)^{\delta^B_{l,n}} \cdot \\
&\quad \prod_{j=1}^{n-\delta^B_{l,n}} \exp\left\{ - \int_{\tau_{j-1}}^{\tau_j} r^{B,l_B}(\tau \mid T^A_{l_A} \geq \tau_{j-1}, H_{l-1} = h_{l-1}, L \geq l)\, d\tau \right\}
\end{aligned}
$$

Diese Darstellung gilt für beliebige Einteilungen der stetigen Zeitachse in Intervalle. Die für die beiden Teilprozesse angegebenen konditionalen Ereigniswahrscheinlichkeiten können also unmittelbar durch auf einer stetigen Zeitachse definierte Übergangsraten approximiert werden. Insofern braucht die Frage, unter welchen Stetigkeitsannahmen in der Darstellung (4.39) ein Grenzübergang zu beliebig kleinen Intervallen durchgeführt werden kann, nicht überlegt zu werden. Versteht man die Verwendung eines stetigen Übergangsratenmodells als eine Approximation an einen empirisch erfaßbaren Prozeß, kann von der Frage, ob der reale Prozeß stetig oder diskret verläuft, abgesehen werden.

4.2.3 Zeitabhängige Kovariablen

Die Ausführungen in den Abschnitten 4.2.1 und 4.2.2 haben gezeigt, wie man für zwei parallel ablaufende Prozesse ein Modell finden kann, in dem mögliche Interdependenzen der beiden Teilprozesse sichtbar werden. Ausgangspunkt war das in Abschnitt 4.1.3 diskutierte Prinzip der konditionalen Unabhängigkeit. Hiervon ausgehend kann ein Modell für den gemeinsamen Prozeß in zwei Teilmodelle zerlegt werden.

Wichtig ist, daß die beiden Teilmodelle separat geschätzt werden können. Denn die Wahrscheinlichkeit, daß im gemeinsamen Prozeß ein Ereignis stattfindet, ergibt sich multiplikativ aus den Wahrscheinlichkeiten für ein Ereignis in den beiden Teilprozessen, jeweils konditional auf die

[29] Wenn $\delta^A_{l,n} = 0$ bzw. wenn $\delta^B_{l,n} = 0$, kann der in (4.38) verbleibende Ausdruck dem in (4.34) angegebenen Produkt hinzugefügt werden.

gemeinsame Vorgeschichte des Prozesses. Für jeden Teilprozeß kann infolgedessen ein separates Modell konstruiert werden, bei dem der jeweils andere Prozeß in der Form zeitabhängiger Kovariablen berücksichtigt wird.[30] Es ist bei dieser Betrachtungsweise nicht erforderlich, eine Exogenitätsbedingung, wie sie in Abschnitt 4.1.4 diskutiert worden ist, anzunehmen. Vielmehr wird jeweils die gemeinsame Vorgeschichte als eine exogene Bedingung für den Episodenverlauf in jedem Teilprozeß angesehen.

Die Modellbildung kann unmittelbar an (4.26) für eine diskrete Zeitachse und an (4.39) für eine stetige Zeitachse anknüpfen. Praktisch wird man allerdings in der Regel so vorgehen, daß für jede im A-Prozeß und im B-Prozeß mögliche Episode ein separates Modell spezifiziert und geschätzt wird. Um Einsichten in den parallel ablaufenden Prozeß zu gewinnen, ist dies in den meisten Fällen ausreichend. Eine simultane Betrachtung aller möglichen Teilmodelle – für jede Episode innerhalb jedes Teilprozesses – ist nur erforderlich, wenn die Modellbildung, zum Beispiel im Rahmen einer Simulationsstudie, verwendet werden soll, um einen datengenerierenden Prozeß für den gemeinsamen Prozeßverlauf zu gewinnen.

Beispiel: Nicht-eheliche Lebensgemeinschaften

Als Beispiel betrachte ich einige Informationen, die das SOEP über die Entwicklung von nicht-ehelichen Lebensgemeinschaften liefert.[31] Die Fragestellung richtet sich auf die Dauer solcher Lebensgemeinschaften und ihren möglichen Übergang in eine Ehe. Es wird also ein Bedingungszusammenhang von zwei Ereignissen betrachtet: $E1$ ist der Beginn einer nicht-ehelichen Lebensgemeinschaft, E_2 ist eine möglicherweise dann stattfindende Heirat.

Die Datengrundlage bezieht sich auf alle Personen aus der Teilstichprobe A des SOEP, die an der ersten Welle (1984) teilgenommen haben. Die

[30] Tuma und Hannan [1984, S. 268] haben im Hinblick auf parallele Prozesse eine Unterscheidung von zwei Arten der Interdependenz eingeführt. Einerseits *Interdependenz auf der Ebene der Zustände*; damit ist gemeint, daß die Übergangsrate im Teilprozeß A davon abhängig ist, welcher Zustand im Teilprozeß B eingenommen wird; analog für die Abhängigkeit des Teilprozesses B vom Teilprozeß A. Andererseits *Interdependenz auf der Ebene der Übergangsraten*, womit gemeint ist, daß die Übergangsrate im Teilprozeß A (auch) von der Rate abhängig ist, mit der sich die Zustände im Teilprozeß B verändern; analog für die Abhängigkeit des Teilprozesses B vom Teilprozeß A. Es ist jedoch bemerkenswert, daß diese Unterscheidung in der in den Abschnitten 4.2.1 und 4.2.2 dargestellten Zerlegung eines gemeinsamen Prozesses in zwei Teilprozesse nicht auftritt. Die Wahrscheinlichkeit für eine Zustandsänderung im Teilprozeß A (bzw. B) ist *auf beliebige Weise* abhängig von der gesamten Vorgeschichte des *gemeinsamen* Prozesses. Damit kann nicht nur eine mögliche Abhängigkeit von den jeweils unmittelbar zuvor eingenommenen Zuständen erfaßt werden, sondern gleichermaßen eine mögliche Abhängigkeit von den Übergangsraten im Teilprozeß B (bzw. A), soweit sie durch die jeweilige Vorgeschichte gegeben sind. Vgl. auch Pötter [1993], der zu einer entsprechenden Schlußfolgerung kommt.

[31] Das Beispiel sowie die wesentlichen Ideen zur Modellierung zeitabhängiger Kovariablen entstammen der Kooperation mit H.-P. Blossfeld, vgl. Blossfeld et al. [1994].

Personen werden für maximal 6 Jahre beobachtet; sobald sie aus dem SOEP ausscheiden, wird ihre Beobachtung als rechts zensiert angenommen.[32] Zeitpunkt für Ereignisse können auf einer in Monaten definierten Zeitachse erfaßt werden. Ich verwende eine Prozeßzeitachse, die mit dem Eintritt in eine nicht-eheliche Lebensgemeinschaft beginnt.

Insgesamt können in diesem Beobachtungsfenster 431 Lebensgemeinschaften beobachtet werden. Davon enden 131 mit einer Heirat, 37 Episoden enden mit einer Trennung, und 300 Episoden sind rechts zensiert. Abbildung

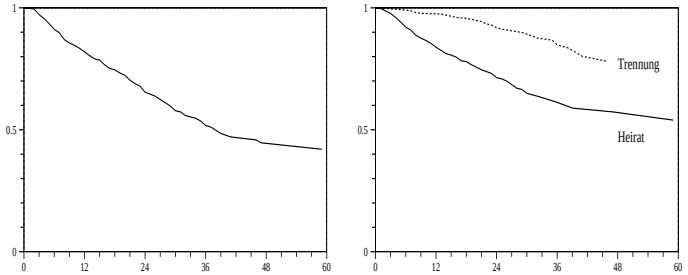


Abbildung 4.2.1 Mit dem Kaplan-Meier-Verfahren geschätzte Survivorfunktionen für die Dauer nicht-ehelicher Lebensgemeinschaften, insgesamt (links) und differenziert nach Übergängen in Heirat oder Trennung (rechts). Datengrundlage: 431 Personen aus der Teilstichprobe A des SOEP. Abszisse: Episodendauer in Monaten.

Im folgenden beschränke ich mich auf das mögliche Ereignis *Heirat*. Eine Analyse der Daten zeigt, daß es mindestens zwei wichtige Bedingungen gibt. Erstens das Alter, in dem die Lebensgemeinschaft begonnen wird; zweitens ob während der Lebensgemeinschaft eine Schwangerschaft eintritt bzw. ein Kind geboren wird.[33] Das Alter zum Episodenbeginn kann als ein Aspekt der Vorgeschichte der Episode angesehen werden, das Ereignis *Schwangerschaft* kann jedoch zu irgendeinem Zeitpunkt während des Episodenverlaufs eintreten, es sollte also als ein den Ereigniszusammenhang von E_1 und E_2 vermittelndes Ereignis E_3 betrachtet werden. Grundsätzlich ist davon auszugehen, daß das Ereignis E_3 sowohl durch das Ereignis E_1 bedingt wird als auch selbst eine Bedingung für das Ereignis E_2 sein kann. Es liegt also eine Situation paralleler Prozesse vor. Der A-Prozeß beschreibt die Entwicklung der Lebensgemeinschaft, der B-Prozeß beschreibt Schwangerschaften und Geburten von Kindern.

Nach der in den Abschnitten 4.2.1 und 4.2.2 gegebenen Analyse kann

[32] Der Datensatz wurde bereits von Blossfeld et al. [1993] analysiert.

[33] Da alle Lebensgemeinschaften erst während des von 1984 bis 1989 reichenden Beobachtungsfensters beginnen, ist eine zusätzliche Differenzierung nach Geburtskohorten nicht erforderlich.

ein Modell für diesen parallelen Prozeß in zwei Teilmodelle zerlegt werden. Das erste Modell beschreibt, wie der Zusammenhang der Ereignisse E_1 und E_2 durch das Ereignis E_3 vermittelt wird, das zweite Modell beschreibt, wie das Ereignis E_3 durch die Ereignisse E_1 und E_2 bedingt wird. Ich konzentriere mich im folgenden auf die erste Fragestellung.

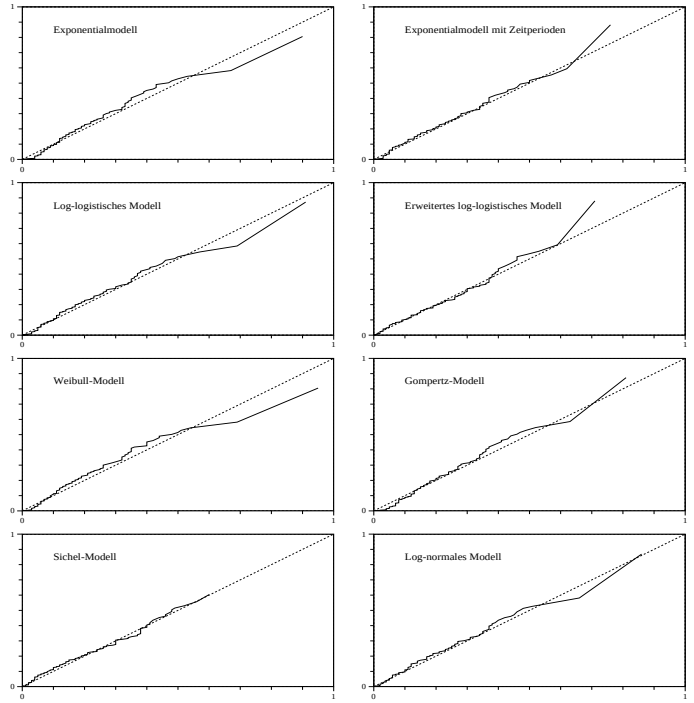


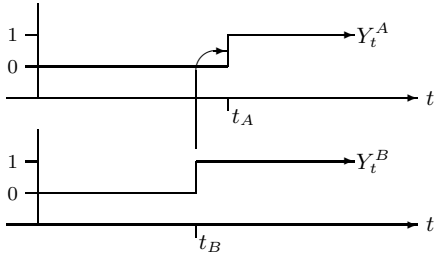
Abbildung 4.2.2 Darstellung verallgemeinerter Residuen für einige Standardmodelle zur Beschreibung der Dauer von nicht-ehelichen Lebensgemeinschaften für 431 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1984 – 1989 eine nicht-eheliche Lebensgemeinschaft begonnen haben. Abszisse: Residuen, Ordinate: Minus Logarithmus der Survivorfunktion der Residuen. Schätzung mit dem Kaplan-Meier-Verfahren.

Zunächst muß eine geeignete Modellklasse gefunden werden, um den Zusammenhang der Ereignisse $E_1 \rightarrow E_2$ zu beschreiben. Grundsätzlich steht eine große Anzahl unterschiedlicher Modelle zur Verfügung. Um einen Hinweis auf ein angemessenes Modell zu finden, betrachte ich für eine Reihe unterschiedlicher Modelle verallgemeinerte Residuen (vgl. Abschnitt 3.7). Abbildung 4.2.2 zeigt, daß das Sichel-Modell am besten erscheint. Es wird deshalb im folgenden verwendet.

Um ein konkretes Modell zu spezifizieren, muß überlegt werden, wie die Abhängigkeit von den möglichen Bedingungen zum Ausdruck gebracht

werden soll. Für das Alter beim Episodenbeginn ist dies einfach, denn dieser Sachverhalt ist zum Episodenbeginn gegeben und kann sich während des Episodenverlaufs nicht verändert. Das Alter kann also wie eine einfache Klassifizierungsvariable behandelt werden. Da jedoch die Abhängigkeit der Übergangsrate nicht linear sein kann, ist es zweckmäßig, sowohl das Alter als auch das Quadrat des Alters zu berücksichtigen.

Schwieriger ist es, eine geeignete Spezifikation für eine mögliche Abhängigkeit des Episodenverlaufs vom Ereignis *Schwangerschaft* zu finden.[34] Folgende Abbildung illustriert das Problem.

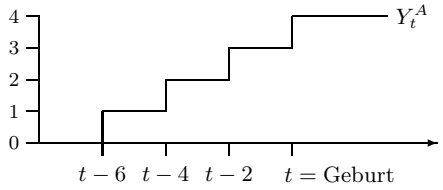


t_A ist der Zeitpunkt für den (möglichen) Übergang in eine Heirat, t_B ist der Zeitpunkt für eine Schwangerschaft. Die Idee ist, daß die Schwangerschaft eine mögliche Bedingung für die Heirat ist. Es stellen sich jedoch zwei Fragen. Ob und ggf. in welcher Art eine zeitliche Verzögerung zwischen den beiden Ereignissen angenommen werden soll. Zweitens stellt sich die Frage, ob es überhaupt angemessen ist, eine Schwangerschaft als ein Ereignis zu betrachten, das auf der zugrundeliegenden Zeitachse zu einem gewissen Zeitpunkt stattfindet. Da die Zeitachse als eine Folge von Monaten bestimmt wurde, kann man die Schwangerschaft selbst als eine Episode ansehen, die während des B-Prozesses stattfindet. Wie mit dieser zweiten Frage umgegangen werden sollte, kann nicht generell entschieden werden, sondern hängt vom jeweils zu betrachtenden Ereigniszusammenhang ab. Zwar könnte man es sich für das vorliegende Beispiel leicht machen und, statt der Schwangerschaft, die Geburt eines Kindes als das maßgebliche Ereignis ansehen. Aber wie sich zeigen wird, würde dies eine wesentliche Einsicht in die Form des Vermittlungszusammenhangs verschleiern.

Eine einfache Möglichkeit, um den möglichen Einfluß der Schwangerschaft, die selbst eine Episode ist, auf das Heiratsverhalten sichtbar zu machen, besteht darin, den B-Prozeß durch eine Serie von Variablen zu repräsentieren; zum Beispiel für jeden Monat der Schwangerschaft eine eigene Variable. Da in unserem Datensatz nur verhältnismäßig wenige Schwangerschaften auftreten, verwenden wir nur vier Variablen, die jeweils als 0/1-Variablen definiert werden. Folgende Abbildung veranschaulicht, wie

[34] Die im folgenden verwendeten Ideen entstammen einer gemeinsamen Arbeit mit H.-P. Blossfeld und E. Klijzing, vgl. Blossfeld et al. [1994].

diese Variablen gebildet werden.



Wenn zum Zeitpunkt t eine Geburt stattfindet, nimmt die erste Variable den Wert 1 während des 6. bis 4. Monats vor der Geburt an, die zweite Variable während des 4. bis 2. Monats vor der Geburt, die dritte Variable in den zwei Monaten vor der Geburt, und die vierte Variable nimmt zum Zeitpunkt der Geburt den Wert 1 an.

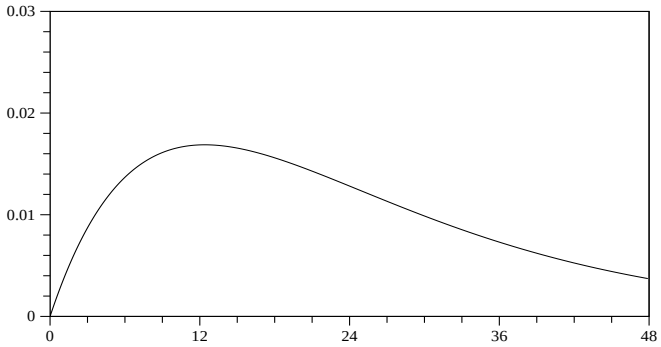


Abbildung 4.2.3 Übergangsrate für die Dauer nicht-ehelicher Lebensgemeinschaften bis zum Ereignis *Heirat*, geschätzt mit einem Sichel-Modell für 431 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1984 – 1989 eine nicht-eheliche Lebensgemeinschaft begonnen haben. Die Übergangsrate bezieht sich auf Personen, die zum Beginn der Lebensgemeinschaft 25 Jahre alt sind und auf Episoden, bei denen keine Schwangerschaft eintritt. Abszisse in Monaten seit Beginn der Lebensgemeinschaft.

Mit dieser Spezifikation von Kovariablen kann dann ein Sichel-Modell geschätzt werden. Abbildung 4.2.3 zeigt den Verlauf der geschätzten Übergangsrate für Personen, die zum Beginn der Lebensgemeinschaft 25 Jahre alt sind. Außerdem wird in dieser Abbildung angenommen, daß während der Lebensgemeinschaft keine Schwangerschaft stattfindet.

Abbildung 4.2.4 illustriert, wie eine Schwangerschaft den Übergang der Lebensgemeinschaft in eine nachfolgende Ehe beeinflußt. Für diese Abbildung wurde angenommen, daß 14 Monate nach dem Beginn der Lebensgemeinschaft eine Geburt stattfindet. Man erkennt, daß in der Regel bereits während der Schwangerschaft eine Heirat stattfindet; ein Sachverhalt, der

nicht erkannt werden könnte, wenn nur der Zeitpunkt der Geburt als eine zeitabhängige Kovariable verwendet worden wäre.

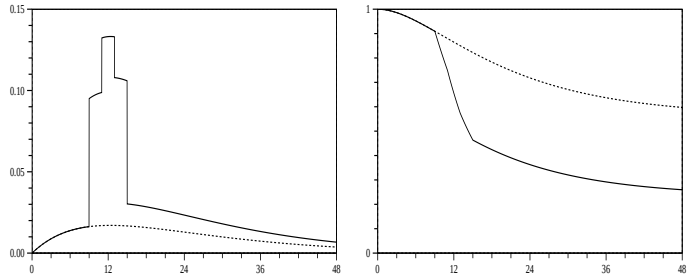


Abbildung 4.2.4 Übergangsraten (links) und Survivorfunktionen (rechts) für die Dauer nicht-ehelicher Lebensgemeinschaften bis zum Ereignis *Heirat*, geschätzt mit einem Sichel-Modell für 431 Personen aus der Teilstichprobe A des SOEP, die im Zeitraum 1984 – 1989 eine nicht-eheliche Lebensgemeinschaft begonnen haben. Übergangsraten und Survivorfunktionen beziehen sich auf Personen, die zum Beginn der Lebensgemeinschaft 25 Jahre alt sind. Gestrichelt: keine Schwangerschaft während der Lebensgemeinschaft, durchgezogen: 14 Monate nach dem Beginn der Lebensgemeinschaft wird ein Kind geboren. Abszisse in Monaten seit Beginn der Lebensgemeinschaft.

Das Beispiel illustriert, wie zeitabhängige Kovariablen dazu verwendet werden können, um sichtbar zu machen, wie ein Prozeß (das Heiratsverhalten) von einem anderen Prozeß (Schwangerschaften bzw. Geburt von Kindern) bedingt wird. Wichtig erscheinen vor allem zwei Aspekte. Erstens kann diese Art der Modellbildung vorgenommen werden, obwohl angenommen werden muß, daß sich beide Prozesse wechselseitig bedingen. Bei der Frage, wie ein A-Prozeß durch einen B-Prozeß bedingt wird, kann sinnvoll davon abstrahiert werden, wie der B-Prozeß selbst zustande gekommen ist. Dies schließt natürlich nicht aus, daß auch für die Bedingtheit des B-Prozesses ein Modell konstruiert werden kann. Zweitens zeigt das Beispiel, wie das statistische Konzept zeitabhängiger Kovariablen nicht nur verwendet werden kann, um zeitpunktbezogene Ereignisse zu erfassen, sondern auch dazu dienen kann, die soziologische Vorstellung, daß Verhaltensweisen durch soziale Situationen bedingt werden, ereignisanalytisch zu rekonstruieren und empirisch zugänglich zu machen. Die Idee ist, eine soziale Situation selbst als einen Prozeß zu repräsentieren, so daß sichtbar wird, *wie* das individuelle Verhalten durch soziale Situationen bedingt wird.

Kapitel 5

Zufallsstichproben und statistische Inferenz

In dieser Arbeit wird von einem deskriptiven Verständnis statistischer Modelle für die Beschreibung von Lebensverläufen ausgegangen. Sie beziehen sich auf die empirische Verteilung gewisser Zufallsvariablen, die in einem deskriptiven Wahrscheinlichkeitsraum für eine endliche Grundgesamtheit von Individuen definiert sind. Typischerweise wird eine *vereinfachte* Darstellung der empirischen Verteilung angestrebt, so daß das Modell als ein Hilfsmittel für theoretische Deutungen des zugrundeliegenden Sachverhalts dienen kann. Bei der Diskussion solcher Modelle in den Kapiteln 3 und 4 sind wir davon ausgegangen, daß sich das Schätzproblem in zwei Teilprobleme zerlegen läßt. Erstens in die Frage, wie bei der Modellschätzung berücksichtigt werden kann, daß in der Regel nur unvollständige und ungenaue Daten verfügbar sind. Zweitens in die Frage, wie berücksichtigt werden kann, daß in der Regel nur Informationen aus einer Stichprobe aus der eigentlich interessierenden Grundgesamtheit verfügbar sind. Einige Aspekte dieses zweiten Problems, von dem bisher vollständig abstrahiert wurde, sollen in diesem Kapitel erörtert werden. Das Problem ist leider kompliziert und dementsprechend umstritten.[1] Eine kurze Erörterung ist jedoch im Rahmen der vorliegenden Arbeit hauptsächlich aus zwei Gründen sinnvoll. Erstens weicht die in dieser Arbeit vertretene Auffassung statistischer Modelle in einige Aspekten von der üblichen Lehrbuchkonzeption ab. Es ist deshalb sinnvoll, auch einige der Implikationen zu erörtern, die sich daraus für das Problem der Modellschätzung mit Stichprobendaten ergeben. Zweitens spielt das Inferenzproblem in der praktischen Anwendung statistischer Methoden eine zentrale, jedoch in seiner Bedeutung of nur mangelhaft wahrgenommene Bedeutung. Auch eine kurze und zweifellos unzureichende Diskussion dieses Problems kann infolgedessen dazu beitragen, ein besseres Problembewußtsein zu gewinnen.

5.1 Bemerkungen zur Problemstellung

Ausgangspunkt ist ein deskriptiver Wahrscheinlichkeitsraum für eine endliche Grundgesamtheit $\Omega = \{\omega_1, \ldots, \omega_N\}$. In diesem Rahmen wird eine

[1] Eine einführende Diskussion der verschiedenen Standpunkte zum statistischen Inferenzproblem geben Gigerenzer et al. [1989, Kap. 3].

(ggf. mehrdimensionale) Zufallsvariable

$$Z : \Omega \longrightarrow \mathcal{Z} \tag{5.1}$$

betrachtet. Es wird angenommen, daß es sich um eine diskrete Zufallsvariable mit einem endlichen Wertebereich handelt.

Die übliche Problemformulierung besteht darin, daß mithilfe einer Stichprobe aus der Grundgesamtheit Ω Informationen über die Verteilung der Zufallsvariable Z gewonnen werden sollen. Um diese Aufgabe zugänglich zu machen, wird für die Verteilung von Z – oder für gewisse Aspekte ihrer Verteilung – ein parametrisches Modell angenommen, dessen unbekannte Parameter mithilfe der Daten aus einer Stichprobe geschätzt werden können. Diese Problemformulierung erscheint zum Beispiel angemessen, wenn unmittelbar gewisse Merkmale der Verteilung von Z, etwa ihr Mittelwert, ihre Varianz oder ihre Quantile, geschätzt werden sollen; sie ist jedoch der Aufgabe, deskriptive statistische Modelle zu schätzen, nicht vollständig angemessen. Folgende Abbildung kann zur Verdeutlichung dienen.

Links oben befindet sich die Grundgesamtheit und die in ihr realisierte Verteilung $F(z)$ der Zufallsvariable Z. Rechts oben befindet sich eine Klasse möglicher Modelle $\tilde{F}(z;\theta)$ $(\theta \in \Theta)$. Wie bereits diskutiert worden ist, kann ein optimales Modell $\tilde{F}(z;\theta^\circ)$ definiert werden, wenn man ein geeignetes Distanzmaß voraussetzt und wenn die Verteilung von Z bekannt ist. Wichtig ist jedoch, daß die Entscheidung des Modellkonstrukteurs für eine bestimmte Modellklasse nicht ohne weiteres als eine statistische Hypothese über die Verteilung von Z betrachtet werden kann. Es kann *nicht* angenommen werden, daß es ein $\theta' \in \Theta$ gibt, so daß $F(z) = \tilde{F}(z;\theta')$ ist; insbesondere kann nicht angenommen werden, daß $F(z) = \tilde{F}(z;\theta^\circ)$. Denn diese Annahme würde der Zielsetzung für die Modellbildung widersprechen: *vereinfachende* Beschreibungen für die Verteilung von Z zu gewinnen.

Als erste Frage stellt sich infolgedessen, wie mit dieser besonderen Situation bei der Schätzung statistischer *Modelle* sinnvoll umgegangen werden kann. Mehrere unterschiedliche Betrachtungsweisen sind möglich.

a) Eine erste Möglichkeit besteht darin, sich auf die in Form einer Stichprobe gegebenen Daten zu beschränken und die Frage, ob die anhand der gegebenen Daten gewonnenen Einsichten für eine Grundgesamtheit verallgemeinerbar sind, offen zu lassen. Praktisch bedeutet dies, daß bei

der Modellschätzung die Stichprobe so behandelt wird, als ob es sich um eine Grundgesamtheit handelt. Als Ergebnis der Modellbildung erhält man dann eine vereinfachende Beschreibung der in der Stichprobe realisierten empirischen Verteilung der Zufallsvariable Z.

Diese Vorgehensweise kann keineswegs von vornherein verworfen werden. Sie entspricht nicht nur einer in der empirischen Sozialforschung vielfach üblichen Praxis der Modellbildung, sie kann auch durch Argumente gestützt werden. Denn mit den Daten einer Stichprobe können keine sicheren Aussagen über die Verfassung einer Grundgesamtheit getroffen werden. Jede Form einer induktiven Verallgemeinerung der mit den Daten eine Stichprobe gewonnenen Aussagen muß an subjektive, bestenfalls partiell rationalisierbare Vorstellungen appellieren. Zwar ist die soziologische Forschung nicht an Aussagen über „zufällig“ zustande gekommene Mengen von Individuen interessiert, sondern an Aussagen über gesellschaftliche Verhältnisse. Dieses Interesse begründet die soziologische Beschäftigung mit Daten, und auch wenn nur Daten aus einer Stichprobe zur Verfügung stehen, besteht stets ein – ggf. impliziter – Anspruch, daß den gewonnenen Ergebnissen eine allgemeinere, d.h. für bestimmte gesellschaftliche Verhältnisse verallgemeinerbare Bedeutung zukommt. Die Frage ist jedoch nicht, ob auf diesen Anspruch verzichtet wird, sondern wie er vertreten wird. Es erscheint durchaus möglich, den Standpunkt zu vertreten, daß es in der empirischen Sozialforschung nicht auf „statistische Signifikanz“ ankomme, vielmehr auf die „substantielle Bedeutung“ der erzielten Ergebnisse.[2] Es ist zwar schwer, das hiermit Gemeinte zu präzisieren (darauf wird weiter unten näher eingegangen); ein wichtiger Aspekt dieser Position liegt aber sicherlich darin, daß sie das Vertrauen in die induktiv gewonnenen Ansichten in erster Linie von der Masse und Konsistenz der verfügbaren, jeweils partiellen Informationen über die Verfassung gesellschaftlicher Verhältnisse abhängig machen möchte, im wesentlichen unabhängig von der „statistischen Signifikanz“ jeder einzelnen Teilinformation.

b) Obwohl eine Reihe von Argumenten für die unter (a) skizzierte Haltung spricht, schließt sie natürlich eine Ergänzung durch inferenzstatistische Überlegungen nicht aus. Da es in jedem Fall darum geht, Einsichten in die Grundgesamtheit zu gewinnen, sollte auch die Frage gestellt werden, welcher Art die Vermutungen sind, die mit den Daten einer Stichprobe über die intendierte Grundgesamtheit gewonnen werden können, und welche Rationalisierungschancen es für die Bildung solcher Vermutungen gibt. Es muß dann allerdings präzisiert werden, an welcher Stelle im Prozeß der Modellbildung das inferenzstatistische Problem auftritt.

Eine Möglichkeit besteht darin, das Problem gewissermaßen in zwei Stufen zu zerlegen. In einem ersten Schritt wird die aus der Stichprobe

[2] Vgl. zum Beispiel Reynolds [1969]. Eine breite Dokumentation der im Rahmen der Soziologie (und Psychologie) geführten „Signifikanz-Test-Kontroverse“ findet sich bei Morrison und Henkel [1970].

verfügbare Information genutzt, um zu einer Schätzung der Verteilung von Z in der Grundgesamtheit zu gelangen. In einem zweiten Schritt wird dann die Modellbildung auf die für die Grundgesamtheit geschätzte Verteilung von Z bezogen. Diese beiden Stufen entsprechen nicht der praktischen Vorgehensweise bei der Modellkonstruktion, die zunächst immer auf der Grundlage der verfügbaren Daten durchgeführt wird.[3] Die Zerlegung in zwei Stufen dient vielmehr einer nachträglichen Rationalisierung der inferenzstatistischen Aspekte der Modellkonstruktion. Entscheidend ist, daß diese Betrachtungsweise es nicht erfordert, das zu konstruierende Modell als eine statistische Hypothese über die Verteilung von Zufallsvariablen in einer Grundgesamtheit anzusehen; die deskriptive Aufgabe der Modellbildung wird gewissermaßen von der inferenzstatistischen Problematik losgelöst.

Ich nenne diese Betrachtungsweise *deskriptive Modellschätzung*. Eine in vieler Hinsicht ähnliche Betrachtungsweise wird in der englischsprachigen Literatur *design-based model estimation* genannt,[4] allerdings wird sie fast immer mit einer spezifischen Konzeption für die Interpretation des statistischen Inferenzproblems verbunden. Die Bezeichnung *deskriptive Modellschätzung* ist demgegenüber neutral; sie impliziert nur, daß der Sinn der Modellbildung in einer approximativen Beschreibung von Merkmalen bzw. Merkmalsrelationen in einer endlichen Grundgesamtheit von Individuen gesehen wird und daß infolgedessen das inferenzstatistische *Problem* darin besteht, wie man mithilfe einer Stichprobe Aufschluß über die Merkmalsverteilungen in einer Grundgesamtheit gewinnen kann.[5]

Eine deskriptive Modellschätzung läßt sich unmittelbar als eine Ergänzung der unter (a) beschriebenen Betrachtungsweise ansehen. Denn die beste Schätzung für eine Verteilungsfunktion in der Grundgesamtheit liefert ihre durch die Stichprobe gegebene empirische Verteilungsfunktion. Bei der Modellschätzung erhält man infolgedessen die gleichen Ergebnisse.

[3] Dagegen wird manchmal eingewendet, daß statistische Signifikanztests nur bei solchen Hypothesen sinnvoll angewendet werden können, die unabhängig von den zum Test zu verwendenden Daten gebildet worden sind. Auf der Grundlage des üblichen Selbstverständnisses statistischer Signifikanztests ist dieser Einwand berechtigt; er zeigt jedoch nur, daß die Vorstellung, daß Hypothesen getestet werden, der Praxis der empirischen Sozialforschung nicht gerecht wird.

[4] Vgl. u.a. Kalton [1983], Hansen et al. [1983] sowie die im Anschluß an diesen Beitrag [ebda.] geführte Diskussion.

[5] In der Literatur wird gelegentlich zwischen *deskriptiven* und *analytischen* Modellen unterschieden. Die Unterscheidung ist jedoch uneinheitlich. Einige Autoren scheinen sich auf die Art des Modells zu beziehen und bezeichnen bereits jedes Modell für eine bedingte Wahrscheinlichkeitsverteilung als ein analytisches Modell. Andere Autoren scheinen analytische Modelle mit Superpopulationsmodellen zu identifizieren; zum Beispiel sagt Royall [1976, S. 463]: „A useful criterion for determining whether a particular study is descriptive or analytic is based on the standard error assigned to an estimate. If the standard error would be zero if all the population units where observed without error, then the study is descriptive.“ In diesem Sinne wird auch hier die Formulierung *deskriptive Modellschätzung* verwendet.

Der Vorteil liegt jedoch darin, daß man einen theoretischen Rahmen gewinnt, um die statistisch rationalisierbaren Aspekte des Induktionsproblems zu erörtern. Insbesondere kann die Frage überlegt werden, ob und ggf. wie die Art und Weise, wie die Stichprobe zustande gekommen ist, bei der Modellschätzung berücksichtigt werden sollte.

c) Eine Alternative zur deskriptiven Modellschätzung liegt darin, die Interpretation des deskriptiven Anspruchs der Modellbildung $\tilde{F}(z;\theta)$ für die Verteilung von Z in der Grundgesamtheit zu verändern. Bei der Definition dieses Anspruchs sind wir bisher davon ausgegangen, daß die in der Grundgesamtheit realisierten Werte der Zufallsvariable Z den durch die Realität fest vorgegebenen Gegenstand der Beschreibung bilden. Eine alternative Betrachtungsweise ergibt sich dadurch, daß man sie als Realisierungen einer durch Modellbildung konstruierbaren Zufallsvariable $\tilde{Z}$ ansieht. Die Grundgesamtheit und mithin jede aus ihr gebildete Stichprobe besteht dann aus Realisierungen dieser Zufallsvariable $\tilde{Z}$; und $\tilde{F}(z;\theta)$ kann – in der üblichen Weise – als eine parametrische Klasse von Hypothesen über die Verteilung von $\tilde{Z}$ betrachtet werden.

Werden statistische Modelle auf diese Weise interpretiert, wird in der Literatur häufig von *Superpopulationsmodellen* gesprochen. Eine Schwierigkeit liegt offenbar darin, der Vorstellung, daß die in der Grundgesamtheit realisierten Merkmale der Individuen „zufällig“ entstehen, eine klare Bedeutung zu geben. Denn es ist nicht ohne weiteres verständlich, welche Bedeutung Wahrscheinlichkeitsaussagen über $\tilde{Z}$ haben können.[6] Dieses Problem hängt eng mit der Frage zusammen, welches Erkenntnisinteresse mit der Modellkonstruktion verfolgt wird. In dieser Arbeit wird von einem deskriptiven Verständnis ausgegangen: statistische Modelle sollen einer vereinfachenden, gleichwohl angemessenen und informativen Beschreibung von Merkmalen und ihrer Zusammenhänge bei den Individuen in einer Grundgesamtheit dienen. Dieses Verständnis der Modellbildung schließt es nicht aus, die Frage nach der Angemessenheit eines Modells zu stellen und bei ihrer Reflexion sowohl statistische (auf den *Modellfit* bezogene) als auch, vor allem, theoretische Erwägungen zu berücksichtigen. Das Verhältnis zwischen Realität (der in einer Grundgesamtheit realisierten Merkmalsverteilungen) und Modell ist jedoch konstruktiv, und es macht bei dieser Betrachtungsweise keinen Sinn, von *wahren* Modellen zu sprechen oder dies auch nur hypothetisch zu unterstellen.

Mit Superpopulationsmodellen wird jedoch ein weitergehender Anspruch verbunden. Sie implizieren die Behauptung einer spezifischen Relation zwischen Modell und Realität: daß die Realität durch einen „zufälligen“ (nicht „informativen“) Prozeß aus einer durch die Modellbildung er-

[6] Man kann natürlich das Modell verwenden, um einen Zufallsgenerator zu konstruieren, der dem Modell entsprechende Daten erzeugt. Aber der dadurch installierte „datengenerierende Prozeß“ hat mit der Frage, wie die Daten bzw. die durch die Daten erfaßten Sachverhalte in der Realität zustandegekommen sind, zunächst nichts zu tun.

faßbaren Menge möglicher Realitäten entstanden ist. Abgesehen von der spekulativen Komponente dieses Gedankengangs entsteht die Frage, ob und ggf. wie geprüft werden kann, daß die stets vorhandene Differenz zwischen Modell und Realität nicht informativ ist und *deshalb* als „zufällig“ interpretiert werden kann. Denn nur ein Teil dieses Problems kann mit statistischen Hilfsmitteln (zum Beispiel durch eine statistische Analyse von Residuen) reflektiert werden. Der wesentliche Teil des Problem resultiert vielmehr daraus, daß es keine objektiven Kriterien für die Abstraktionen gibt, auf denen jede Modellbildung beruht. Dies wiederum ist nicht in erster Linie eine Frage der Verfügbarkeit von Daten, sondern der Modellkonstrukteur muß entscheiden, welche Merkmale der Individuen in seinem Modell eine (potentielle) Bedeutung haben sollen. Modellbildung beruht insofern stets auf einer Abstraktion. Diejenigen Aspekte des zu modellierenden Sachverhalts, von denen bei der Modellbildung abgesehen wird, sind per Definition nicht-informativ. Natürlich kann in jedem Einzelfall die Modellbildung infrage gestellt werden, insbesondere durch eine Kritik der jeweils vorgenommenen Abstraktionen. Bezieht sich die Modellbildung auf gewisse Aspekte einer (mehrdimensionalen) Zufallsvariable Z, könnte zum Beispiel eingewendet werden, daß die Modellbildung zusätzlich eine Zufallsvariable Z' zu berücksichtigen habe. Aber die bloße Tatsache, daß Z' mit Z korreliert ist (wovon in der Realität fast immer ausgegangen werden kann), liefert für sich genommen noch keine hinreichende Begründung dafür, daß Z' in die Modellbildung einbezogen werden muß, obwohl in diesem Fall Z' für die Verteilung von Z „informativ“ und insofern nicht „zufällig“ ist.

In den folgenden Überlegungen gehe ich von der Betrachtungsweise einer deskriptiven Modellschätzung aus; Superpopulationsmodelle werden nur gelegentlich erwähnt. Die Überlegungen betreffen hauptsächlich zwei Fragen:

1. In welcher Weise können inferenzstatistische Überlegungen helfen, um den Erkenntnisanspruch von Modellen, die auf der Grundlage von Stichproben gewonnen worden sind, einschätzbar zu machen?

2. Ist es wichtig, wie die Stichprobe, auf die die Modellbildung gegründet wird, zustande gekommen ist? Ergeben sich daraus Konsequenzen dafür, wie statistische Modelle zu konstruieren bzw. zu schätzen sind?

Um es gleich vorweg zu sagen: Ich kann keine dieser beiden Fragen befriedigend beantworten; schon deshalb nicht, weil sie in der statistischen Literatur immer noch sehr kontrovers diskutiert werden.[7] Ich glaube jedoch, daß es im Kontext der vorliegenden Arbeit erforderlich ist, diese Fragestellungen zu erläutern und damit zugleich auf die Grenzen auch aller übrigen Ausführungen hinzuweisen.[8]

[7] In der üblichen Lehrbuchliteratur wird dieser Sachverhalt selten sichtbar gemacht. Vgl. dazu die kritischen Anmerkungen von Gigerenzer et al. [1989, insb. S 106ff].

[8] Eine Bemerkung Fishers [1953, S. 1] kann als zusätzliche Entschuldigung dienen: „The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation.“

5.2 Zufallsstichproben

Es ist intuitiv einleuchtend, daß die Frage, wie eine Menge von Daten zustande gekommen ist, wichtig ist, um die mithilfe der Daten möglichen Schlußfolgerungen einschätzbar zu machen. Aus einer statistischen Perspektive lassen sich im wesentliche zwei Formen der Datengewinnung unterscheiden: intentional vorgenommene Datenauswahl und Zufallsstichproben. Welche der beiden Formen der Datengewinnung besser ist, kann nicht generell entschieden werden, sondern hängt vom Anwendungsfall ab.[9]

Für die Gewinnung von Individualdaten für die empirische Sozialforschung hat sich, etwa seit den 30er Jahren, das Verfahren der Zufallsstichproben weitgehend durchgesetzt.[10] Zur Begründung werden hauptsächlich zwei Gründe angeführt, in einer Formulierung von T. M. F. Smith [1983, S. 394]: „The arguments for randomization are twofold. The first, and most important for science, is that randomization eliminates personal choice and hence eliminates the possibility of subjective selection bias. The second is that the randomization distribution provides a basis for statistical inference." Smith fügt hinzu: „The question for scientists is whether such statistical inferences are relevant for scientific inference."

Zunächst beschränke ich mich auf das erste Argument. Seine Bedeutung ist unmittelbar plausibel. Sucht man sich gezielt eine Reihe von Untersuchungseinheiten aus, besteht stets die Gefahr, daß die verwendeten (oft impliziten) Auswahlkriterien zu „verzerrten" Ergebnissen führen. Soweit ist das Argument überzeugend, jedoch rein negativ. Denn warum möchten wir subjektive Selektionsverfahren ausschließen? Doch deshalb, um schließlich eine Stichprobe zu erhalten, die möglichst repräsentativ für die Grundgesamtheit ist, über die man Informationen gewinnen möchte. Damit stellt sich jedoch die Frage, ob dieses Ziel damit erreicht werden kann, daß man die Auswahl der in die Stichprobe aufzunehmenden Individuen einem Zufallsgenerator überläßt. Dadurch können zwar *subjektive*, individuell zurechenbare Selektionsmechanismen ausgeschaltet werden; aber liefert ein solches Verfahren zugleich repräsentative Stichproben?

Offensichtlich besteht ein Problem dieser Frage darin, daß unklar ist, was genau mit dem Wort „repräsentativ" gemeint ist. Das übliche Vorverständnis geht davon aus, daß eine Stichprobe für eine Grundgesamtheit repräsentativ ist, wenn sie ihr „ähnlich" ist, d.h. wenn die Verteilung gewisser Zufallsvariablen in der Stichprobe ihrer Verteilung in der Grundgesamtheit „ähnlich" ist. Dies dient als intuitive Begründung dafür, daß mithilfe der Stichprobe näherungsweise zutreffende Aussagen über die Grundgesamtheit erreicht werden können. Die Frage, ob Zufallsstichproben

[9] Vgl. die von Basu [1971] gegebenen Beispiele und ihre anschließende Diskussion [ebda.].

[10] Interessante Ausführungen zur Geschichte der sich wandelnden Auffassungen über Datengewinnung durch Stichproben geben Kruskal und Mosteller [1979-80].

repräsentativ sind, ist insofern nur eine zunächst naive Formulierung für die Frage, ob bzw. in welcher Weise Zufallsstichproben für das statistische Inferenzproblem wichtig sind.

Um das Problem zu verdeutlichen, erscheint es mir gleichwohl sinnvoll, zunächst von diesem naiven Vorverständnis repräsentativer Stichproben auszugehen. Dann ist evident, daß Zufallsstichproben nicht unbedingt repräsentativ sind; denn als Ergebnis der Verwendung eines Zufallsgenerators kann im Prinzip jede mögliche Stichprobe auftreten.[11] Es besteht stets die Möglichkeit, daß eine Stichprobe realisiert wird, die für die zu untersuchende Grundgesamtheit nicht repräsentativ ist. Zum Beispiel ist es durchaus möglich, daß ein Zufallsgenerator eine Stichprobe aus der bundesrepublikanischen Bevölkerung liefert, die mehr als 80 Prozent Männer enthält.

Man könnte einwenden, daß eine solche Stichprobenziehung sehr unwahrscheinlich ist. Der Einwand ist sicherlich richtig; die Frage ist jedoch, ob er das Problem löst.[12] Daß bei der Verwendung eines Zufallsgenerators gewisse Ereignisse unwahrscheinlich sind, kann man so verstehen, daß in langen Serien wiederholter Zufallsexperimente (mit diesem Zufallsgenerator) ein Ereignis selten auftritt. Dann stellt sich jedoch die Frage, ob die Aussage auch dann eine Bedeutung hat, wenn ein Zufallsexperiment *nur einmal* durchgeführt wird. Denn dies ist typischerweise der Fall bei der Erhebung von Stichproben für die empirische Sozialforschung; und dies ist ein bemerkenswerter Unterschied zu anderen Anwendungsfällen statistischer Methoden. Man kann sich zwar *hypothetisch* eine große Anzahl wiederholter Stichprobenziehungen vorstellen, schließlich hat man jedoch immer *nur eine* realisierte Stichprobe, über deren Zustandekommen man nur weiß, daß sie von einem Zufallsgenerator „ausgewählt" worden ist.[13] Insofern ist es in gewisser Weise irreführend, die Gewinnung von Daten für die empirische Sozialforschung als Durchführung von Zufallsexperimenten zu betrachten. Denn von einem Experiment, auch von einem Zufallsexperiment, kann eigentlich nur gesprochen werden, wenn es tatsächlich wiederholt werden kann. Stichprobenerhebungen für die empirische Sozialforschung können jedoch in der Regel nur hypothetisch, nicht tatsächlich wiederholt werden. Zum Beispiel ist niemand in der Lage, die für das Sozio-

[11] Natürlich liegt es im Ermessen des Anwenders, den Bereich der möglichen Stichproben einzuschränken; damit wird dann jedoch ein intentionales Element in die Stichprobenziehung eingeführt.

[12] Denn auch unwahrscheinliche Ereignisse können eintreten. Insbesondere bei Stichprobenziehungen ist stets mit der Möglichkeit zu rechnen, daß die gezogene Stichprobe nicht repräsentativ ist. Man vgl. die von Royall [1983] gegebenen Beispiele und Hinweise auf weitere Literatur.

[13] Ich vernachlässige hier die zahlreichen Aspekte der Feldarbeit, die durchaus nicht immer „zufällig" sind und einen oft erheblichen Einfluß auf die schließlich verfügbaren Daten haben. Am Beispiel des Sozio-ökonomischen Panels wurde dies sehr informativ von Rendtel [1993] gezeigt.

ökonomische Panel verwendete Stichprobenziehung zu wiederholen. Selbst wenn man ausnahmsweise versuchen würde, eine Stichprobenerhebung zu wiederholen, hätte sich dann bereits die Grundgesamtheit verändert, auf die die Stichprobenziehung Bezug nimmt. Auf dieser Vorstellung beruht jedenfalls die übliche Interpretation von Daten aus unterschiedlichen Stichproben. Vergleicht man zum Beispiel die auf der Grundlage von zwei Zufallsstichproben zu den Zeitpunkten t_1 und t_2 ermittelten Einkommensverteilungen, gibt es zwei mögliche Gründe für ggf. feststellbare Unterschiede. Einerseits kann sich die Einkommensverteilung in der Grundgesamtheit verändert haben; andererseits können die Unterschiede eine Folge dessen sein, daß zwei unterschiedliche Stichproben gezogen worden sind. Eine Interpretation der beiden Stichprobenziehungen als Wiederholungen des gleichen Zufallsexperiments würde von vornherein die Möglichkeit ausschließen, daß sich die Einkommensverteilung in der Grundgesamtheit verändert haben könnte.

Indem die Auswahl von Stichprobenmitgliedern einem Zufallsgenerator überlassen wird, kann man zwar subjektiv gefärbte Selektionen ausschließen, man verzichtet jedoch zugleich (insoweit es sich um eine Zufallsstichprobe handelt) auf jede Kontrolle über das Resultat. Man kann in gewisser Weise nur hoffen, daß der Zufallsgenerator eine repräsentative Stichprobe ausgewählt hat, und gelegentlich versuchen, sich nachträglich davon zu überzeugen, daß die gezogene Stichprobe nicht allzu unglücklich ausgefallen ist.[14]

Diese Überlegung beruht darauf, daß es nicht gleichgültig ist, was für eine Stichprobe durch den Zufallsgenerator erzeugt worden ist, daß wir an Stichproben interessiert sind, die in gewisser Weise für die Grundgesamtheit repräsentativ sind. Ein Problem besteht offenbar darin, eine hinreichend genaue Vorstellung darüber zu gewinnen, was eine repräsentative Stichprobe ist. Leider gibt es auf diese Frage keine vollständig befriedigende Antwort. Manche Statistiker vertreten die Meinung, daß eine repräsentative Stichprobe dadurch *definiert* werden sollte, daß es sich um eine Zufallsstichprobe handelt, bei der jedes Individuum aus der Grundgesamtheit die gleiche Chance hat, in die Stichprobe aufgenommen zu werden.[15] Diese Definition impliziert jedoch die unbefriedigende Annahme, daß jede Zufallsstichprobe (ggf. nach einer geeigneten Berücksichtigung des Stichprobendesigns) als repräsentativ angesehen werden kann und daß jede darüber hinausgehende Frage nach der Repräsentativität der *tatsächlich* realisierten Stichprobe überflüssig ist.[16]

[14] Natürlich kann immer nur versucht werden, die vorliegenden Stichprobendaten mit anderen Informationen zu vergleichen, die man bereits über die Grundgesamtheit hat; in der Regel wiederum nur Informationen aus Stichproben. Vgl. exemplarisch Blossfeld [1987], Papastefanou [1990, S. 46ff], Hartmann und Schimpl-Neimanns [1992].

[15] Dies ist gewissermaßen der Prototyp einer Zufallsstichprobe, an dem sich auch die statistische Reflexion komplexer Stichproben mit differenzierten Ziehungswahrscheinlichkeiten orientiert; darauf wird weiter unten näher eingegangen.

Bei Moore [1991, S. 19] findet sich folgendes Beispiel: „The advice columnist Ann Landers once asked her readers: 'If you had it to do over again, would you have children?' She received nearly 10,000 responses, almost 70 % saying 'No!' [...] Now this is an egregious example of voluntary response. How egregious was suggested by a professional nationwide random sample commissioned by *Newsday*. That sample polled 1373 parents and found that 91 % *would* have children again. It is, you see, quite possible for a voluntary response sample to give 70 % 'No' when the truth about the population is close to 91 % 'Yes'."

Sicherlich würden in diesem Fall die meisten Menschen der *Newsday*-Umfrage mehr vertrauen als der von Ann Landers. Aber warum? Weil in diesem Fall die Art und Weise, wie Ann Landers ihre Stichprobe gewonnen hat, einen *offensichtlichen* Grund für die Annahme liefert, daß ihre Ergebnisse verzerrt sind. Eine Zufallsstichprobe ist demgegenüber gerade dadurch charakterisiert, daß es keine *offensichtlichen* Gründe gibt, warum sie zu verzerrten Ergebnissen führen sollte.[17] Dies ist jedoch eine rein negative Feststellung, aus der Moore's Schlußfolgerung, daß die

[16] Zu dieser Auffassung neigen, wenn ich es richtig verstanden habe, Rendtel und Pötter [1993] in ihrer Kritik an einem Versuch von Hartmann und Schimpl-Neimanns [1992], die Repräsentativität der ALLBUS-Daten zu prüfen. Ich stimme zwar Rendtel und Pötter darin zu, daß die Representativität von Stichproben nicht mithilfe „klassischer" Signifikanztests geprüft werden kann; damit wird jedoch weder die von Hartmann und Schimpl-Neimanns verfolgte Intention, Hinweise auf die Repräsentativität der ALLBUS-Daten zu gewinnen, noch die von ihnen verwendete Methode sinnlos. Ihr χ^2-Maß, mit dem sie die ALLBUS-Daten mit denen des Mikrozensus vergleichen, muß nicht unbedingt als eine „klassische" Teststatistik interpretiert werden, sondern kann auch als ein deskriptives Distanzmaß für den Vergleich empirischer Verteilungen angesehen werden. Findet man ähnliche Verteilungen, wird dies – vermutlich bei allen, deren statistisches Weltbild sich nicht auf das Neyman-Pearson'sche Testparadigma beschränkt – zum Vertrauen in *beide* Stichproben beitragen. Findet man große Unterschiede, wird man dementsprechend an der Repräsentativität beider Stichproben Zweifel bekommen. (Sollte man dann wiederum einen Zufallsgenerator befragen, welcher der beiden Stichproben ein größeres Vertrauen geschenkt werden sollte?) Schließlich können auch Rendtel und Pötter keine Lösung des Problems anbieten. Am Schluß ihrer Kritik sagen sie: „Das zentrale Problem von Umfragen besteht darin, daß die Ausfälle durch Nichterreichbarkeit und Antwortverweigerung 40 Prozent und mehr betragen. [...] Der Kern des Problems der Analyse von Umfragedaten besteht nicht in dem untauglichen Versuch, einer Stichprobe 'Repräsentativität' zu bescheinigen, sondern in der Frage, inwieweit der Ausfallprozeß sinnvolle Schätzungen zuläßt. Nicht 'Repräsentativität' ist das Problem, sondern die Aufdeckung bisher nicht berücksichtigter Selektionsprozesse!" (S. 357) Ich stimme dem zu, sehe jedoch keinerlei Begründung dafür, daß sich das Problem der Repräsentativität durch das Problem der Stichprobenausfälle *substitutieren* läßt; im Gegenteil, Stichprobenausfälle bilden einen (zweifellos wesentlichen) Aspekt des Problems der Repräsentativität. Jeder Versuch, mithilfe einer Stichprobe Einsichten in die Verfassung einer Grundgesamtheit zu gewinnen, beruht auf einem stets problematischen Vertrauen in die Repräsentativität der Stichprobe.

[17] Bei dieser Aussage wird natürlich nur auf die zufällige Auswahl von Individuen für die Stichprobe Bezug genommen, nicht auf die zahlreichen Verzerrungen, die durch die sich anschließenden Interviews entstehen können.

Newsday-Umfrage ein korrektes Ergebnis geliefert hat, strenggenommen nicht abgeleitet werden kann. Zu jeder Stichprobe, wie immer sie zustande gekommen sein mag, läßt sich ein Zufallsgenerator konstruieren, der sie hätte erzeugen können. Vielleicht hätte das in diesem Beispiel für die *Newsday*-Umfrage verwendete Stichprobendesign auch die von Ann Landers verwendeten Daten erzeugen können. Die Bedeutung dieses Arguments liegt darin, daß die bloße Tatsache, daß für die Stichprobenziehung ein Zufallsmechanismus verwendet worden ist, noch keine repräsentative Stichprobe garantiert; und daß deshalb die Frage, ob es sich um eine repräsentative Stichprobe handelt, nicht mit dem Hinweis abgetan werden kann, daß sie durch einen Zufallsmechanismus zustandegekommen ist.

Es ist allerdings bemerkenswert, daß die Frage nach der Repräsentativität von Stichproben aus der statistischen Literatur weitgehend verschwunden ist.[18] Das zugrundeliegende Problem ist damit jedoch nicht verschwunden; es erscheint nur in einer neuen Form, nämlich als die Frage, ob und ggf. wie mithilfe von Stichproben sinnvolle Aussagen über die intendierte Grundgesamtheit gewonnen werden können. Die Vernachlässigung des Problems der Repräsentativität ist insofern eine Folge dessen, daß viele Statistiker glauben, daß die Randomisierung bei der Stichprobenziehung eine *hinreichende* Grundlage für inferenzstatistische Aussagen liefert. Stimmt man dieser Auffassung zu, braucht man sich natürlich über die Repräsentativität der jeweils gezogenen Stichprobe keine Gedanken zu machen.[19] Allerdings ist diese Randomisierungskonzeption zur Lösung des statistischen Inferenzproblems umstritten; und sobald man diese Kontroversen zur Kenntnis nimmt, stößt man erneut auf das Repräsentativitätsproblem.

Zur Definition von Zufallsstichproben

Um die Diskussion fortsetzen zu können, ist es zweckmäßig, zunächst einen geeigneten begrifflichen Rahmen zu definieren. Dabei sollte berücksichtigt werden, daß die Art des statistischen Inferenzproblems von der jeweils vorliegenden Situation abhängt. Im folgenden gehe ich davon aus, daß es

[18] Der Vorgang ist wissenschaftsgeschichtlich interessant. In einem grundlegenden Aufsatz zur Begründung der Verwendung von Zufallsstichproben, um statistische Inferenz zu ermöglichen, hatte Neyman [1934] sich noch ausdrücklich auf die damalige Diskussion über die Repräsentativität von Stichproben bezogen und vorgeschlagen, Zufallsstichproben *per Definition* als repräsentativ anzusehen (insb. S. 585f).

[19] Zum Beispiel heißt es bei Schnell et al. [1992, S. 314]: „Zufallsstichproben stellen die einzige Gewähr dafür dar, daß aus Ergebnissen einer Stichprobe in bezug auf die Verteilung aller Merkmale (innerhalb bestimmter statistischer Fehlergrenzen) auf die Verteilung dieser Merkmale in der Grundgesamtheit geschlossen werden kann. Ein solcher *'Repräsentationsschluß'* kann also nur gezogen werden, wenn der Auswahlmechanismus eine Zufallsauswahl ist. Die Bezeichnung einer Stichprobe als 'repräsentativ' ist somit nur im Sinne des Prinzips der Zufallsauswahl zu verstehen: beide Begriffe sind im obigen Sinn synonym."

eine endliche Grundgesamtheit $\Omega = \{\omega_1, \ldots, \omega_N\}$ gibt, für die wie in (5.1) eine diskrete Zufallsvariable Z definiert ist. Die Frage ist dann zunächst, was unter einer Zufallsstichprobe aus der Grundgesamtheit Ω zu verstehen ist.

Aus statistischer Sicht besteht das grundlegende Erfordernis darin, daß man bei einer Zufallsstichprobe sinnvoll von der Wahrscheinlichkeit sprechen kann, mit der die unterschiedlichen möglichen Stichproben gezogen werden können. Es muß also ein Wahrscheinlichkeitsraum für die Stichprobenziehung konstruiert werden. Um dies zu erreichen, kann man unterschiedlich vorgehen. Die Vorgehensweise hängt im wesentlichen davon ab, welche Informationen über die Grundgesamtheit bereits vor der Stichprobenziehung verfügbar sind. Eine vergleichsweise einfache Darstellung erhält man, wenn man von der Annahme ausgeht, daß die vorab verfügbare Information mindestens in der Kenntnis einer Liste von Identifikationsnummern für alle Individuen in der Grundgesamtheit Ω besteht. Man verfügt dann über eine Liste $\{1, 2, \ldots, N\}$ und kann – zum Beispiel mit einer Urne oder mit einem Computer – ein Verfahren zur Ziehung von Zufallsstichproben aus dieser Liste definieren. Die auf diese Weise gewinnbaren Indexmengen bestehen aus Teilmengen von Identifikationsnummern, in formaler Schreibweise:[20]

$$S = \{i_1, i_2, \ldots, i_n\} \subseteq \{1, 2, \ldots, N\} \tag{5.2}$$

Hat man das Verfahren zur Gewinnung solcher Indexmengen festgelegt, gibt es für jede Indexmenge S aus der Menge aller möglichen Indexmengen eine bestimmte Wahrscheinlichkeit $\mathcal{P}(S) \geq 0$, mit der sie gezogen werden kann.[21] Die Menge aller möglichen Indexmengen, im folgenden mit $\mathcal{S}$ bezeichnet, kann als Menge aller Teilmengen der Liste $\{1, 2, \ldots, N\}$ angenommen werden; die nicht-realisierbaren Indexmengen haben dann eine Ziehungswahrscheinlichkeit $= 0$.

Die Festlegung von $(\mathcal{S}, \mathcal{P})$ wird als *Stichprobendesign* bezeichnet. Diese Festlegung liefert eine Kenntnis aller möglichen Indexmengen, außerdem eine Kenntnis der Wahrscheinlichkeit, mit der jede der möglichen Indexmengen gezogen werden kann. Man kann mithilfe des Stichprobendesigns insbesondere sog. *Inklusionswahrscheinlichkeiten*

$$\mathcal{P}(i \in S) = \sum_{S \ni i} \mathcal{P}(S)$$

berechnen, d.h. die Wahrscheinlichkeit dafür, daß ein bestimmtes Individuum $\omega_i \in \Omega$ in einer mit dem Stichprobendesign gezogenen Stichprobe

[20] Ich beschränke mich hier auf Stichproben, bei denen jedes Element der Grundgesamtheit höchstens einmal vorkommen kann; dann können Stichproben als Mengen charakterisiert werden.

[21] Es sei betont, daß es sich *in diesem Kontext* um wiederholbare Zufallsexperimente handelt, so daß auf eine sinnvolle Weise von Wahrscheinlichkeiten gesprochen werden kann.

enthalten ist.[22] Üblicherweise wird das Stichprobendesign so festgelegt, daß sich für alle Individuen der Grundgesamtheit eine positive Inklusionswahrscheinlichkeit ergibt; die Summe der Inklusionswahrscheinlichkeiten entspricht dann dem Umfang der Grundgesamtheit

Das Stichprobendesign liefert zunächst nur ein Verfahren zur Auswahl von Individuen aus der Grundgesamtheit Ω, es liefert noch keinerlei Informationen über diese Individuen.[23] Um solche Informationen zu gewinnen, muß man in einem zweiten Schritt die ausgewählten Individuen beobachten oder befragen. Typischerweise gelingt dies bei den für die empirische Sozialforschung relevanten Datenerhebungen nur unvollständig. Die daraus resultierenden Probleme sollen jedoch zunächst nicht betrachtet werden, d.h. es wird vorläufig angenommen, daß nach der Auswahl einer Indexmenge $S \in \mathcal{S}$ die gewünschten Informationen über die Zufallsvariable Z für alle durch S ausgewählten Individuen ermittelt werden können.

Eine Stichprobe kann dann durch zwei Bestandteile charakterisiert werden. Erstens gibt es eine Indexmenge S, die die Identifikationsnummern der in der Stichprobe enthaltenen Individuen enthält; zweitens gibt es für jedes dieser Individuen einen Wert der Zufallsvariable Z. Nimmt man an, daß die Reihenfolge der Individuen in der Stichprobe keine (für das zugrundeliegende Erkenntnisinteresse) wesentliche Information enthält, kann man folgende formale Darstellung verwenden:

$$\{(i, z_i) \mid i \in S\} \quad \text{wobei} \quad z_i = Z(\omega_i)$$

Die nächste Frage ist, ob bzw. in welcher Weise Wahrscheinlichkeitsaussagen über Stichproben gemacht werden können. Diese Frage ist berechtigt, denn die Festlegung eines Stichprobendesigns liefert zunächst nur die Möglichkeit, Wahrscheinlichkeitsaussagen über das Ziehen von Indexmengen zu treffen. Es ist jedoch evident, daß die Ziehungswahrscheinlichkeiten für die Indexmengen S und die mit ihnen verbundenen Stichproben $\{(i, z_i) \mid i \in S\}$ identisch sind. Der Grund liegt in der oben getroffenen Annahme, daß die Kenntnis einer Indexmenge die Kenntnis der zugehörigen Werte der Zufallsvariable Z impliziert. Solange an dieser Annahme festgehalten wird, kann also auf einfache Weise ein Wahrscheinlichkeitsraum für die Stichprobenziehung konstruiert werden.

Zunächst muß die Menge aller möglichen Stichproben definiert werden. Dabei ist zu berücksichtigen, daß die Notation nicht bereits eine Kenntnis der in der Grundgesamtheit realisierten Werte der Zufallsvariable Z voraussetzen darf. Dies kann durch folgende Definition erreicht werden:

$$\Pi = \left\{ \pi = \{(i, z'_i) \mid i \in S\} \,\middle|\, S \in \mathcal{S}, z'_i \in \mathcal{Z} \right\}$$

[22] Die Schreibweise $S \ni i$ soll eine Summation über alle Stichproben $S \in \mathcal{S}$ bedeuten, die einen Index für das Individuum ω_i enthalten.

[23] Abgesehen natürlich von Informationen, die bereits vor der Stichprobenziehung vorhanden sind und ggf. zur Gestaltung des Stichprobendesigns verwendet werden können.

Hier und im folgenden dient π zur Bezeichnung von Stichproben, und Π bezeichnet die Menge aller möglichen Stichproben, die ohne eine Kenntnis der Grundgesamtheit definierbar sind. Eine in diesem Sinne mögliche Stichprobe besteht also zunächst aus einer der möglichen Indexmengen $S \in \mathcal{S}$ sowie aus einer Zuordnung beliebiger Werte der Zufallsvariable Z zu den durch S erfaßten Individuen. Diese Definition erlaubt es, von einer Ziehungswahrscheinlichkeit für jede Stichprobe $\pi \in \Pi$ zu sprechen, wobei natürlich auf das zugrundeliegende Stichprobendesign Bezug genommen wird. Wir verwenden dafür die Notation

$$\mathcal{P}\left(\pi = \{(i, z'_i) \mid i \in S\}\right)$$

zu lesen als: die Wahrscheinlichkeit, mit der – bei einer Stichprobenziehung auf der Grundlage des Stichprobendesigns $(\mathcal{S}, \mathcal{P})$ – die Stichprobe $\{(i, z'_i) \mid i \in S\}$ realisiert wird.

Es ist jedoch evident, daß nur solche Stichproben realisiert werden können, bei denen die in der Stichprobe realisierten Werte der Zufallsvariable Z ihren in der Grundgesamtheit gegebenen Werten entsprechen.[24] Daraus folgt für die Wahrscheinlichkeit von Stichproben die Formulierung

$$\mathcal{P}\left(\pi = \{(i, z'_i) \mid i \in S\}\right) \quad = \quad \begin{cases} \mathcal{P}(S) \text{ wenn } z'_i = Z(\omega_i) \text{ für alle } i \in S \\ 0 \text{ andernfalls} \end{cases}$$

Das Symbol $\mathcal{P}$ wird verwendet, um darauf hinzuweisen, daß es einen grundsätzlichen Unterschied zwischen Wahrscheinlichkeitsaussagen über Zufallsstichproben und Aussagen über die Verteilung von Zufallsvariablen in einer (endlichen) Grundgesamtheit gibt. Letztere bezeichnen wir mit dem Symbol $\mathrm{P}(\cdot)$; zum Beispiel bedeutet $\mathrm{P}(Z = z)$ den Anteil der Individuen aus Ω, bei denen die diskrete Zufallsvariable Z den Wert z annimt. Wahrscheinlichkeitsaussagen dieser Art sind (im Kontext der in dieser Arbeit vertretenen Interpretation) deskriptive Aussagen über eine Grundgesamtheit. Dagegen bedeutet $\mathcal{P}(\pi)$ die Wahrscheinlichkeit, mit der auf der Grundlage eines bestimmten Stichprobendesigns die Stichprobe π realisiert wird. Dies ist keine deskriptive Aussage über eine Grundgesamtheit, sondern eine probabilistische Aussage über den für die Stichprobenziehung verwendeten Zufallsgenerator.

Einfache und komplexe Zufallsstichproben

Offenbar liegt die Festlegung eines Stichprobendesigns im Ermessen derjenigen, die eine Datenerhebung vornehmen. Grundsätzlich sind beliebig viele unterschiedliche Stichprobendesigns vorstellbar, und in der empirischen Sozialforschung werden tatsächlich zahlreiche unterschiedliche Stichprobendesigns verwendet. Infolgedessen ist eine allgemeine Diskussion der

[24] Strenggenommen beruht dies auf der Annahme, daß die Werte der Zufallsvariable Z fehlerfrei ermittelt werden können. Diese Annahme wird hier und im folgenden stets vorausgesetzt.

Frage, ob und ggf. wie das Stichprobendesign bei der Modellschätzung zu berücksichtigen ist, kaum möglich. Jeder kennt den gewissermaßen prototypischen Begriff einer *einfachen* Zufallsstichprobe: Man zieht mit gleicher Wahrscheinlichkeit und unabhängig voneinander die Identifikationsnummern von n Individuen aus einer vorgegebenen Grundgesamtheit. Dies ist der übliche Ausgangspunkt für die Darstellung statistischer Methoden und die Behandlung statistischer Inferenzprobleme. Denn man kann sich dann eine Stichprobe als Realisierung von n unabhängigen Zufallsvariablen vorstellen, die im Wahrscheinlichkeitsraum für die Stichprobenziehung definiert sind und die alle die gleiche Verteilung haben wie die in der Grundgesamtheit Ω interessierende Zufallsvariable.

Die in der empirischen Sozialforschung tatsächlich verwendeten Stichprobendesigns weichen jedoch fast immer von diesem Ideal einer einfachen Zufallsstichprobe ab. In der Regel werden *komplexe* Stichprobendesigns verwendet, wobei es hauptsächlich drei Leitideen gibt: (a) Die Stichprobenziehung wird in mehreren Stufen konzipiert; es werden zunächst primäre Stichprobeneinheiten ausgewählt (z.B. Regierungsbezirke), dann werden innerhalb der primären Stichprobeneinheiten sekundäre Einheiten ausgewählt usw., bis man schließlich auf der Ebene der Individuen ist, die man in die Stichprobe aufnehmen möchte. (b) Die Grundgesamtheit wird in Schichten eingeteilt, und es wird für jede Schicht ein im Prinzip eigenständiges Stichprobendesign entwickelt. (c) Es werden sog. Cluster („Klumpen“) gebildet, d.h. Mengen von Stichprobeneinheiten mit der Eigenschaft, daß die Auswahl eines Elements aus dem Cluster dazu führt, daß alle ihre Elemente in die Stichprobe aufgenommen werden. Komplikationen ergeben sich insbesondere daraus, daß alle drei Leitideen zur Bildung komplexer Stichprobendesigns kombiniert werden können. Infolgedessen ist eine schwer überschaubare Situation entstanden.[25]

Komplexe Stichproben werden meistens damit begründet, daß mit ihrer Hilfe ein besseres Verhältnis zwischen den Kosten der Datengewinnung und dem resultierenden Informationsgehalt erreicht werden kann.[26] Allerdings liegt dem in der Regel eine spezifische Auffassung über den Informationsgehalt einer Stichprobe zugrunde; man möchte möglichst effizient gewisse einfache deskriptive Parameter der Grundgesamtheit schätzen, zum Beispiel Mittelwerte oder die Anzahl von Personen mit gewissen Merkmalen. Das Interesse an solchen einfachen Parameterschätzungen durchzieht und beherrscht fast die gesamte Literatur zur Stichprobentheorie.[27] Die aus der

[25] Die meisten Bücher zur Stichprobentheorie enthalten umfangreiche Übersichten über die verschiedenen in der Praxis gebräuchlichen Stichprobendesigns. Eine Einführung findet sich auch bei Schnell et al. [1992, Kap. 6].

[26] „The purpose of sampling theory is to make sampling more efficient. It attempts to develop methods of sample selection and of estimation that provide, at the lowest possible cost, estimates that are precise enough for our purpose.“ (Cochran [1977, S. 8]).

[27] Vgl. als Beispiel das neue Lehrbuch von Särndal et al. [1992].

Verwendung komplexer Stichproben möglicherweise resultierenden Nachteile bei der Schätzung statistischer Modelle (für das Verhalten von Individuen in einer Gesellschaft) sind demgegenüber noch kaum untersucht worden.[28] Insbesondere ist die Frage, welche Folgen sich aus komplexen Stichprobendesigns für Modellschätzungen ergeben, noch verhältnismäßig wenig untersucht und diskutiert worden. Ich beschränke mich hier deshalb auf zwei Stichprobendesigns. Erstens auf einfache Zufallsstichproben, zweitens auf geschichtete Stichproben, bei denen es in jeder Schicht eine einfache Zufallsstichprobe gibt. In den folgenden beiden Beispielen wird dies näher erläutert.

Als Beispiel betrachten wir zunächst ein *einfaches* Stichprobendesign, das durch drei Merkmale charakterisiert werden kann: (a) Es gibt einen fest vorgegebenen Stichprobenumfang, (b) alle Individuen der Grundgesamtheit haben die gleiche Inklusionswahrscheinlichkeit, (c) die Auswahl der Individuen für die Stichprobe erfolgt unabhängig voneinander. Es gibt unterschiedliche Möglichkeiten, um Stichproben zu erzeugen, die (näherungsweise) diese drei Bedingungen erfüllen. Man kann sich zum Beispiel vorstellen, daß sich die Identifikationsnummern der Individuen aus der Grundgesamtheit in einer Urne befinden und daß man sukzessive (ohne Zurücklegen) n dieser Nummern herauszieht. Dann erhält man näherungsweise eine einfache Zufallsstichprobe des Umfangs n mit den Inklusionswahrscheinlichkeiten n/N, die für alle Individuen der Grundgesamtheit identisch sind.[29]

Wichtig für inferenzstatistische Überlegungen ist insbesondere die dritte Eigenschaft, die formal folgendermaßen geschrieben werden kann:

$$\mathcal{P}(i \in S, j \in S) \;=\; \mathcal{P}(i \in S)\, \mathcal{P}(j \in S)$$

Diese Bedingung kann allerdings nur als eine ideale Forderung aufgefaßt werden, die bei der praktischen Stichprobenziehung immer nur nährungsweise erfüllt werden kann. Sie gilt zum Beispiel bereits dann nur näherungsweise, wenn man sich vorstellt, daß die Stichprobe durch Ziehen ohne Zurücklegen aus einer Urne erfolgt. In der Praxis bedient man sich zur Stichprobenziehung meist maschinell erzeugter Zufallszahlen, bei denen man immer nur näherungsweise von einer Unabhängigkeit sprechen kann.

Der theoretische Vorteil einfacher Zufallsstichproben liegt darin, daß man sie durch unabhängige und identisch verteilte Zufallsvariablen darstellen kann. Bezeichnet Z die für die Grundgesamtheit definierte Variable mit dem Wertebereich $\mathcal{Z}$, und ist n der Stichprobenumfang, kann man folgende Definition vornehmen:

$$(Z_1, \ldots, Z_n) : \mathcal{S} \longrightarrow \mathcal{Z}$$

[28] Solche Nachteile sind schon deshalb zu erwarten, weil für den typischen Sekundärforscher die erforderlichen Informationen über das Stichprobendesign zumeist nur begrenzt zur Verfügung stehen.

[29] Vgl. Särndal et al. [1992, S. 31].

Diese Zufallsvariablen sind unabhängig voneinander und identisch verteilt. Allerdings ist zunächst nicht ohne weiteres klar, in welchem Verhältnis ihre Verteilung zur Verteilung von Z in der Grundgesamtheit Ω steht. Obwohl es intuitiv einleuchtend erscheint, sie zu identifizieren:

$$\mathcal{P}(Z_i = z) \equiv \mathrm{P}\,(Z = z) \qquad \text{für} \quad z \in \mathcal{Z} \quad \text{und} \quad i = 1, \ldots, n \qquad (5.3)$$

muß doch beachtet werden, daß es sich um ganz unterschiedliche Wahrscheinlichkeitsaussagen handelt.[30]

Als zweites Beispiel soll ein einfaches geschichtetes Stichprobendesign betrachtet werden. Dies setzt voraus, daß es genügend Vorab-Informationen über die Individuen aus Ω gibt, so daß sie in Schichten eingeteilt werden können. Die dafür verwendbaren Merkmale sind im Prinzip beliebig; zum Beispiel gibt es im Sozio-ökonomischen Panel eine Schichtung nach der Nationalität (der Haushaltsvorstände). Formal kann eine Schichtung als eine Zerlegung der Grundgesamtheit Ω in sich wechselseitig nicht überschneidende Teilklassen dargestellt werden, also bei K Schichten:

$$\Omega = \Omega_1 \cup \cdots \cup \Omega_K$$

Eine *einfache geschichtete Stichprobe* kann dann dadurch definiert werden, daß aus jeder Schicht Ω_k eine einfache Zufallsstichprobe des Umfangs n_k gezogen wird, und zwar unabhängig voneinander. Bezeichnet S_k die Indexmenge für die Stichprobe aus Ω_k, erhält man als Gesamtstichprobe

$$\{(i, z_i) \mid i \in S\} = \{(i, z_i) \mid i \in S_1\} \cup \cdots \cup \{(i, z_i) \mid i \in S_K\}$$

Wegen der Unabhängigkeitsbedingung gilt

$$\mathcal{P}\left(\{(i, z_i) \mid i \in S\}\right) = \prod_{k=1}^{K} \mathcal{P}\left(\{(i, z_i) \mid i \in S_k\}\right) = \prod_{i \in S} \mathcal{P}\left(\{(i, z_i)\}\right)$$

Der einzige Unterschied zu einer einfachen Zufallsstichprobe liegt darin, daß die Inklusionswahrscheinlichkeiten zwischen den Schichten variieren. Ist N_k die (als bekannt vorausgesetzte) Anzahl der Individuen in Ω_k, und wird in jeder Schicht eine einfache Zufallsstichprobe durch Ziehen ohne Zurücklegen gebildet, erhält man n_k/N_k als Inklusionswahrscheinlichkeit für die Individuen aus Ω_k.

Auch einfache geschichtete Zufallsstichproben können durch Stichprobenvariablen dargestellt werden, in diesem Fall:

$$(Z_{11}, \ldots, Z_{1n_1}, \ldots, Z_{K1}, \ldots, Z_{Kn_K}) : \mathcal{S} \longrightarrow \mathcal{Z}$$

[30] Häufig wird die Unterscheidung nicht explizit getroffen, sondern davon ausgegangen, daß die Verteilung der Stichprobenvariablen den Gegenstand bildet, über den man probabilistische Aussagen treffen möchte. Ich glaube jedoch, daß es für eine Darstellung inferenzstatistischer Probleme sinnvoll ist, den unterschiedlichen logischen Status der beiden Arten von Wahrscheinlichkeitsmaßen zu betonen.

Alle diese Zufallsvariablen sind unabhängig, aber nur innerhalb jeder Schicht sind sie identisch verteilt. Definiert man eine Zufallsvariable U über Ω, die für jedes Individuum den Index der Schicht angibt, der es angehört, kann man analog zu (5.3) folgende Korrespondenz zwischen Wahrscheinlichkeitsaussagen im Stichprobenraum und Wahrscheinlichkeitsaussagen über die Grundgesamtheit herstellen:

$$\mathcal{P}(Z_{ki} = z) \equiv \mathrm{P}\,(Z = z \mid U = k) \qquad z \in \mathcal{Z},\ i \in S_k,\ k = 1, \ldots, K$$

Um den Zusammenhang zur unbedingten Verteilung $\mathrm{P}(Z = z)$ herstellen zu können, kann man von der Darstellung

$$\mathrm{P}\,(Z = z) = \sum_{k=1}^{K} \mathrm{P}\,(Z = z \mid U = k)\,\mathrm{P}\,(U = k)$$

ausgehen. Wegen $\mathrm{P}(U = k) = N_k/N$ erhält man

$$\mathrm{P}\,(Z = z) \equiv \sum_{k=1}^{K} \frac{N_k}{N}\,\mathcal{P}(Z_{ki_k} = z) \qquad i_k \in S_k$$

In diesem Fall korrespondiert also keine der Stichprobenvariablen unmittelbar mit der Zufallsvariable Z. Die Verteilung von Z in der Grundgesamtheit Ω kann jedoch durch eine Gewichtung aus den Verteilungen der Stichprobenvariablen rekonstruiert werden.

5.3 Die Maximum-Likelihood-Methode

Komplizierte, nicht-lineare statistische Modelle werden fast immer mit der Maximum-Likelihood-Methode geschätzt; dies gilt insbesondere für die in den Kapiteln 3 und 4 beschriebenen Übergangsratenmodelle. Im folgenden soll deshalb diese Methode etwas genauer dargestellt werden.

Zunächst läßt sich die ML-Methode durch einen Rückgriff auf das in Abschnitt 3.2 eingeführte (informationstheoretische) KL-Distanzmaß darstellen. Als ein rechentechnisches Verfahren betrachtet, besteht die ML-Methode tatsächlich nur darin, ein im Hinblick auf dieses Distanzmaß optimales Modell für die gegebenen Daten einer Stichprobe zu berechnen. Es ist zweckmäßig, dies kurz zu vergegenwärtigen.

Die formale Darstellung hängt davon ab, ob es sich um Modelle für bedingte oder unbedingte Wahrscheinlichkeitsverteilungen handelt. Da Modelle für bedingte Wahrscheinlichkeitsverteilungen für praktische Anwendungen weitaus bedeutsamer sind (und da man nicht-konditionale Modelle als einen Spezialfall ansehen kann), gehe ich im folgenden hiervon aus. Es wird also angenommen, daß die in (5.1) definierte Zufallsvariable Z aus zwei, wiederum möglicherweise mehrdimensionalen Zufallsvariablen X und Y besteht, d.h. $Z = (X, Y)$, und daß ein Modell für die bedingte Wahrscheinlichkeitsverteilung $\mathrm{P}(Y = y \mid X = x)$ gebildet werden soll. Dafür sei eine Modellklasse $\tilde{g}(x, y; \theta)$ gegeben, wobei der Parametervektor θ in einer Parametermenge Θ variieren kann.[31]

Ist dieser Rahmen vorgegeben, liefert das KL-Distanzmaß folgende Regel zur Berechnung eines optimalen Modells:

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{P}\,(X = x, Y = y)\, \log \left\{ \frac{\mathrm{P}\,(Y = y \mid X = x)}{\tilde{g}(x, y; \theta)} \right\} \longrightarrow \min \tag{5.4}$$

Ob es (genau) ein optimales Modell gibt, hängt sowohl von der Verteilung von (X, Y) in der Grundgesamtheit als auch von der vorausgesetzten Modellklasse ab. Hier und im folgenden wird stets angenommen, daß ein eindeutiges optimales Modell existiert; es wird mit $\tilde{g}(x, y; \theta^\circ)$ bezeichnet.

Da die in (5.4) auftretenden Wahrscheinlichkeiten feste Größen, d.h. durch die Grundgesamtheit fest vorgegeben sind, erhält man ein formal äquivalentes Kriterium zur Berechnung eines optimalen Modells durch

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{P}\,(X = x, Y = y)\, \log \{\tilde{g}(x, y; \theta)\} \longrightarrow \max \tag{5.5}$$

In dieser Darstellung verläuft die Summierung über alle möglichen Merkmalsausprägungen von X und Y. Stattdessen kann man auch über die individuellen Werte dieser Zufallsvariablen summieren und erhält dann folgende zu (5.5) äquivalente Formulierung:

$$\sum_{i=1}^{N} \log \{\tilde{g}(X(\omega_i), Y(\omega_i); \theta)\} \longrightarrow \max \tag{5.6}$$

In dieser Form wird unmittelbar deutlich, wie die Modellberechnung mit den Daten einer Stichprobe vorgenommen werden kann; einfach dadurch, daß die Summation in (5.6) auf diejenigen Individuen eingeschränkt wird, für die durch die Stichprobe die Werte der Variablen X und Y bekannt sind. Wenn also die Stichprobe durch $\{(i, x_i, y_i) \mid i \in S\}$ gegeben ist, kann das Kriterium zur Modellberechnung folgendermaßen geschrieben werden:

$$\sum_{i \in S} \log \{\tilde{g}(x_i, y_i; \theta)\} \longrightarrow \max \tag{5.7}$$

Die Verwendung dieses Kriteriums zur Modellberechnung wird als *Maximum-Likelihood-Schätzung* eines Modells bezeichnet. Der in (5.7) angegebene Ausdruck wird *Log-Likelihood* des Modells $\tilde{g}(x, y; \theta)$ – für die Daten der gegebenen Stichprobe – genannt. Die korrespondierende *Likelihood* ist

$$\prod_{i \in S} \tilde{g}(y_i, x_i; \theta) \tag{5.8}$$

denn offenbar erhält man (5.7) durch Logarithmieren von (5.8). Der Parametervektor $\hat{\theta}$, den man durch die Maximierung der Likelihood oder – äquivalent dazu – der Log-Likelihood erhält, wird als *Maximum-Likelihood-Schätzung* des durch (5.4) definierten optimalen Parametervektors θ° bezeichnet.[32]

Diese Darstellung führt zu einem intuitiv plausiblen Verständnis der ML-Methode: man wählt aus einer vorgegebenen Modellklasse dasjenige Modell aus, das im Sinne des KL-Distanzmaßes am besten zur empirischen Verteilung der gegebenen Stichprobendaten paßt. Diese Darstellung liefert allerdings noch kein Verständnis der inferenzstatistischen Aspekte der ML-Methode. Um dies zu erreichen, sollen drei weitergehende Fragen überlegt werden.

(a) Welche Möglichkeiten gibt es, um die ML-Methode als eine inferenzstatistische Methode, d.h. $\tilde{g}(x, y; \hat{\theta})$ als eine *Schätzung* von $\tilde{g}(x, y; \theta^\circ)$ zu interpretieren?

[31] Wie bisher verwenden wir das Tilde-Zeichen, um auf Modelle zu verweisen.

[32] Wiederum gilt, daß diese Methode nur angewendet werden kann, wenn sich $\hat{\theta}$ mit den Daten einer Stichprobe eindeutig berechnen läßt. Ob dies der Fall ist, hängt sowohl von der Modelldefinition als auch von den zur Modellschätzung verfügbaren Daten ab und muß in jedem Anwendungsfall geprüft werden. Im folgenden wird stets vorausgesetzt, daß eine eindeutige Berechnung von $\hat{\theta}$ möglich ist.

(b) Ist es erforderlich, und ggf. in welcher Weise ist es erforderlich, das Stichprobendesign bei der Anwendung der ML-Methode zu berücksichtigen?

(c) Welche Möglichkeiten gibt es, um die Güte von ML-Schätzungen statistischer Modelle beurteilen zu können. Kann insbesondere zwischen Zufallsfehlern, die aus der Stichprobenziehung resultieren, und dem approximativen Charakter statistischer Modelle sinnvoll unterschieden werden?

Zu allen drei Fragen gibt es in der statistischen Literatur eine ausgedehnte Diskussion und zahlreiche unterschiedliche Auffassungen. Da es in dieser Arbeit nicht möglich ist, dieser Diskussion in ihre zahlreichen Nuancen zu folgen, beschränke ich mich auf eine kurze Darstellung von zwei unterschiedlichen Auffassungen, die sich in gewisser Weise als prototypisch ansehen lassen; ich nenne sie die *Randomisierungskonzeption* und die *Likelihoodkonzeption*.[33]

Die in der statistischen Literatur geführte Diskussion dieser beiden Auffassungen beschränkt sich nicht auf die ML-Methode, sondern bezieht sich generell auf Probleme der statistischen Inferenz. Im folgenden wird jedoch nur auf die ML-Methode Bezug genommen. Gewisse Besonderheiten ergeben sich auch daraus, daß ich im folgenden von der Problemstellung einer deskriptiven Modellschätzung ausgehe. Demgegenüber werden in der Literatur häufig unterschiedliche Auffassungen zum statistischen Inferenzproblem unmittelbar mit unterschiedlichen Auffassungen über die zu schätzenden Modelle verknüpft; in der Regel so, daß das Randomisierungskonzept mit einem Interesse an der Beschreibung endlicher Grundgesamtheiten verbunden wird, das Likelihoodkonzept demgegenüber mit einem Interesse an kausal interpretierbaren Superpopulationsmodellen. Deutlich wird dies zum Beispiel in folgender Diskussionsbemerkung von T. M. F. Smith [1983a, S. 801] zum Beitrag von Hansen, Madow und Tepping [1983]:

> To examine the inference problem more closely we must ask ourselves what it is we are trying to achieve when we make an inference. In scientific applications the position is fairly clear. There is a random phenomenon in nature that generates data that the scientist is trying to explain. A model is proposed, parameters are estimated, the fit is tested, if necessary the model is changed, and finally a working model is adopted. The overall objective is to find a model that represents the mechanism that generates the data. Nobody believes that the chosen model represents the absolute truth; it is at best a good approximation to the underlying natural process. We may call this exercise scientific inference.
>
> In the randomization approach to sample surveys the position appears to be quite different. No attempt is made to model the process that generates the values in the population and by implication there is no underlying random phenomenon. Probability is introduced only by the statistician's choice of a random sampling scheme, which is then completely known. There are no unknown parameters in this probability distribution and no need for tests of fit. Inferences are made with respect to repeated applications of the design, that is, with respect to the distribution generated by random sampling.

[33] Gelegentlich werden nicht nur zwei, sondern (mindestens) drei Konzeptionen unterschieden. Zum Beispiel unterscheidet Cox [1978] einen „sampling theory approach" (der in etwa der hier sog. Randomisierungskonzeption entspricht), einen „pure likelihood approach" (der in etwa der hier sog. Likelihoodkonzeption entspricht) und einen „Bayesian approach", der im folgenden nicht gesondert diskutiert wird. Cox hat mehrfach für eine *eklektizistische* Haltung zu den verschiedenen Auffassungen über statistische Inferenz plädiert. Ich stimme dem weitgehend zu; gerade deshalb sollte allerdings eine Erörterung im Hinblick auf das jeweilige Anwendungsfeld erfolgen.

In der Terminologie von Smith beschäftige ich mich im folgenden also nicht mit *„scientific inference"*, sondern mit der Frage, wie *deskriptive* Modelle für endliche Grundgesamtheiten von Individuen geschätzt werden können. Nicht nur die Randomisierungskonzeption, auch die Likelihoodkonzeption wird im Hinblick auf *diese* Aufgabe diskutiert.

5.3.1 Die Randomisierungskonzeption

Der grundlegende Gedanke der Randomisierungskonzeption kann darin gesehen werden, daß als wesentliche Basis für statistische Inferenz die Randomisierung der Stichprobenziehung angenommen wird. Nach dieser Auffassung sind statistisch begründbare Aussagen über eine Grundgesamtheit möglich, weil und insoweit die Stichprobenziehung durch einen Zufallsgenerator vorgenommen wird.

Um diese Auffassung etwas genauer charakterisieren zu können, ist es zweckmäßig, den Begriff einer *Schätzfunktion* einzuführen. Anknüpfend an die bisher verwendeten Notationen besteht der Ausgangspunkt in der Annahme einer Modellklasse $\tilde{g}(x, y; \theta)$ für die Zufallsvariablen (X, Y) in einer Grundgesamtheit Ω. Es wird außerdem angenommen, daß es ein optimales Modell $\tilde{g}(x, y; \theta^\circ)$ gibt, wobei in der üblichen Interpretation davon ausgegangen wird, daß es sich um eine statistische Hypothese über die Verteilung von (X, Y) handelt.[34] Die Aufgabe besteht dann darin, θ° mit den aus einer Stichprobe verfügbaren Daten zu schätzen. Ihre Lösung besteht darin, mit den Daten einer Stichprobe einen Schätzwert $\hat{\theta}$ für den unbekannten Parametervektor θ° zu berechnen. Die Methode, nach der diese Berechnung erfolgt, wird als eine *Schätzfunktion* bezeichnet. Handelt es sich zum Beispiel darum, den in einer Grundgesamtheit realisierten Mittelwert einer Variable zu schätzen, besteht eine einfache Schätzfunktion in folgender Anweisung: Berechne den Mittelwert aus der vorliegenden Stichprobe und bezeichne ihn als Schätzwert für den Mittelwert in der Grundgesamtheit.

Das Randomisierungskonzept beruht darauf, daß Schätzfunktionen als

[34] Auf dieser Betrachtungsweise beruht die verbreitete Redeweise, daß $\tilde{g}(x, y; \theta^\circ)$ ein *wahres* Modell für die Grundgesamtheit ist.

Zufallsvariablen innerhalb eines Wahrscheinlichkeitsraums betrachtet werden, der durch das jeweils zugrundeliegende Stichprobendesign definiert wird. In formaler Schreibweise:

$$T : \Pi \longrightarrow \Theta$$

Hier ist T die Schätzfunktion, die jeder möglichen Stichprobe $\pi \in \Pi$ einen Schätzwert $\hat{\theta} = T(\pi)$ für den unbekannten Parametervektor θ° zuordnet. Da T auf der Menge aller möglichen Stichproben definiert ist, und da es eine Wahrscheinlichkeitsverteilung $\mathcal{P}$ für die Ziehung von Zufallsstichproben $\pi \in \Pi$ gibt, kann also auch T als eine Zufallsvariable aufgefaßt werden, d.h. man kann Wahrscheinlichkeitsaussagen der folgenden Art betrachten:

$$\mathcal{P}(T(\pi) = \theta) \qquad \text{oder} \qquad \mathcal{P}(T(\pi) \in A) \quad \text{wobei} \quad A \subseteq \Theta$$

zu lesen als: die Wahrscheinlichkeit, daß die Schätzfunktion T bei der Ziehung einer Zufallsstichprobe π den Wert θ bzw. einen Wert in der Menge A annimmt.

Diese Betrachtungsweise liefert den für das Randomisierungskonzept wesentlichen Zugang zum Schätzproblem. Die zentrale Behauptung lautet, daß sinnvolle Urteile über die Sicherheit und die Qualität einer statistischen Schätzung durch eine Betrachtung der Wahrscheinlichkeitsverteilung der verwendeten Schätzfunktion gewonnen werden können.

Ausgehend von dieser Betrachtungsweise sind zahlreiche unterschiedliche Kriterien für die Beurteilung von Schätzfunktionen entwickelt worden. Bei einfachen deskriptiven Schätzfunktionen werden üblicherweise zwei Kriterien als besonders wichtig angesehen. Erstens sollen gute Schätzfunktionen *erwartungstreu* sein, d.h. ihr Erwartungswert soll gleich dem zu schätzenden Parameter sein. Dabei ist der Erwartungswert einer Schätzfunktion T folgendermaßen durch eine Bezugnahme auf das Stichprobendesign $(\mathcal{S}, \mathcal{P})$ definiert:

$$\mathrm{E}\{T\} = \sum_{\pi \in \Pi} T(\pi)\, \mathcal{P}(\pi)$$

Der Erwartungswert ist also das bei allen möglichen Stichprobenziehungen im Durchschnitt zu erwartende Schätzergebnis.

Wichtiger für die Beurteilung der Güte einer Schätzfunktion (auf der Grundlage der Randomisierungskonzeption) ist der *mittlere quadratische Fehler*, der folgendermaßen definiert ist:

$$\mathrm{MSE}\{T\} = \mathrm{E}\left\{(T - \theta^\circ)^2\right\} = \sum_{\pi \in \Pi} (T(\pi) - \theta^\circ)^2\ \mathcal{P}(\pi)$$

d.h. die bei allen möglichen Stichprobenziehungen im Durchschnitt zu erwartende quadratische Abweichung des jeweils erzielten Schätzergebnisses vom zu schätzenden Parameter θ°.[35] Man kann dann sagen: Je kleiner der mittlere quadratische Fehler einer Schätzfunktion ist, desto größer ist die Wahrscheinlichkeit, daß sie einen Schätzwert in der Nähe des zu schätzenden Parameters liefert.

Allerdings ist zu beachten, daß sich diese Kriterien auf Schätzfunktionen, nicht auf Schätzergebnisse beziehen. Auch wenn eine Schätzfunktion erwartungstreu ist und einen relativ kleinen mittleren quadratischen Fehler aufweist, garantiert dies noch keineswegs, daß man eine gute Schätzung erhält. Denn die Schätzung, die man tatsächlich erhält, hängt von der Stichprobe ab, die „zufällig“ realisiert worden ist. Wenn die Stichprobe nicht repräsentativ ist, liefern auch Schätzfunktionen, die im Sinne des Randomisierungskonzepts gut oder sogar optimal sind, nur schlechte Ergebnisse. Insofern stößt man auch an dieser Stelle wieder auf das bereits diskutierte Problem der Repräsentativität von Zufallsstichproben. Der Lösungsvorschlag des Randomisierungskonzepts besteht darin, Kriterien für Regeln zu entwickeln, so daß man, *wenn man diesen Regeln oft folgt*, im Durchschnitt gute Ergebnisse erwarten kann.[36] Diese Betrachtungsweise erscheint sinnvoll, wenn es sich um Zufallsexperimente handelt, die tatsächlich sehr oft durchgeführt werden (können). Das Problem für den empirischen Sozialforscher besteht jedoch darin, daß er es immer nur mit einer, nicht wiederholbaren Zufallsstichprobe zu tun hat. Insofern helfen ihm Regeln, deren Sinnvoraussetzung in der Wiederholbarkeit ihrer Anwendung liegt, nur wenig.[37]

Dies ist zugleich der wesentliche Grund dafür, warum sich das Problem der Repräsentativität von Stichproben bei der Anwendung inferenzstatistischer Verfahren in der empirischen Sozialforschung immer wieder neu stellt. In einer Situation, in der man es mit (beliebig) wiederholbaren Experimenten oder Beobachtungen zu tun hat, erscheint es plausibel, gewissermaßen eine Problemverschiebung vorzunehmen: die wesentliche Grundlage für verallgemeinerbare Schlußfolgerungen nicht im singulären Beobachtungsresultat zu sehen, sondern in der – in diesem Fall realisierbaren – Möglichkeit, das Experiment oder die Beobachtung wiederholen zu

[35] Wenn es sich um einen Vektor handelt, kann der quadratische Fehler komponentenweise berechnet werden.

[36] Sehr deutlich hat Neyman [1934, S. 624] betont: „This circumstance, that in the problem of confidence intervals the probability statements concern the results of our behaviour, not the populations and that they relate to this given rule of behaviour, not to the properties of samples to which this rule is being applied, is very important.“

[37] Vgl. Fishers [1955] Kritik an der entscheidungstheoretischen Deutung des statistischen Inferenzproblems. Neyman [1977, S. 109] hat demgegenüber festzustellen versucht, daß eine Wiederholbarkeit *des gleichen* Zufallsexperiments keine notwendige Sinnvoraussetzung darstellt. Dieser Hinweis Neymans ist wichtig, weil durch ihn eine zusätzliche Reflexionsmöglichkeit entsteht, um den Sinn gewisser Regeln für das Inferenzproblem zu verstehen. Er löst jedoch nicht das grundsätzliche Problem, mit welchem Vertrauen die Schätzergebnisse in jedem Einzelfall betrachtet werden können.

können. Infolge dieser Problemverschiebung ist es dann gleichgültig (bzw. nur eine Frage der Effizienz), ob man in jedem Einzelfall eine repräsentative Beobachtung gemacht hat; wichtig ist vielmehr die im Durchschnitt erzielte Qualität der Schlußfolgerungen. Bei dieser Betrachtungsweise liefert die Verwendung von Zufallsgeneratoren für die Stichprobenziehung tatsächlich ein zusätzliches Argument. Sie liefert nicht nur das rein negative Argument, daß damit subjektiv induzierte Verzerrungen bei der Stichprobenauswahl vermieden werden können, sondern begründet die Vorstellung, daß es sich um ein wiederholbares Zufallsexperiment handelt. Bei einer *praktisch* nicht-wiederholbaren Stichprobenziehung reduziert sich die Bedeutung der Tatsache, daß sie durch einen Zufallsmechanismus zustandegekommen ist, jedoch auf das Selektivitätsargument. Man ist dann zwar bei der Erzeugung der Stichprobe *einer Regel gefolgt*; aber da man der Regel nur einmal folgen kann, ist es in gewisser Weise gleichgültig, daß es sich um eine *Regel* handelt, von entscheidender Bedeutung ist vielmehr, welches Resultat erzielt worden ist.

Natürlich liefert eine Bezugnahme auf Regeln argumentative Vorteile. Man kann dadurch nicht nur erklären, wie man die Daten erzeugt hat; man kann auch mithilfe von Analogieschlüssen die Vor- und Nachteile unterschiedlicher Regeln diskutieren. Insofern hilft es der Rationalisierbarkeit des inferenzstatistischen Problems, wenn man auf Regeln verweisen kann, nach denen man die jeweils verwendeten Daten gewonnen hat. Insofern kann man auch der Aussage zustimmen, „daß eine *Auswahl aufs Geratewohl* keine Zufallsauswahl ist: Sie garantiert in keiner Weise, daß jede Einheit die gleiche Chance hat, in die Stichprobe einbezogen zu werden.“[38] Dem mit dieser Aussage üblicherweise verbundenen Umkehrschluß kann man jedoch nur zustimmen, wenn und insoweit es sich um wiederholbare Zufallsstichproben handelt. Denn in jedem Einzelfall hätte der Zufallsgenerator auch die „aufs Geratewohl gezogene“ Stichprobe liefern können. Die Unterscheidbarkeit der beiden Formen der Stichprobenziehung liegt nur darin, daß man bei Zufallsstichproben einer *wiederholbaren* Regel folgt.[39] Wenn jedoch die Regel nur hypothetisch wiederholbar ist, verliert sie an Bedeutung und wichtig wird vielmehr das (mit ihr) erzielte Resultat, d.h. die Frage, ob eine repräsentative Stichprobe erzielt worden ist.

[38] Pfanzagl [1983, S. 173].

[39] Dies ist bereits impliziert, wenn man den Vorteil von Zufallsstichproben darin sieht, daß sie auf *bekannten* Inklusionswahrscheinlichkeiten für die Individuen der Grundgesamtheit beruhen. Denn andernfalls könnte man sich für jede Stichprobe, auch für die „aufs Geratewohl gezogene“, bekannte Inklusionswahrscheinlichkeiten verschaffen; es wäre nur erforderlich, einen Zufallsgenerator zu konstruieren, der die tatsächlich realisierte Stichprobe hätte erzeugen können. Eine Analyse dieses Zufallsgenerators würde dann die Inklusionswahrscheinlichkeiten liefern.

5.3.2 ML-Schätzfunktionen

Im Rahmen des Randomisierungskonzepts wird auch die ML-Methode als eine Methode zur Konstruktion von Schätzfunktionen betrachtet.[40] Daß diese Interpretation möglich ist, ist unmittelbar einsichtig. Es treten dann jedoch einige zusätzliche Schwierigkeiten auf, denn die einfachen Kriterien, die sich auf den Erwartungswert und den mittleren quadratischen Fehler von Schätzfunktionen beziehen, sind nicht ohne weiteres anwendbar. Einerseits sind ML-Schätzfunktionen im allgemeinen nicht erwartungstreu; andererseits ist es in den meisten Fällen praktisch unmöglich, ihren mittleren quadratischen Fehler zu berechnen.

Es ist deshalb im Rahmen des Randomisierungskonzepts üblich, asymptotische Eigenschaften von ML-Schätzfunktionen zu betrachten.[41] Der Grundgedanke besteht darin, hypothetisch anzunehmen, daß der Stichprobenumfang beliebig groß werden kann. Dann kann man die Eigenschaften von Schätzfunktionen in Abhängigkeit vom Stichprobenumfang untersuchen. Leider ist eine genaue Darstellung dieser Betrachtungsweise mathematisch ziemlich kompliziert. Das Hauptproblem besteht jedoch darin, daß die Untersuchung asymptotischer Eigenschaften von Schätzfunktionen nicht ohne weiteres sinnvoll auf die Aufgabe bezogen werden kann, deskriptive Modelle für endliche Grundgesamtheiten von Individuen zu schätzen.

Strenggenommen ist eine asymptotische Betrachtung der ML-Methode nur sinnvoll, wenn von einem Superpopulationsmodell ausgegangen wird. Dann kann man hypothetisch annehmen, daß es eine unendliche Folge von unabhängigen und identisch verteilten Zufallsvariablen gibt, die der Verteilung des unterstellten Modells entsprechen, also

$$Z_1, Z_2, Z_3, \ldots, Z_n, \ldots \overset{\text{iid}}{\sim} \tilde{F}(z, \theta^\circ)$$

Ist dann $\tilde{g}(z, \theta)$ eine Modellklasse für $\tilde{F}(z, \theta^\circ)$, kann man für jede endliche Teilfolge $Z_1, \ldots, Z_n$ die ML-Schätzfunktion $T_n(Z_1, \ldots, Z_n)$ definieren, die jeder möglichen Realisierung dieser Zufallsvariablen den durch sie (wenn möglich) aus der Maximierung von

$$\sum_{i=1}^{n} \log\left(\tilde{g}(Z_i, \theta)\right)$$

berechenbaren ML-Schätzwert für den Parametervektor θ zuordent. Man erhält dann für $n \to \infty$ eine Folge von Schätzfunktionen und kann deren

[40] Zum Beispiel bezeichnet Schaich [1990, S. 167] die ML-Methode „als eine Methode zur Ermittlung von Schätzfunktionen mit wünschenswerten Eigenschaften für explizite Parameter von Verteilungen“.

[41] „The last refuge for the Randomization Principle is in asymptotics.“ bemerkt Royall [1983, S. 796] in einer Kritik an dem Beitrag von Hansen et al. [1983].

Eigenschaften untersuchen. Durch diese Betrachtungsweise können (unter sehr schwachen Voraussetzungen) insbesondere zwei wichtige Eigenschaften von Folgen von ML-Schätzfunktionen bewiesen werden. Erstens, daß die Folge der Schätzfunktionen $T_n(Z_1, \ldots, Z_n)$ im Hinblick auf den zu schätzenden Parametervektor θ° konsistent ist, d.h. im Sinne eines Gesetzes der großen Zahlen probabilistisch gegen θ° konvergiert. Zweitens, daß die Verteilung von $T_n(Z_1, \ldots, Z_n)$ mit wachsendem n probabilistisch gegen eine Normalverteilung konvergiert.[42]

Diese beiden Eigenschaften hypothetischer Folgen von ML-Schätzfunktionen (die sich noch durch asymptotische Aussagen über ihre Effizienz ergänzen lassen) werden häufig als wesentliche Begründung für die ML-Methode angesehen. Es ist jedoch offensichtlich schwer, der zugrundeliegenden asymptotischen Betrachtungsweise eine sinnvolle Bedeutung zu geben, wenn es sich darum handelt, deskriptive Modelle für endliche Grundgesamtheiten von Individuen zu schätzen.[43]

Gelegentlich ist versucht worden, die asymptotische Theorie der Schätzfunktionen dadurch für Schätzprobleme in endlichen Grundgesamtheiten nutzbar zu machen, daß man von der Vorstellung einer Folge endlicher Grundgesamtheiten mit jeweils wachsendem Umfang ausgeht.[44] Diese Betrachtungsweise ist jedoch nicht weniger spekulativ als die Schätzung von Superpopulationsmodellen, konzeptionell sogar noch weniger plausibel.[45] Will man den spekulativen Übergang zur Betrachtung von Superpopulationsmodellen oder von Folgen endlicher Grundgesamtheiten vermeiden, gibt es, soweit ich sehen kann, nur eine unmittelbare Möglichkeit, um auf der Grundlage einer endlichen Grundgesamtheit eine unendliche Folge von Zufallsvariablen zu konstruieren: indem man vom Urnenmodell *mit Zurücklegen* ausgeht. Dann wird jedoch, wie das folgende Beispiel zeigt, die asymptotische Betrachtungsweise in gewisser Weise trivial.

Wir stellen uns vor, daß sich die Identifikationsnummern der N Individuen aus der endlichen Grundgesamtheit Ω in einer großen Urne befinden,

[42] Beide Aussagen setzen natürlich voraus, daß zuvor ein geeigneter Wahrscheinlichkeitsraum konstruiert wird, innerhalb desssen probabilistische Aussagen über Folgen von Schätzfunktionen formuliert werden können.

[43] Kritische Überlegungen finden sich u.a. bei Kempthorne [1969, S. 674f].

[44] Vgl. zu dieser Betrachtungsweise zum Beispiel Hansen et al. [1983, S. 777], Särndal et al. [1992, S. 167].

[45] Vermutlich ist dies der Grund, warum diese Betrachtungsweise bisher nur wenige Anhänger gefunden hat. Smith [1983, S. 801] bemerkt in seiner Kritik an Hansen et al. [1983]: „Although the randomization distribution is well defined, the method for making inferences is not assumption free. The authors argue that the inferences should be based on two criteria, consistency and asymptotic normality. Neither of these criteria is defined for the given finite population; both require the construction of a hypothetical sequence of finite populations of increasing size. This sequence is just as much a statistical model as a regression model relating two variables Y and X, the main distinction being that the construction of the sequence is purely arbitrary and can never be tested against data."

aus der man unabhängig voneinander und jeweils mit der gleichen Wahrscheinlichkeit $1/N$ eine Identifikationsnummer herausziehen kann. Wenn man nach jedem Zug die gezogene Nummer wieder zurücklegt, kann man dieses Zufallsexperiment hypothetisch beliebig oft wiederholen. Werden auf diese Weise n Identifikationsnummern gezogen, kann man das Ergebnis folgendermaßen beschreiben:

$$a_1 \times z_1, \ldots, a_N \times z_N \quad \text{mit} \quad \sum_{i=1}^{N} a_i = n$$

Hierbei sind $z_1, \ldots, z_N$ die in der Grundgesamtheit realisierten Werte der Zufallsvariable Z; $a_i \in \{0, 1, 2, \ldots\}$ ist die Häufigkeit, mit der das Individuum ω_i bzw. der Merkmalswert z_i gezogen wird. Das einzige stochastische Element besteht in diesen Häufigkeiten. Betrachtet man die relativen Häufigkeiten a_i/n, kann man sie als Realisationen von unabhängigen und identisch verteilten Zufallsvariablen $A_{i,n}$ ansehen. Also kann man eine Folge von ML-Schätzfunktionen $T_n(A_{1,n}, \ldots, A_{N,n})$ definieren, die jeder Realisierung dieser Zufallsvariablen den durch (wenn möglich) aus der Maximierung von

$$\sum_{i=1}^{N} A_{i,n} \, \log\left(\tilde{g}(z_i, \theta)\right)$$

berechenbaren ML-Schätzwert für den Parametervektor θ zuordnet. Diese Folge ist offensichtlich konsistent im Hinblick auf die Schätzung des für die Grundgesamtheit deskriptiven Modells $\tilde{g}(z, \theta^\circ)$, das sich durch das Standardkriterium

$$\sum_{i=1}^{N} \frac{1}{N} \log\left(\tilde{g}(z_i, \theta)\right) \longrightarrow \max$$

für die Grundgesamtheit definieren läßt. Die Konsistenz besteht einfach darin, daß man mit einer einfachen Zufallsstichprobe, wenn man sie hinreichend groß macht, die endliche Grundgesamtheit fast sicher vollständig ausschöpfen kann und daß dann die Stichprobe (probabilistisch) mit der Grundgesamtheit identisch ist.[46]

Das Problem, asymptotische Eigenschaften von ML-Funktionen für Modellschätzungen in endlichen Grundgesamtheiten nutzbar zu machen,

[46] Dies entspricht Cochrans [1977, S. 21] Definition konsistenter Schätzfunktionen: „a method of estimation is called *consistent* if the estimate becomes exactly equal to the population value when $n = N$, that is, when the sample consists of the whole population." In gewisser Weise entspricht dies auch Fishers [1922, S. 316] ursprünglicher Definition: „The common-sense criterion employed in problems of estimation may be stated thus: That when applied to the whole population the derived statistic should be equal to the parameter." Allerdings bezieht sich Fishers Definition auf seine *grundsätzliche* Voraussetzung einer hypothetisch-unendlichen Grundgesamtheit.

zeigt sich insbesondere in der – im Rahmen des Randomisierungskonzepts – üblichen Praxis, die Berechnung von Signifikanzniveaus und Konfidenzintervallen auf die Annahme zu gründen, daß die ML-Schätzfunktionen asymptotisch normalverteilt sind. Der zugrundeliegende Gedankengang scheint etwa folgender zu sein: Wenn eine Schätzfunktion $T_n(Z_1, \ldots, Z_n)$ gewisse asymptotische Eigenschaften hat, d.h. Eigenschaften für $n \to \infty$, dann kann man auch mit einer gewissen Berechtigung annehmen, daß sie *näherungsweise* gelten, wenn n *hinreichend groß* ist.[47] Hierbei stellen sich jedoch zwei Probleme. Erstens entsteht natürlich die Frage, wann eine Stichprobe *hinreichend groß* ist; darüber ist tatsächlich nur wenig bekannt. Zweitens stellt sich ein nur selten beachtetes Problem, daß der Grenzübergang $n \to \infty$ aus mathematischen Gründen erforderlich ist, um asymptotische Aussagen über die Verteilung von ML-Schätzfunktionen formulieren zu können. Um eine angemessene Vorstellung des approximativen Gehalts asymptotischer Aussagen zu gewinnen, muß dagegen davon ausgegangen werden, daß nur eine endliche Anzahl von Beobachtungen zur Verfügung steht und infolgedessen kein Grenzübergang vollzogen werden kann.

Um dies zu illustrieren, soll kurz dargestellt werden, wie man zu der Vorstellung kommen kann, daß ML-Schätzfunktionen *näherungsweise* normalverteilt sind. Dabei wird, im Unterschied zur üblichen Betrachtungsweise, von einem Stichprobendesign für eine endliche Grundgesamtheit des Umfangs N ausgegangen. Gegeben sei also ein Stichprobendesign $(\mathcal{S}, \mathcal{P})$, mit dem endliche Stichproben $S \in \mathcal{S}$ mit den Wahrscheinlichkeiten $\mathcal{P}(S)$ erzeugt werden können. Z sei die in der Grundgesamtheit interessierende Zufallsvariable; ihre in der Grundgesamtheit realisierten Werte werden mit z_i $(i = 1, \ldots, N)$, das mit der ML-Methode zu schätzende Modell für Z wird mit $\tilde{g}(z, \theta)$ bezeichnet. Unsere Fragestellung folgt dem Randomisierungskonzept, es soll also überlegt werden, welche Aussagen über die durch das Stichprobendesign definierte Stichprobenverteilung einer ML-Schätzfunktion für den Modellparameter θ möglich sind.

Um die Fragestellung zu vereinfachen, wird angenommen, daß es sich um ein einfaches Stichprobendesign handelt, d.h. alle Stichproben haben den gleichen Umfang n, und die Aufnahme der Individuen in die Stichproben erfolgt unabhängig voneinander mit gleichen Inklusionswahrscheinlichkeiten. Das Stichprobendesign kann dann durch eine Folge von N Indikatorvariablen

$$I_1, \ldots, I_N \; : \; \mathcal{S} \longrightarrow \{0, 1\}$$

repräsentiert werden. Für jede Stichprobe $S \in \mathcal{S}$ bedeutet $I_i(S) = 1$, daß das Individuum $\omega_i \in \Omega$ in der Stichprobe enthalten, daß also die Information z_i verfügbar ist, und $I_i(S) = 0$ bedeutet, daß das Individuum ω_i nicht in der Stichprobe enthalten ist. Es handelt sich also um bzgl. des Stichprobendesigns definierte Zufallsvariablen. Sie sind unabhängig und identisch

[47] Vgl. zum Beispiel die Formulierungen bei Diekmann und Mitter [1984, S. 78].

verteilt; Erwartungswert und Varianz sind (näherungsweise) durch

$$\mathrm{E}\{I_i\} = \frac{n}{N} \quad \text{und} \quad \mathrm{Var}\{I_i\} = \frac{n}{N}\,\frac{N-n}{N}$$

gegeben. Geschätzt werden soll ein deskriptives Modell für die Grundgesamtheit, also $\tilde{g}(z, \theta^\circ)$, definiert durch die Maximierung der für die Grundgesamtheit gebildeten Log-Likelihood

$$\sum_{i=1}^{N} \log(\tilde{g}(z_i, \theta)) \tag{5.9}$$

Jede Stichprobe $S \in \mathcal{S}$ liefert einen ML-Schätzwert $\hat{\theta} = \hat{\theta}(S)$.[48] Der Zusammenhang mit dem Stichprobendesign kann durch folgende Stichprobenfunktion dargestellt werden:

$$L(\theta) = \sum_{i=1}^{N} I_i \, \log(\tilde{g}(z_i, \theta))$$

Jede Stichprobenziehung liefert eine Realisierung der Funktion $L(\theta)$, aus deren Maximierung sich dann der aus dieser Stichprobe zu gewinnende ML-Schätzwert $\hat{\theta}$ ergibt. (Zur Vereinfachung der Notation wird die Abhängigkeit von der jeweils gezogenen Stichprobe $S \in \mathcal{S}$ nicht explizit angeführt.)

Wie können nun approximative Aussagen über die Stichprobenverteilung der ML-Schätzwerte $\hat{\theta}$ gewonnen werden. Zunächst bietet sich die Stichprobenfunktion $L(\theta)$ an; denn sie besteht aus einer Summe unabhängiger Zufallsvariablen, und man könnte argumentieren, daß sie – für jedes fest vorgegebene θ – näherungsweise normalverteilt ist. Diese Betrachtungsweise führt jedoch nicht unmittelbar zum Ziel. Denn $L(\theta)$ ist eine (in θ) nicht-lineare Funktion, so daß aus Aussagen über die Verteilung von $L(\theta)$ nicht unmittelbar Aussagen über die Verteilung von θ gewonnen werden können. An dieser Stelle setzt deshalb die erste Approximation ein: es wird eine lineare Approximation an den Gradienten von $L(\theta)$ betrachtet.

Zunächst sollen die erforderlichen Begriffe definiert werden. Wir benötigen die erste und zweite Ableitung der Log-Likelihoodfunktionen $L(\theta)$. Die erste Ableitung wird als *Gradient* bezeichnet. Da θ im allgemeinen ein Vektor ist, ist auch der Gradient ein Vektor. Wir nehmen an, daß der Parametervektor aus m Komponenten besteht, also $\theta = (\theta_1, \ldots, \theta_m)$. Die

[48] Im allgemeinen sind stets einige Stichproben zu erwarten, mit denen die gewünschte ML-Schätzung nicht durchgeführt werden kann. Strenggenommen müßte also die folgende Betrachtung auf eine Teilmenge des Stichprobenraums konditioniert werden, für die die ML-Schätzungen berechnet werden können. Von dieser Komplikation wird hier jedoch abgesehen.

j.te Komponente des Gradienten ist dann

$$G_j(\theta) \;=\; \sum_{i=1}^{N} I_i \,\frac{\partial}{\partial\theta_j}\log(\tilde{g}(z_i,\theta)) \qquad j=1,\ldots,m \tag{5.10}$$

Der Gradient ist ein Spaltenvektor, der aus diesen Komponenten besteht, also

$$G(\theta) \;=\; \begin{bmatrix} G_1(\theta) \\ \vdots \\ G_m(\theta) \end{bmatrix} \tag{5.11}$$

Die zweite Ableitung von $L(\theta)$ nach θ ergibt eine Matrix; sie wird als *Hesse-Matrix* bezeichnet. Ihre Komponenten sind folgendermaßen definiert:

$$H_{jj'}(\theta) \;=\; \sum_{i=1}^{N} I_i \,\frac{\partial^2}{\partial\theta_j\,\partial\theta_{j'}}\log(\tilde{g}(z_i,\theta)) \qquad j,j'=1,\ldots,m$$

Daraus ergibt sich die Hesse-Matrix, offensichtlich eine symmetrische (m,m)-Matrix:

$$H(\theta) \;=\; \begin{bmatrix} H_{11}(\theta) & \cdots & H_{1m}(\theta) \\ \vdots & & \vdots \\ H_{m1}(\theta) & \cdots & H_{mm}(\theta) \end{bmatrix}$$

Der ML-Schätzwert $\hat{\theta}$ ergibt sich aus einer Maximierung der Log-Likelihood $L(\theta)$. Wenn wir annehmen, daß ein eindeutiges (möglicherweise jedoch nur lokales) Maximum existiert, impliziert dies zwei wesentliche Bedingungen. Erstens muß der Gradient an der Stelle des Maximums den Wert 0 annehmen, also

$$G(\hat{\theta}) \;=\; 0 \tag{5.12}$$

Zweitens muß die Hesse-Matrix in einer Umgebung des Maximums negativ definit sein, insbesondere also invertierbar.

Jetzt kann die erste wesentliche Approximation vorgenommen werden, indem der Gradient durch eine lineare Funktion approximiert wird. Das erscheint zunächst trivial, wenn man sich dabei auf eine kleine Umgebung für die Nullstelle $\hat{\theta}$ bezieht. Dann kann man den Satz von Taylor verwenden und erhält

$$G(\hat{\theta}+\delta) \;=\; G(\hat{\theta}) + H(\hat{\theta})\,\delta + R \tag{5.13}$$

Hier ist δ ein „kleiner“ Vektor der Dimension m, so daß sich $\hat{\theta}+\delta$ in einer kleinen Umgebung zu $\hat{\theta}$ befindet. R ist das Restglied, das bei der linearen Approximation vernachlässigt wird.

Nun ist jedoch der Gradient $G(\theta)$ eine Stichprobenfunktion, d.h. ein Zufallsvektor; und das Problem besteht hier nicht darin, die Eigenschaften dieses Gradienten als eine Funktion von θ bei einer gegebenen Stichprobe zu untersuchen. Das Problem besteht vielmehr darin, eine Approximation für die Verteilung von $\hat{\theta}$ bei allen möglichen Stichprobenziehungen zu finden. Das bedeutet, daß $\hat{\theta}$ mit dem Erwartungswert $\mathrm{E}\{\hat{\theta}\}$ oder mit dem zu schätzenden Parametervektor θ° verglichen werden muß. Beides ist möglich; entscheidend ist jedoch, daß wir eine lineare Approximation benötigen, die im Prinzip über den gesamten Raum möglicher Parameterschätzwerte reicht, der durch das Stichprobendesign erzeugt wird. Dieser Gedankengang zeigt, daß die zunächst harmlos klingende lineare Approximation des Gradienten in ihren möglichen Folgen praktisch nicht einschätzbar ist.[49]

Wir folgen hier der üblichen Vorgehensweise, bei der $\hat{\theta}$ mit dem zu schätzenden Parameter θ° verglichen wird. Indem man in (5.13) θ° an die Stelle von $\hat{\theta}$ und $\delta = \hat{\theta}-\theta^\circ$ setzt, erhält man

$$G(\hat{\theta}) \;=\; G(\theta^\circ) + H(\theta^\circ)\,(\hat{\theta}-\theta^\circ) + R$$

woraus dann, wenn man (5.12) berücksichtigt und das Restglied R vernachlässigt, folgende Approximation resultiert:

$$H(\theta^\circ)\,(\theta^\circ-\hat{\theta}) \;\approx\; G(\theta^\circ) \tag{5.14}$$

Diese durchaus problematische Approximation kann nun verwendet werden, um approximative Aussagen über die Verteilung von $\hat{\theta}$ zu gewinnen. Die Überlegung erfolgt in zwei Schritten.

In einem ersten Schritt wird die rechte Seite von (5.14) betrachtet, also $G(\theta^\circ)$. Erinnert man sich an (5.10) bzw. (5.11), sieht man, daß es sich – da der Parametervektor θ° fest vorgegeben ist – um eine Summe unabhängiger Zufallsvektoren handelt. Also kann man, gestützt auf eine geeignete Variante des zentralen Grenzwertsatzes annehmen, daß bei hinreichend großer durchschnittlicher Stichprobengröße die *Stichprobenverteilung* von $G(\theta^\circ)$ näherungsweise einer multivariaten Normalverteilung folgt.[50] Natürlich kann keine allgemeine Aussage darüber getroffen werden, wie gut diese Approximation ist, denn dies hängt auf eine praktisch undurchschaubare Weise vom Stichprobendesign und von der mathemati-

[49] Es sei betont, daß dieses Problem daraus resultiert, daß wir hier eine strikt endliche Betrachtungsweise zugrundelegen; ihre mathematisch strenge Lösung *beruht* hingegen darauf, daß ein Grenzübergang vollzogen werden kann.

[50] Die Überlegungen, die zu dieser Aussage führen, sind ziemlich kompliziert und sollen hier deshalb nicht näher dargestellt werden. In unserem Fall haben wir es mit einer Summe unabhängiger (verallgemeinerter) Binomialverteilungen zu tun. Überlegungen, wie die Grenzwertsatze der mathematischen Statistik auf diese Situation angewendet werden können, finden sich zum Beispiel bei Fisz [1976, S. 246f].

schen Form des zu schätzenden Modells ab.[51]

Unabhängig von dieser Approximation läßt sich jedoch auf einfache Weise zeigen, daß $G(\theta^\circ)$ den Erwartungswert Null hat. Um die Notation zu vereinfachen, führen wir folgende Abkürzung ein:

$$U_{ij}(\theta) \;=\; \frac{\partial}{\partial \theta_j}\, \log(\tilde{g}(z_i, \theta)) \tag{5.15}$$

Außerdem sei $U_i(\theta)$ der Spaltenvektor der Dimension m, der aus diesen Komponenten besteht, also $U_i(\theta) = [U_{i1}(\theta), \ldots, U_{im}(\theta)]'$.[52] Unter Verwendung dieser Abkürzungen kann man den Erwartungswert von $G(\theta^\circ)$ folgendermaßen schreiben:

$$\mathrm{E}\left\{G(\theta^\circ)\right\} \;=\; \sum_{i=1}^{N} \mathrm{E}\{I_i\}\, U_i(\theta^\circ) \;=\; \frac{n}{N} \sum_{i=1}^{N} U_i(\theta^\circ)$$

Die ganz rechts stehende Summe ist jedoch 0, denn dies ist eine notwendige Bedingung für die Definition von θ° als Maximum von (5.9). Also gilt für den Erwartungswert $\mathrm{E}\left\{G(\theta^\circ)\right\} = 0$. Bezeichnen wir die Kovarianzmatrix von $G(\theta^\circ)$ mit $J^*(\theta^\circ)$, kann als Zwischenergebnis festgehalten werden:

$$G(\theta^\circ) \;\sim\; \mathcal{N}\big(0, J^*(\theta^\circ)\big)$$

zu lesen als: $G(\theta^\circ)$ ist näherungsweise (multivariat) normalverteilt mit dem Erwartungswert 0 und der Kovarianzmatrix $J^*(\theta^\circ)$. Bevor auf die Frage eingegangen wird, wie diese Kovarianzmatrix näherungsweise berechnet werden kann, soll zunächst überlegt werden, was mit diesem Zwischenergebnis über die Verteilung von $(\theta^\circ - \hat{\theta})$ ausgesagt werden kann, denn darin liegt schließlich das Ziel der ganzen Betrachtung. Ausgehend von (5.14) liegt es nahe, $(\theta^\circ - \hat{\theta})$ als Ergebnis einer linearen Transformation von $G(\theta^\circ)$ zu betrachten, also

$$(\theta^\circ - \hat{\theta}) \;\approx\; H(\theta^\circ)^{-1}\, G(\theta^\circ) \tag{5.16}$$

denn dann folgt aus den Eigenschaften der multivariaten Normalverteilung, daß auch $(\theta^\circ - \hat{\theta})$ näherungsweise normalverteilt ist, mit dem Erwartungswert Null und einer aus $H(\theta^\circ)^{-1}$ und $J^*(\theta^\circ)$ berechenbaren Kovarianzmatrix. Allerdings stellen sich zwei Probleme. Erstens ist nicht sicher, daß die Hesse-Matrix $H(\theta^\circ)$ invertierbar ist; zweitens handelt es sich um eine Zufallsmatrix, d.h. jede Stichprobe liefert eine unterschiedliche Hesse-Matrix. Beide Probleme hängen zusammen. Wir wissen zwar (aufgrund der vorausgesetzten Annahmen), daß $H(\theta)$ in einer Umgebung von $\hat{\theta}$ negativ definit und infolgedessen invertierbar ist. Aber θ° kann bei einer

[51] Eine Einführung in bisherige Untersuchungen dieses Approximationsproblems gibt Spanos [1986, S. 202ff].

[52] Das Zeichen $'$ bezeichnet hier die Transposition eines Vektors oder einer Matrix.

gegebenen Stichprobe beliebig weit von $\hat{\theta}$ entfernt sein; insbesondere kann $H(\theta^\circ)$ nicht invertierbar sein. Wiederum handelt es sich um ein Problem, das sich bei einem Grenzübergang $n \to \infty$ nicht stellt. Aber auch wenn wir annehmen, daß die Invertierbarkeit gewährleistet ist, bleibt noch das Problem, daß es sich um eine Zufallsmatrix handelt; das heißt, (5.16) zeigt zunächt nur, daß $(\theta^\circ - \hat{\theta})$ das Produkt aus einer Zufallsmatrix und einem Zufallsvektor ist. Ohne einen Grenzübergang zu vollziehen, läßt sich über die resultierende Verteilung keine handhabbare Aussage treffen.

Bei einer asymptotischen Betrachtungsweise kann man an dieser Stelle den stochastischen Grenzwert von $H(\theta^\circ)$ betrachten. Hält man sich jedoch an das vorgegebene Stichprobendesign, muß ein anderer Weg gefunden werden, um aus $H(\theta^\circ)$ eine deterministische Matrix zu machen. Eine naheliegende Möglichkeit besteht darin, ihren Erwartungswert zu betrachten. Es ist jedoch klar, daß dadurch wiederum eine praktisch nicht kontrollierbare Approximation eingeführt wird.

Läßt man sich auf diesen Gedankengang ein, erhält man

$$H^*(\theta^\circ)\,(\theta^\circ - \hat{\theta}) \;\approx\; G(\theta^\circ)$$

wobei $H^*(\theta^\circ) = \mathrm{E}\{H(\theta^\circ)\}$ den Erwartungswert von $H(\theta^\circ)$ bezeichnet. Setzt man außerdem die Invertierbarkeit von $H^*(\theta^\circ)$ voraus, erhält man

$$(\theta^\circ - \hat{\theta}) \;\sim\; \mathcal{N}\big(0, H^*(\theta^\circ)^{-1}\, J^*(\theta^\circ)\, H^*(\theta^\circ)^{-1}\big)$$

das heißt: die ML-Schätzungen $\hat{\theta}$ sind näherungsweise normalverteilt mit dem Mittelwert θ° und der Kovarianzmatrix

$$\Sigma \;=\; H^*(\theta^\circ)^{-1}\, J^*(\theta^\circ)\, H^*(\theta^\circ)^{-1}$$

Nach all diesen praktisch nicht kontrollierbaren Approximationen bleibt schließlich noch die Frage, wie die Kovarianzmatrix Σ berechnet werden kann. Auch dies kann nur näherungsweise erfolgen, denn Σ nimmt Bezug auf Erwartungswerte (definiert durch das zugrundeliegende Stichprobendesign), für eine Berechnung steht jedoch nur die tatsächlich realisierte Stichprobe zur Verfügung.

Bei der praktischen Anwendung der asymptotischen ML-Theorie im Rahmen des Randomisierungskonzepts werden an dieser Stelle tatsächlich mehrere wesentliche Vereinfachungen und Abstraktionen vorgenommen. Betrachten wir zunächst $J^*(\theta^\circ)$. Da $\mathrm{E}\left\{G(\theta^\circ)\right\} = 0$, kann man, unter Verwendung der in (5.15) eingeführten Notation, das Element (j, j') dieser Kovarianzmatrix zunächst folgendermaßen schreiben:

$$J^*_{jj'}(\theta^\circ) \;=\; \mathrm{E}\left\{G_j(\theta^\circ) G_{j'}(\theta^\circ)\right\} \;=\; \sum_{i=1}^{N} \mathrm{E}\{I_i\}\, U_{ij}(\theta^\circ) \sum_{i'=1}^{N} \mathrm{E}\{I_{i'}\}\, U_{i'j'}(\theta^\circ)$$

Da $I_i^2 = I_i$ ist, kann man das Produkt der beiden Summen folgendermaßen

umformen:

$$J^*_{jj'}(\theta^\circ) \;=\; \frac{n}{N}\sum_{i=1}^{N} U_{ij}(\theta^\circ)\,U_{ij'}(\theta^\circ) + \frac{n^2}{N^2}\sum_{i\neq i'} U_{ij}(\theta^\circ)\,U_{i'j'}(\theta^\circ)$$

Hieraus ergibt sich die erste Approximation: es wird nur der erste Teil der Summe betrachtet, da der zweite Teil um einen Faktor n/N kleiner ist, also

$$J^*_{jj'}(\theta^\circ) \;\approx\; \frac{n}{N}\sum_{i=1}^{N} U_{ij}(\theta^\circ)\,U_{ij'}(\theta^\circ)$$

Allerdings kennt man θ° nicht, und man verfügt nur über die Daten aus einer Stichprobe S. Also kann man nur die Daten aus dieser Stichprobe verwenden und muß θ° durch einen mit der gegebenen Stichprobe berechenbaren Schätzwert $\hat{\theta}$ ersetzen. Dies liefert die in der Praxis verwendete Approximation

$$J^*_{jj'}(\theta^\circ) \;\approx\; \sum_{i\in S} U_{ij}(\hat{\theta})\,U_{ij'}(\hat{\theta}) \tag{5.17}$$

Zweitens enthält die Kovarianzmatrix Σ die Matrix $H^*(\theta^\circ)$. Ganz analog wird auch sie durch eine aus der gegebenen Stichprobe berechenbare Matrix approximiert. Zunächst kann man analog zu $U_{ij}(\theta)$ die Schreibweise

$$V_{i,jj'}(\theta) \;=\; \frac{\partial^2}{\partial\theta_j\,\partial\theta_{j'}}\,\log(\tilde{g}(z_i,\theta))$$

einführen. Dann erhält man für ein Element (j,j') der Matrix $H^*(\theta^\circ)$:

$$H^*_{jj'}(\theta^\circ) \;=\; \mathrm{E}\left\{H_{jj'}(\theta^\circ)\right\} \;=\; \sum_{i=1}^{N} \mathrm{E}\{I_i\}\,V_{i,jj'}(\theta^\circ) \;=\; \frac{n}{N}\sum_{i=1}^{N} V_{i,jj'}(\theta^\circ)$$

Durch eine Beschränkung auf die verfügbare Stichprobe S und eine Substitution von θ° durch den mit dieser Stichprobe berechenbaren Schätzwert $\hat{\theta}$ erhält man

$$H^*_{jj'}(\theta^\circ) \;\approx\; \sum_{i\in S} V_{i,jj'}(\hat{\theta}) \tag{5.18}$$

Aus (5.17) und (5.18) können Näherungen für $J^*(\theta^\circ$ und $H^*(\theta^\circ)$ berechnet werden; beides Zusammen liefert eine Näherung für Σ.[53]

[53] Es sei erwähnt, daß man unter gewissen vereinfachenden Annahmen bei der Schätzung eines Superpopulationsmodells die asymptotisch gültige Beziehung $J^*(\theta^\circ) = -H^*(\theta^\circ)$ beweisen kann; vgl. zum Beispiel White [1982]. Infolgedessen begnügt man sich bei der praktischen Anwendung der ML-Methode meistens damit, eine Kovarianzmatrix für die ML-Schätzwerte durch eine Näherung für $J^*(\theta^\circ)$ oder $-H^*(\theta^\circ)$ zu berechnen. Die genannte Beziehung gilt jedoch bei der ML-Schätzung deskriptiver Modelle im Rahmen des Randomisierungskonzepts in der Regel nicht.

5.3.3 Die Likelihoodkonzeption

Im vorangegangenen Abschnitt wurden einige der Probleme deutlich gemacht, die bei der Interpretation der ML-Methode auftreten, wenn man von der Randomisierungskonzeption ausgeht. Abgesehen von diesen Problemen kann man die Randomisierungskonzeption, d.h. die Auffassung, daß die Randomisierung bei der Stichprobenziehung eine geeignete *Grundlage* für statistische Inferenz liefert, auch grundsätzlich infrage stellen. Insbesondere zwei Argumente erscheinen mir wichtig.

Erstens ist es fragwürdig, ob die Randomisierung der Stichprobenziehung in irgendeiner Weise eine zusätzliche, für die inferenzstatistische Aufgabe bedeutsame Information liefert, die über den Informationsgehalt der jeweils gezogenen Stichprobe hinausgeht. Sehr deutlich wurde dies Problem von D. Basu formuliert:

> Let θ be the unknown state of nature and suppose there are k experiments $E_1, E_2, ..., E_k$ each of which may be performed to gain a meaningful quantum of information on θ. The scientist allots probabilities $\Pi_1, \Pi_2, \ldots, \Pi_k$ $(\sum \Pi_i = 1)$ (the Π_i's do not depend on θ) to the k experiments and thereby selects an experiment say E_s. He then performs the experiment E_s thereby generating the data x. It seems axiomatic to me that at the time of analyzing the data x, the scientist should refer the data to the experiment E_s that he has actually performed and forget about the other experiments that he might have performed. It follows that the selection probabilities $\Pi_1, \Pi_2, \ldots, \Pi_k$ have nothing to do with the analysis of the data.[54]

Zweitens ist darauf hingewiesen worden, daß Wahrscheinlichkeitsaussagen (über ein konstruiertes Stichprobendesign) und induktive Aussagen (über die Verteilung von Merkmalen in einer endlichen Grundgesamtheit) grundsätzlich unterschieden werden sollten.[55] Auf diesen Punkt wurde bereits zu Beginn des vorangegangenen Abschnitts hingewiesen. Auch wenn man im Sinne der Randomisierungskonzeption optimale Schätzfunktionen verwendet, hängt es schließlich von der tatsächlich verfügbaren Stichprobe ab, ob man zu näherungsweise korrekten Aussagen über die Grundgesamtheit gelangt. Oder anders gesagt: Regeln, die bei einer langen Serie von Zufallsexperimenten im Durchschnitt optimal sind, liefern nicht unbedingt in jedem einzelnen Anwendungsfall optimale Ergebnisse. In der empirischen Sozialforschung bezieht sich das statistische Inferenzproblem dagegen stets auf einen einzelnen Anwendungsfall.

[54] Diskussionsbemerkung zu Rao [1971, S. 190].

[55] Zum Beispiel bemerkt Royall [1983, S. 794]: „The reason why randomization is not a panacea is found in the distinction between probability and statistical inference. Randomization can guarantee probabilistic validiy. But a calculation can be probabilistically correct and inferentially wrong. Validity in one sense does not imply validity in the other.“

Eine zum Randomisierungskonzept alternative Auffassung des statistischen Inferenzproblems kann als Likelihoodkonzept bezeichnet werden. Der Kern dieser Konzeption besteht in dem sogenannten *Likelihoodprinzip*, das in einer Formulierung von Edwards [1992, S. 31] folgendermaßen lautet:

> Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other.

Dieses Likelihoodprinzip zeigt deutlich den Gegensatz zum Randomisierungskonzept. Während bei diesem die wesentliche Grundlage für inferenzstatistische Erwägungen in einer durch das Stichprobendesign definierten Wahrscheinlichkeitsverteilung liegt, beschränkt sich die Likelihoodkonzeption auf die tatsächlich realisierten bzw. verfügbaren Daten; die bei einer randomisierten Stichprobenziehung hypothetisch erzielbaren, jedoch nicht realisierten Stichproben enthalten demgegenüber keine inferenzstatistisch relevante Information.

Um die Likelihoodkonzeption zu verstehen, benötigt man vor allem ein Verständnis des Likelihoodbegriffs. Seine genaue Definition ist allerdings umstritten. Dies betrifft bereits den logischen Status des Begriffs. Für Bayesianisch orientierte Statistiker ist der Likelihoodbegriff identisch mit dem Begriff einer (subjektiven) Wahrscheinlichkeit. Dagegen betonen Statistiker, die mehr in der Tradition R. A. Fishers stehen, daß es einen grundsätzlichen Unterschied zwischen dem Likelihoodbegriff und dem Wahrscheinlichkeitsbegriff gibt: der Likelihoodbegriff bezieht sich auf statistische Hypothesen und liefert ein relatives Maß für das Vertrauen, das man in die Richtigkeit einer statistischen Hypothese setzen darf; der Wahrscheinlichkeitsbegriff bezieht sich dagegen (idealisierend) auf relative Häufigkeiten für das Auftreten von Ereignissen.[56]

Wichtiger ist die Frage, wie eine *operationale* Definition des Likelihoodbegriffs gegeben werden kann. Diese Frage ist kompliziert, weil sie unmittelbar mit dem Problem verknüpft ist, auf welche Art von Daten das Likelihoodprinzip sinnvoll anwendbar ist. Im vorliegenden Diskussionszusammenhang kann jedoch davon ausgegangen werden, daß eine randomisierte Stichprobenziehung als Rahmen vorgegeben ist. Der Likelihoodbegriff bezieht sich infolgedessen auf statistische Hypothesen (bzw. deskriptive Modelle) für eine endliche Grundgesamtheit; und die Daten, auf die mit diesem Begriff Bezug genommen wird, sind das Ergebnis einer Stichprobenziehung, die mit einem vorgegebenen Stichprobendesign durchgeführt worden ist. Dann kann, wie in Abschnitt 5.2 ausgeführt worden ist, von

[56] Eine ausführliche Darstellung der an Fisher orientierten Likelihoodkonzeption, verbunden mit kritischen Anmerkungen zum Bayesianischen Konzept, gibt Edwards [1992].

der Wahrscheinlichkeit von Stichproben gesprochen werden, und der Likelihoodbegriff muß sinnvollerweise auf diesen Wahrscheinlichkeitsbegriff bezogen werden.

Um den Likelihoodbegriff zu definieren, gehen wir hier also davon aus, daß ein Stichprobendesign $(\mathcal{S}, \mathcal{P})$ gegeben ist und daß in diesem Rahmen sinnvoll über die Wahrscheinlichkeit von Stichproben gesprochen werden kann. Die Hypothesen, auf die sich der Likelihoodbegriff bezieht, werden zunächst als parametrische Hypothesen über die Verteilung einer Zufallsvariable $Z = (X, Y)$ angesehen, die in der üblichen Weise für eine endliche Grundgesamtheit von Individuen definiert ist. Ihre Form entspricht also der eines parametrischen Modells $\tilde{g}(z; \theta)$; jedes $\theta \in \Theta$ entspricht einer bestimmten Hypothese über die Verteilung von Z.

Eine *vollständige* parametrische Hypothese kann dadurch definiert werden, daß ihre Kenntnis, zusammen mit einer Kenntnis des Stichprobendesigns, hypothetische Wahrscheinlichkeitsverteilungen für die Stichprobenziehung berechenbar macht. Ich verwende dafür die Schreibweise:

$$\mathcal{P}[\theta]\,(\,\pi = \{(i, z_i) \mid i \in S\}\,) \tag{5.19}$$

zu lesen als: die Wahrscheinlichkeit, mit der die Stichprobe $\pi = \{(i, z_i) \mid i \in S\}$ realisiert wird, wenn (a) die Stichprobenziehung auf der Grundlage des Stichprobendesigns $(\mathcal{S}, \mathcal{P})$ erfolgt, und wenn (b) die Hypothese $\tilde{g}(z; \theta)$ für die Verteilung von Z in Ω zutreffen würde. Ich wähle diese Notation (im Unterschied zur häufig verwendeten Schreibweise in der Form einer bedingten Wahrscheinlichkeit), um deutlich zu machen, daß $\mathcal{P}[\theta]$ ein *hypothetisches* Wahrscheinlichkeitsmaß ist, sich also vom Wahrscheinlichkeitsmaß $\mathcal{P}$ unterscheidet, das konstruktiv durch die Wahl eines Stichprobendesigns festgelegt und infolgedessen vollständig bekannt ist.

In diesem begrifflichen Rahmen kann dann die Likelihood der Hypothese θ (bzw. $\tilde{g}(z; \theta)$) folgendermaßen formal definiert werden:

$$\mathcal{L}(\theta; \pi) \;\propto\; \mathcal{P}[\theta]\,(\,\pi = \{(i, z_i) \mid i \in S\}\,) \tag{5.20}$$

Das heißt, die einer Hypothese θ durch die Daten π gegebene Likelihood ist proportional ($\propto$) zur hypothetischen Wahrscheinlichkeit für die Daten, wenn die Hypothese zutreffend wäre. Nimmt man diese Definition als Ausgangspunkt, folgt aus dem Likelihoodprinzip, daß eine Hypothese θ_1 durch die gegebenen Stichprobendaten π genau dann besser gestützt wird als eine Hypothese θ_2, wenn $\mathcal{L}(\theta_1, \pi) > \mathcal{L}(\theta_2, \pi)$.

Ob und wie eine Likelihoodkonzeption begründet werden kann, ist umstritten. Ausführliche Überlegungen zur Begründung wurden u.a. von Hacking [1965] und Edwards [1992] gegeben.[57] Sicherlich berechtigt ist

[57] Eine sehr umfassende Darstellung der Diskussionen zum Likelihoodprinzip haben Berger und Wolpert [1988] gegeben.

darauf hingewiesen worden, daß das Likelihoodprinzip nicht evident ist.[58] Insbesondere ist umstritten, ob und ggf. wie das Likelihoodkonzept eine Randomisierung bei der Datengewinnung erfordert. Denn es ist klar, daß Likelihoodbetrachtungen ganz unabhängig davon durchgeführt werden können, wie die jeweils verwendeten Daten zustandegekommen sind. Im vorliegenden Kontext kann jedoch davon ausgegangen werden, daß die Daten in Form einer Zufallsstichprobe gegeben sind. Insofern kann sich unsere Diskussion auf einen Teilaspekt des Problems beschränken: ob und ggf. wie das Stichprobendesign bei der Anwendung des Likelihoodprinzips berücksichtigt werden sollte. Darauf wird im nächsten Abschnitt näher eingegangen. In diesem Abschnitt soll zunächst der Zusammenhang zwischen der Likelihoodkonzeption und der ML-Methode betrachtet werden. Dabei gehen wir von der Aufgabenstellung einer deskriptiven Modellschätzung aus. Die zunächst wichtige Frage besteht also darin, welche Aussagen mithilfe einer Stichprobe über die Verteilung von Zufallsvariablen in der Grundgesamtheit erreicht werden können.

Ob und wie das erreicht werden kann, hängt in gewisser Weise von der Art der statistischen Hypothese ab. Es ist, auch zum Verständnis der Likelihoodkonzeption, aufschlußreich, sich dies kurz zu vergegenwärtigen.[59] Der Gedankengang beruht darauf, daß eine Hypothese über die Verteilung der Zufallsvariable Z in der endlichen Grundgesamtheit Ω folgendermaßen geschrieben werden kann:

$$\theta = (\theta_1, \ldots, \theta_N) \tag{5.21}$$

zu lesen als: $Z(\omega_i) = \theta_i$ für $i = 1, \ldots, N$. Hier werden also die unbekannten Merkmalsausprägungen bei den Individuen in der Grundgesamtheit als Parameter einer statistischen Hypothese aufgefaßt.

Um die Likelihood einer Hypothese dieser Art auf der Grundlage eines Stichprobendesigns $(\mathcal{S}, \mathcal{P})$ zu berechnen, muß man – für jede mögliche Stichprobe $\pi \in \Pi$ – die Wahrscheinlichkeit $\mathcal{P}[\theta](\pi)$ berechnen. Schreibt man die Stichproben, wie bisher, in der Form $\pi = \{(i, z_i) \mid i \in S\}$, kann man zu jeder möglichen Stichprobe eine Menge von mir ihr verträglichen Hypothesen definieren:

$$\Theta_\pi = \{(\theta_1, \ldots, \theta_N) \mid \theta_i = Z(\omega_i) \text{ für } i \in S\}$$

Dann erhält man

$$\mathcal{L}(\theta; \pi) \propto \mathcal{P}[\theta](\pi) = \begin{cases} \mathcal{P}(S) & \text{wenn } \theta \in \Theta_\pi \\ 0 & \text{andernfalls} \end{cases}$$

Dies besagt: Wenn eine Stichprobe $\pi = \{(i, z_i) \mid i \in S\}$ gegeben ist, haben

[58] Vgl. Kendall [1949, S. 111] sowie die ausführliche Diskussion in Kendall [1940].

[59] Der folgende Gedankengang stützt sich in erster Linie auf die Arbeiten von Basu [1969, 1971].

alle Hypothesen über die Verteilung von Z in der Grundgesamtheit, die mit der Stichprobe verträglich sind, die gleiche Likelihood. Die Art der Hypothesenformulierung, wie sie in (5.21) angegeben wurde, führt also nicht zu einem eindeutigen Ergebnis.

Man kann dies zunächst als eine Folge dessen ansehen, daß mit dieser Art der Hypothesenformulierung versucht wird, aus einer Stichprobe Informationen zu gewinnen, die nicht in ihr enthalten sind, nämlich Informationen über identifizierbare Individuen in der Grundgesamtheit. Dies ist intuitiv plausibel: Betrachtet man eine Grundgesamtheit als eine Menge identifizierbarer Individuen, liefert eine Stichprobe genau über den Teil der Individuen Informationen, die in der Stichprobe enthalten sind; über alle anderen Individuen liefert sie keinerlei Informationen.

Man kann jedoch einen weiteren Schritt tun, wenn man sich vergegenwärtigt, daß das Ziel der Analyse und der Modellbildung gar nicht darin besteht, Aussagen über identifizierbare Individuen zu machen, sondern in Wahrscheinlichkeitsaussagen über die *Verteilung* von Merkmalen in einer Grundgesamtheit (wobei an dieser Stelle deskriptive Wahrscheinlichkeitsaussagen gemeint sind). Dann folgt, daß die Hypothesenformulierung in (5.21) der Aufgabenstellung nicht angemessen ist. Das Ziel besteht vielmehr darin, Informationen über die unbekannten Wahrscheinlichkeiten $\mathrm{P}(Z = z)$ zu gewinnen.

Man kann darin ein wesentliches Merkmal *statistischer* Hypothesen sehen: Sie beziehen sich auf die Verteilung von Merkmalen und abstrahieren dadurch von der individuellen Zurechenbarkeit dieser Merkmale. Insofern liefert die in (5.21) angegebene Formulierung strenggenommen keine statistische Hypothese. Eine statistische Hypothese über die Verteilung der Zufallsvariable Z kann demgegenüber etwa folgendermaßen formuliert werden:

$$\theta = \left\{ \theta_z \;\middle|\; z \in \mathcal{Z}, 0 \leq \theta_z \leq 1, \sum_{z \in \mathcal{Z}} \theta_z = 1 \right\} \tag{5.22}$$

wobei θ_z die Hypothese über $\mathrm{P}(Z = z)$ ist. Es ist leicht zu sehen, daß Hypothesen dieser Art mit dem Likelihoodprinzip diskriminiert werden können.

Das Likelihoodprinzip impliziert, daß Stichprobeninformationen nur von Bedeutung sind, soweit sie die Likelihood der zu vergleichenden Hypothesen beeinflussen. Im Hinblick auf die in (5.22) formulierten Hypothesen zeigt sich, daß die relativen Häufigkeiten $h(z)$, mit denen die Werte der Zufallsvariable Z in der Stichprobe angenommen werden, eine hinreichende Information liefern.[60] Dem entspricht folgende Reformulierung der Stichprobendarstellung:

$$\pi = \{(i, z_i) \mid i \in S\} \longrightarrow \pi' = \{h(z) \mid z \in \mathcal{Z}\}$$

[60] In der statistischen Fachsprache wird von einer *suffizienten Statistik* gesprochen.

Schließlich bleibt nur noch zu überlegen, wie die Likelihood für die in (5.22) formulierten Hypothesen berechnet werden kann. Dies hängt im allgemeinen vom zugrundeliegenden Stichprobendesign ab. Das folgende Beispiel zeigt zunächst, wie die Likelihood für eine einfache Zufallsstichprobe berechnet werden kann.

Es soll die hypothetische Wahrscheinlichkeit $\mathcal{P}[\theta](\pi')$ für eine Stichprobe $\pi' = \{h(z) \mid z \in \mathcal{Z}\}$ berechnet werden, wobei θ entsprechend (5.22) eine Hypothese über die in der Grundgesamtheit gegebenen Wahrscheinlichkeiten $\mathrm{P}(Z = z)$ ist. Unter der Voraussetzung eines einfachen Stichprobendesigns kann man die Stichprobe durch unabhängige Zufallsvariable $Z_1, \ldots, Z_n$ darstellen und annehmen, daß die Verteilung jeder dieser Zufallsvariablen mit der Verteilung von Z in der Grundgesamtheit Ω identisch ist. Jede mögliche Realisation dieser Zufallsvariablen liefert relative Häufigkeiten $h(z)$, die angeben, wie oft der Wert $z \in \mathcal{Z}$ bei den insgesamt n Werten vorkommt. Also kann für jede vorgegebene Menge $\{h(z) \mid z \in \mathcal{Z}\}$ berechnet werden, mit welcher Wahrscheinlichkeit die in ihr angegebenen Häufigkeiten $h(z)$ realisiert werden. Da die Verteilung der Zufallsvariablen $Z_1, \ldots, Z_n$ durch die Hypothese vollständig spezifiziert wird, handelt es sich um eine rein deduktiv lösbare Aufgabe. In diesem Beispiel erhält man eine Multinomialverteilung:[61]

$$\mathcal{P}[\theta](\pi') \;=\; \frac{n!}{\prod_{z\in\mathcal{Z}} h(z)!} \prod_{z\in\mathcal{Z}} \theta_z^{h(z)}$$

Aus dieser hypothetischen Wahrscheinlichkeit folgt nun unmittelbar die Likelihood von θ. Sie ist, vgl. die Definition (5.20), proportional zu $\mathcal{P}[\theta](\pi')$, also kann der konstante Faktor in der Likelihoodformulierung unberücksichtigt bleiben, und die Log-Likelihood kann in der Form

$$\ell(\theta; \pi') \;=\; \sum_{z\in\mathcal{Z}} h(z)\, \log\left(\theta_z\right)$$

geschrieben werden. Die Maximierung dieser Log-Likelihood liefert die Lösungen

$$\hat{\theta}_z = h(z)$$

D.h. unter den genannten Voraussetzungen sind die aus der Stichprobe ermittelbaren relativen Häufigkeiten $h(z)$ mit den Maximum-Likelihood-Schätzwerten für die Wahrscheinlichkeiten $\mathrm{P}(Z = z)$ identisch.

Der Gedankengang hat zunächst gezeigt, daß bei der Formulierung statistischer Hypothesen von der Identität der Individuen abstrahiert werden muß.[62] Er zeigt darüber hinaus, daß jeder Versuch, zu einer partiellen

[61] Vgl. zum Beispiel Feller [1957, S. 157].

[62] Es ist allerdings eine offene Frage, ob die in einer Stichprobe ggf. enthaltenen Identifikationsnummern der Individuen eine für statistische Hypothesen relevante Informationsquelle darstellen können; vgl. hierzu den Beitrag von Rao [1971].

Rationalisierung des statistischen Inferenzproblems zu gelangen, irgendein Prinzip benötigt, um die Informationen aus einer Stichprobe mit den gewünschten Aussagen über die Grundgesamtheit zu verknüpfen. Bei der Randomisierungskonzeption liegt dies Prinzip darin, daß die Verteilung einer Zufallsvariable Z in der Grundgesamtheit Ω mit einer aus dem Stichprobendesign ableitbaren Verteilung von Stichprobenvariablen identifiziert wird. Liefert ein Stichprobendesign die unabhängigen Stichprobenvariablen $Z_1, \ldots, Z_n$, kann man demzufolge von der Beziehung

$$\mathrm{P}\left(Z = z\right) \equiv \mathcal{P}(Z_i = z) \qquad \text{für} \quad z \in \mathcal{Z} \quad \text{und} \quad i = 1, \ldots, n \qquad (5.23)$$

ausgehen. Diese Beziehung ist intuitiv einleuchtend und zunächst unabhängig von dem Gegensatz zwischen Randomisierungs- und Likelihoodkonzept. Der Gegensatz beginnt erst bei der Frage, wie man zu Aussagen über $\mathrm{P}(Z = z)$ gelangen kann. Die Randomisierungskonzeption versucht, Aussagen über $\mathrm{P}(Z = z)$ auf Aussagen über $\mathcal{P}(Z_i = z)$ zurückzuführen. Dies erscheint sinnvoll, wenn und insoweit empirische Einsichten in $\mathcal{P}(Z_i = z)$ gewonnen werden können; also in Situationen, in denen ein Zufallsexperiment sehr oft wiederholt werden kann. Wenn dies nicht der Fall ist, liefert die Beziehung (5.23) jedoch keine Information über $\mathrm{P}(Z = z)$, und jeder Versuch, sie gleichwohl zur Grundlage von Aussagen über die Verteilung von Z zu machen, wird problematisch.

Die Likelihoodkonzeption erscheint demgegenüber gerade dann sinnvoll, wenn der datengenerierende Prozeß nicht wiederholbar ist, also insbesondere bei nur einmalig realisierbaren Zufallsstichproben. Die Beziehung (5.23) hilft dann nicht weiter, sondern Aussagen über $\mathrm{P}(Z = z)$ können nur auf die tatsächlich verfügbaren Daten, die jeweils realisierte Stichprobe – oder, bei hinreichender Vergleichbarkeit, die bisher realisierten Stichproben – gegründet werden. Das Likelihoodprinzip trägt dieser Situation Rechnung, indem es statistische Hypothesen auf der Grundlage einer gegebenen Menge an Daten vergleichbar macht.

Das Likelihoodprinzip ist jedoch nicht evident; insbesondere ist nicht unmittelbar klar, wie es die Daten einer gegebenen Stichprobe mit der zugrundeliegenden Grundgesamtheit verknüpft. Als Maximum-Likelihoodprinzip wird es manchmal so interpretiert, daß es auf der Annahme beruht, „that the most likely event has happened“ (Kendall [1949, S. 111]). Kendall fügt jedoch sogleich hinzu, daß es sich – in dieser Form – um eine ziemlich unsinnige Annahme handelt. Denn bereits der Begriff einer Wahrscheinlichkeitsverteilung impliziert, daß nicht immer dasjenige Ereignis eintritt, das der größten Wahrscheinlichkeitsdichte korrespondiert. Um das Likelihoodprinzip für statistische Inferenzprobleme in endlichen Grundgesamtheiten sinnvoll interpretieren zu können, kommt man, wie ich glaube, ohne die Vorstellung der Repräsentativität von Stichproben nicht aus. Akzeptiert man diesen Begriff, obwohl eine genaue Definition unmöglich zu sein scheint, kann man eine zu (5.23) analoge

Basisannahme für die Likelihoodkonzeption folgendermaßen formulieren:

$$\mathrm{P}\,(Z = z) \;\approx\; \mathrm{P}\,(Z = z \mid S) \qquad z \in \mathcal{Z} \tag{5.24}$$

Hierdurch wird angenommen, daß die Verteilungen von Z in der Grundgesamtheit und in der jeweils vorliegenden Stichprobe S ähnlich sind, daß die Stichprobe repräsentativ ist. Es handelt sich um eine Annahme, da sie (trotz ihrer unpräzisen Formulierung) infrage gestellt und in gewissen Grenzen auch geprüft werden kann.

Zur Begründung, daß eine Annahme dieser Art erforderlich ist, kann noch einmal auf das oben (S. 306) gegebene Beispiel Bezug genommen werden. Dort wurde gezeigt, daß bei einer einfachen Zufallsstichprobe die relativen Häufigkeiten von Z in der Stichprobe, also $h(z)$, zugleich die ML-Schätzwerte für die Wahrscheinlichkeiten $\mathrm{P}(Z = z)$ sind. Das Likelihoodprinzip sagt uns jedoch nur, daß infolgedessen die relativen Häufigkeiten $h(z)$ die – in der Terminologie dieses Prinzips – am besten gestützten Hypothesen über die Wahrscheinlichkeiten $\mathrm{P}(Z = z)$ sind. Der bloße Glaube an das Likelihoodprinzip liefert allerdings keine Gründe, um den am besten gestützten Hypothesen auch Vertrauen zu schenken. Solche Gründe können schließlich nur darin liegen, daß mit ihnen die in (5.24) formulierte Annahme begründet werden kann.

Man könnte einwenden, daß die Annahme (5.24) das Likelihoodprinzip überflüssig machen würde. Das wäre jedoch ein Mißverständnis. Aus meiner Sicht liefert (5.24) zunächst eine partielle Rationalisierung der Verwendung des Likelihoodprinzips für Inferenzprobleme in endlichen Grundgesamtheiten (einen Ersatz für das Likelihoodprinzip kann diese Annahme schon deshalb nicht liefern, weil sie keine hinreichend präzise Formulierung zuläßt). Darüber hinaus macht (5.24) auf eine offene Flanke des Likelihoodprinzips aufmerksam. Sie besteht darin, daß es für das Likelihoodprinzip gleichgültig ist, wie die Daten zustandegekommen sind. Oder genauer gesagt: Das Likelihoodprinzip kann die Frage, wie die Daten zustandegekommen sind, nur dadurch und insoweit berücksichtigen, wie dies in den mithilfe der Daten zu vergleichenden Hypothesen formuliert wird. Es liefert insbesondere keinerlei Anhaltspunkte für die Frage, ob und ggf. wie das Stichprobendesign bei der Hypothesenformulierung bzw. bei der Anwendung der ML-Methode zu berücksichtigen ist. Der Hinweis auf (5.24) macht demgegenüber darauf aufmerksam, daß wir (bei Inferenzproblemen für endliche Grundgesamtheiten) nicht an Hypothesen über einen nur spekulativ definierbaren datengenerierenden Prozeß interessiert sind, sondern an Hypothesen über die Grundgesamtheit, aus der die Daten kommen.

5.3.4 Likelihoodprinzip und ML-Methode

Wie kommt man vom Likelihoodprinzip zur ML-Methode? In gewisser Weise ist dies trivial, denn die ML-Methode sagt, daß man zu einer optimalen Modellschätzung durch die Maximierung der Likelihood des Modells

auf der Grundlage einer gegebenen Stichprobe gelangt. Insofern liefert die ML-Methode die im Sinne des Likelihoodprinzips durch die Daten am besten gestützte Hypothese über einen unbekannten Parametervektor. Es ist jedoch zweckmäßig, den Zusammenhang noch etwas genauer zu betrachten.

a) Die ML-Methode ergibt sich als eine unmittelbare Schlußfolgerung aus dem Likelihoodprinzip zunächst dann, wenn zwei Bedingungen erfüllt sind. Die erste Bedingung bezieht sich auf das Stichprobendesign, mit dem die für die Modellschätzung zu verwendende Stichprobe erzeugt worden ist. Es muß gelten, daß es sich um eine *einfache* Zufallsstichprobe handelt; d.h. die Wahrscheinlichkeit, in die Stichprobe aufgenommen zu werden, ist für alle Individuen der Grundgesamtheit gleich groß, und die Ziehung der Individuen erfolgt unabhängig voneinander. Die zweite Bedingung besteht darin, daß das zu schätzende Modell $\tilde{g}(z;\theta)$ als eine *vollständige* statistische Hypothese über die Zufallsvariable Z angesehen werden kann. Damit ist gemeint, daß das Modell eine mit Ausnahme des unbekannten Parametervektors θ vollständige Beschreibung der Verteilung von Z liefert. Wenn diese beiden Bedingungen erfüllt sind, erhält man zunächst für jedes beliebige Individuum $\omega_i \in \Omega$ die Likelihoodformulierung

$$\tilde{g}(z_i;\theta) \;\propto\; \mathcal{P}[\theta]\,(\{(i, z_i)\}\,)$$

Da insbesondere vorausgesetzt worden ist, daß die Stichprobenziehung der Individuen unabhängig voneinander erfolgt, ergibt sich daraus für jede Stichprobe $\{(i, z_i) \mid i \in S\}$ die Äquivalenz der in (5.20) und (5.8) angegebenen Likelihoodformulierungen:

$$\mathcal{L}(\theta;\pi) \;=\; \prod_{i \in S} \tilde{g}(z_i;\theta) \;\propto\; \mathcal{P}[\theta]\,(\{(i, z_i) \mid i \in S\}\,)$$

Diese Überlegung zeigt, daß die für die ML-Methode übliche Likelihoodformulierung, wie sie in (5.8) angegeben worden ist, darauf beruht, daß eine einfache Zufallsstichprobe verfügbar ist. Genauer gesagt: Nur unter dieser Voraussetzung kann die übliche Likelihoodformulierung aus der Likelihoodkonzeption abgeleitet werden (die allgemeinen Prinzipien der Likelihoodkonzeption sind natürlich nicht von dieser Voraussetzung abhängig). Es ist infolgedessen eine durchaus sinnvolle Frage, ob und ggf. wie die für die ML-Methode übliche Likelihoodformulierung modifiziert werden sollte, wenn keine einfache Zufallsstichprobe vorliegt.

b) Ein zweites Problem betrifft die statistischen Hypothesen. In der Regel liefert die Modellspezifikation keine vollständige Charakterisierung der Verteilung der Zufallsvariablen $Z = (X, Y)$, sondern betrifft nur einen Aspekt dieser Verteilung. Dies ist insbesondere immer dann der Fall, wenn sich das Modell auf eine bedingte Wahrscheinlichkeitverteilung bezieht. Denn hat man ein Modell $\tilde{g}(x, y;\theta)$ für $\mathrm{P}(Y = y \,|\, X = x)$, kann daraus

nicht ohne weiteres das für die Likelihoodformulierung erforderliche hypothetische Wahrscheinlichkeitsmaß $\mathcal{P}[\theta]$ berechnet werden.

Das Likelihoodprinzip bietet jedoch eine einfache Lösung für dieses Problem, da unmittelbar die Likelihoods für bedingte Wahrscheinlichkeiten verglichen werden können. Man sieht dies folgendermaßen. Angenommen, die Modellbildung geht von folgender Zerlegung aus:

$$\mathrm{P}\,(X = x, Y = y) \;=\; \mathrm{P}\,(Y = y \mid X = x)\,\mathrm{P}\,(X = x)$$

Das Modell $\tilde{g}(x, y; \theta)$, das man schätzen möchte, bezieht sich auf die bedingte Wahrscheinlichkeit $\mathrm{P}(Y = y \mid X = x)$. Dann kann man für $\mathrm{P}(X = x)$ ein beliebiges Modell annehmen, etwa $\tilde{g}'(x)$, und erhält als Likelihood für den Modellparameter θ, bei einer gegebenen Stichprobe π, den Ausdruck

$$\mathcal{L}(\theta, \pi) \;\propto\; \prod_{i \in S} \tilde{g}(x_i, y_i; \theta) \prod_{i \in S} \tilde{g}'(x_i)$$

Wenn das hilfsweise eingeführte Modell $\tilde{g}'(x)$ nicht von dem Parametervektor θ abhängt, ist offensichtlich die relative Likelihood verschiedener Hypothesen über θ, und mithin die ML-Schätzung dieses Parameters, vollständig unabhängig davon, welches Modell man für $\mathrm{P}(X = x)$ gewählt hat.

c) Schließlich ist das bereits mehrfach erwähnte Problem zu berücksichtigen, daß sich bei der ML-Schätzung deskriptiver statistischer Modelle (im Unterschied zu Superpopulationsmodellen) die vorausgesetzte Modellklasse nicht unmittelbar als eine Klasse statistischer Hypothesen über die Verteilung von Zufallsvariablen in der Grundgesamtheit auffassen läßt. Im Rahmen der Likelihoodkonzeption kann diesem Problem auf einfache Weise Rechnung getragen werden. Die Grundidee besteht, wie in Abschnitt 5.1 bereits beschrieben wurde, in einer zweistufigen Rekonstruktion des Schätzproblems. Ausgangspunkt ist die Definition des zu schätzenden Modells für die endliche Grundgesamtheit, zum Beispiel durch die in (5.5) in Abschnitt 5.3 gegebene Formulierung:

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{P}\,(X = x, Y = y)\, \log\{\tilde{g}(x, y; \theta)\} \;\longrightarrow\; \max \tag{5.25}$$

Um das Modell mit den Daten einer Stichprobe zu schätzen, werden dann zunächst die Wahrscheinlichkeiten $\mathrm{P}(X = x, Y = y)$ geschätzt, dann wird das Kriterium (5.25) zur Modellberechnung auf der Grundlage der geschätzten Wahrscheinlichkeiten verwendet.

Bei einfachen Zufallsstichproben liefern, wie oben exemplarisch gezeigt wurde, die in der Stichprobe realisierten relativen Häufigkeiten $h(x, y)$ optimale Schätzwerte für die Wahrscheinlichkeiten $\mathrm{P}(X = x, Y = y)$. Die Modellberechnung erfolgt dann also mit dem Kriterium

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} h(x, y)\, \log\{\tilde{g}(x, y; \theta)\} \;\longrightarrow\; \max$$

Dies ist offensichtlich äquivalent zur üblichen Likelihoodformulierung auf der Grundlage von einfachen Zufallsstichproben, wie sie in Abschnitt 5.3 dargestellt wurde. Es bleibt jedoch die Frage, ob diese Likelihoodformulierung verändert werden muß, wenn es sich nicht um eine einfache Zufallsstichprobe handelt.

5.3.5 ML-Methode und Stichprobendesign

Ein möglicher Zugang zu diesem Problem kann folgendermaßen dargestellt werden.[63] Gegeben sei ein Stichprobendesign $(\mathcal{S}, \mathcal{P})$ zur Erfassung von Informationen über die Verteilung einer Zufallsvariable Z in der endlichen Grundgesamtheit Ω. Es sei angenommen, daß die mit diesem Stichprobendesign ziehbaren Stichproben folgendermaßen repräsentiert werden können:

$$\pi \;=\; \big(S, n(S) = n, Z_1 = z_1, \ldots, Z_n = z_n\big)$$

Hierbei ist $S \in \mathcal{S}$ die Indexmenge für die in π erfaßten Individuen, $n(S)$ ist eine Zufallsvariable für den Stichprobenumfang, und $z_1, \ldots, z_n$ sind die beobachteten Werte der Stichprobenvariablen $Z_1, \ldots, Z_n$. Dann kann zunächst rein formal folgende Darstellung der Stichprobenwahrscheinlichkeit betrachtet werden:

$$\mathcal{P}(\pi) = \mathcal{P}(Z_1 = z_1, \ldots, Z_n = z_n \mid S, n(S) = n)\, \mathcal{P}(S, n(S) = n)$$

Es liegt nahe, dies mit folgender Interpretation zu verbinden: *Wenn* die Stichprobenziehung, d.h. die Auswahl der Indexmenge S, unabhängig von der Verteilung der Zufallsvariable Z in der Grundgesamtheit erfolgt, kann das Stichprobendesign bei der ML-Schätzung eines Modells für diese Variable ignoriert werden. Die Frage ist jedoch, in welcher Weise an dieser Stelle von *Unabhängigkeit* gesprochen werden kann. Es ist zweckmäßig, zunächst ein Beispiel zu betrachten.

Es wird eine einfache geschichtete Stichprobe betrachtet. Die Grundgesamt gliedert sich in K Schichten, $\Omega = \Omega_1 \cup \cdots \cup \Omega_K$, und die Indexmenge der Stichprobe dementsprechend in $S = S_1 \cup \cdots \cup S_K$. N_k sei die Anzahl der Individuen in Ω_k, so daß $\sum_k N_k = N$. n_k sei der fest vorgegebene Stichprobenumfang von S_k, so daß die gesamte Stichprobe $n = \sum_k n_k$ Individuen umfaßt. Die Stichprobe liefert beobachtete Werte $z_i = Z(\omega_i)$ $(i \in S)$ einer in der Grundgesamtheit definierten Zufallsvariable Z. Es soll ein deskriptives Modell $\tilde{g}(z; \theta^\circ)$ geschätzt werden, das für die Grundgesamtheit durch

$$\sum_{z \in \mathcal{Z}} \mathrm{P}\,(Z = z)\, \log\{\tilde{g}(z; \theta)\} \;\longrightarrow\; \max$$

[63] Die Ausführungen stützen sich in vieler Hinsicht auf die Beiträge von Hoem [1985, 1989], allerdings gelange ich zu teilweise anderen Schlußfolgerungen.

definiert ist. Um dieses Modell mit den Daten der geschichteten Stichprobe zu schätzen, gibt es unterschiedliche Möglichkeiten.

Die erste Möglichkeit besteht darin, von der üblichen Form der Log-Likelihood auszugehen, d.h. von

$$\sum_{z\in\mathcal{Z}} h(z)\,\log\{\tilde{g}(z;\theta)\} \;\longrightarrow\; \max$$

wobei $h(z)$ die in der Stichprobe realisierten relativen Häufigkeiten der Werte von Z bezeichnet. Es liegt jedoch nahe, gegen diesen Ansatz der ML-Schätzung einzuwenden, daß er zu „verzerrten“ Schätzergebnissen führt, weil er nicht berücksichtigt, daß es sich um eine geschichtete Stichprobe handelt. Die Frage ist, wie mit diesem Einwand umgegangen werden soll.

Zunächst sieht man sofort, daß dieser Einwand auf einer Annahme beruht, nämlich auf der Annahme, daß sich die bedingten Wahrscheinlichkeiten $\mathrm{P}(Z=z\,|\,\Omega_k)$ zwischen den Teilgesamtheiten, aus denen die Teilstichproben gezogen worden sind, unterscheiden. Wenn diese Annahme richtig ist, gibt es wiederum unterschiedliche Möglichkeiten.

Eine Möglichkeit besteht darin, von unterschiedlichen Modellen für die Teilgesamtheiten auszugehen, im einfachsten Fall durch eine schichtspezifische Differenzierung der Parametervektoren. Dann ist also $\tilde{g}(z;\theta_k)$ ein Modell für die Verteilung von $\mathrm{P}(Z=z)$ in Ω_k, und man könnte versuchen, jedes dieser Modelle mit den zugehörigen Stichprobendaten zu schätzen. Da für jede Teilgesamtheit eine einfache Zufallsstichprobe angenommen wurde, kann dies mit der üblichen Likelihoodformulierung erreicht werden.

Was jedoch, wenn man – unabhängig von der Frage, ob die genannte Annahme richtig ist oder nicht – daran festhalten möchte, das Modell $\tilde{g}(z;\theta^\circ)$ zu schätzen? Es liegt dann nahe, eine *gewichtete* ML-Schätzung durchzuführen, wobei die inversen Inklusionswahrscheinlichkeiten als Gewichte in der Log-Likelihood verwendet werden. Dies folgt unmittelbar aus unserem bisherigen deskriptiven Verständnis der ML-Schätzung, nämlich:

$$\begin{aligned}
&\sum_{z\in\mathcal{Z}} \mathrm{P}\,(Z=z)\,\log\{\tilde{g}(z;\theta)\} = \\
&\quad \sum_{k=1}^{K} \frac{N_k}{N} \sum_{z\in\mathcal{Z}} \mathrm{P}\,(Z=z\mid\Omega_k)\,\log\{\tilde{g}(z;\theta)\} \;\equiv \\
&\quad \sum_{k=1}^{K} \frac{N_k}{N} \sum_{z\in\mathcal{Z}} h_k(z)\,\log\{\tilde{g}(z;\theta)\} = \\
&\quad \sum_{k=1}^{K} \frac{1}{N}\,\frac{1}{n_k/N_k} \sum_{i\in S_k} \log\{\tilde{g}(z_i;\theta)\} \;\propto \\
&\quad \sum_{i\in S} \frac{1}{p_i}\,\log\{\tilde{g}(z_i;\theta)\}
\end{aligned}$$

Hierbei bezeichnet $h_k(z)$ die relative Häufigkeit des Merkmalswerts z in der k.ten Teilstichprobe, und p_i ist die Inklusionswahrscheinlichkeit für das i.te Individuum. In unserem Beispiel einer einfachen geschichteten Stichprobe haben alle Individuen aus Ω_k die gleiche Inklusionswahrscheinlichkeit n_k/N_k. Die Ableitung zeigt jedoch, daß der Gedankengang für beliebige Inklusionswahrscheinlichkeiten verallgemeinert werden kann.

Das Beispiel läßt, wie ich glaube, die Schlußfolgerung zu, daß es bei einem deskriptiven Verständnis der Modellschätzung im allgemeinen sinnvoll ist, Stichprobengewichte (inverse Inklusionswahrscheinlichkeiten) zu verwenden. Dies liefert eine optimale Schätzung desjenigen Modells, das man berechnen könnte, wenn die Daten für die endliche Grundgesamtheit vollständig verfügbar wären.

Man könnte einwenden, daß es nicht erforderlich ist, Stichprobengewichte zu verwenden, wenn das zu schätzende Modell „korrekt spezifiziert“ ist. Dieser Einwand wird typischerweise dann gemacht, wenn sich das Interesse auf die Schätzung eines Superpopulationsmodells richtet. Der Einwand erscheint insofern berechtigt, als es dann darauf ankommt, ein Modell zu finden, daß möglichst alle relevanten Faktoren berücksichtigt, also auch diejenigen, die bei der Konzeption des Stichprobendesigns eine Rolle gespielt haben (sofern sie für den zu modellierenden Sachverhalt von Bedeutung sind). Aber selbst dann, wenn das Ziel in der Schätzung eines Superpopulationsmodells besteht, bleibt es stets eine nicht vollständig entscheidbare Frage, ob das Modell „korrekt spezifiziert“ worden ist. Einen Hinweis auf diese Frage kann dann ein Vergleich der ungewichteten mit gewichteten Schätzergebnissen liefern.

Für den in dieser Arbeit verfolgten Versuch, statistische Modelle für Lebensverlaufsdaten als Beschreibungen gewisser Regeln zu interpretieren, denen die Lebensverläufe in einer Gesamtheit von Individuen folgen,

ist diese Bezugnahme auf die Schätzproblematik von Superpopulationsmodellen jedoch nicht wesentlich. Denn wie in Abschnitt 4.1.7 zu begründen versucht wurde, wird bei einem deskriptiven Verständnis der Modellbildung in diesem Kontext von vielen möglicherweise wichtigen Bedingungen abstrahiert, d.h. es wird bewußt mit „fehlspezifizierten“ Modellen operiert.

Kapitel 6

Zusammenfassung

Thema der vorliegenden Arbeit sind die statistischen Begriffe, Methoden und Modelle, wie sie in der soziologischen Lebensverlaufsforschung verwendet werden. Es wurde versucht, sie in ihrem soziologischen Verwendungszusammenhang zu diskutieren, genauer gesagt: aus der Sicht einer Sozialstrukturanalyse, die Einsichten in gesellschaftliche Bedingungen individueller Lebensverläufe gewinnen möchte. Zum Abschluß sollen die wichtigsten Überlegungen noch einmal zusammengefaßt werden. Ich versuche dies in der Form eines „argumentativen Leitfadens“, der es ermöglichen soll, den Zusammenhang der Überlegegungen der vorausgegangenen Kapitel sichtbar zu machen.

1. Lebensverlaufsforschung bezieht sich auf die in einer Gesellschaft realisierten Lebensverläufe von Individuen. Lebensverläufe werden als zeitliche Abfolgen von Zuständen beschrieben, die durch Ereignisse (Zustandswechsel) zustande kommen. Als formaler Beschreibungsrahmen dient die Vorstellung eines Biographieschemas, das die (aus der jeweiligen Beobachtungsperspektive) möglichen Lebensverläufe definiert.

2. Individuelle Lebensverläufe können aus unterschiedlichen Perspektiven thematisiert werden. Eine dieser möglichen Perspektiven kann durch das Stichwort „Sozialstrukturanalyse“ charakterisiert werden, zunächst ganz allgemein definierbar durch ein Erkenntnisinteresse, das auf Einsichten in gesellschaftliche Verhältnisse als Bedingungen individueller Lebensverläufe zielt. Die vorliegende Arbeit geht von dieser Perspektive aus.

3. Jeder Versuch, soziale Bedingungen individueller Lebensverläufe sichtbar zu machen, benötigt gewisse Basisvorstellungen darüber, wie Lebensverläufe zustande kommen. In dieser Arbeit wird davon ausgegangen, daß zumindest einige der aus soziologischer Sicht relevanten Aspekte von Lebensverläufen als (transitorische) Resultate von Verhaltensweisen ihrer Subjekte verstanden werden können. Anders formuliert: individuelle Lebensverläufe werden als das Ergebnis individuell getroffener Entscheidungen angesehen. Damit wird weder unterstellt, daß sich alle Aspekte individueller Lebensverläufe durch diese Betrachtungsweise angemessen erfassen lassen (was sicherlich nicht der Fall ist), noch wird angenommen, daß es sich bei den Subjekten von Lebensverläufen um (in irgendeinem faßbaren Sinne des Wortes) „rationale“ Akteure handelt. Es wird nur davon ausgegangen, daß ein angemessenes Verständnis individueller Lebensverläufe nur erreicht werden kann, wenn – wie es in der alltäglichen Kommunikation geschieht – Handlungssubjekte unterstellt werden, die über Handlungsalternativen verfügen.

4. Aus soziologischer Sicht kann man an dieser Stelle einen weiteren Schritt tun und sagen, daß Handlungen, als deren Folge sich Lebensverläufe entwickeln, in sozialen Situationen stattfinden und durch sie bedingt werden. Damit stellt sich allerdings die Frage, wie diese gängige Vorstellung präzisiert und wie ein empirischer Zugang zu ihrem Verständnis gefunden werden kann. Das Problem hat mehrere Aspekte.

5. Die erste Frage ist, ob man sich vorstellen soll, daß das Verhalten der Individuen durch die sozialen Situationen, in denen sie sich jeweils befinden, determiniert wird. Ich gehe in dieser Arbeit davon aus, daß diese Vorstellung unzweckmäßig ist, weil sie dem Selbstverständnis der sozialen Akteure (und mithin den potentiellen Addressaten des soziologisch zu gewinnenden Wissens) widerspricht. Mit der in dieser Arbeit verwendeten Bezeichnung, daß individuelle Lebensverläufe *kontingent* sind, ist im wesentlichen gemeint, daß nicht von der Vorstellung ausgegangen wird, daß die sie generierenden Verhaltensweisen durch soziale Situationen determiniert werden. Man könnte vielleicht einwenden, daß der Eindruck einer solchen Kontingenz nur eine Folge mangelhafter Einsichten in die tatsächlichen Bedingungszusammenhänge menschlicher Verhaltensweisen sei. Das mag sein, aber ich glaube, daß es ein gewissermaßen methodologisches Gegenargument gibt: daß jeder sinnvolle Versuch, Bedingungen, insbesondere soziale Bedingungen von Verhaltensweisen sichtbar zu machen, in der Form einer Vergegenständlichung dieser Bedingungen vorgenommen werden muß, so daß infolgedessen ein Verhalten gegenüber oder im Hinblick auf diese Bedingungen vorstellbar wird. – Abgesehen von dieser eher philosophischen Frage erscheint jedenfalls im Hinblick auf die in der soziologischen Lebensverlaufsforschung typischerweise betrachteten sozialen Situationen die Annahme, daß sie das in diesen Situationen stattfindende Verhalten der Individuen determinieren, nicht sinnvoll.

6. Damit stellt sich für die soziologische Lebensverlaufsforschung die Frage, wie gleichwohl von sozialen Bedingungen individueller Lebensverläufe gesprochen werden kann. In dieser Arbeit versuche ich, einen Zugang zu dieser Frage durch die Vorstellung zu gewinnen, daß die Individuen in ihren Verhaltensweisen – und mithin in ihren Lebensverläufen – zumindest partiell sozialen Regeln folgen. Dies erscheint sinnvoll, um aus soziologischer Sicht einen nicht-deterministischen Bedingungszusammenhang vorstellbar zu machen. Diese Vorstellung ist damit vereinbar, daß das Verhalten der Individuen durch die sozialen Situationen bedingt wird, in denen sie sich jeweils befinden. Denn soziale Regeln sind typischerweise situationsspezifisch; sie charakterisieren mögliche Verhaltensweisen in sozialen Situationen. Dabei lasse ich ausdrücklich offen, ob man sich soziale Regeln als durch soziale Akteure internalisierte „soziale Normen“ vorstellen kann, was vielleicht bei einigen sozialen Regeln sinnvoll erscheint. Um von dieser Frage absehen zu können, spreche ich davon, daß sich Individuen an sozialen Regeln *orientieren*, wodurch zugleich betont wird, daß niemand durch soziale Regeln dazu gezwungen wird, ihnen zu folgen.

7. Diese Vorüberlegungen bilden den theoretischen Rahmen, um die potentielle Bedeutung statistischer Verfahren und Modelle für die soziologische Lebensverlaufsforschung lokalisieren zu können. Ich interpretiere sie in dieser Arbeit als Hilfsmittel, um *empirische*, d.h. auf einer systematischen Sammlung von Beobachtungen beruhende Einsichten in soziale Bedingungen individueller Lebensverläufe zu gewinnen.

8. Ein zentrales Problem für jeden Versuch, empirische Einsichten in die soziale Bedingtheit individueller Lebensverläufe zu finden, besteht in deren Kontingenz. Eine Möglichkeit, um sich zu den daraus resultierenden Schwierigkeiten zu verhalten. besteht darin, Lebensverläufe so zu beschreiben, *als ob* sie nicht kontingent wären. Diese Auffassung könnte vertreten werden, wenn es nur darum geht, die Entwicklung von Lebensverläufen konditional prognostizierbar zu machen. Denn bei dieser Zielsetzung ist die Kontingenz individueller Lebensverläufe nur ein Sachverhalt, der ihrer Voraussagbarkeit Grenzen setzt. Dieser Als-ob-Betrachtungsweise wird in der vorliegenden Arbeit nicht gefolgt; stattdessen wird nach einer Betrachtungsweise gesucht, die es erlaubt, von der Kontingenz individueller Lebensverläufe zu abstrahieren. Eine primäre Bedeutung statistischer Verfahren für die Lebensverlaufsforschung wird darin gesehen, daß mir ihrer Hilfe eine solche Abstraktion vollzogen werden kann.

9. Aus dieser Überlegung ergibt sich der in dieser Arbeit verfolgte Versuch, statistische Aussagen über Lebensverläufe als deskriptive Wahrscheinlichkeitsaussagen über endliche Gesamtheiten von Individuen zu interpretieren. Wie im Verlauf der Arbeit durchgängig betont worden ist, entsteht daraus ein Bedeutungswandel für Aussagen über Lebensverläufe. An die Stelle von Aussagen über die Lebensverläufe konkreter, jeweils bestimmter Individuen treten Aussagen über Eigenschaften von Gesamtheiten von Individuen. Dies entspricht einer soziologischen Betrachtungsweise, die an Aussagen über gesellschaftliche Verhältnisse interessiert ist, nicht an Aussagen über identifizierbare Individuen. Damit ist jedoch keineswegs zugleich eine Abstraktion von der Vielfalt der individuell realisierten Lebensverläufe verbunden. Die Verwendung deskriptiver Wahrscheinlichkeitsaussagen impliziert weder Quetelets *l'homme moyen* noch verpflichtet sie zur Konstruktion „typischer“ Lebensverläufe. Im Gegenteil, man erhält zunächst Aussagen über die *Verteilung* von Lebensverlaufsmerkmalen in einer endlichen Gesamtheit von Individuen.

10. Es bleibt die Frage, wie man zu empirischen Aussagen über soziale Bedingungen von Lebensverläufen gelangen kann. Die Statistik bietet den Begriff einer bedingten Wahrscheinlichkeitsverteilung an. Dieser Begriff liefert allerdings zunächst nur einen formalen Rahmen, um Aussagen über Bedingungen formulieren zu können. Tatsächlich liefert die Statistik keinerlei Kriterien, um zwischen sinnvollen und nicht sinnvollen Aussagen über Bedingungen unterscheiden zu können. Solche Kriterien können nur aus dem jeweils intendierten theoretischen Verwendungszusammenhang der statistischen Aussagen gewonnen werden. Allerdings liefert die

Statistik bzw. die sie begründende Wahrscheinlichkeitstheorie einen geeigneten formalsprachlichen Rahmen, um die potentielle theoretische Bedeutung von Aussagen über bedingte Wahrscheinlichkeitsverteilungen reflektieren zu können.

11. An dieser Stelle gewinnt die Tatsache eine entscheidende Bedeutung, daß Lebensverläufe in der Zeit ablaufende Prozesse sind. Infolgedessen kann ein einfaches, aber folgenreiches Prinzip zur Bildung bedingter Wahrscheinlichkeitsverteilungen zur Beschreibung von Lebensverläufen begründet werden: Nur Sachverhalte, die in der jeweiligen Vergangenheit eines zeitlich sich entwickelnden Prozesses vorhanden gewesen sind, können sinnvoll als Bedingungen seiner Entwicklung in der jeweiligen Gegenwart verwendet werden. Die elementare Form einer bedingten Wahrscheinlichkeitsverteilung, die diesem Prinzip genügt, liefert der Begriff der Übergangsrate. Dieser Begriff steht deshalb im Zentrum einer statistischen Beschreibung von Lebensverläufen. Dieser Begriff liefert zugleich eine – zunächst noch sehr allgemeine – Möglichkeit, um die statistische Beschreibung von Lebensverläufen mit soziologischen Vorstellungen über ihre soziale Bedingtheit zu verknüpfen. Denn das Konzept der Übergangsrate konditioniert explizit auf eine Ausgangssituation, den Beginn einer Episode bei einer Gesamtheit von Individuen; soziologisch interpretiert handelt es sich um eine soziale Situation, die den Ausgangspunkt für die weitere Entwicklung von Lebensverläufen definiert. Mithilfe des Begriffs zustandsspezifischer Übergangsraten kann dann beschrieben werden, wie sich ausgehend von dieser sozialen Situation die Lebensverläufe der Individuen im Hinblick auf mögliche neue Situationen entwickeln. Eine Beschreibung dieser Art liefert somit einen Ausgangspunkt für eine soziologische Interpretation sozialer Situationen als Bedingungen für die in dieser Situation sich entwickelnden Lebensverläufe.

12. Wichtig ist, daß bei einer statistischen Beschreibung von Lebensverläufen durch zustandsspezifische Übergangsraten zwar von der individuellen Zurechenbarkeit der jeweils individuell realisierten Lebensverläufe abstrahiert wird, nicht jedoch von ihrer Vielfalt. Beschrieben wird nicht, wie der Prozeß bei einem „durchschnittlichen“ oder „typischen“ Individuum verläuft, sondern wie er sich bei einer Gesamtheit von Individuen entwickelt. Daraus resultiert, wie ich zu zeigen versucht habe, eine wesentliche Differenz zwischen Übergangsratenmodellen und herkömmlichen Regressionsmodellen.

13. Wichtig erscheint auch, daß die Beschreibung von Lebensverläufen durch Übergangsraten einen einfachen Zugang zur Rekonstruktion der Vorstellung liefert, daß sich in der Entwicklung von Lebensverläufen zahlreiche Zustände und Ereignisse „wechselseitig bedingen“. Bei Modellansätzen, die nicht explizit eine Zeitdimension berücksichtigen, erzeugt diese Vorstellung äußerst komplizierte Probleme. Bei der statistischen Beschreibung von Prozessen kann sie jedoch auf einfache Weise berücksichtigt werden. Als formaler Rahmen dient die Vorstellung parallel ablaufender Prozesse. Aus dem Prinzip, daß nur die jeweils vergangene Prozeßentwicklung sinnvoll als Bedingung für die jeweils gegenwärtige Prozeßentwicklung angesehen werden kann, folgt unmittelbar ein Prinzip der „lokalen konditionalen Unabhängigkeit“. Mithilfe dieses Prinzips kann eine Form der Prozeßbeschreibung begründet werden, die separat jeden Teilprozeß in seiner Bedingtheit durch die übrigen Prozesse sichtbar macht.

14. Es bleibt schließlich die Frage, wie die Vorstellung, daß Lebensverläufe durch gesellschaftliche Verhältnisse bedingt werden, präziser gefaßt werden kann. Aus soziologischer Sicht gibt es, wie in der Einleitung ausgeführt wurde, zwei komplementäre Strategien. Eine Strategie versucht, hermeneutische Deutungen der sozialen Regeln zu erreichen, an denen sich soziale Akteure orientieren. Der Versuch zielt, allgemein gesprochen, darauf, einen subjektiv zurechenbaren Sinn darin sehen zu können, daß gewissen Regeln gefolgt wird. Eine komplementäre Strategie versucht, gesellschaftliche Verhältnisse als ein situationsbezogenes Geflecht von Handlungschancen zu beschreiben. Eine ereignisanalytisch orientierte statistische Beschreibung von Lebensverläufen kann beiden Strategien dienen, insofern der dabei verwendete Begriff bedinger Wahrscheinlichkeitsverteilungen für unterschiedliche theoretische Deutungen genutzt werden kann. Daraus ergibt sich, wie ich zu zeigen versucht habe, eine sinnvolle Abgrenzung statistischer Modelle und soziologischer Theoriebildung. Mit Hilfe statistischer Modell kann empirisches, also deskriptives Wissen darüber gewonnen werden, *wie* sich Lebensverläufe situationsspezifisch, also bedingt durch die jeweilige Situation, entwickeln.

Zweck dieser Zusammenfassung der vorliegenden Arbeit war, den Zusammenhang ihrer Leitgedanken zu verdeutlichen. Bei der Verfolgung dieser Leitgedanken treten natürlich zahlreiche Detailprobleme auf, teilweise bloß technische, teilweise aber auch grundsätzliche Probleme. Auf einige dieser Probleme wurde in den vorangegangenen Kapiteln mehr oder weniger ausführlich eingegangen. Hauptsächlich handelt es sich um Fragen, die daraus resultieren, daß für die empirische Lebensverlaufsforschung in der Regel nur Daten aus Stichproben verfügbar sind, die außerdem meisten unvollständig und ungenau sind. Die statistischen Inferenzprobleme, die daraus entstehen, sind kompliziert, und ihre Diskussion in der vorliegenden Arbeit konnte nur anstreben, sie als Probleme für die soziologische Lebensverlaufsforschung sichtbar zu machen.

Literaturverzeichnis

Aalen, O. O. (1987). Dynamic Modelling and Causality. Scandinavien Actuarial Journal 1987, 177 – 190

Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories. In: Sociological Methodology 1982, ed. by S. Leinhardt. San Francisco: Jossey-Bass 1982, 61 – 98

Allison, P. D. (1984). Event History Analysis. Regression for Longitudinal Event Data. Newbury Park: Sage 1984

Allison, P. D. (1985). Survival Analysis of Backward Recurrence Times. Journal of the American Statistical Association 80 (1985), 315 – 322

Anderson, O. (1965). Probleme der statistischen Methodenlehre in den Sozialwissenschaften. Würzburg: Physica 1965

Andreß, H.-J. (1985). Multivariate Analyse von Verlaufsdaten. Mannheim: ZUMA (Methodentexte, Band 1) 1985

Andreß, H.-J. (1992). Einführung in die Verlaufsdatenanalyse. Köln: Zentrum für historische Sozialforschung 1992

Aranda-Ordaz, F. J. (1983). An Extension of the Proportional-Hazards Model for Grouped Data. Biometrics 39 (1983), 109 – 117

Arminger, G. (1990). Testing against Misspecification in Parametric Rate Models. In: Event History Analysis in Life Course Research, ed. by K. U. Mayer, N. B. Tuma. Madison: University of Wisconsin Press 1990, 253 – 266

Ayer, A. J. (1956). The Problem of Knowledge. London: Penguin Books 1990 (Reprint)

Ayer, A. J. (1972). Probability and Evidence. London: Macmillan 1972

Bartholomew, D. J. (1977). The Analysis of Data arising from Stochastic Processes. In: The Analysis of Survey Data (Vol. 2), ed. by C. A. O'Muircheartaigh, C. Payne. New York: Wiley 1977, 145 – 174

Bartlett, N. R. (1978). A Survival Model for a Wood Preservative Trial. Biometrics 34 (1978), 673 – 679

Basu, D. (1969). Role of Sufficiency and Likelihood Principles in Sample Survey Theory. Sankhya 31 (1969), 441 – 454

Basu, D. (1971). An Essay on the Logical Foundations of Survey Sampling (Part I). In: Foundations of Statistical Inference, ed. by V. P. Godambe, D. A. Sprott. Toronto: Holt, Rinehart and Winston 1971, 203 – 242

Bauer, H. (1978). Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie. Berlin: de Gruyter 1978

Baydar, N., White, M. (1988). A Method for Analyzing Backward Recurrence Time Data on Residential Mobility. In: Sociological Methodology 1988, ed. by C. C. Clogg. San Francisco: Jossey-Bass 1988, 105 – 135

Becker, H. A. (1990). Dynamics of Life Histories and Generations Research. In: Life Histories and Generations. Proceedings of a Symposium held on 22 and 23 June 1989 at the Netherlands Institute for Advanced Studies in the Humanities and Social Sciences, at Wassenar. Ed. by H. A. Becker. University of Utrecht: ISOR / Faculty of Social Science 1990, 1 – 55

Benenson, F. C. (1984). Probability, Objectivity and Evidence. London: Routledge & Kegan Paul 1984

Berger, J. O., Wolpert, R. L. (1988). The Likelihood Principle (2nd Edition). Hayward/Cal.: Institute of Mathematical Statistics 1988

Berger, P. A., Hradil, S. (Hg.) (1990). Lebenslagen, Lebensläufe, Lebensstile (= Soziale Welt Sonderband 7). Göttingen: Schwartz 1990

Berger, P. L. (1963). Einladung zur Soziologie. Eine humanistische Perspektive. München: DTV 1977

Bergman, L. R., Eklund, G., Magnusson, D. (1991). Studying Individual Development: Problems and Methods. In: Problems and Methods in Longitudinal Research: Stability and Change. Ed. by D. Magnusson, L. R. Bergman, G. Rudinger, B. Törestad. Cambridge: University Press 1991, 1 – 27

Berk, R. A. (1983). An Introduction to Sample Selection Bias in Sociological Data. American Sociological Review 48 (1983), 386 – 398

Bernard, H. R., Killworth, P., Kronenfeld, D., Sailer, L. (1984). The Problem of Informant Accuracy: The Validity of Retrospective Data. Annual Review of Anthropology 13 (1984), 495 – 517

Birnbaum, A. (1977). The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; with a Criticism of the Lindley-Savage Argument for Bayesian Theory. Synthese 36 (1977), 19 – 49

Blossfeld, H.-P. (1987). Zur Repräsentativität der Sfb-3-Lebensverlaufsstudie. Ein Vergleich mit Daten aus der amtlichen Statistik. Allgemeines Statistisches Archiv 71 (1987), 126 – 144

Blossfeld, H.-P. (1989). Kohortendifferenzierung und Karriereprozeß. Eine Längsschnittstudie über die Veränderung der Bildungs- und Berufschancen im Lebenslauf. Frankfurt: Campus 1989

Blossfeld, H.-P. (1994). Causal Modelling in Event History Analysis. Paper prepared for the XIII World Congress of Sociology (Bielefeld, July 1994). Ms. Bremen 1994

Blossfeld, H.-P., Hamerle, A. (1989). Using Cox Models to Study Multiepisode Processes. Sociological Methods & Research 17 (1989), 432 – 448

Blossfeld, H.-P., Hamerle, A., Mayer, K. U. (1986). Ereignisanalyse. Statistische Theorie und Anwendungen in den Sozialwissenschaften. Frankfurt: Campus 1986

Blossfeld, H.-P., Hamerle, A., Mayer, K. U. (1989). Event History Analysis. Statistical Theory and Applications in the Social Sciences. Hillsdale, New Jersey: Lawrence Erlbaum 1989

Blossfeld, H.-P., Hamerle, A., Mayer, K. U. (1991). Event-history Models in Social Mobility Research. In: Problems and Methods in Longitudinal Research: Stability and Change. Ed. by D. Magnusson, L. R. Bergman, G. Rudinger, B. Törestad. Cambridge: University Press 1991, 212 – 235

Blossfeld, H.-P., Huinink, J. (1989). Die Verbesserung der Bildungs- und Berufschancen von Frauen und ihr Einfluß auf den Prozeß der Familienbildung. Zeitschrift für Bevölkerungswissenschaft 15 (1989), 383 – 404

Blossfeld, H.-P., Jaenichen, U. (1990). Bildungsexpansion und Familienbildung. Soziale Welt 41 (1990), 454 – 476

Blossfeld, H.-P., Klijzing, E., Rohwer, G. (1994). Modelling Parallel Processes in Demography. Paper prepared for the European Population Conference (Milano, September 1995). Ms. Bremen 1994

Blossfeld, H.-P., Manting, D., Rohwer, G. (1993). Patterns of Change in Family Formation in the Federal Republic of Germany and the Netherlands. PDOD-Paper No. 18. Amsterdam 1993

Blossfeld, H.-P., Nuthmann, R. (1990). Transition from Youth to Adulthood as a Cohort Process in the Federal Republic of Germany. In: Life Histories and Generations. Proceedings of a Symposium held on 22 and 23 June 1989 at the Netherlands Institute for Advanced Studies in the Humanities and Social Sciences, at Wassenar. Ed. by H. A. Becker. University of Utrecht: ISOR / Faculty of Social Science 1990, 183 – 217

Blossfeld, H.-P., Rohwer, G. (1994). Analysis of Interacting Life Course Processes. Ms. Bremen 1994

Bohman, J. (1991). New Philosophy of Social Science. Problems of Indeterminacy. Cambridge: Polity Press 1991

Bonß, W. (1982). Die Einübung des Tatsachenblicks. Zur Struktur und Veränderung empirischer Sozialforschung. Frankfurt: Suhrkamp 1982

Bonß, W. (1991). Unsicherheit und Gesellschaft – Argumente für eine soziologische Risikoforschung. Soziale Welt 42 (1991), 258 – 277

Boudon, R. (1979). Generating Models as a Research Strategy. In: Qualitative and Quantitative Social Research. Papers in Honor of Paul F. Lazarsfeld, ed. by R. K. Merton, J. S. Coleman, P. H. Rossi. New York: Free Press 1979, 51 – 64

Boudon, R. (1983). Individual Action and Social Change: A No-Theory of Social Change. British Journal of Sociology 34 (1983), 1 – 18

Boudon, R. (1984). Theories of Social Change. A Critical Appraisal. Cambridge: Polity Press 1986

Brose, H.-G. (1990). Berufsbiographien im Umbruch. Erwerbsverlauf und Lebensführung von Zeitarbeitnehmern. In: Lebensverläufe und sozialer Wandel, hrsg. von K. U. Mayer. Opladen: Westdeutscher Verlag 1990, 179 – 211

Brown, C. C. (1975). On the Use of Indicator Variables for Studying the Time-Dependence of Parameters in a Response-Time Model. Biometrics 31 (1975), 863 – 872

Brückner, E. (1990). Die retrospektive Erhebung von Lebensverläufen. In: Lebensverläufe und sozialer Wandel, hrsg. von K. U. Mayer. Opladen: Westdeutscher Verlag 1990, 374 – 403

Brüderl, J., Diekmann, A. (1994a). The Log-Logistic Rate Model: Two Generalizations with an Application to Demographic Data. München / Bern: mimeo 1994

Brüderl, J., Diekmann, A. (1994b). Bildung, Geburtskohorte und Heiratsalter. Zeitschrift für Soziologie 23 (1994), 56 – 73

Cancian, F. (1976). Norms and Behavior. In: Explorations in General Theory on Social Sciences, ed. by J. J. Loubser, R. C. Baum, A. Effrat, V. Meyer Lidz. Vol. I, 354 - 366. New York: Free Press 1976

Carlsson, G. (1951). Sampling, Probability and Causal Inference. Theoria 17 (1951), 139 – 154

Carnap, R. (1945). The Two Concepts of Probability. Philosophy and Phenomenological Research 5 (1945), 513 – 532

Cochran, W. G. (1977). Sampling Techniques. New York: Wiley 1977

Coleman, J. S. (1968). The Mathematical Study of Change. In: Methodology in Social Research, ed. by H. M. Blalock, A. B. Blalock. New York: McGraw Hill 1968, 428 – 478

Coleman, J. S. (1981). Longitudinal Data Analysis. New York: Basic Books 1981

Courgeau, D., Lelièvre, E. (1988). Estimation of Transition Rates in Dynamic Household Models. In: Modelling Household Formation and Dissolution, ed. by N. Keilman, A. Kuijsten, A. Vossen. Oxford: Clarendon 1988, 160 - 176

Courgeau, D., Lelièvre, E. (1992). Event History Analysis in Demography. Oxford: Clarendon 1992

Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society 34 (1972), 187 – 220

Cox, D. R. (1978). Foundations of Statistical Inference: The Case for Eclecticism. Australian Journal of Statistics 20 (1978), 43 – 59

Cox, D. R., Oakes, D. (1984). Analysis of Survival Data. London: Chapman and Hall 1984

Cox, D. R., Snell, E.J. (1968). A General Definition of Residuals, Journal of the Royal Statistical Society B 30 (1968), 248 – 275

Cox, D. R., Snell, E.J. (1981). Applied Statistics. Principles and Examples. London: Chapman and Hall 1981

Danto, A. C. (1965). Analytische Philosophie der Geschichte. Frankfurt: Suhrkamp 1980

Dex, S. (1991). Life and Work History Analyses. In: Life and Work History Analyses: Qualitative and Quantitative Developments, ed. by S. Dex. London: Routledge 1991, 1 – 19

Diekmann, A. (1981). Ein einfaches stochastisches Modell zur Analyse von Häufigkeitsverteilungen abweichenden Verhaltens. Zeitschrift für Soziologie 10 (1981), 319 – 325

Diekmann, A. (1990). Diffusion and Survival Models for the Process of Entry into Marriage. In: Event History Analysis in Life Course Research, ed. by K. U. Mayer, N. B. Tuma. Madison: University of Wisconsin Press 1990, 170 – 183

Diekmann, A. (1991). Mathematische Modelle des Heiratsverhaltens und Ehescheidungsrisikos. In: Modellierung sozialer Prozesse, hrsg. von H. Esser, K. G. Troitzsch. Bonn: Informationszentrum Sozialwissenschaften 1991, 589 – 620

Diekmann, A., Mitter, P. (1984). Methoden zur Analyse von Zeitverläufen. Stuttgart: Teubner 1984

Diekmann, A., Mitter, P. (1990). Stand und Probleme der Ereignisanalyse. In: Lebensverläufe und sozialer Wandel, hrsg. von K. U. Mayer. Opladen: Westdeutscher Verlag 1990, 404 – 441

Diekmann, A., Weick, S. (Hg.) (1993). Der Familienzyklus als sozialer Prozess. Bevölkerungssoziologische Untersuchungen mit den Methoden der Ereignisanalyse. Berlin: Duncker & Humblot 1993

Dierckx, P. (1975). An Algorithm for Smoothing, Differentiation and Integration of Experimental Data Using Spline Functions. Journal of Computational and Applied Mathematics 1 (1975), 165 – 184

Donagan, A. (1964). The Popper-Hempel Theory Reconsidered. In: Philosophical Analysis and History, ed. by W. H. Dray. New York 1966, Reprint Westport (Connecticut): Greenwood 1978, 127 – 159

Dretske, F. I., Snyder, A. (1972). Causal Irregularity. Philosophy of Science 39 (1972), 69 – 71

Dummett, M. (1960). A Defense of McTaggart's Proof of the Unreality of Time. Philosophical Review 59 (1960), 497 – 504. Dt. Übers. in Zimmerli und Sandbothe [1993].

Duncan, G. J., Hill, M. S. (1985). Conceptions of Longitudinal Households. Fertile or Futile? Journal of Economic and Social Measurement 13 (1985), 361 – 375

Edwards, A. W. F. (1992). Likelihood. Expanded Edition. London: Johns Hopkins University Press 1992

Eells, E. (1991). Probabilistic Causality. Cambridge: Cambridge University Press 1991

Elbers, C., Ridder, G. (1982). True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model. Review of Economic Studies 49 (1982), 403 – 409

Elder, G. H. (1975). Age Differentiation and the Life Course. Annual Review of Sociology 1 (1975), 165 – 190

Elder, G. H. (1985). Perspectives on the Life Course. In: Life Course Dynamics. Trajectories and Transitions, 1968 - 1980. Ed. by G. H. Elder. Ithaca / London: Cornell University Press 1985, 23 – 49

Elder, G. H., Caspi, A. (1990). Persönliche Entwicklung und sozialer Wandel. Die Entstehung der Lebensverlaufsforschung. In: Lebensverläufe und sozialer Wandel, hrsg. von K. U. Mayer. Opladen: Westdeutscher Verlag 1990, 22 – 57

Elster, J. (1989). The Cement of Society. A Study of Social Order. Cambridge: University Press 1989

Esser, H. (1987). Zum Verhältnis von qualitativen und quantitativen Methoden in der Sozialforschung, oder: Über den Nutzen methodologischer Regeln bei der Diskussion von Scheinkontroversen. In: Methoden der Biographie- und Lebenslaufforschung, hrsg. von W. Voges. Opladen: Leske + Budrich 1987, 87 – 101

Esser, H. (1989). Verfällt die 'soziologische Methode'? Soziale Welt 40 (1989), 57 – 75

Esser, H. (1993). Soziologie. Allgemeine Grundlagen. Frankfurt: Campus 1993

Featherman, D. L. (1980). Retrospective Longitudinal Research: Methodological Considerations. Journal of Economics and Business 32 (1980), 152 – 169

Feller, W. (1957). An Introduction to Probability Theory and Its Applications, Vol. I. New York: Wiley 1957

Ferriss, A. L. (1970). An Indicator of Marriage Dissolution by Marriage Cohort. Social Forces 48 (1970), 356 – 365

Fine, T. L. (1973). Theories of Probability. An Examination of Foundations. New York: Academic Press 1973

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London. Series A, Vol. 222 (1922), 309 – 368

Fisher, R. A. (1925). Theory of Statistical Estimation. Proceedings of the Cambridge Philosophical Society, No. 22 (1925), 700 – 725

Fisher, R. A. (1953). The Design of Experiments (6th Edition). Edinburgh: Oliver and Boyd 1953

Fisher, R. A. (1955). Statistical Methods and Scientific Induction. Journal of the Royal Statistical Society B 17 (1955), 69 – 78

Fisher, R. A. (1956). Statistical Methods and Scientific Inference. Edinburgh: Oliver and Boyd 1956

Fisher, R. A. (1970). Statistical Methods for Research Workers (14th Edition). Edinburgh: Oliver and Boyd 1970

Fisz, M. (1976). Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin: Deutscher Verlag der Wissenschaften 1976

Friedman, M. (1982). Piecewise Exponential Models for Survival Data with Covariates. Annals of Statistics 10 (1982), 101 – 113

Gail, M. (1975). A Review and Critique of Some Models Used in Competing Risk Analysis. Biometrics 31 (1975), 209 – 222

Galler, H. P. (1985). Übergangsratenmodelle bei intervalldatierten Ereignissen. Sonderforschungsbereich 3: Arbeitspapier 164. Frankfurt / Mannheim 1985

Galler, H. P. (1988). Ratenmodelle mit stochastisch abhängigen konkurrierenden Risiken. Sonderforschungsbereich 3, Arbeitspapier Nr. 261. Frankfurt/Mannheim 1988

Gardner, W., Griffin, W. A. (1986). A Structural-Causal Model for Analyzing Parallel Streams of Continuously Recorded Discrete Events. Unpubl. Paper, University of Washington 1986

Gerchak, Y. (1984). Durations in Social States: Concepts of Inertia and Related Comparisons in Stochastic Models. In: Sociological Methodology 1983-84, ed. by S. Leinhardt. San Francisco: Jossey-Bass 1984, 194 – 224

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L. (1989). The Empire of Chance. How Probability Changed Science and Everyday Life. Cambrdige: Cambridge University Press 1989

Glenn, N. D. (1977). Cohort Analysis. Beverly Hills und London: Sage 1977

Goldthorpe, J. H. (1991). The Uses of History in Sociology: Reflections on Some Recent Tendencies. British Journal of Sociology 42 (1991), 211 – 230

Gottinger, H. W. (1980). Elements of Statistical Analysis. Berlin–New York: de Gruyter 1980

Granger, C. W. J (1982). Generating Mechanisms, Models, and Causality. In: Advances in Econometrics, ed. by W. Hildenbrand. Cambridge: Cambridge University Press 1982, 237 – 253

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness. American Journal of Sociology 91 (1985), 481 – 510

Guo, G. (1993). Event-History Analysis for Left-Truncated Data. Sociological Methodology, Vol. 23 (1993), ed. by P. V. Marsden. San Francisco: Jossey-Bass 1993, 217 – 242

Hacking, I. (1965). Logic of Statistical Inference. Cambridge: Cambridge University Press 1965

Hacking, I. (1990). The Taming of Chance. Cambrige: University Press 1990

Hagestad, G. (1991). Trends and Dilemmas in Life Course Research: An International Perspective. In: Theoretical Advances in Life Course Research, ed. by W. R. Heinz. Weinheim: Deutscher Studienverlag 1991, 23 – 57

Hamerle, A. (1989). Multiple-Spell Regression Models for Duration Data. Applied Statistics 38 (1989), 127 – 138

Hamerle, A. (1991). On the Treatment of Interrupted Spells and Initial Conditions in Event History Analysis. Sociological Methods & Research 19 (1991), 388 – 414

Hamerle, A., Tutz, G. (1989). Diskrete Modelle zur Analyse von Verweildauer und Lebenszeiten. Frankfurt: Campus 1989

Hanefeld, U. (1984). Das Sozio-ökonomische Panel. Eine Längsschnittsstudie für die Bundesrepublik Deutschland. DIW-Vierteljahreshefte zur Wirtschaftsforschung, Heft 4, 1984, 391 - 406

Hanefeld, U. (1986). Das Sozio-ökonomische Panel. Konzeption und ausgewählte erhebungsmethodische Ergebnisse. Allgemeines Statistisches Archiv 69 (1986), 399 – 410

Hanefeld, U. (1987). Das Sozio-ökonomische Panel. Grundlagen und Konzeption. Frankfurt: Campus 1987

Hansen, M. H., Madow, W. G., Tepping, B. J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys (with Discussion). Journal of the American Statistical Association 78 (1983), 776 – 805

Harding, S. G. (ed.) (1976). Can Theories be Refuted? Essays on the Duhem-Quine Thesis. Dordrecht: Reidel Publ. 1976

Hartwig, H. (1956). Naturwissenschaftliche und sozialwissenschaftliche Statistik. Zeitschrift für die gesamte Staatswissenschaft 112 (1956), 252 – 266

Heckman, J. J., Singer, B. (1982). Population Heterogeneity in Demographic Models. In: Multidimensional Mathematical Demography, ed. by K.C. Land, A. Rogers. New York: Academic Press 1982, 567 – 599

Heckman, J., Singer, B. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. Econometrica 52 (1984), 271 – 320

Heckman, J. J. (1990). Selection Bias and Self-Selection. In: The New Palgrave: Econometrics. Ed. by J. Eatwell, M. Milgate, P. Newman. New York: Norton 1990, 201 – 224

Helsen, K. (1990). New Developments in Duration Time Modeling with Applications to Marketing. University of Pennsylvania Dissertation 1990

Hempel, C. G. (1965). Scientific Explanation. Essays in the Philosophy of Science. New York: Free Press 1965

Henkel, R. E. (1976). Tests of Significance. Beverly Hills: Sage 1976

Hoem, J. M. (1985). Weighting, Misclassification, and Other Issues in the Analysis of Survey Samples of Life Histories. In: Longitudinal Analysis of Labor Market Data, ed. by J.J. Heckman, B. Singer. Cambridge: Cambridge University Press 1985, 249 – 293

Hoem, J. M. (1989). The Issue of Weights in Panel Surveys of Individual Behavior. In: Panel Surveys, ed. by D. Kasprzyk, G. Duncan, G. Kalton, M. P. Singh. New York: Wiley 1989, 539 – 565

Hogan, D. P. (1978). The Variable Order of Events in the Life Course. American Sociological Review 43 (1978), 573 – 586

Holford, T. R. (1976). Life Tables with Concomitant Information. Biometrics 32 (1976), 587 – 597

Holford, T. R. (1980). The Analysis of Rates and of Suvivorship Using Log-Linear Models. Biometrics 36 (1980), 299 – 305

Huckfeldt, R. R., Kohfeld, C. W., Likens, T. W. (1982). Dynamic Modeling. An Introduction. Newbury Park – London: Sage 1982

Huinink, J. (1992). Die Analyse interdependenter Lebensverlaufsprozesse. Zum Zusammenhang von Familienbildung und Erwerbstätigkeit bei Frauen. In: Theorie, Daten, Methoden. Neue Modelle und Verfahrensweisen in den Sozialwissenschaften. Hrsg. von H. J. Andress u.a., München: Oldenbourg 1992, 343 – 366

Hutchison, D. (1987). Methods of Dealing with Grouped Data. An Application to Drop Out from Apprenticeship. In: Longitudinal Data Analysis, ed. by R. Crouchley. Aldershout: Avebury 1987, 205 – 234

International Statistical Institute (1986). Declaration on Professional Ethics. International Statistical Review 54 (1986), 227 – 242

Janson, C.-G. (1990). Retrospective Data, Undesirable Behavior, and the Longitudinal Perspective. In: Data Quality in Longitudinal Research, ed. by D. Magnusson, L. R. Bergman. Cambridge: Cambridge University Press 1990, 100 – 121

Johnson, N. L., Kotz, S. (1977). Urn Models: A Useful Tool in Applied Statistics. In: Applications of Statistics, ed. by P. R. Krishnaiah, Amsterdam: North-Holland 1977, 249 – 264

Kalbfleisch, J. D., Lawless, J. F. (1988). Estimation of Reliability in Field-Performance Studies. Technometrics 30 (1988), 365 – 378

Kalbfleisch, J. D., Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data. New York: Wiley 1980

Kalton, G. (1983). Models in the Practice of Survey Sampling. International Statistical Review 51 (1983), 175 – 188

Kaplan, E. L., Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 53 (1958), 457 – 481

Kapur, J. N. (1989). Maximum-Entropy Models in Science and Engineering. New York: Wiley 1989

Keilman, N., Keyfitz, N. (1988). Recurrent Issues in Dynamic Household Modelling. In: Modelling Household Formation and Dissolution, ed. by N. Keilman, A. Kuijsten, A. Vossen. Oxford: Clarendon 1988, 254 – 285

Kempthorne, O. (1969). Some Remarks on Statistical Inference in Finite Sampling. In: New Developments in Survey Sampling, ed. by N. L. Johnson, H. Smith. New York: Wiley 1969, 671 – 692

Kempthorne, O. (1971). Probability, Statistics, and Knowledge Business. In: Foundations of Statistical Inference, ed. by V. P. Godambe, D. A. Sprott. Toronto: Holt, Rinehart and Winston 1971, 470 – 499

Kendall, M. G. (1940). On the Method of Maximum Likelihood. Journal of the Royal Statistical Society 103 (1940), 388 – 399

Kendall, M. G. (1949). On the Reconciliation of Theories of Probability. Biometrika 36 (1949), 101 – 116

Kienzle, B. (Hg.) (1994). Zustand und Ereignis. Frankfurt: Suhrkamp 1994

Klein, T. (1988). Zur Abhängigkeit zwischen konkurrierenden Mortalitätsrisiken. Allgemeines Statistisches Archiv 72 (1988), 248 – 258

Kohli, M. (Hg.) (1978). Soziologie des Lebenslaufs. Neuwied 1978

Kohli, M. (1985). Die Institutionalisierung des Lebenslaufs. Kölner Zeitschrift für Soziologie und Sozialpsychologie 37 (1985), 1 – 29

Kohli, M. (1986). Social Organization and Subjective Construction of the Life Course. In: Human Development and the Life Course: Multidisciplinary Perspectives, ed. by A. B. Sørensen, F. E. Weinert, L. R. Sherrod. Hillsdale: Lawrence Erlbaum 1986, 271 – 292

Kolmogoroff, A. N. (1933). Grundbegriffe der Wahrscheinlichkeitsrechnung. Berlin: Springer 1933

Kruskal, W., Mosteller, F. (1979-80). Representative Sampling, I – IV. International Statistical Review 47 (1979), 13 – 24, 111 – 127, 245 – 265, und 48 (1980), 169 – 195

Kullback, S., Leibler, R. A. (1951). On Information and Sufficiency. Annals of Mathematical Statistics 22 (1951), 79 – 86

Kyburg, H. E. (1974). Propensities and Probabilities. British Journal for the Philosophy of Science 25 (1974), 358 – 375

Kyburg, H. E. (1980). Statistical Statements: Their Meaning, Acceptance, and Use. In: Science, Belief and Behaviour, ed. by D. H. Mellor. Cambridge: Cambridge University Press 1980, 161 – 177

Laird, N., Olivier, D. (1981). Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. Journal of the American Statistical Association 76 (1981), 231 – 240

Lancaster, T. (1979). Econometric Methods for the Duration of Unemployment. Econometrica 47 (1979), 939 – 956

Lancaster, T. (1985). Generalised Residuals and Heterogeneous Duration Models. Journal of Econometrics 28 (1985), 155 – 169

Lancaster, T. (1990). The Econometric Analysis of Transition Data. Cambridge: University Press 1990

Lancaster, T., Chesher, A. (1987). Residuals, Tests and Plots with a Job Matching Illustration. In: Longitudinal Data Analysis, ed. by R. Crouchley. Aldershout: Avebury 1987, 59 – 87

Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data. New York: Wiley 1982

Lexis, W. (1875). Einleitung in die Theorie der Bevölkerungsstatistik. Strassburg 1875

Lexis, W. (1903). Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik. Jena: Fischer 1903

Lieberson, S. (1985). Making It Count. The Improvement of Social Research and Theory. Berkeley: University of California Press 1985

Lindsey, J. K. (1973). Inferences from Sociological Survey Data. A Unified Approach. Amsterdam: Elsevier 1973

Lorenzen, P. (1974). Konstruktive Wissenschaftstheorie. Frankfurt: Suhrkamp 1974

Lorenzen, P. (1978). Eine konstruktive Deutung des Dualismus in der Wahrscheinlichkeitstheorie. Zeitschrift für allgemeine Wissenschaftstheorie 9 (1978), 256 – 275

Ludwig-Mayerhofer, W. (1992). Fakt und Artefakt in der Analyse von Arbeitslosigkeitsverläufen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44 (1992), 124 – 133

Mackie, J. L. (1973). Truth, Probability and Paradox. Studies in Philosophical Logic. Oxford: Clarendon 1973

Maddala, G. S. (1987). Limited Dependent Variable Models Using Panel Data. Journal of Human Resources 22 (1987), 307 – 338

Maistrov, L. E. (1974). Probability Theory. A Historical Sketch. New York: Academic Press 1974

Mandelbaum, M. (1955). Societal Facts. In: Modes of Individualism and Collectivism, ed. by J. O'Neill. London: Heinemann 1973, 221 – 234

Manting, D. (1994). Dynamics in Marriage and Cohabitation. An Inter-Temporal, Life Course Analysis of First Union Formation and Dissolution. Amsterdam: Thesis Publishers 1994

Manton, K. G., Singer, B., Woodbury, M. A. (1992). Some Issues in the Quantitative Characterization of Heterogeneous Populations. In: Demographic Applications of Event History Analysis, ed. by J. Trussell, R. Hankinson, J. Tilton. Oxford: Clarendon 1992, 9 – 37

Mayer, K. U. (1986). Structural Constraints on the Life Course. Human Development 29 (1986), 163 – 170

Mayer, K. U. (1987). Lebenslaufforschung. In: Methoden der Biographie- und Lebenslaufforschung, hrsg. von W. Voges. Opladen: Leske+Budrich 1987, 51 – 73

Mayer, K. U. (1990). Lebensverläufe und sozialer Wandel. Anmerkungen zu einem Forschungsprogramm. In: Lebensverläufe und sozialer Wandel, hrsg. von K. U. Mayer. Opladen: Westdeutscher Verlag 1990, 7 – 21

Mayer, K. U., Blossfeld, H.-P. (1990). Die gesellschaftliche Konstruktion sozialer Ungleichheit im Lebensverlauf. In: Lebenslagen, Lebensläufe, Lebensstile (= Soziale Welt Sonderband 7), hrsg. von P. A. Berger, S. Hradil. Göttingen: Schwartz 1990, 297 – 318

Mayer, K. U., Huinink, J. (1990a). Age, Period, and Cohort in the Study of the Life Course: A Comparison of Classical APC-Analysis with Event History Analysis. In: Data Quality in Longitudinal Research, ed. by D. Magnusson, L. R. Bergman. Cambridge: Cambridge University Press 1990, 211 – 232

Mayer, K. U., Huinink, J. (1990b). Alters-, Perioden- und Kohorteneffekte in der Analyse von Lebensverläufen oder: Lexis ade? In: Lebensverläufe und sozialer Wandel, hrsg. von K. U. Mayer. Opladen: Westdeutscher Verlag 1990, 442 – 459

Mayer, K. U., Müller, W. (1986). The State and the Structure of the Life Course. In: Human Development and the Life Course: Multidisciplinary Perspectives, ed. by A. B. Sørensen, F. E. Weinert, L. R. Sherrod. Hillsdale: Lawrence Erlbaum 1986, 217 – 245

Mayr, G. (1877). Die Gesetzmäßigkeit im Gesellschaftsleben. Statistische Studien. München: Oldenbourg 1877

McPherson, G. (1990). Statistics in Scientific Investigation. Its Basis, Application, and Interpretation. New York: Springer 1990

McTaggart, J. M. E. (1908). The Unreality of Time. Mind 17 (1908), 457 – 474. Dt. Übers. in Zimmerli und Sandbothe [1993].

Mellor, D. H. (1971). The Matter of Chance. Cambridge: Cambridge University Press 1971

Menges, G. (1959). Stichproben aus endlichen Gesamtheiten. Theorie und Technik. Frankfurt: Klostermann 1959

Meyer, J. W. (1986). The Self and the Life Course: Institutionalization and Its Effects. In: Human Development and the Life Course: Multidisciplinary Perspectives, ed. by A. B. Sørensen, F. E. Weinert, L. R. Sherrod. Hillsdale, N.J.: Lawrence Erlbaum 1986, 199 – 216

Mises, R. von (1928). Probability, Statistics and Truth. New York: Dover 1981 (Reprint)

Moore, D. S. (1991). Statistics. Concepts and Controversies (3rd Edition). New York: Freeman & Comp. 1991

Morrison, D. E., Henkel, R. E. (eds.) (1970). The Significance Test Controversy – A Reader. Chicago: Aldine Publ. 1970

Mueller, J. H., Schuessler, K. F., Costner, H. L. (1970). Statistical Reasoning in Sociology (2nd edition). New York: Houghton Mifflin 1970

Namboodiri, K., Suchindran, C. M. (1987). Life Table Techniques and their Applications. New York: Academic Press 1987

Nelder, J. A. (1984). The Role of Models in Official Statistics. In: Recent Developments in the Analysis of Large-Scale Data Sets. Eurostat News, Special Number 1984, 15 – 22

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection (with Discussion). Journal of the Royal Statistical Society 97 (1934), 558 – 625

Neyman, J. (1950). First Course in Probability and Statistics. London: Constable and Company 1950

Neyman, J. (1952). Lectures and Conferences on Mathematical Statistics and Probability. Washington: US Department of Agriculture (Graduate School) 1952

Noura, A. A., Read, K. L. O. (1990). Proportional Hazards Changepoint Models in Survival Analysis. Applied Statistics 39 (1990), 241 – 253

Nowak, S. (1977). Methodology of Sociological Research. Dordrecht: Reidel 1977

Nussbaum, M., Sen, A. (eds.) (1993). The Quality of Life. Oxford: Clarendon 1993

Petersen, T. (1990). Analyzing Event Histories. In: Statistical Methods in Longitudinal Research, Vol. II (Time Series and Categorical Longitudinal Data), ed. by A. von Eye. New York: Academic Press 1990, 259 – 288

Petersen, T. (1991). The Statistical Analysis of Event Histories. Sociological Methods & Research 19 (1991), 270 – 323

Pfanzagl, J. (1983). Allgemeine Methodenlehre der Statistik I. Berlin: de Gruyter 1983

Pfeiffer, P. E. (1978). Concepts of Probability Theory. New York: Dover 1978

Pierce, D. A., Stewart, W. H., Kopecky, K. J. (1979). Distribution-Free Regression Analysis of Grouped Survival Data. Biometrics 35 (1979), 785 – 793

Pötter, U. (1993). Models for Interdependent Decisions over Time. In: Applied Stochastic Models and Data Analysis, ed. by J. Janssen, C. H. Skiadas. World Scientific Publ. 1993, 767 – 779

Popper, K. R. (1960). The Propensity Interpretation of Probability. British Journal for the Philosophy of Science 10 (1960), 25 – 42

Popper, K. R. (1965). Das Elend des Historizismus. Tübingen: Mohr 1979 (5. Aufl.)

Porter, T. M. (1986). The Rise of Statistical Thinking 1820 – 1900. Princeton University Press 1986

Porter, T. M. (1987). Lawless Society: Social Science and the Reinterpretation of Statistics in Germany. In: The Probabilistic Revolution, ed. by L. Krüger, L. J. Daston, M. Heidelberger. Vol. I, 351 – 376. Cambridge: MIT Press 1987

Pratt, J. W., Gibbons, J. D. (1981). Concepts of Nonparametric Theory. New York: Springer 1981

Prentice, R. L., Gloeckler, L. A. (1978). Regression Analysis of Grouped Survival Data with Application to Beast Cancer Data. Biometrics 34 (1978), 57 – 67

Rao, C. R. (1971). Some Aspects of Statistical Inference in Problems of Sampling from Finite Populations. In: Foundations of Statistical Inference, ed. by V. P. Godambe, D. A. Sprott. Toronto: Holt, Rinehart and Winston 1971, 177 – 202

Reichenbach, H. (1949). Wahrscheinlichkeitslehre. Gesammelte Werke Band 7. Hrsg. von A. Kamlah und M. Reichenbach. Braunschweig: Vieweg 1994

Rendtel, U. (1990). Teilnahmebereitschaft in Panelstudien: Zwischen Beeinflussung, Vertrauen und sozialer Selektion. Kölner Zeitschrift für Soziologie und Sozialpsychologie 42 (1990), 280 – 299

Rendtel, U. (1993). Über die Repräsentativität von Panelstichproben. Eine Analyse der feldbedingten Ausfälle im Sozio-ökonomischen Panel (SOEP). DIW Diskussionspapier Nr. 70. Berlin: Deutsches Institut fuer Wirtschaftsforschung 1993

Rendtel, U. (1994). Die Analyse von Paneldaten unter Berücksichtigung der Panelmortalität. Theorie und Empirie am Beispiel des Sozio-ökonomischen Panels (sc soep). Berlin: Deutsches Institut für Wirtschaftsforschung 1994

Rendtel, U., Pötter, U. (1992). Über Sinn und Unsinn von Repräsentativitätsstudien. DIW-Diskussionspapier Nr. 61, Berlin: Deutsches Institut für Wirtschaftsforschung 1992

Rendtel, U., Pötter, U. (1993). „Empirie“ ohne Daten. Kritische Anmerkungen zu einer Repräsentativitätsstudie über den Allbus. Kölner Zeitschrift für Soziologie und Sozialpsychologie 45 (1993), 350 – 358

Reynolds, R. R. (1969). Replication and Substantive Import: A Critique on the Use of Statistical Inference in Social Research. Sociology and Social Research 53 (1969), 299 – 310

Ridder, G. (1984). The Distribution of Single-Spell Duration Data. In: Studies in Labor Market Dynamics, ed. by G. R. Neumann, N. C. Westergard-Nielsen. New York: Springer 1984, 45 – 73

Riley, M. W. (1986). Overview and Highlights of a Sociological Perspective. In: Human Development and the Life Course: Multidisciplinary Perspectives, ed. by A. B. Sørensen, F. E. Weinert, L. R. Sherrod. Hillsdale: Lawrence Erlbaum 1986, 153 – 175

Rosen, D. A. (1983). A Critique of Deterministic Causality. Philosophical Forum 14 (1983), 101 – 130

Rozeboom, W. W. (1960). The Fallacy of the Null-Hypothesis Significance Test. Psychological Bulletin 57 (1960), 416 – 428

Rowe, N. (1989). Rules and Institutions. New York: Philip Allan 1989

Royall, R. M. (1976). Current Advances in Sampling Theory: Implications for Human Observational Studies. American Journal of Epidemiology 104 (1976), 463 – 477

Royall, R. M. (1983). Comment. Journal of the American Statistical Association 78 (1983), 794 – 796

Ryder, N. B. (1964). Notes on the Concept of a Population. American Journal of Sociology 68 (1964), 447 - 463

Ryder, N. B. (1965). The Cohort as a Concept in the Study of Social Change. American Sociological Review 30 (1965), 843 – 861

Särndal, C.-E., Swensson, B., Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer 1992

Salmon, W. C. (1966). The Foundations of Scientific Inference. University of Pittsburgh Press 1966

Savage, L. J. (1962). The Foundations of Statistical Inference. A Discussion. London: Methuen 1962

Schaich, E. (1990). Schätz- und Testmethoden für Sozialwissenschaftler. München: Vahlen 1990

Schneider, H. (1991). Verweildaueranalyse mit GAUSS. Frankfurt: Campus 1991

Schneider, H. (1992). Zur Berücksichtigung links abgeschnittener Spells bei der Schätzung parametrischer Verweildauermodelle. Mimeo 1992

Schnell, R., Hill, P. B., Esser, E. (1992). Methoden der empirischen Sozialforschung (3. Auflage). München: Oldenbourg 1992

Scriven, M. (1966). Causes, Connections and Conditions in History. In: Philosophical Analysis and History, ed. by W. H. Dray. New York 1966, Reprint Westport (Connecticut): Greenwood 1978, 238 – 264

Sherrod, L. R., Brim, O. G. (1986). Epilogue: Retrospective and Prospective Views of Life-Course Research on Human Development. In: Human Development and the Life Course: Multidisciplinary Perspectives, ed. by A. B. Sørensen, F. E. Weinert, L. R. Sherrod. Hillsdale: Lawrence Erlbaum 1986, 557 – 580

Singer, B., Spilerman, S. (1976). Some Methodological Issues in the Analysis of Longitudinal Surveys. Annals of Economic and Social Measurement 5 (1976), 447 – 473

Sklar, L. (1970). Is Probability a Dispositional Property? Journal of Philosophy 67 (1970), 355 – 366

Smith, T. M. F. (1983). On the Validity of Inferences from Non-random Samples. Journal of the Royal Statistical Society A 146 (1983), 394 – 403

Smith, T. M. F. (1983a). Comment. Journal of the American Statistical Association 78 (1983), 801 – 802

Sørensen, A. B. (1977). Estimating Rates from Retrospective Questions. In: Sociological Methodology 1977, ed. by D. R. Heise. San Francisco: Jossey-Bass 1977, 209 – 223

Sørensen, A. B., Weinert, F. E., Sherrod, L. R. (eds.) (1986). Human Development and the Life Course: Multidisciplinary Perspectives. Hillsdale: Lawrence Erlbaum 1986

Spanos, A. (1986). Statistical Foundations of Econometric Modelling. Cambridge: University Press 1986

Stegmüller, W. (1971). Das Problem der Induktion: Humes Herausforderung und modernde Antworten. Darmstadt: Wissenschaftliche Buchgesellschaft 1991

Stegmüller, W. (1973). Personelle und statistische Wahrscheinlichkeit. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band IV. Berlin: Springer 1973

Stegmüller, W. (1983). Erklärung, Begründung, Kausalität. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band I. Berlin: Springer 1983

Teachman, J. D. (1983). Analyzing Social Processes: Life Tables and Proportional Hazards Models. Social Science Research 12 (1983), 263 – 301

Theil, H. (1972). Statistical Decomposition Analysis. Amsterdam: North Holland Publ. 1972

Thompson, W. A. (1977). On the Treatment of Grouped Observations in Life Studies. Biometrics 33 (1977), 463 – 470

Trussell, J., Richards, T. (1985). Correcting for Unmeasured Heterogeneity in Hazard Models Using the Heckman-Singer Procedure. In: Sociological Methodology 1985, San Francisco: Jossey-Bass 1985, 242 – 276

Tsiatis, A. (1975). A Nonidentifiability Aspect of the Problem of Competing Risks. Proceedings of the National Academy of Science (USA) 72 (1975), 20 – 22

Tuma, N. B., Hannan, M. T. (1979). Approaches to the Censoring Problem in Analysis of Event Histories. In: Sociological Methodology 1977, ed. by K. F. Schuessler. San Francisco: Jossey-Bass 1979, 209 – 240

Tuma, N. B., Hannan, M. T., Groeneveld, L. P. (1979). Dynamic Analysis of Event Histories. American Journal of Sociology 84 (1979), 820 – 854

Tuma, N. B., Hannan, M.T. (1984). Social Dynamics. Models and Methods. New York: Academic Press 1984

Vaupel, J. W., Yashin, A. I. (1985). Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics. The American Statistician 39 (1985), 176 – 185

Venn, J. (1888). The Logic of Chance (3rd edition). New York: Chelsea Publ. 1962 (Reprint)

Voges, W. (Hg.) (1987). Methoden der Biographie- und Lebenslaufforschung. Opladen: Leske+Budrich 1987

Watkins, J. W. N. (1952). Ideal Types and Historical Explanation. In: Modes of Individualism and Collectivism, ed. by J. O'Neill. London: Heinemann 1973, 143 – 165

Watkins, J. W. N. (1957). Historical Explanation in the Social Sciences. In: Modes of Individualism and Collectivism, ed. by J. O'Neill. London: Heinemann 1973, 166 – 178

Weber, M. (1976). Wirtschaft und Gesellschaft. 5., rev. Auflage (Studienausgabe), hrsg. von J. Winckelmann. Tübingen: Mohr 1976

White, H. (1982). Maximum Likelihood Estmation of Misspecified Models. Econometrica 50 (1982), 1 – 25

Willekens, F. (1988). A Life Course Perspective on Household Dynamics. In: Modelling Household Formation and Dissolution, ed. by N. Keilman, A. Kuijsten, A. Vossen. Oxford: Clarendon 1988, 87 – 107

Winch, P. (1958). Die Idee der Sozialwissenschaft und ihr Verhältnis zur Philosophie. Frankfurt: Suhrkamp 1966

Wright, G. H. von (1963). Norm und Handlung. Eine logische Untersuchung. Königstein: Scriptor 1979

Wright, G. H. von (1974). Causality and Determinism. New York: Columbia Press 1974

Wu, L. L. (1990). Simple Graphical Goodness-of-Fit Tests for Hazard Rate Models. In: Event History Analysis in Life Course Research, ed. by K. U. Mayer, N. B. Tuma. Madison: University of Wisconsin Press 1990, 184 – 199

Wu, L. L., Tuma, N. B. (1991). Local Hazard Models. In: Sociological Methodology 1990, ed. by C. C. Clogg. Oxford: Blackwell 1990, 141 – 180

Wu, L. L., Tuma, N. B. (1991a). Assessing Bias and Fit of Global and Local Hazard Models. Sociological Methods & Research 19 (1991), 354 – 387

Wurzel, E. (1988a). Distribution of Single-Spell Durations in Sample Designs with Time Aggregated Data. Discussion Paper No. A-162. Sonderforschungsbereich 303 (Information und die Koordination wirtschaftlicher Aktivitäten). Bonn 1988

Wurzel, E. (1988b). Unemployment Duration in West Germany – An Analysis of Grouped Data. Discussion Paper No. 88/2. Universität Bonn, Institut für Stabilisierungs- und Strukturpolitik. Bonn 1988

Yamaguchi, K. (1986). Alternative Approaches to Unobserved Heterogeneity in the Analysis of Repeatable Events. In: Sociological Methodology 1986, ed. by N.B. Tuma. San Francisco: Jossey-Bass 1986, 213 – 249

Yamaguchi, K. (1991). Event History Analysis. Newbury Park: Sage 1991

Zimmerli, W. C., Sandbothe, M. (Hg.) (1993). Klassiker der modernen Zeitphilosophie. Darmstadt: Wissenschaftliche Buchgesellschaft 1993