Statistische Modellierung von Verlaufsdaten in der Analyse sozialer Probleme, Teil I: Grundlagen[1]

von Wolfgang Ludwig-Mayerhofer

Abstract

Although „dynamic" notions like the concept of career have a long history in the sociology of social problems, still most research is static, referring to the number and the characteristics of people belonging to problem groups at a given point in time. Methods of „event history analysis" (also called „survival analysis" or „analysis of failure time") provide flexible tools for analyses which take into account the changes that take place on the individual level. They model processes where we have information (a) about the time individuals spend in a given social state, and (b) about the state they occupy after a change has occurred, if any. This paper provides an introduction to the basic concepts of event history analysis, that is, survivor, density, and hazard function, and illustrates these by using example data from the Socio-Economic Panel (SOEP). A companion paper, to appear in the next volume, will give an overview on various types of multivariate models and related issues.

Zusammenfassung

Obwohl dynamische Konzepte wie der Karriere-Begriff schon lange in der Soziologie sozialer Probleme etabliert sind, werden vielfach immer noch statische Untersuchungen durchgeführt, die sich auf die Anteile und die Merkmale von Personen in bestimmten sozialen Lagen zu bestimmten Zeitpunkten beziehen. Methoden der Analyse von Verlaufsdaten („Ereignisanalyse", „Survivalanalyse") erlauben es, in flexibler Weise Veränderungen auf der Individualebene zu untersuchen. Sie modellieren Prozesse, zu denen Informationen (a) über die Zeit, die Individuen in einer sozialen Lage verbringen, und (b) über die Lage, die sie im Anschluß einnehmen (falls eine Veränderung geschehen ist), vorliegen. Diese Arbeit führt in die Grundbegriffe dieser Modelle ein (Survivor-, Dichte- und Hazardfunktion) und erläutert sie anhand eines Beispieles aus dem Sozio-ökonomischen Panel (SOEP). Eine weitere Arbeit, die für die nächste Ausgabe der Zeitschrift vorgesehen ist, wird verschiedene Möglichkeiten der multivariaten Auswertung solcher Daten diskutieren.

1. Zur Problemstellung

In der sozialwissenschaftlichen Analyse sozialer Probleme herrscht häufig eine statische Sichtweise vor. Untersucht wird der *Umfang* eines sozialen Problems, also die Anzahl von Armen, Kranken, Straffälligen, Arbeitslosen usw., zu einem gegebenen *Zeitpunkt*, und wenn nach relevanten *Merkmalen* der jeweiligen Personen(gruppen) gefragt wird, bezieht sich dies dementsprechend auf die zu einem be-

stimmten Zeitpunkt Betroffenen. In letzter Zeit ist jedoch zunehmend deutlich geworden, daß derartige auf Bestandsdaten bezogene Analysen nur *einen* Aspekt der Problematik erfassen und die sehr beachtliche „Dynamik" des untersuchten Sachverhalts auf der Individualebene vernachlässigen. Denn während der Anteil der Armen, Arbeitslosen usf. an der Population häufig über Jahre konstant bleibt oder sich nur langsam ändert, ist auf der Ebene der betroffenen Personen eine oftmals sehr ausgeprägte „Mobilität" zu beobachten. Viele soziale Lagen wie Armut, Arbeitslosigkeit usw. sind reversibel; sie werden im Zeitverlauf zwar einerseits von sehr viel mehr Personen eingenommen, als die Bestandsdaten nahelegen, andererseits jedoch auch wesentlich häufiger und schneller wieder verlassen.[2]

Im Grunde sollte dies für Sozialwissenschaftler gar nicht überraschend sein. So gehört zum gängigen Vokabular in der Analyse sozialer Probleme der „Karriere"-Begriff, also die Vorstellung eines sich stufenweise aufschaukelnden, möglicherweise aber auch reversiblen *Prozesses*, in dem ein Individuum zu einem „sozialen Problemfall" werden, aber unter Umständen seine problematische Lebenslage auch wieder verlassen kann.[3] Ganz grundsätzlich läßt sich festhalten, daß viele Fragestellungen in der Analyse sozialer Probleme den *Aspekt der Zeit, also eine dynamische Sichtweise* thematisieren: Wie lange dauert die Arbeitslosigkeit von Individuen? Wann ist mit dem - erstmaligen oder erneuten - Auftreten einer Krankheit, wann ist u.U. mit einer sozialen Marginalisierung aufgrund dieser Krankheit zu rechnen? Wie schnell werden sanktionierte Straftäter erneut straffällig?

In den letzten 20 Jahren wurde die Entwicklung statistischer Modelle vorangetrieben, mit deren Hilfe solche Untersuchungsfragen angemessen analysiert werden können.[4] In vielen Bereichen - etwa der Arbeitsmarkt- und Mobilitätsforschung oder der Demographie - gehören sie zum etablierten Methodenbestand. Doch obwohl diese Verfahren bereits relativ gut entwickelt und erprobt sind, haben sie noch wenig Aufnahme in die Standard-Methodenlehrbücher gefunden. Das erschwert sowohl ihre Anwendung im Forschungsprozeß als auch die Rezeption einschlägiger Ergebnisse. In dieser Arbeit sollen die Grundlagen und Anwendungsmöglichkeiten dieser Verfahren anhand von einfachen Beispielen dargestellt werden. Das Ziel der Arbeit ist didaktisch; es geht darum, die grundlegende Logik der Verfahren zu erläutern und intuitiv verständlich zu machen, während ein Anspruch auf Originalität oder Innovation nicht erhoben wird. Der statistische Hintergrund der Verfahren soll nur ganz rudimentär behandelt werden. Das geschieht nicht, weil dieser Hintergrund irrelevant wäre, sondern in der Absicht, zunächst einen einfachen Zugang zu ermöglichen, der dann entsprechend vertieft werden kann. Dafür versuche ich zumindest gelegentlich, Fragen des praktischen Vorgehens bei der Datenanalyse anzusprechen. Die Arbeit hat zwei Teile. In diesem Teil werden die elementaren Grundlagen und Konzepte dargestellt; in einem zweiten Teil, der im folgenden Heft der *Sozialen Probleme* erscheinen soll, werden die wichtigsten für Sozialwissenschaftler relevanten Techniken der Datenanalyse vorgestellt und diskutiert.

2. Zur Unangemessenheit vertrauter statistischer Modelle

„Karrieren“ oder „Verläufe“ sind zeitliche Phänomene. Sie lassen sich, vereinfacht formuliert, anhand der *Dauer* beschreiben, die Untersuchungseinheiten (Personen, Familien etc.) in bestimmten Zuständen verbleiben, z.B. in Arbeitslosigkeit oder Krankheit, oder umgekehrt in Beschäftigung oder Gesundheit (ausführlicher dazu siehe Abschnitt 3.1). Man könnte daher zu der Annahme verleitet werden - und dies geschieht offensichtlich immer noch häufig, so beispielsweise bei Jacobs/ Ringbeck (1992) -, daß diese Datenkonstellation problemlos mit dem klassischen statistischen Verfahren der linearen Regression auf der Basis des Kleinst-Quadrat-Schätzers („OLS“) analysiert werden kann, da es sich bei der untersuchten Zielvariablen, der Dauer, um ein metrisches, intervallskaliertes Merkmal zu handeln scheint. Tatsächlich sprechen gravierende Gründe gegen die Anwendung der linearen Regression auf Verlaufsdaten.

Der zweifellos wichtigste Grund ist darin zu sehen, daß die interessierende Dauer *im allgemeinen nicht bei allen Untersuchungseinheiten gemessen werden kann*. Denn diese Dauer ist erst vollständig bekannt, wenn ein Individuum den betreffenden Zustand wieder verläßt, wenn also ein Zustandswechsel geschieht. Im Regelfall tritt aber bei einer kleineren oder größeren Anzahl der Fälle kein Zustandswechsel auf (Personen bleiben arbeitslos, werden nicht [erneut] straffällig, usw.). Ferner ist häufig der Beobachtungszeitraum faktisch begrenzt, so daß selbst dann, wenn ein Zustandswechsel untersucht wird, der bei jedem Individuum eintritt - etwa der Tod -, dieser vom Forscher nicht in allen Fällen beobachtet werden kann. Schließlich ist damit zu rechnen, daß bei einer Längsschnittuntersuchung ein Teil der Personen schon vor dem geplanten Untersuchungsende ausscheidet. In all diesen Fällen ist zwar bekannt, daß ein Individuum sich bis zu einem bestimmten Zeitpunkt in dem untersuchten Ausgangszustand befand, es fehlt aber die „positive“ Information über den „Zustandswechsel“ oder „Übergang“ in einen neuen Zustand. Man spricht hier von *(rechts-)zensierten Daten* oder einfach „Zensierungen“.

Liegen zensierte Daten vor, so ist einsichtig, daß das lineare Regressionsmodell inadäquat ist. Weder kann man die zensierten Fälle einfach weglassen, noch kann man so tun, als handele es sich bei den zensierten Beobachtungsdauern um vollständige Beobachtungen. In beiden Fällen werden nämlich die wahren Dauern unterschätzt, und zwar *in systematischer Weise*.[5] Aber noch weitere Gründe sprechen gegen das lineare Regressionsmodell: Die abhängige Variable weicht oft sehr stark von der Normalverteilung ab, und sie ist manchmal nicht exakt (stetig), sondern nur annäherungsweise, nämlich in diskreten Intervallen gemessen worden.

Sofern keine Untersuchungseinheiten vorzeitig ausgefallen sind, mithin die Beobachtungsdauer bei allen Untersuchungseinheiten identisch ist und Zensierungen - nicht eingetretene Zustandswechsel - nur am Ende der Beobachtungsdauer auftreten, wäre als alternative Auswertungsstrategie an ein Regressionsmodell für binäre abhängige Variablen (z.B. Logit- oder Probit-Modell) zu denken, dessen abhängige Variable als „Zustandswechsel bis zum Ende der Beobachtungsdauer eingetreten:

ja/nein“ formuliert werden kann. Trotzdem vergibt man sich auch hier wichtige Analysemöglichkeiten, da nur der Anteil der Zustandswechsel bis zu einem einzigen, möglicherweise recht willkürlich gewählten Zeitpunkt berücksichtigt wird. Sehr oft ist jedoch die „Zeitabhängigkeit“ des untersuchten Prozesses von Interesse: Ist das „Rückfallrisiko“ von Straffälligen zu jedem Zeitpunkt gleich, oder gibt es zunächst einen Abschreckungseffekt, der mit zunehmendem Abstand nachläßt? Hält die Wirkung einer ärztlichen Behandlung eine Zeit lang an, um dann abrupt aufzuhören, ist eine kontinuierliche, oder vielleicht gar keine Abnahme der Wirksamkeit zu beobachten? Wenn solche Fragen von Bedeutung sein können, was sich beim gegenwärtigen Wissensstand nur selten ausschließen läßt, verschenkt eine rein zeitpunkt-orientierte Betrachtung sehr gewichtige Informationen (vgl. hierzu insgesamt Breslow 1991 oder Yamaguchi 1991: 9).

Alle geschilderten Probleme lassen sich jedoch (zumindest im Prinzip) durch statistische Modelle lösen, die in den letzten Jahren unter Begriffen wie „Event History Analysis“ (Tuma/Hannan 1984), „Verlaufsdatenanalyse“ (Andreß 1992), „Ereignisanalyse“ (Blossfeld/Hamerle/Mayer 1986) oder „Survivalanalyse“ (Lee 1980) bekannt geworden sind. Die Vorzüge dieser Verfahren lassen sich nicht nur negativ, in Abgrenzung von ungeeigneten Verfahren beschreiben, es ist auch zu betonen, daß damit ganz neue Untersuchungsmöglichkeiten eröffnet werden. Zum einen kann, wie schon erwähnt, der „Zeitabhängigkeit“ der Prozesse - im Sinne von Änderungen der Wahrscheinlichkeit, den untersuchten Ausgangszustand zu verlassen - häufig durch die Wahl einer entsprechenden Verteilungsannahme für die abhängige Variable Rechnung getragen werden. Zweitens und vor allem ist auf die Möglichkeit zu verweisen, auch *solche erklärenden Variablen zu berücksichtigen, deren Werte sich selbst im Zeitverlauf ändern* (vgl. Andreß 1992: 16 ff.).

Bevor ich fortfahre, möchte ich einige Hinweise auf weiterführende Literatur geben. Als wesentliche *Lehrbuch-Einführungen* sind zu nennen: Allison (1984), Andreß (1992), Blossfeld/Hamerle/Mayer (1986), Diekmann/Mitter (1984a), Kalbfleisch/Prentice (1980), Lawless (1982), Lee (1980), Namboodiri/Suchindran (1987) und Yamaguchi (1991), unter denen die Arbeit von Lee mit Abstand am leichtesten verständlich (und mit vielen Beispielen versehen), allerdings nicht mehr ganz auf dem neuesten Stand ist. Unter den deutschsprachigen Lehrbüchern ist dasjenige von Blossfeld/Hamerle/Mayer (1986) am ausführlichsten und grundsätzlich auch sehr anwendungsorientiert, jedoch sollten unbedingt Hinweise auf neuere EDV-Programme beachtet werden, wie sie sich bei Andreß (1992) und in dieser Arbeit finden. - In den folgenden Ausführungen verweise ich jeweils auf *Fundstellen in den drei deutschen Lehrbüchern* mit den Abkürzungen *A* für Andreß, *BHM* für Blossfeld/Hamerle/Mayer und *DM* für Diekmann/Mitter. - Das Buch von Tuma/Hannan (1984), das sich ebenso wie die Arbeiten von Tuma/Hannan/Groeneveld (1979) und Tuma (1982) als Einführung versteht, ist dagegen teilweise relativ anspruchsvoll und nicht sehr eingängig. Die Monographie von Lancaster (1990) und die Sammelbände von Crouchley (1987) und Diekmann/Mitter (1984b) sind ebenfalls nur mit recht weit fortgeschrittenen Kenntnissen zugänglich. - Unter den

nicht-monographischen Arbeiten seien als Einführungen Carroll (1983), Hutchison (1988a), Kiefer (1988), Teachman (1983) sowie Kap. 6 in Toutenburg (1992) genannt, als Übersichtsartikel (teilweise auch zu neueren und komplexeren Verfahren bzw. Problemen) Diekmann (1988), Diekmann/Mitter (1990, 1993), Hutchison (1988b), Meinken (1992) sowie Petersen (1990, 1991a, 1993); in sehr mathematisch-straffer Form führt auch Arminger (1984, 1988) in alle wesentlichen Begriffe und Modelle ein. Ebenfalls sehr hilfreich ist das Manual des Programms TDA (Rohwer 1993), mit dem auch die in dieser Arbeit vorgestellten Beispiele berechnet wurden.

3. Grundkonzepte der Verlaufsdatenanalyse

In diesem Abschnitt erläutere ich zunächst die Art der Daten, die der Verlaufsdatenanalyse zugrunde liegen. Anschließend stelle ich die wichtigsten Grundkonzepte an Beispielen dar mit dem Ziel, ein intuitives Verständnis dieser Konzepte zu ermöglichen und aufzuzeigen, wie mit dem Problem der rechtszensierten Daten umgegangen werden kann.

3.1 Vorbemerkungen zur Datenstruktur

Eine Beschreibung von Prozessen setzt sich aus zwei Aspekten zusammen. Erstens benötigt man Informationen über die sozialen Lagen, Positionen usw. - allgemein: *Zustände* -, die Individuen in diesem Prozeß einnehmen können, etwa „arbeitslos - beschäftigt“. Die Gesamtheit der für eine bestimmte Untersuchungsfrage relevanten Zustände wird als *Zustandsraum* bezeichnet. Dabei interessiert man sich vor allem für die *Zustandswechsel* - oder „Übergänge“ oder „Ereignisse“[6] - von einem Ausgangs- in einen Zielzustand. Zweitens benötigt man Angaben über die *Verweildauer* im Ausgangszustand, anders gesagt, über die Dauer bis zum Eintreten eines Zustandswechsels (man spricht daher auch von Wartezeiten [im Ausgangszustand] oder Ankunftszeiten [im Zielzustand]). Beide Aspekte, Zustandsraum und Verweildauer, charakterisieren eine *Episode* (englisch „spell“, was auch manchmal in deutschsprachigen Arbeiten gebraucht wird) (vgl. A: 45 ff.; BHM: 27 ff.; DM: 33 ff.). Wenn, wie es häufig der Fall ist, nur je ein einziger Ausgangs- und Endzustand vorliegt (bzw. untersucht wird), genügt auch die Information, ob die Episode mit einem Zustandswechsel endete oder nicht (also rechtszensiert ist).

Wie beim Beispiel einer arbeitslosen Person unmittelbar deutlich wird, sind manche Prozesse, etwa der Erwerbsverlauf, durch mehr als zwei Zustände gekennzeichnet - im genannten Beispiel kann die Arbeitslosigkeit nicht nur in eine Beschäftigung münden, sondern auch in eine berufliche Erstausbildung, eine berufliche Weiterbildung, Umschulung, längere Krankheit und insbesondere bei Frauen auch in Unterbrechungen der Erwerbstätigkeit, schließlich in den vorzeitigen oder altersgemäßen Ruhestand. Allgemein spricht man, wenn aus einem Zustand ein Übergang in mehrere andere Zustände möglich ist, davon, daß die Personen *kon-*

kurrierenden Risiken (englisch: *competing risks)* ausgesetzt sind. Auf die Behandlung dieses Problems gehe ich in Teil II kurz ein und beschränke mich hier auf Zwei-Zustands-Modelle, also Modelle, bei denen ein Wechsel nur aus einem Ausgangszustand in *einen* anderen Zielzustand stattfinden kann.

Auch unter dieser vereinfachenden Voraussetzung bleibt noch als weiteres Problem, daß viele Episoden im Leben eines Individuums *mehrfach auftreten* können. Es ist aber ohne weiteres einsichtig, daß wiederholte Episoden möglicherweise anderen Bedingungen unterliegen als erste. Beispielsweise mag jemand, der zum zweiten Mal oder noch häufiger arbeitslos wurde, bei der Arbeitssuche im Vorteil sein, weil er bereits früher Erfahrungen mit dem Arbeitsamt und bei der Stellensuche machte; auf der anderen Seite könnte allein die Tatsache einer wiederholten Arbeitslosigkeit etwa potentielle Arbeitgeber mißtrauisch werden lassen. Dieses Beispiel verdeutlicht, daß die verschiedenen Episoden eines Individuums untereinander *abhängig* sein können, was in der Datenauswertung berücksichtigt werden muß. Es ist daher im allgemeinen ratsam, zwischen ersten und weiteren Episoden zu unterscheiden und bei Vorliegen mehrerer Episoden entsprechende Mehr-Episoden-Modelle zu schätzen (vgl. Teil II).

Anzumerken ist vielleicht auch noch, daß die Definition eines „Zustandes" grundsätzlich eine *inhaltliche* Frage ist. So könnte man es durchaus für sinnvoll erachten, einen „Zustandswechsel" anzunehmen, wenn auf einer eigentlich kontinuierlichen Skala (etwa dem Einkommen) ein bestimmter Schwellenwert über- oder unterschritten wird, etwa eine wie auch immer definierte „Armutsgrenze". Ob dies sinnvoll ist oder nicht, kann natürlich nicht mit Hilfe der Statistik, sondern nur aus den jeweiligen substantiell-inhaltlichen Annahmen einer Fachdisziplin entschieden werden.[7]

Nun einige Bemerkungen zum Aspekt der *Verweildauern*, genauer ihrer *Messung*. Die meisten der später diskutierten Modelle basieren auf der Annahme stetig gemessener Zeiten. Nun muß man in der Forschungspraxis häufig Abstriche von dieser Annahme machen. So können *erstens* manche Zustandswechsel gar nicht jederzeit, sondern nur zu bestimmten Zeitpunkten stattfinden: Politische Wahlen finden in mehrjährigem Abstand statt, in der Schule wird einmal im Jahr über die Versetzung in die nächst höhere Klasse entschieden, usf. Für solche *diskreten Zeitpunkte* wurden bereits einige Verfahren entwickelt (näheres in Teil II). *Zweitens* liegen auch bei im Prinzip stetiger Zeit oft nicht sehr genaue Messungen vor, so daß man etwa nur weiß, in welcher Woche oder in welchem Monat, allgemein: in welchem Zeitraum oder Intervall ein Zustandswechsel eingetreten ist; man spricht von sog. *gruppierten* oder *aggregierten* Verweildauern. In beiden Fällen werden in aller Regel oft mehrere Zustandswechsel zum gleichen Zeitpunkt bzw. im gleichen Zeitraum auftreten (sog. „Ties"), was für manche Verfahren im Prinzip gleichfalls unerwünscht ist (siehe z.B. Hutchison 1988a).[8]

Nun ist zu konstatieren, daß die Probleme teilweise auf der abstrakten Ebene größer aussehen als in der Praxis. Das beginnt mit der Frage, wann eine Messung exakt genug ist, um als „stetig" bzw. „nicht aggregiert" aufgefaßt werden zu kön-

nen. Soll die Zeit bis zum Auftreten eines Zustandswechsels in Sekunden, Minuten, Stunden, Tagen, Wochen, Monaten oder gar Jahren gemessen werden? Diese Frage muß sicherlich unterschiedlich beantwortet werden, je nachdem um welchen Zustand es sich handelt (die Dauer von Ehen ist im Durchschnitt wesentlich länger als die von Arbeitslosigkeitsepisoden). Es geht also offensichtlich vor allem um die *relative Genauigkeit der Messung*. Nach einigen neueren Arbeiten (z.B. Arminger 1984; Galler 1986; Petersen/Koput 1992) kann man davon ausgehen, daß zumindest bei in der Zeit konstanten Risiken (vgl. Abschnitt 3.2 und Teil II) einigermaßen valide Ergebnisse erzielt werden können, wenn die Zeitskala, auf der die Dauer gemessen wird, maximal die Hälfte des Medians der Verweildauer (vgl. Abschnitt 3.2) ausmacht. Wie sich die Dinge bei zeitlich variablen Risiken verhalten, wie sie in der Praxis häufiger auftreten, ist dagegen weniger eindeutig zu sagen.

Für gruppierte oder aggregierte Verweildauern gibt es ein exploratives Verfahren, den sog. Life-Table-Schätzer, dessen wir uns weiter unten ausführlich bedienen werden. Für eine multivariate Analyse solcher Daten existieren dagegen nur wenige Ansätze, die zudem in den verfügbaren Statistikprogrammen kaum implementiert sind (vgl. als Überblick Heijtan 1989). Sofern man von einer hinreichenden relativen Meßgenauigkeit ausgehen kann, werden daher in der Praxis zumeist Verfahren angewendet, die sich eigentlich auf kontinuierliche Verweildauern beziehen. Für diesen Fall wurde aufgrund theoretischer Überlegungen (Hujer/Schneider 1986, 1989) wie von Simulationsstudien (Petersen 1991b, 1993; Petersen/Koput 1992) vorgeschlagen, von den gemessenen Dauern eine halbe Zeiteinheit abzuziehen. Die Überlegung ist einfach: Wenn man annimmt, daß die Zustandswechsel sich in etwa gleichmäßig auf das Zeitintervall verteilen, so kann der Mittelpunkt des Intervalls als beste Schätzung des Eintritts des Zustandswechsels gelten. Ersichtlich gilt dies nur dann, wenn die Annahme gleicher Verteilung zutrifft, technisch gesprochen, wenn die Hazardfunktion im Zeitverlauf konstant ist (s. unten). Ist sie das nicht, müssen u.U. andere Transformationen der beobachteten Dauern vorgenommen werden (Petersen/Koput 1992) - wobei sich natürlich die Katze ein wenig in den Schwanz beißt, da sich die Zeitabhängigkeit der Hazardfunktion in der Regel erst anhand der empirischen Auswertung zeigt.[9] Offensichtlich können solche Transformationen die Ergebnisse gerade im Hinblick auf die Zeitabhängigkeit der Verweildauern nicht unbeträchtlich beeinflussen (vgl. Ludwig-Mayerhofer 1992).

Nun wird im allgemeinen angenommen, daß zwar - nicht zuletzt aufgrund des Problems gruppierter Messungen - Aussagen über die Zeitabhängigkeit von Prozessen oft nicht eindeutig zu treffen sind, daß aber die Einflüsse von erklärenden Variablen von diesem Problem weitgehend unberührt bleiben (Galler/Pötter 1992). Wie wir in Teil II sehen werden, muß auch diese Annahme, so richtig sie im Kern sein dürfte, im Einzelfall nicht unbedingt zutreffen.

Ganz kurz sei auf einen weiteren Aspekt hingewiesen. Im allgemeinen untersucht man einen konkreten Prozeß - etwa den Prozeß der Arbeitslosigkeit, der Drogenabhängigkeit, der Gesundung - ab seinem Beginn. Das heißt, alle Individuen

beginnen den Prozeß zum Zeitpunkt „Null“, etwa beim Eintritt in die Arbeitslosigkeit, beim ersten Drogenkonsum, bei Beginn der Erkrankung oder der Behandlung. Wir sprechen hier von einer Analyse der *Prozeßzeit*, und dieses Verfahren wird in der weitaus größten Zahl der Untersuchungen zugrundegelegt. Es ließen sich jedoch andere Auffassungen von der relevanten Zeit denken. Man könnte den Prozeß auf einen anderen „Null-Zeitpunkt“ beziehen, etwa auf den Beginn des Erwerbslebens, auf die Geburt, usf. (vgl. DM: 25; Blossfeld/Hamerle 1989). Hierauf kann an dieser Stelle nicht weiter eingegangen werden, und im folgenden gehe ich immer von Analysen der Prozeßzeit aus.

Schließlich ist noch auf das Problem der *Zensierungen* einzugehen. Oben wurde unterstellt, daß die untersuchten Individuen *von Anfang an* beobachtet wurden, also ab dem Zeitpunkt, zu dem sie den Ausgangszustand eingenommen haben; nur das *Ende* wird unter Umständen nicht beobachtet (weswegen man von *rechts*zensierten Daten spricht). Dabei werden häufig verschiedene Zensierungstypen unterschieden.[10] Der wesentliche Gesichtspunkt läßt sich jedoch relativ einfach formulieren: Die Zensierungen sollen zufällig erfolgen, was häufig auch so formuliert wird, daß Zensierungen und „Ereignisse“, also Zustandswechsel, voneinander unabhängig sein sollen. In der Praxis läßt sich diese Annahme allerdings oft nicht überprüfen, sondern nur mehr oder wenig glaubhaft vertreten.[11]

Wenn dagegen eine Episode nicht von Anfang an beobachtet werden konnte, d.h., wenn zu Beginn der Beobachtung nicht bekannt ist, wie lange die Untersuchungseinheiten sich bereits im Ausgangszustand befinden, spricht man von *linkszensierten Daten*. Ein Beispiel wäre eine Untersuchung von Personen im Arbeitslosenbestand, bei denen nicht erhoben wird, wie lange sie bereits arbeitslos sind, und somit nur die Arbeitslosigkeitsdauer ab Beginn der Untersuchung (und nicht ab Beginn der Arbeitslosigkeit) gemessen werden kann.[12] Liegen Linkszensierungen vor, sollte der Datensatz *nicht* oder nur mit größter Vorsicht ausgewertet werden, da hier ja grundsätzlich keine Angabe zur Dauer der jeweiligen Episode gemacht werden kann, ganz unabhängig davon, ob ihr Ende - der Übergang in den Zielzustand - beobachtet werden kann oder nicht (A: 94 f.; BHM: 26). Allenfalls wenn gut begründete Annahmen vorliegen, daß das Ausmaß der Linkszensierung nur sehr geringfügig ist und/oder daß die „Geschichte“ des Prozesses bis zum Beginn der Beobachtung ohne Einfluß auf den weiteren Verlauf ist, könnte an einen Versuch der Auswertung gedacht werden. Gerade letztere Annahme ist aber in der Regel eher zweifelhaft.

3.2 Survivor-, Dichte- und Hazardfunktion

Wie im zweiten Abschnitt verdeutlicht wurde, ist es wegen der rechtszensierten Daten nicht sinnvoll, die Dauer der beobachteten Episoden unmittelbar als abhängige Variable zu verwenden. Stattdessen werden einige andere Funktionen von Verweildauern herangezogen, die sich auch beim Vorliegen von Rechtszensierungen schätzen lassen. Ich will im folgenden *zuerst* ein *fiktives Beispiel* bringen, in welchem keine Zensierungen auftreten, und zeigen, daß die zu erläuternden Funk-

tionen sich in diesem Fall tatsächlich ganz einfach auf die beobachteten Daten zurückführen lassen. Gleichzeitig soll aber gezeigt werden, wie man diese Funktionen in einer Art und Weise schätzen kann, die auch bei rechtszensierten Daten anwendbar ist. Dies wird im Anschluß anhand eines *zweiten* Beispiels mit Daten aus dem „*Sozio-ökonomischen Panel" (SOEP)* (vgl. Projektgruppe „Das sozio-ökonomische Panel" 1990) noch einmal ausführlich dargestellt.

Die nachfolgenden Beispiele wenden *eine spezifische Form* der Schätzung der relevanten Größen an, die sog. „Life Table-Schätzung" (A: 139 ff.; BHM: 42 f., 116 ff.; DM: 60 ff.).[13] Es handelt sich hierbei um die einfachste Form der Verlaufsdatenanalyse, aber gerade deshalb ist sie für eine Einführung am sinnvollsten.[14] Denn dieses Verfahren bezieht sich auf gruppierte Verweildauern (wobei stetige Verweildauern in gruppierte Form gebracht werden), während bei stetigen Dauern auf weniger anschauliche Grenzwertbetrachtungen zurückgegriffen werden muß.

Darstellung 1: *Grunddaten (fiktive Arbeitslosigkeitsdauern, n = 10, Dauern in Wochen)*

Person	Dauer	Person	Dauer
A	1	F	6
B	2	G	7
C	3	H	8
D	4	I	9
E	5	J	10

Wir beginnen mit einem ganz einfachen, konstruierten Beispiel von zehn Untersuchungspersonen (vgl. *Darstellung 1*).[15] Auch wenn es grundsätzlich beliebig ist, welcher Prozeß untersucht wird, sei der besseren Verständlichkeit halber davon ausgegangen, daß es sich um die Dauer von Arbeitslosigkeitsepisoden handeln soll. Da wir in diesem Beispiel zensierte Daten unberücksichtigt lassen wollen, sollen die gemessenen Verweildauern die tatsächliche Dauer bis zum Verlassen der Arbeitslosigkeit repräsentieren. Wie gesagt, betrachten wir Zeitdauern, die gruppiert oder aggregiert sind, und zwar auf der Basis von Wochen. Eine Arbeitslosigkeitsdauer von 1 Woche[16] soll bedeuten, daß die Arbeitslosigkeit *während der ersten Woche* beendet wurde, usw.[17] Ersichtlich verläßt jede Woche genau eine Person die Arbeitslosigkeit, so daß nach 10 Wochen sämtliche Personen nicht mehr arbeitslos sind.

In *Darstellung 2* werden nun einige Funktionen vorgestellt, mit denen sich die Daten aus Darstellung 1 unter dem Aspekt des zeitlichen Verlaufs charakterisieren lassen. Dabei wird nun sozusagen nicht mehr die einzelne Untersuchungsperson, sondern die *Zeit* zum Untersuchungsgegenstand, denn die einzelnen Zeilen beziehen sich jetzt auf die zehn Wochen, die der hier untersuchte Prozeß insgesamt dauerte (vgl. Spalte 1).[18] In der zweiten Spalte ist angegeben, bei wievielen Personen

(Untersuchungseinheiten) *zu Beginn* dieses Intervalls *noch kein Zustandswechsel (Ereignis)* stattgefunden hatte; man kann auch von der „Risikomenge“ sprechen, nämlich der Zahl der Personen, die jeweils noch dem „Risiko“ (statistisch gesprochen!) ausgesetzt sind, ein Ereignis zu „erleiden“, d.h., die Arbeitslosigkeit zu verlassen. Notwendigerweise steht hier in der ersten Zeile die Gesamtzahl der Untersuchungseinheiten. Die dritte Spalte zeigt, bei wievielen Personen *während* des Intervalls *ein Zustandswechsel eingetreten ist*. Der Wert in der zweiten Spalte abzüglich des Wertes der dritten Spalte ergibt demzufolge den Wert der zweiten Spalte in der darauffolgenden Woche.

Darstellung 2: *Life-Table-Schätzung für die fiktiven Beispieldaten aus Darstellung 1*

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0 bis <1	10	1	0.100	0.900	1.000	0.100	0.105
1 bis <2	9	1	0.111	0.889	0.900	0.100	0.118
2 bis <3	8	1	0.125	0.875	0.800	0.100	0.133
3 bis <4	7	1	0.143	0.857	0.700	0.100	0.155
4 bis <5	6	1	0.167	0.833	0.600	0.100	0.182
5 bis <6	5	1	0.200	0.800	0.500	0.100	0.222
6 bis <7	4	1	0.250	0.750	0.400	0.100	0.286
7 bis <8	3	1	0.333	0.667	0.300	0.100	0.400
8 bis <9	2	1	0.500	0.500	0.200	0.100	0.667
9 bis <10	1	1	1.000	0.000	0.100	0.100	2.000
(Ende letztes Intervall)					0.000		

Erläuterungen zu Darstellung 2: (vgl. auch Text)

(1): Intervall (Woche)
(2): n zu Beginn des Intervalls
(3): Ereignisse (Übergänge aus der Arbeitslosigkeit) während des Intervalls
(4): Bedingte „Sterbewahrscheinlichkeit“ für das Intervall
(5): Bedingte „Überlebenswahrscheinlichkeit“ p_i für das Intervall
(6): Survivorfunktion $S(t_i)$
(7): Dichtefunktion $f(t_i)$
(8): Hazardfunktion $r(t_i)$

Die Werte der vierten Spalte geben an, wie groß der *Anteil* der Personen mit einem Ereignis während dieses Intervalls ist; die fünfte Spalte zeigt den Komplementärwert hierzu, den Anteil der „Überlebenden“ (der Personen ohne Zustandswechsel, hier also der immer noch Arbeitslosen) am Ende des Intervalls. Beides gilt bezogen auf die zweite Spalte, also die Personen, die zu Beginn des Intervalls noch arbeitslos waren. So waren nach dem ersten Intervall, also am Ende der ersten Woche, neun von zehn, also 90 Prozent der Personen, die zu Beginn der Woche arbeitslos waren, immer noch arbeitslos, am Ende der zweiten Woche waren es acht von neun, also 88,9 Prozent. Anders formuliert: Die fünfte Spalte stellt nicht die Wahrscheinlichkeit schlechthin, statistisch gesprochen: die unbedingte Wahrscheinlich-

keit dar, das betreffende Intervall zu überleben (die Wahrscheinlichkeit bezogen auf alle Personen überhaupt), sondern die Überlebenswahrscheinlichkeit unter der Bedingung, daß man bis zum Beginn des Intervalls „überlebt" hatte. Man spricht daher von der *„bedingten Überlebenswahrscheinlichkeit"*, die ich mit p_i - Wahrscheinlichkeit p für das i-te Intervall - bezeichnen will.

In den Spalten 6 bis 8 sind nun die entscheidenden Größen eingetragen. Spalte 6 enthält die *Survivor-Funktion S(t)*. Diese gibt die Wahrscheinlichkeit an, daß ein Individuum bis zum Zeitpunkt t „überlebt", d.h., daß bis zum Zeitpunkt t (hier: dem Beginn des betreffenden Intervalls) das Individuum sich noch im Ausgangszustand befindet, anders formuliert: daß noch kein Zustandswechsel oder „Ereignis" stattgefunden hat.[19] Formal schreiben wir:

$$S(t) = P(T \geq t) \text{ .}^{20} \tag{1}$$

Zu Beginn des Prozesses befinden sich definitionsgemäß alle Personen im Ausgangszustand, daher hat S(t) zu Beginn des Prozesses, also $S(t_1)$,[21] immer einen Betrag von 1. In unserem Beispiel nun geht jede Woche genau eine Person aus der Arbeitslosigkeit ab, d.h. jeweils 10 Prozent der Ausgangsstichprobe. Die „Überlebenswahrscheinlichkeit" nimmt also pro Woche um 0,1 ab, anders gesagt, die Survivorfunktion verringert sich in jedem Zeitintervall um den Betrag von 0,1. Die Wahrscheinlichkeit, mindestens eine Woche arbeitslos zu bleiben, beträgt also 0,9 (oder 90 Prozent), die Wahrscheinlichkeit, mindestens zwei Wochen arbeitslos zu bleiben, beträgt 0,8 (oder 80 Prozent), usw.

Daß die Survivorfunktion sich hier unmittelbar aus den Daten ablesen läßt, liegt natürlich daran, daß bewußt ein einfaches Beispiel ohne Zensierungen gewählt wurde. Daher soll sogleich eine allgemeinere Möglichkeit betrachtet werden, S(t) zu berechnen (A: 142; BHM: 43 f.; DM: 65; etwas ausführlicher bei Kiefer 1988: 648 f.; Petersen 1991a: 274 f.). Die bedingte Überlebenswahrscheinlichkeit p_i für das erste Intervall, also p_1, beträgt 0,9, und dies ist offenkundig auch die Wahrscheinlichkeit, den Beginn des zweiten Intervalls zu „erleben", also $S(t_2)$. $S(t_3)$, die Wahrscheinlichkeit, auch das zweite Intervall zu „überleben" und mithin zu Beginn des dritten Intervalls immer noch arbeitslos zu sein, ergibt sich aus der Wahrscheinlichkeit, das erste Intervall *und* das zweite Intervall zu überleben, was nach den Gesetzen der Wahrscheinlichkeitsrechnung durch die Multiplikation ausgedrückt wird, also $p_1 \cdot p_2$ oder $S(t_2) \cdot p_2$. $S(t_4)$, die Wahrscheinlichkeit, daß eine Person mindestens 3 Wochen oder bis zum Beginn der vierten Woche arbeitslos bleibt, ergibt sich aus $p_1 \cdot p_2 \cdot p_3$ - die Person muß die erste Woche *und* die zweite Woche *und* die dritte Woche „überleben" - oder $S(t_3) \cdot p_3$. Allgemein ergibt sich S(t) also jeweils aus dem Wert von $S(t_{i-1})$ multipliziert mit p_{i-1}, formal:

$$\hat{S}(t_i) = \hat{S}(t_{i-1}) \cdot \hat{p}_{i-1} \text{ ,} \tag{2}$$

oder, da ja $S(t_{i-1})$ sich wiederum aus Survivorfunktion und bedingter Überlebenswahrscheinlichkeit des vor-vorherigen Intervalls berechnen läßt,

$$\hat{S}(t_i) = \prod_{j=1}^{i-1} \hat{p}_j \ . \tag{3}$$

Die nächste Größe (Spalte 7), die *Dichteverteilung der Verweildauern f(t)*, ist häufig nicht direkt von Interesse, sie wird jedoch benötigt, um den zentralen Begriff der Hazardfunktion verständlich zu machen, der nachfolgend erörtert wird. Die (diskrete) Dichteverteilung ist die (unbedingte) Wahrscheinlichkeit, daß in einem Intervall ein Zustandswechsel (Ereignis) eintritt. Formal läßt sich dies so schreiben:

$$f(t) = P(t \le T < t+1) \ . \tag{4}$$

Auch sie läßt sich in unserem einfachen Beispiel aus der Ausgangstabelle ablesen: In jeder Woche tritt genau ein Ereignis ein, was bezogen auf die Gesamtstichprobe einer Wahrscheinlichkeit von 0,1 entspricht.

Allgemein ergibt sich aber f(t) aus S(t) (bzw. umgekehrt). Dazu muß man sich nur vor Augen halten, daß S(t) für ein beliebiges Intervall nichts anderes ist als die Wahrscheinlichkeit, bis zum vorherigen Intervall überlebt zu haben, abzüglich der Wahrscheinlichkeit, in diesem Intervall zu „sterben“, also in den Zielzustand zu wechseln. Umgekehrt läßt sich also für f(t) - bezogen auf ein konkretes Zeitintervall - schreiben:

$$\hat{f}(t_i) = \hat{S}(t_i) - \hat{S}(t_{i+1}) \ .^{22} \tag{5}$$

Nur der Vollständigkeit halber sei noch kurz erwähnt, daß statt S(t), der Überlebenswahrscheinlichkeit, ebensogut deren Gegenteil betrachtet werden kann, die (kumulierte) „Sterbewahrscheinlichkeit“ F(t), also die Wahrscheinlichkeit, bis zum Intervall t_i *nicht* zu überleben (bis dahin in den Zielzustand zu wechseln). F(t) ist offensichtlich der Komplementärwert zu S(t), formal:

$$F(t) = 1 - S(t) = P(T \le t) \ . \tag{6}$$

F(t) zu Beginn eines Intervalls ist nichts anderes als die Summe der bis dahin aufgetretenen intervallspezifischen Wahrscheinlichkeiten f(t), es gilt also:

$$F(t_i) = \sum_{j=1}^{i-1} f(t_j) \ . \tag{7}$$

Zentral für die Verlaufsdatenanalyse ist die *Hazardfunktion r(t)* (Spalte 8 in Darstellung 2).[23] Folgende Überlegung kann diese verständlich machen: Wenn man betrachtet, wieviele Personen in einem Zeitintervall den Ausgangszustand verlassen, so sollte man auch berücksichtigen, wieviele Personen *bis dahin überhaupt noch im Ausgangszustand verblieben waren*. Alle anderen Personen sind ja gar nicht mehr arbeitslos und nicht mehr dem „Risiko" ausgesetzt, eine Beschäftigung zu finden oder die Arbeitslosigkeit anderweitig zu verlassen. Im Beispiel: In der ersten Woche verläßt ebenso eine Person die Arbeitslosigkeit wie in der achten oder neunten Woche; im ersten Fall handelt es sich jedoch um ein Zehntel der zu diesem Zeitpunkt noch Arbeitslosen, in den anderen beiden Fällen sind es ein Drittel bzw. die Hälfte! Offensichtlich ist also im ersteren Zeitraum die Chance, die Arbeitslosigkeit zu verlassen, wesentlich geringer als im letzteren Zeitraum - *bezogen auf die jeweils noch Arbeitslosen*. Abstrakter: Wir können die (unbedingte) Wahrscheinlichkeit, die Arbeitslosigkeit in einem bestimmten Intervall zu verlassen, auch beziehen auf die Wahrscheinlichkeit, bis dahin überhaupt arbeitslos geblieben zu sein, also auf die Survivorfunktion:

$$r(t) = P(t \le T < t+1 \mid T \ge t) = \frac{f(t)}{S(t)} \quad .^{24} \tag{8}$$

Die Größe r(t) ist die einzige, die sich in unserem Beispiel nicht unmittelbar aus den Daten ablesen läßt. Bei ihrer Schätzung ist zu bedenken, daß wir von wöchentlichen Intervallen ausgehen und S(t) sich immer auf den Anfang des Intervalls bezieht. Es wäre aber sinnvoll, die Wahrscheinlichkeit, die Arbeitslosigkeit zu verlassen, auf die Survivorfunktion „während" des Intervalls zu beziehen, wofür sich hier der Mittelwert von $S(t_i)$ und $S(t_{i+1})$ anbietet. Daher beträgt die Hazardfunktion für das erste Intervall nicht einfach 0,1 sondern 0,1 / ((1 + 0,9) · 0,5) = 0,1 / 0,95 = 0,105, und in der letzten Zeile von Darstellung 2 hat die Hazardfunktion sogar einen Betrag von 2, da $f(t_{10})$ durch den Durchschnitt von 0,1 (=$S(t_{10})$) und 0 (= $S(t_{Ende})$) dividiert wird.[25] Für den hier dargestellten Life-Table-Schätzer ergibt sich also:

$$\hat{r}(t_i) = \frac{\hat{f}(t_i)}{(\hat{S}(t_i) + \hat{S}(t_{i+1}))/2} \quad . \tag{9}$$

Die Hazardfunktion kann man sich vorstellen als die *momentane Neigung oder Tendenz zu einem Zustandswechsel*, und insofern ist sie aus inhaltlichen Gründen gut geeignet, die Dynamik des Prozesses auszudrücken. Insbesondere sei noch einmal darauf hingewiesen, daß die Hazardfunktion sich im Verlauf des untersuchten Prozesses ändern kann. In unserem fiktiven Beispiel nimmt sie kontinuierlich zu, was bedeutet, daß die Wahrscheinlichkeit, die Arbeitslosigkeit zu verlassen, von Woche zu Woche größer wird. Es sind aber auch ganz andere Verläufe

denkbar. Weiter unten wird gezeigt, daß die Hazardfunktion auch aus „formaler" Sicht wichtig ist, da sich alle hier dargestellten Größen auf sie zurückführen lassen. Die Tatsache, daß die Hazardfunktion im Grunde eine unanschauliche (und unbeobachtbare) Größe ist, sollte nicht weiter irritieren. Die gleiche Feststellung gilt auch für die „momentane Geschwindigkeit" eines bewegten Objekts, und doch lassen sich hieraus sehr sinnvolle Aussagen ableiten, z.B. über die Zeit, die benötigt wird, um eine bestimmte Entfernung zurückzulegen. Und auch wer in die Geheimnisse der Infinitesimalrechnung, anhand derer sich ein Konzept wie die „momentane Geschwindigkeit" verstehen ließe, nicht eingeweiht ist, kennt den grundlegenden Sachverhalt, daß eine höhere Geschwindigkeit impliziert, daß man schneller am Ziel ist. Genau das gleiche gilt aber auch für die Hazardrate.[26]

Nachdem nunmehr die zentralen Größen S(t), f(t) und r(t) eingeführt sind, sollen sie im folgenden noch einmal für den Fall rechtszensierter Beobachtungen dargestellt werden.

Wir greifen hier auf Daten aus dem Sozio-ökonomischen Panel zurück *(Darstellung 3)*. Es handelt sich um die jeweils erste (nicht linkszensierte) Arbeitslosigkeitsepisode der *männlichen* Erwerbspersonen aus der *deutschen* Teilstichprobe;[27] verwendet werden die Daten aus den ersten sieben Erhebungswellen. Die Untersuchungspersonen sind nach dem Alter zu Beginn der Arbeitslosigkeit in drei Gruppen eingeteilt: Bis zu 30 Jahre, 31 bis 50 Jahre und über 50 Jahre. Später werden wir darauf eingehen, wie mögliche Unterschiede zwischen diesen Gruppen analysiert werden können. Aus Platzgründen verzichte ich auf eine Wiedergabe der Rohdaten, die sich ja leicht „zurückrechnen" lassen. Außerdem werden für die drei Altersgruppen nur die Daten der ersten 12 Monate dargestellt, tatsächlich sind die beobachteten Arbeitslosigkeitsdauern teilweise länger. Es werden nur die Übergänge in eine Vollzeitbeschäftigung als Ereignis gewertet. Es handelt sich um 630 Episoden, von denen 418 in eine Vollzeitbeschäftigung münden. Auch hier haben wir gruppierte Daten, wobei wegen der großen Stichprobe in den meisten Zeitintervallen mehrere Übergänge in die Beschäftigung stattfinden.

In diesem Beispiel treten nun rechtszensierte Daten auf (vgl. Spalte (3) in Darstellung 3).[28] Erstens haben einige Personen die entsprechenden Daten nicht vollständig angegeben, außerdem sind einige Personen während des Untersuchungszeitraumes ausgeschieden (Teilnahmeverweigerung, Nicht-Anwesenheit zum Befragungszeitpunkt, Umzug ohne Angaben über neuen Wohnort), und ferner waren einige Personen am Ende der siebten Erhebungswelle arbeitslos.[29] Schließlich gibt es Personen, die aus der Arbeitslosigkeit in einen anderen Zielzustand als eine Vollzeitbeschäftigung übergehen. Auch diese werden bei einer Analyse, die sich auf die Vollzeitbeschäftigung konzentriert, wie rechtszensierte Beobachtungen behandelt.

Darstellung 3: *Life-Table-Schätzung zur Arbeitslosigkeitsdauer von Männern*

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Alter bis 30 Jahre									
0 bis < 1	355	22	46	344.0	0.134	0.866	1.000	0.134	0.143
1 bis < 2	287	18	50	278.0	0.180	0.820	0.866	0.156	0.198
2 bis < 3	219	12	49	213.0	0.230	0.770	0.710	0.163	0.260
3 bis < 4	158	6	22	155.0	0.142	0.858	0.547	0.078	0.153
4 bis < 5	130	10	21	125.0	0.168	0.832	0.469	0.079	0.183
5 bis < 6	99	11	10	93.5	0.107	0.893	0.391	0.042	0.113
6 bis < 7	78	4	13	76.0	0.171	0.829	0.349	0.060	0.187
7 bis < 8	61	5	7	58.5	0.120	0.880	0.289	0.035	0.127
8 bis < 9	49	2	4	48.0	0.083	0.917	0.255	0.021	0.087
9 bis <10	43	1	4	42.5	0.094	0.906	0.233	0.022	0.099
10 bis <11	38	3	2	36.5	0.055	0.945	0.211	0.012	0.056
11 bis <12	33	3	3	31.5	0.095	0.905	0.200	0.019	0.100
Alter 31 bis 50 Jahre									
0 bis < 1	191	5	25	188.5	0.133	0.867	1.000	0.133	0.142
1 bis < 2	161	1	18	160.5	0.112	0.888	0.867	0.097	0.119
2 bis < 3	142	4	34	140.0	0.243	0.757	0.770	0.187	0.276
3 bis < 4	104	2	17	103.0	0.165	0.835	0.583	0.096	0.180
4 bis < 5	85	1	15	84.5	0.178	0.822	0.487	0.086	0.195
5 bis < 6	69	3	8	67.5	0.119	0.881	0.400	0.047	0.126
6 bis < 7	58	1	7	57.5	0.122	0.878	0.353	0.043	0.130
7 bis < 8	50	4	5	48.0	0.104	0.896	0.310	0.032	0.110
8 bis < 9	41	0	5	41.0	0.122	0.878	0.278	0.034	0.130
9 bis <10	36	0	1	36.0	0.028	0.972	0.244	0.007	0.028
10 bis <11	35	0	4	35.0	0.114	0.886	0.237	0.027	0.121
11 bis <12	31	1	5	30.5	0.164	0.836	0.210	0.034	0.179
Alter über 50 Jahre									
0 bis < 1	84	2	0	83.0	0.000	1.000	1.000	0.000	0.000
1 bis < 2	82	1	3	81.5	0.037	0.963	1.000	0.037	0.038
2 bis < 3	78	0	4	78.0	0.051	0.949	0.963	0.049	0.053
3 bis < 4	74	2	2	73.0	0.027	0.973	0.914	0.025	0.028
4 bis < 5	70	2	3	69.0	0.043	0.957	0.889	0.039	0.044
5 bis < 6	65	2	1	64.0	0.016	0.984	0.850	0.013	0.016
6 bis < 7	62	1	1	61.5	0.016	0.984	0.837	0.014	0.016
7 bis < 8	60	1	0	59.5	0.000	1.000	0.823	0.000	0.000
8 bis < 9	59	5	0	56.5	0.000	1.000	0.823	0.000	0.000
9 bis <10	54	2	0	53.0	0.000	1.000	0.823	0.000	0.000
10 bis <11	52	2	0	51.0	0.000	1.000	0.823	0.000	0.000
11 bis <12	50	8	0	46.0	0.000	1.000	0.823	0.000	0.000

Erläuterungen zu Darstellung 3 (vgl. auch Text):

(1): Intervall
(2): n zu Beginn des Intervalls
(3): (Rechts-)Zensierungen c_i während des Intervalls
(4): Ereignisse d_i während des Intervalls
(5): Risikomenge n' für das Intervall
(6): Bedingte „Sterbewahrscheinlichkeit“ für das Intervall
(7): Bedingte „Überlebenswahrscheinlichkeit“ p_i für das Intervall
(8): Survivorfunktion $S(t_i)$
(9): Dichtefunktion $f(t_i)$
(10): Hazardfunktion $r(t_i)$

Quelle: Das Sozio-ökonomische Panel, eigene Datenaufbereitung (jeweils erste Arbeitslosigkeitsepisode aus dem „Erwerbskalender“, Welle 1 bis 7).

Beim Vorliegen von Rechtszensierungen können die entscheidenden Größen zur Charakterisierung des Prozesses nicht mehr, wie in unserem ersten Beispiel, direkt aus den beobachteten Verweildauern abgelesen werden. Die „Lösung“ des Problems basiert auf der schon oben dargestellten Überlegung, den Beobachtungszeitraum in - möglichst kleine - Intervalle zu zerlegen, für jedes dieser Intervalle die bedingte Überlebenswahrscheinlichkeit p_i und daraus die relevanten Größen zu berechnen. D.h., wir nehmen - wie sich ja schon bei der „Transformation“ von Darstellung 1 in Darstellung 2 gezeigt hat - nicht mehr direkt auf die Verweildauer einer jeden Untersuchungsperson Bezug, sondern fragen umgekehrt für jede Zeiteinheit (jedes Intervall), welchen Beitrag jede Person zur Definition der relevanten Größen, insbesondere der Risikomenge und damit der bedingten Überlebenswahrscheinlichkeit, leistet.[30]

Hierzu geht man von folgender Überlegung aus. p_i ist der Anteil der „überlebenden“ Personen im i-ten Intervall, und zwar bezogen auf die Zahl der Personen, die während dieses Intervalls überhaupt dem „Risiko“ ausgesetzt waren, die Arbeitslosigkeit zu verlassen oder allgemeiner: in den Zielzustand überzugehen. Zu dieser Risikomenge gehören *auch* die Personen mit rechtszensierten Beobachtungsdauern, solange die Zensierung noch nicht eingetreten ist. Wenn man nun die Annahme, daß die Zensierungen zufällig erfolgen, im konkreten Fall für gültig erachtet, dann sollte der Anteil p_i in den einzelnen Intervallen nicht davon beeinflußt werden, wieviele Dauern gerade in diesem Intervall zensiert wurden; es kommt nur darauf an, eine angemessene Formulierung für die Risikomenge zu finden. Dabei ist zu bedenken, daß die Zensierungen (ebenso wie die „Ereignisse“) vermutlich nicht erst am Ende des Intervalls stattfanden, sondern irgendwann während des jeweiligen Intervalls erfolgt sein können, daß also die zensierten Fälle nicht während des gesamten Intervalls dem „Risiko“ ausgesetzt waren, die Arbeitslosigkeit zu verlassen. Als „Risikomenge“ kann daher nicht einfach die Anzahl der Personen aufgeführt werden, die bis zu Beginn des Intervalls noch „überlebt“ haben (weder die Arbeitslosigkeit verlassen haben noch zensiert waren - vgl. Spalte 2). Üblicherweise wird man davon ausgehen, daß „im Durchschnitt“ die zensierten Fälle während

der Hälfte des Intervalls zur Risikomenge zählten; es wird also von der Zahl der Personen, die bis zu Beginn des Intervalls „überlebten“, die Hälfte der während des Intervalls durch Zensierung ausgeschiedenen Personen abgezogen (vgl. Spalte 2, 3 und 5 in Darst. 3).

Die „bedingte Überlebenswahrscheinlichkeit“ für das i-te Intervall, p_i, wird also im Falle von Rechtszensierungen folgendermaßen geschätzt:

$$\hat{p}_i = \frac{Anzahl\ der\ Überlebenden}{Umfang\ der\ Risikomenge} .$$

Die Risikomenge im i-ten Intervall (Spalte 5 in Darst. 3) - wir bezeichnen sie mit n'_i im Unterschied zur Gesamtzahl n_i der Personen zu Beginn des Intervalls - läßt sich nach den obigen Erläuterungen berechnen, indem man von n_i die Hälfte der Zensierungen des betreffenden Intervalls - als c_i bezeichnet - abzieht:

$$n'_i = n_i - c_i / 2 . \tag{10}$$

Die Anzahl der Überlebenden, also die Zahl der Personen, die *kein* Ereignis „erlitten“, wird ebenfalls auf die Risikomenge bezogen. Wenn wir mit d_i die Anzahl der „Ereignisse“ während des Intervalls bezeichnen (Spalte 4 in Darst. 3), so wird die Anzahl der Überlebenden durch (n'_i - d_i) geschätzt (und nicht durch n_i - d_i). Damit ergibt sich (vgl. Spalte 7 in Darst. 3):

$$\hat{p}_i = \frac{n'_i - d_i}{n'_i} . \tag{11}$$

Mit dieser Größe lassen sich nun die Survivor-, Dichte- und Hazardfunktion, die in den Spalten 8 bis 10 von Darstellung 3 wiedergegeben sind, nach den Formeln (3), (5) und (9) berechnen. Die *wesentlichen Ergebnisse für die Beispielsdaten* sollen hier kurz zusammengefaßt werden:

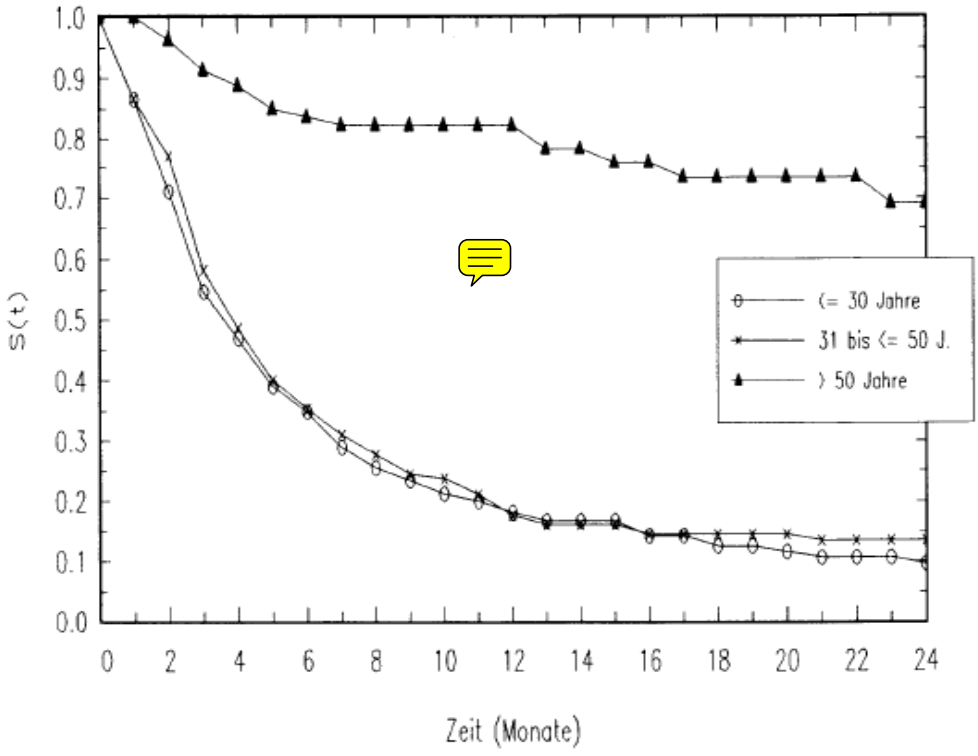
1. Am leichtesten verständlich sind sicherlich die Aussagen, die sich auf die Survivorfunktion beziehen. Wir können z.B. feststellen, daß nach 6 Monaten (also zu Beginn des 7. Intervalls!) *in den beiden jüngeren Altersgruppen* nur mehr ca. 40 Prozent in der Arbeitslosigkeit verblieben sind, es haben bis dahin also bereits 60 Prozent aller Übergänge in die Vollzeitbeschäftigung stattgefunden. Nach 9 Monaten beträgt der Anteil der noch arbeitslosen Personen in diesen beiden Gruppen nur mehr knapp 25 Prozent, über drei Viertel der Abgänge in eine Vollzeitbeschäftigung sind also bis zu diesem Zeitpunkt erfolgt. Die Unterschiede zwischen den beiden jüngeren Altersgruppen sind nur ganz geringfügig; die Survivorfunktion der 31- bis 50jährigen liegt ganz dicht bei oder nur wenig über der der jüngsten Altersgruppe. - Sehr große Unterschiede im Vergleich hierzu sind bei den *über 50jährigen* zu beobachten, bei denen nach 6 wie nach 9 Monaten über 80 Prozent noch keinen Übergang in eine Vollzeitbeschäftigung vollzogen haben.

2. Die Hazardfunktion gibt Auskunft über die „Risiken“ (inhaltlich natürlich: Chancen), in den einzelnen Monaten eine Vollzeitbeschäftigung aufzunehmen. Die wesentlich niedrigeren Verbleibsquoten in der Arbeitslosigkeit in den beiden jüngeren Altersgruppen schlagen sich in dementsprechend höheren Werten in der Hazardfunktion nieder. Von Interesse ist auch der zeitliche Verlauf der Hazardfunktion, die offenbar zunächst ansteigt und dann wieder abnimmt.

Besonders eingängig lassen sich die Ergebnisse aber *graphisch* darstellen, und ich will ihre Kommentierung anhand der folgenden Abbildungen noch einmal aufnehmen.[31]

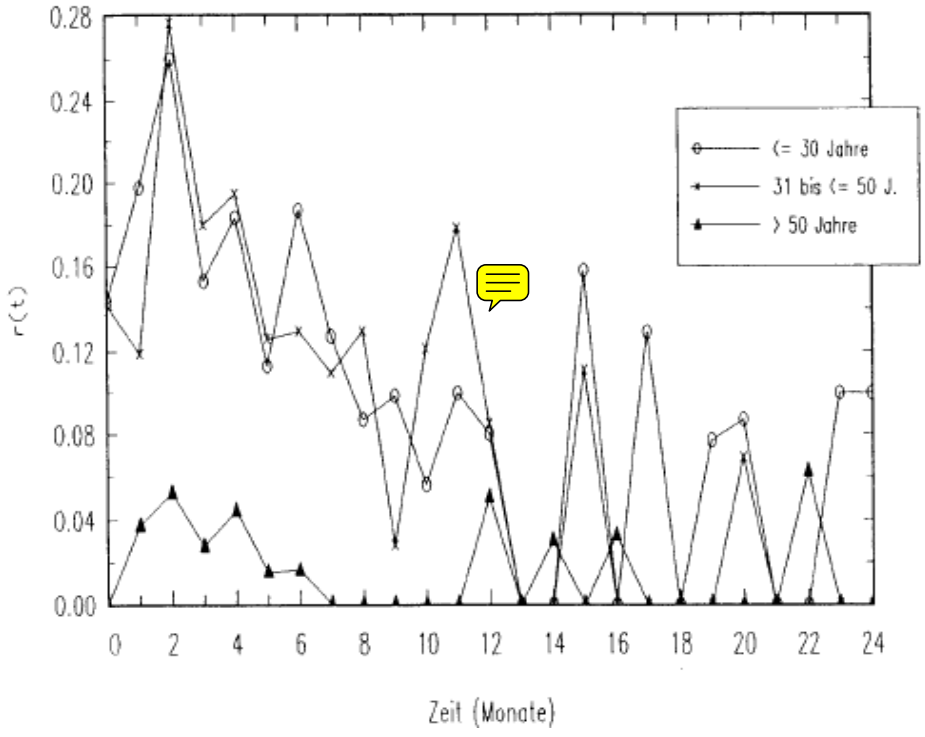
Die Survivorfunktion *(Darstellung 4)* verdeutlicht sehr plastisch, ob bzw. welche Unterschiede zwischen den Gruppen vorhanden sind; im vorliegenden Beispiel läßt sich weitaus schneller als durch Inspektion der Ausgangstabelle erkennen, daß die jüngste und die mittlere Altersgruppe sich in ihren Beschäftigungschancen kaum unterscheiden, während die Gruppe der über 50jährigen schlechtere Beschäftigungschancen aufweist.

Darstellung 4: *Life-Table-Schätzer der Survivor-Funktionen für die Beispielsdaten aus Darstellung 3 (Arbeitslosigkeit, Männer)*



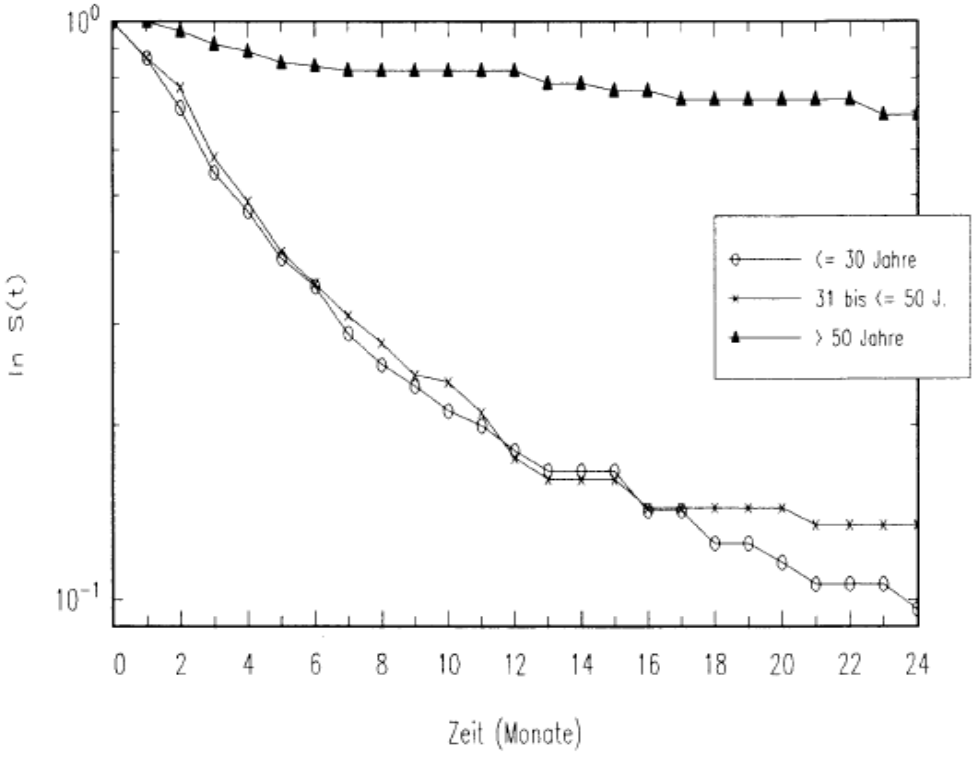
Ganz allgemein kann man eine Survivorfunktion wie eine Kreuztabelle lesen, die zeigt, welcher Anteil der Stichprobe bzw. der einzelnen Gruppen zu einem Zeitpunkt die Arbeitslosigkeit bereits verlassen hat und welcher Anteil noch nicht (wobei ersterer Wert - es handelt sich dabei um die oben erwähnte kumulierte „Sterbewahrscheinlichkeit“! - üblicherweise nicht ausgegeben wird, da er sich aus der Differenz von 1 und S(t) ergibt) - dies aber eben nicht nur für einen einzelnen Zeitpunkt, sondern für viele Zeitpunkte bzw. Zeitintervalle. Darüber hinaus lassen sich auch sehr einfach - wenn auch nicht ganz genau - Angaben zu den Verweildauern ablesen. So wird man sich häufig für die mittlere Verweildauer interessieren. Hierzu verwendet man üblicherweise nicht den Mittelwert, da dieser oft durch einige wenige lange Verweildauern beeinflußt wird, sondern den Median, also den Zeitpunkt, zu dem genau die Hälfte der Untersuchungspersonen ein Ereignis hatte, anders gesagt: zu dem S(t) den Wert von 0,5 hat. Dieser liegt in den beiden jüngeren Altersgruppen etwa bei 4 Monaten.[32] Für die älteste Gruppe läßt sich der Median allerdings gar nicht schätzen, weil auch zum Ende des Beobachtungszeitraumes gerade erst 30 Prozent der Männer über 50 Jahre die Arbeitslosigkeit verlassen haben. Grundsätzlich läßt sich aber nicht nur der Median, sondern der Wert eines jeden beliebigen Perzentils zum Vergleich heranziehen.

Darstellung 5: *Life-Table-Schätzer der Hazard-Funktionen für die Beispielsdaten aus Darstellung 3 (Arbeitslosigkeit, Männer)*



Die Hazardfunktion wird in *Darstellung 5* gezeigt. Ersichtlich ist diese wegen der häufigen starken Sprünge etwas verwirrender als die Survivorfunktion.[33] Die wesentlich besseren Chancen der beiden jüngeren Gruppen im Vergleich zu den über 50jährigen sind aber auch hier sehr gut zu erkennen. Ebenfalls deutlich wird, daß die Werte in den ersten 8 bis 10 Monaten höher liegen als im weiteren Verlauf; zumindest läßt sich dies für die beiden jüngeren Gruppen sagen, während bei der älteren Gruppe die Werte von Anfang an so niedrig liegen, daß ein Abfallen kaum möglich ist.[34] Auffällig ist auch, daß die höchsten Werte in allen drei Gruppen im dritten Intervall liegen, d.h., bis dahin steigen die Werte der Hazardfunktion sogar an, um dann allmählich zurückzugehen. Mit anderen Worten: Im dritten Monat der Arbeitslosigkeit sind die (relativen) Chancen, eine Beschäftigung aufzunehmen, am höchsten. Alles in allem zeigt sich jedenfalls an dieser Darstellung sehr gut, daß die Hazardfunktion - vorbehaltlich der Einbeziehung weiterer erklärender Variablen - *zeitabhängig*, also nicht während der gesamten Dauer des Prozesses gleich ist.[35]

Darstellung 6: *Logarithmierte Survivor-Funktionen (Life-Table-Schätzer) für die Beispielsdaten aus Darstellung 3 (Arbeitslosigkeit, Männer)*



Um die Zeitabhängigkeit des Prozesses zu untersuchen, wird häufig auch vorgeschlagen, die Survivorfunktion auf einer logarithmischen Skala abzutragen (vgl. *Darstellung 6*). Besteht keine Zeitabhängigkeit, muß sich eine Gerade ergeben, ansonsten werden die Kurven steiler oder flacher, je nachdem ob die Hazardfunktion zu- oder abnimmt. Im vorliegenden Beispiel wird vor allem die Abnahme der Hazardfunktion etwa ab dem achten Monat deutlich, bei genauem Hinsehen läßt sich auch erkennen, daß der Plot im dritten Monatsintervall besonders steil ist.

Zum Abschluß dieses Teils möchte ich noch einmal darauf eingehen, daß die bislang verwendeten Formeln auf einer Betrachtung basieren, die die eigentlich stetige oder kontinuierliche Zeit in Intervalle zerlegt. Man kann sich aber leicht vorstellen, daß man sich einer kontinuierlichen Betrachtung annähert, indem man die Zeitintervalle möglichst klein werden, also tendenziell gegen Null gehen läßt. Damit ergeben sich folgende Definitionen für Dichte- und Hazardfunktion bei stetiger Zeit, die ich hier deshalb anführe, weil sie in den Lehrbüchern oder Übersichtsartikeln manchmal im Vordergrund stehen:

$$f(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \le T < t + \Delta t)}{\Delta t} \text{ ,}^{36} \tag{12}$$

und entsprechend

$$r(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \text{ .} \tag{13}$$

Diese „abstrakten“ Größen können natürlich nicht mehr unmittelbar aus den Daten abgelesen oder einfach berechnet werden; hierzu müssen komplexere Schätzverfahren angewandt werden (vgl. Teil II). An dieser Stelle sei nur kurz gezeigt, daß sich die Größen f(t) und S(t) auch im stetigen Fall aus der Hazardfunktion bzw. Übergangsrate r(t) berechnen lassen, was ein wichtiger Grund dafür ist, daß in den komplexeren Schätzverfahren r(t) im Mittelpunkt steht.

Zunächst ergibt sich aus der Beziehung r(t) = f(t) / S(t) die umgekehrte Beziehung f(t) = r(t) · S(t). Es muß also nur gezeigt werden, daß S(t) sich aus r(t) berechnen läßt. Der Beweis hierfür übersteigt den Charakter einer Einführung (vgl. etwa DM: 195), es läßt sich aber zeigen, daß gilt:

$$S(t) = \exp\left[-\int_0^t r(u)du\right] . \tag{14}$$

Der Ausdruck ∫r(u)du ist das Integral der Hazardfunktion r(t) von 0 bis zum Zeitpunkt t und wird auch als *kumulierte Hazardfunktion* bezeichnet. Intuitiv ver-

ständlich machen läßt sich diese Größe an dem obigen Beispiel mit gruppierten Verweildauern, wenn man daran denkt, daß ein Integral durch eine Summierung - hier also: vom Beginn des Prozesses bis zum Zeitpunkt t - angenähert werden kann. Tatsächlich läßt sich auch in den Beispieldaten zeigen, daß sich S(t) zu jedem Zeitintervall aus der Summe aller bis dahin „angefallenen“ r(t) nach folgender Gleichung ergibt:[37]

$$S(t_i) = \exp\left[-\sum_{j=0}^{i-1} r(t_j)\right]. \qquad (15)$$

Anhang

Statistik-Software zur Verlaufsdatenanalyse

Die hier vorgestellten einfachen Life-Table-Schätzer können mit allen verbreiteten Statistikprogrammen (SPSS, BMDP, SAS, SYSTAT) berechnet werden. Für komplexere Verfahren reichen deren Möglichkeiten (mit Ausnahme von SAS) im allgemeinen nicht aus. Daher sei schon hier auf das Programm TDA (Transition Data Analysis) hingewiesen, welches von Götz Rohwer entwickelt wurde und das umfassendste Programm zur Verlaufsdatenanalyse darstellt, das derzeit verfügbar ist (erhältlich gegen Einsendung von 4 Disketten - 1,44 MB, MS-DOS-formatiert - und eines frankierten Rückumschlags; Anschrift: Universität Bremen, Fachbereich 8, Postfach 330 440, 28334 Bremen). Weitere Hinweise auf Spezialsoftware enthält Teil II.

Anmerkungen

1 Ich danke Günter Albrecht und Götz Rohwer für die kritische Durchsicht des Manuskripts. Für die noch vorhandenen Fehler bin ich natürlich selbst verantwortlich.

2 Vgl. für Einkommensarmut bzw. allgemein Einkommenslagen die Analysen von Bonß/Plum (1990) und Rohwer (1991, 1992); für Sozialhilfe Leisering/Zwick (1990) und Voges/Rohwer (1991); für Arbeitslosigkeit Ludwig-Mayerhofer (1990, 1992).

3 Vgl. beispielsweise zu „Patientenkarrieren“ Gerhardt (1986) und Keupp (1987), zu „Arbeitslosigkeitskarrieren“ (deren Häufigkeit vielfach überschätzt wird) Andreß (1989) und Wagner (1990), für den Bereich der Prostitution Hess (1978). Siehe auch, auf der Grundlage qualitativer Forschung, allerdings (vielleicht etwas zu) skeptisch gegenüber möglichen „karrierehaften“ Verfestigungsprozessen von Abweichung, Mutz/Kühnlein (1992).

4 Eine Warnung vor zu großer Euphorie hinsichtlich der hier vorgestellten Modelle formuliert Esser (1987). - Ich will im übrigen keinesfalls bestreiten, daß sich „Prozesse“, „Karrieren“ usw. auch mit Methoden der „qualitativen“ Forschung untersuchen lassen. Allerdings stellt sich auch hier die Frage, ob die häufig zu findende Vorstellung einer grundsätzlichen Gegensätzlichkeit von qualitativer und quantitativer Forschung nicht jedenfalls teilweise auf Mißverständnissen beruht (vgl. etwa Diekmann 1987 und Ostner 1987).

5 Vgl. die gute Darstellung bei Andreß (1984, S. 250 ff.) Das muß wohl „Die ersten 10 Berufsjahre“ sein! sowie das eindrucksvolle Beispiel bei Schneider (1991, S. 229 ff.). Dem steht nicht entgegen, daß in der Literatur gelegentlich *modifizierte* Verfahren der linearen Regression vorgeschlagen werden, die allerdings kaum Verbreitung gefunden haben.

6 Daher die Bezeichnungen „Analysis of Transition Data“ (Lancaster) oder „Event History Analysis“ (Tuma) bzw. „Ereignisanalyse“ (BHM).

7 Wenn die Annahme diskreter „Zustände“ nicht gerechtfertigt ist, also ein kontinuierlicher Zustandsraum angenommen wird, kann die Modellierung mit stochastischen Differentialgleichungen erfolgen, vgl. Blossfeld/Hannan/Schömann (1988).

8 Ein gravierendes Problem ist hier (wie anderswo) der uneinheitliche Sprachgebrauch. So verwendet Andreß (1992, S. 40) den Begriff „aggregiert“ noch ein einem anderen Sinn als ich es hier - im Anschluß an Petersen (1991b) bzw. Petersen/Koput (1992) - getan habe; Hamerle/Tutz (1989) gebrauchen den Begriff „diskret“ dagegen im Sinne von „gruppiert“ bzw. „aggregiert“. Dementsprechend wird in der Literatur manchmal zwischen den verschiedenen Fällen nicht hinreichend unterschieden.

9 Außerdem ist zu beachten: Hujer/Schneider (1986) schlagen vor, zensierte Zeiten sogar um eine ganze Zeiteinheit zu kürzen, während die Simulationsstudie von Petersen/Koput (1992) davon ausging, daß Zensierungen exakt gemessen wurden - eine nicht ganz realistische Annahme.

10 Einer der wichtigsten Zensierungsmechanismen ist die faktisch beschränkte Beobachtungsdauer von Untersuchungen: Alle Individuen werden über den gleichen Zeitraum beobachtet, einige haben aber am Ende den Zielzustand noch nicht erreicht. Ebenso können aber aus den unterschiedlichsten Gründen - z.B. wegen Panelmortalität, aber auch wegen des Studiendesigns - verschieden lange Beobachtungsdauern vorliegen. Eine ausführliche Diskussion findet sich etwa bei Lawless (1982, S. 31 ff.).

11 Das von DM: 91 vorgeschlagene Verfahren, Zensierungen als Ereignisse zu betrachten und so die Zensierungsmuster verschiedener Gruppen zu vergleichen, ist nur eine ad-hoc-Lösung. Weder ist die Gleichheit der Zensierungsmuster in den Gruppen eine Garantie der Unabhängigkeit von Zensierungen und Ereignissen, noch müssen Unterschiede zwischen den Gruppen notwendig auf eine Abhängigkeit hindeuten. Man kommt also ohne theoretische Überlegungen oder zusätzliche Untersuchungen nicht aus.

12 Allerdings wäre es auch unabhängig von diesem Problem zumeist nicht sinnvoll, beim Arbeitslosenbestand eines bestimmten Zeitpunktes anzusetzen, da in diesem jeweils die längerfristig Arbeitslosen überrepräsentiert sind. Hier wie auch sonst kann je nach Fragestellung eine Zugangsstichprobe sinnvoller sein, die jedoch auch mit Problemen behaftet ist (vgl. Diekmann/Mitter 1990, S. 432; Diekmann/Mitter 1993, S. 53).

13 Deutschen spricht man dagegen von „Sterbetafel“. Diese Ausdrücke verweisen ebenso wie der später eingeführte Begriff der „Survivor-Funktion“ darauf, daß zentrale Impulse zur Entwicklung der hier vorgestellten Verfahren aus dem Bereich der Lebensversicherung stammen.

14 Ähnlich elementare Einführungen finden sich bei Fleiss/Dunner/Stallone/Fieve (1976) sowie bei Peto et al. (1977, Abschnitt 18 ff.).

15 Mit Tukey (1977) halte ich die getrennte Numerierung von Tabellen und Abbildungen für unnötig (und gelegentlich eine Quelle von Konfusion), fasse beides unter dem Oberbegriff „Darstellung“ (Tukey: „exhibit“) zusammen und numeriere durchgängig.

16 In den statistischen Verfahren geht es natürlich um abstrakte Zeiteinheiten; sie „wissen“ nicht, ob sich die Verweildauern wie hier auf Wochen oder aber auf Sekunden, Tage, Jahre oder beliebige andere Zeiteinheiten beziehen. In manchen Programmen - etwa BMDP - läßt sich die verwendete Zeiteinheit explizit angeben, dies ist jedoch nur für die Beschriftung von Tabellen oder Graphiken von Bedeutung.

17 Für die Auswertung müssen die Daten allerdings anders kodiert werden, da die Statistikprogramme eine Dauer von 1 so interpretieren würden, daß die Dauer eine ganze Zeiteinheit betrug,

konkret also, daß die Arbeitslosigkeit *nach einer Woche, d.h. genau zum Beginn der zweiten Woche* verlassen wurde. Für das hier vorgestellte Verfahren muß bei einer Beendigung während des ersten Zeitintervalls ein Wert aus dem Bereich $0 \leq T < 1$ gewählt werden (analog für die folgenden Intervalle). Man beachte also, daß eine Dauer von genau 0 von den meisten Programmen für einfache (non-parametrische) Auswertungen zugelassen und so interpretiert wird, daß das Ereignis zwischen dem Zeitpunkt 0 und dem nächsten Zeitpunkt liegt. Bei komplexeren parametrischen Verfahren (siehe Teil II), wo Angaben zur Verweildauer als exakte Messungen aufgefaßt werden, wäre eine Kodierung mit 0 allerdings nicht sinnvoll.

18 Faktisch fungieren natürlich nach wie vor die Untersuchungspersonen als Analyseeinheit - schließlich sind die Daten in Darstellung 2 nur eine andere Form des Arrangements der Daten aus Darstellung 1.

19 Genauer gesagt handelt es sich um eine Schätzung (hier: der Überlebenswahrscheinlichkeit), insoweit davon ausgegangen wird, daß es sich bei den Untersuchungseinheiten praktisch immer um eine Stichprobe und nicht um die Grundgesamtheit handelt. Dies gilt auch für die anderen Funktionen, die im folgenden erläutert werden.

20 Der Ausdruck ist zu lesen: S(t) ist die Wahrscheinlichkeit, daß die konkrete Verweildauer einer Untersuchungseinheit, T, größer oder gleich einer bestimmten (beliebigen) Zeit t ist. - Manche Autoren definieren S(t) als $P(T > t)$ (Lee 1980, S. 10; Heckman/Singer 1986, S. 1693). Die substantiellen Unterschiede beider Definitionen sind unerheblich. Es ist nur genau darauf zu achten, wie die einzelnen Programme S(t) ausgeben; beispielsweise wird beim Life-Table-Schätzer von SPSS die Survivorfunktion immer für das *Ende* des jeweiligen Intervalls dargestellt. Grundsätzlich ist darauf hinzuweisen, daß die Literatur hinsichtlich der Definition der relevanten Größen, insbesondere des Zeitbezugs, etwas uneinheitlich ist, wodurch sich teilweise unterschiedliche Formeln ergeben.

21 Die tiefergestellten Indices sollen anzeigen, daß es sich jeweils um die Survivorfunktion - oder andere Größen - für ein bestimmtes Intervall handelt. Streng genommen müßte S(t) zu Beginn des Prozesses als $S(t_0)$ bezeichnet werden, da das erste Intervall eben zum Zeitpunkt 0 beginnt und S(t) sich immer auf den Beginn des Intervalls bezieht. Da wir jedoch vom ersten, zweiten Intervall usw. sprechen, würde es Verwirrung stiften, wenn der Zeitpunkt von S(t) immer von der Ordnungszahl des Intervalls abweichen würde.

22 Anzumerken ist, daß die Formulierungen für f(t) und im folgenden für r(t) eine Intervallbreite von 1 unterstellen. Die häufig vorzufindenden komplizierteren Formeln entstehen dadurch, daß man beliebige Intervallbreiten definieren kann und dies entsprechend berücksichtigen muß (besonders mißlich ist dies bei DM: 65 ff., da sie die Intervallbreite h_i zwar in ihren Formeln verwenden, aber nirgends einführen). Im Beispiel: Da eine Woche sieben Tage enthält, läßt sich auch eine auf Tage bezogene Dichtefunktion berechnen, indem man die „wöchentliche" Dichte von 0,1 unseres Beispiels durch 7 dividiert.

23 Ich gebrauche den Begriff Hazardfunktion als Oberbegriff. Im Fall stetiger Zeiten spricht man im allgemein von Hazard*rate*, aber manche Autoren (bzw. Statistikprogramme) gebrauchen diesen Begriff auch bei diskreten oder gruppierten Zeiten. Ferner wird häufig der Begriff „Übergangsrate" gebraucht. Vielfach wird dieser Begriff synonym zum Begriff Hazardrate verwendet (z.B. BHM: 31), andere gebrauchen den Begriff *Übergangs*rate für die Wahrscheinlichkeit eines je spezifischen Übergangs (z.B. Arbeitslosigkeit - Beschäftigung) und den Begriff *Hazard*rate für die Wahrscheinlichkeit, daß irgendein beliebiger Übergang (von möglicherweise mehreren, unterschiedlichen Übergängen) stattfindet (z.B. DM: 51).

24 Der Ausdruck $P(t \leq T < t+1 | T \geq t)$ ist zu lesen als „Wahrscheinlichkeit, daß der Zustandswechsel zwischen t und t+1 liegt (daß also die Verweildauer T eines Individuums in den Zeitraum zwischen t und t+1 fällt), gegeben, daß bis zum Zeitpunkt t noch kein Zustandswechsel stattfand".

25 Daher wird auch häufig betont, daß die Hazardfunktion keine echte Wahrscheinlichkeit ist, denn Wahrscheinlichkeiten können nur Werte zwischen 0 und 1 annehmen. Das unterscheidet sie von der „bedingten Sterbewahrscheinlichkeit“, obwohl natürlich die Ähnlichkeit mit dieser recht groß ist.

26 Vgl. ferner zu diesem Aspekt Petersen (1991a, S. 295). Tatsächlich läßt sich aus den hier dargestellten Größen teilweise die geschätzte durchschnittliche Dauer bis zum Zustandswechsel ableiten, doch ist dies häufig gerade dann nicht möglich, wenn die Hazardrate im Zeitverlauf variiert (vgl. DM: 47).

27 Genauer: Aus allen Haushalten der deutschen Teilstichprobe (sog. „Stichprobe A“) wurden diejenigen Personen ausgewählt, die über alle 7 Wellen eine deutsche Staatsangehörigkeit hatten. - Es wurden auch die Daten zu den weiblichen Arbeitslosen ausgewertet, die Männer sind aber im konkreten Fall „interessanter“ - unter rein *didaktischen* Gesichtspunkten. Eine *inhaltliche* Untersuchung zur Arbeitslosigkeit, bei der es u.a. ganz zentral um den Vergleich von Frauen und Männern geht, befindet sich in Vorbereitung.

28 Im Datensatz befinden sich also 22 Personen im Alter bis 30 Jahre, die nur einen Monat beobachtet werden konnten und am Ende der Beobachtungsdauer noch arbeitslos waren, 46 Personen dieser Altersgruppe mit einer Beobachtungsdauer von 1 Monat, die während dieser Zeit eine Vollzeitbeschäftigung antraten, usw. Die Daten müssen also auf jeden Fall eine dichotome Variable als Zensierungsindikator enthalten, der angibt, ob eine Beobachtungsdauer mit einem Zustandswechsel endete oder nicht. (Auch bei den Daten in Darstellung 1 ist zur Auswertung eine solche Variable notwendig, sie kann jedoch als Konstante jederzeit beliebig erzeugt werden). Bei Mehrzustandsmodellen tritt an die Stelle des dichotomen Zensierungsindikators eine Variable mit mehreren Ausprägungen.

29 Natürlich waren diese nicht sieben Jahre lang arbeitslos. Die meisten der am Ende des siebten Jahres arbeitslos Gebliebenen waren erst während dieses Jahres arbeitslos geworden.

30 Dieser Gedankengang findet sich am deutlichsten bei Lawless (1982, S. 52 ff.), sowie, etwas knapper, bei Petersen (1990, S. 263 ff.; 1991a, S. 274 f.).

31 Die nachfolgend dargestellten Plots werden auch von den einschlägigen Statistikprogrammen ausgegeben. Im Gegensatz zu Darstellung 3 werden die ersten 24 Monate wiedergegeben.

32 Formeln zu exakten Schätzung des Medians finden sich z.B. bei Andreß (1992, S. 143). Im konkreten Fall beträgt die Schätzung 3,61 Monate für die bis 30jährigen und 3,86 für die über 30- bis 50jährigen.

33 Im Programm TDA sind Verfahren zur „Glättung“ der hier gezeigten Funktionen implementiert, wodurch eine übersichtlichere Darstellung möglich wird. Mir geht es hier aber gerade um die Verdeutlichung, wie die ursprünglichen, nicht geglätteten Funktionen aussehen. Die Glättung mag in einem Fall wie dem vorliegenden, wo die Dauern gar nicht exakt gemessen wurden, vielleicht auch die Datenqualität überstrapazieren.

34 Bei genauem Hinsehen läßt sich aber sagen, daß die Werte für die älteste Gruppe in den ersten 6 Monaten - abgesehen vom allerersten Monat! - stets über Null liegen, während danach immer wieder längere oder kürzere Intervalle mit einer Hazardfunktion von Null beobachtet werden, so daß man auch hier davon sprechen kann, daß im Durchschnitt die Hazardfunktion anfangs höhere Werte annimmt.

35 Nur kurz angemerkt sei, daß sich die hier beobachtete Form der Hazardfunktion - Anstieg bis zum dritten Monat und nachfolgender Rückgang - bei *allen* explorativen Analysen der männlichen Arbeitslosen zeigt, welche Variable auch immer man heranziehen mag. Das kann als Indiz dafür gelten, daß es sich hierbei nicht um ein Artefakt handelt (vgl. Ludwig-Mayerhofer 1992 und Klein 1992).

36 Der senkrechte Pfeil im Ausdruck „$\Delta t \downarrow 0$“ soll bedeuten, daß Δt sich dem Grenzwert 0 von oben, also aus dem positiven Wertebereich nähert.

37 Da $S(t_i)$ immer zu Beginn eines Intervalls definiert wurde, dürfen natürlich nur die $r(t_i)$ bis zum Beginn des i-ten Intervalls herangezogen werden. Bei dieser Überlegung ergibt sich ebenfalls $S(t_1) = 1$, da ja die kumulierte Hazardrate zu Beginn des Prozesses 0 ist und exp(-0) = 1. - Wegen Rundungen, und weil Formel (15) nur eine Näherung darstellt, ergeben sich im Beispiel beim Nachrechnen kleine Abweichungen.

Literatur

Allison, P. D., 1984: Event History Analysis. Regression for Longitudinal Event Data. Beverly Hills: Sage.

Andreß, H.-J., 1984: Die ersten zehn Berufsjahre. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (BeitrAB 87).

Andreß, H.-J., 1989: Instabile Erwerbskarrieren und Mehrfacharbeitslosigkeit - ein Vergleich mit der Problemgruppe der Langzeitarbeitslosen. MittAB 22: 17-32.

Andreß, H.-J., 1992: Verlaufsdatenanalyse (Historical Social Research/Historische Sozialforschung, Supplement/Beiheft No. 5). Köln: Zentrum für Historische Sozialforschung.

Arminger, G., 1984: Modelltheoretische und methodische Probleme bei der Analyse von Paneldaten mit qualitativen Variablen. Vierteljahreshefte zur Wirtschaftsforschung: 470-479.

Arminger, G., 1988: Modelle zur Analyse qualitativer Variablen in stetigem Zeitverlauf. S. 77-91 in: Meier, F. (Hrsg.), Prozeßforschung in den Sozialwissenschaften. Anwendungen zeitreihenanalytischer Methoden. Stuttgart, New York: G. Fischer.

Blossfeld, H.-P./Hamerle, A./Mayer, K. U., 1986: Ereignisanalyse. Frankfurt/New York: Campus.

Blossfeld, H.-P./Hamerle, A., 1989: Using Cox Models to Study Multiepisode Processes. Sociological Methods & Research 17: 432-448.

Bonß, W./Plum, W., 1990: Gesellschaftliche Differenzierung und sozialpolitische Normalitätsfiktion. Zeitschrift für Sozialreform 36: 692-715.

Breslow, N., 1991: Use of the Logistic and Related Models in Longitudinal Studies of Chronic Disease Risk. S. 163-197 in: Dwyer, J. H./Feinleib, M./Lippert, P./Hoffmeister, H. (Hrsg.), Statistical Models for Longitudinal Studies of Health. (Monographs in Epidemiology and Biostatistics, Vol. 16). New York, Oxford: Oxford University Press.

Carroll, G. R., 1983: Dynamic Analysis of Discrete Dependent Variables: A Didactic Essay. Quality & Quantity 17: 425-460.

Crouchley, R. (Hrsg.), 1987: Longitudinal Data Analysis. Aldershot: Avebury.

Diekmann, A., 1987: Lebensverläufe und Verlaufsdatenanalyse - Statistische Auswertungsmethoden von Ereignisdaten. S. 171-196 in: Voges, W. (Hrsg.), Methoden der Biographie- und Lebenslaufforschung. Opladen: Leske + Budrich.

Diekmann, A., 1988: Ereignisdatenanalyse - Beispiele, Probleme und Perspektiven. ZUMA- Nachrichten 23: 7-25.

Diekmann, A./Mitter, P., 1984a: Methoden zur Analyse von Zeitverläufen. Stuttgart: Teubner.

Diekmann, A./Mitter, P., 1984b: Stochastic Modelling of Social Processes. Orlando: Academic Press.

Diekmann, A./Mitter, P., 1990: Stand und Probleme der Ereignisanalyse. S. 404-441 in: Mayer, K. U. (Hrsg.), Lebensverläufe und sozialer Wandel. (Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 31). Opladen: Westdeutscher Verlag.

Diekmann, A./Mitter, P., 1993: Methoden der Ereignisanalyse in der Bevölkerungssoziologie: Stand und Probleme. S. 20-65 in: Diekmann, A./Weick, S. (Hrsg.), Der Familienzyklus als sozialer Prozeß. Bevölkerungssoziologische Untersuchungen mit den Methoden der Ereignisanalyse. (Sozialwissenschaftliche Schriften, Heft 26). Berlin: Duncker & Humblot.

Esser, H., 1987: Warum die Routine nicht weiterhilft. Überlegungen zur Kritik an der "Variablen-Soziologie". S. 230-245 in: Müller, N./Stachowiak, H. (Hrsg.), Problemlösungsoperator Sozialwissenschaft, Band 1. Stuttgart: Enke.

Fleiss, J. L./Dunner, D. L./Stallone, F./Fieve, R. R., 1976: The Life Table. A Method for Analyzing Longitudinal Studies. Archives of General Psychiatry 33: 107-112.

Galler, H. P., 1986: Übergangsratenmodelle bei intervalldatierten Ereignissen. Statistische Hefte 27: 1-22.

Galler, H. P./Pötter, U., 1992: Zur Robustheit von Schätzmodellen für Ereignisdaten. S. 379-405 in: Hujer, R./Schneider, H./Zapf, W. (Hrsg.), Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel. Frankfurt/New York: Campus.

Gerhardt, U., 1986: Patientenkarrieren. Frankfurt a.M.: Suhrkamp.

Hamerle, A./Tutz, G., 1989: Diskrete Modelle zur Analyse von Verweildauern und Überlebenszeiten. Frankfurt/New York: Campus.

Heckman, J./Singer, B., 1986: Econometric Analysis of Longitudinal Data. S. 1689-1763 in: Griliches, Z./Intriligator, M. D. (Hrsg.), Handbook of Econometrics, Volume III. Amsterdam: Elsevier.

Heijtan, D. F., 1989: Inference from Grouped Continuous Data: A Review. Statistical Science 4: 163-183.

Hess, H., 1978: Das Karriere-Modell und die Karriere von Modellen. S. 1-30 in: Hess, H./ Störzer, H. U./Streng, F. (Hrsg.), Sexualität und soziale Kontrolle. Heidelberg: Kriminalistik.

Hujer, R./Schneider, H., 1986: Semi-parametrische und parametrische Ratenmodelle. Eine anwendungsbezogene Einführung in die statistischen Grundlagen mit Programmbeispielen. Frankfurt: Sonderforschungsbereich 3, Arbeitspapier Nr. 200.

Hujer, R./Schneider, H., 1989: The Analysis of Labor Market Mobility Using Panel Data. European Economic Review 33: 530-536.

Hutchison, D., 1988a: Event History and Survival Analysis in the Social Sciences I. Background and Introduction. Quality & Quantity 22: 203-219.

Hutchison, D., 1988b: Event History and Survival Analysis in the Social Sciences II. Advanced Applications and Recent Developments. Quality & Quantity 22: 255-278.

Jacobs, H./Ringbeck, A., 1992: Zweiter Zwischenbericht zum Projekt "Hilfen zur Überwindung von Sozialhilfebedürftigkeit" im Auftrag des Bundesministeriums für Familie und Senioren. Köln: ISG (Institut für Sozialforschung und Gesellschaftspolitik).

Kalbfleisch, J. D./Prentice, R. L., 1980: The Statistical Analysis of Failure Time Data. New York: Wiley.

Keupp, H., 1987: Psychisches Leid als gesellschaftlich produzierter Karriereprozeß. S. 341-366 in: Voges, W. (Hrsg.), Methoden der Biographie- und Lebenslaufforschung. Opladen: Leske + Budrich.

Kiefer, N. M., 1988: Economic Duration Data and Hazard Functions. Journal of Economic Literature 26: 646-679.

Klein, T., 1992: Zur Zeitabhängigkeit der Wiederbeschäftigungsrate Arbeitsloser. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44: 134-138.

Lancaster, T., 1990: The Econometric Analysis of Transition Data. Cambridge: Cambridge University Press.

Lawless, J. F., 1982: Statistical Models and Methods for Lifetime Data. New York: John Wiley.

Lee, E. T., 1980: Statistical Methods for Survival Data Analysis. Belmont, CA: Lifetime Learning.

Leisering, L./Zwick, M., 1990: Heterogenisierung der Armut? Alte und neue Perspektiven zum Strukturwandel der Sozialhilfeklientel in der Bundesrepublik Deutschland. Zeitschrift für Sozialreform 36: 715-745.

Ludwig-Mayerhofer, W., 1990: Arbeitslosigkeit im Erwerbsverlauf. Zeitschrift für Soziologie 19: 345-359.

Ludwig-Mayerhofer, W., 1992: Fakt und Artefakt in der Analyse von Arbeitslosigkeitsverläufen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44: 124-133.

Meinken, H., 1992: Die Modellierung zeitstetiger sozialer Prozesse - Untersuchungsmethoden für Lebensverlaufsereignisse. S. 67-88 in: Andreß, H. J./Huinink, J./Meinken, H./Rumianek, D./Sodeur, W./Sturm, G. (Hrsg.), Theorie, Daten, Methoden. Neue Modelle und Verfahrensweisen in den Sozialwissenschaften. München: R. Oldenbourg.

Mutz, G./Kühnlein, I., 1992: Zur Reformulierung des Karriere-Modells: Theoretische Skizze und empirische Ergebnisse einer Fallstudie. mps-texte 1: 37-50.

Namboodiri, N. K./Suchindran, C. M., 1987: Life Table Techniques and Their Applications. Orlando: Academic Press.

Ostner, I., 1987: Scheu vor der Zahl? Die qualitative Erforschung von Lebenslauf und Biographie als Element einer feministischen Wissenschaft. S. 103-124 in: Voges, W. (Hrsg.), Methoden der Biographie- und Lebenslaufforschung. Opladen: Leske + Budrich.

Petersen, T., 1990: Analyzing Event Histories. S. 259-288 in: von Eye, A. (Hrsg.), Statistical Methods in Longitudinal Research, Volume II. San Diego: Academic Press.

Petersen, T., 1991a: The Statistical Analysis of Event Histories. Sociological Methods and Research 19: 270-323.

Petersen, T., 1991b: Time-Aggregation Bias in Continuous-Time Hazard-Rate Models. S. 263-290 in: Marsden, P. V. (Hrsg.), Sociological Methodology 1991. Cambridge, MA: Basil Blackwell.

Petersen, T., 1993: Recent Advances in Longitudinal Methodology. Annual Review of Sociology 19: 425-454.

Petersen, T./Koput, K. W., 1992: Time-Aggregation Bias in Hazard-Rate Models With Covariates. Sociological Methods and Research 21: 25-51.

Peto, R./Pike, M. C./Armitage, P./Breslow, N. E./Cox, D. R./Howard, S. V./Mantel, N./ McPherson, K./Peto, J./Smith, P. G., 1977: Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient II: Analysis and Examples. British Journal of Cancer 35: 2-39.

Projektgruppe „Das Sozio-ökonomische Panel“, 1990: Das Sozio-ökonomische Panel für die Bundesrepublik Deutschland nach fünf Wellen. Vierteljahreshefte für Wirtschaftsforschung: 141-151.

Rohwer, G., 1991: Einkommensmobilität privater Haushalte 1984-1989. S. 379-408 in: Rendtel, U./Wagner, G. (Hrsg.), Lebenslagen im Wandel: Zur Einkommensdynamik in Deutschland seit 1984. Frankfurt/New York: Campus.

Rohwer, G., 1992: Einkommensmobilität und soziale Mindestsicherung. Einige Überlegungen zum Armutsrisiko. S. 367-379 in: Leibfried, S./Voges, W. (Hrsg.), Armut im modernen Wohlfahrtsstaat. (Sonderheft 32 der Kölner Zeitschrift für Soziologie und Sozialpsychologie). Opladen: Westdeutscher Verlag.

Rohwer, G., 1993: TDA Working Papers, Bremen, Ms.

Schneider, H., 1991: Verweildaueranalyse mit GAUSS. Frankfurt/New York: Campus.

Teachman, J. D., 1983: Analyzing Social Processes: Life Tables and Proportional Hazards Models. Social Science Research 12: 263-301.

Toutenburg, H., 1992: Moderne nichtparametrische Verfahren der Risikoanalyse. Heidelberg: Physica.

Tukey, J. W., 1977: Exploratory Data Analysis. Reading, MA: Addison & Wesley.

Tuma, N. B., 1982: Nonparametric and Partially Parametric Approaches to Event-History Analysis. S. 1-60 in: Leinhardt, S. (Hrsg.), Sociological Methodology 1982. San Francisco: Jossey--Bass.

Tuma, N. B./Hannan, M. T./Groeneveld, L. P., 1979: Dynamic Analysis of Event Histories. American Journal of Sociology 84: 820-854.

Tuma, N. B./Hannan, M. T., 1984: Social Dynamics. Orlando: Academic Press.

Voges, W./Rohwer, G., 1991: Zur Dynamik des Sozialhilfebezugs. S. 510-531 in: Rendtel, U./Wagner, G. (Hrsg.), Lebenslagen im Wandel: Zur Einkommensdynamik in Deutschland seit 1984. Frankfurt/New York: Campus.

Wagner, M., 1990: Arbeitslosenkarrieren. Journal für Sozialforschung 30: 5-23.

Yamaguchi, K., 1991: Event History Analysis. Newbury Park: Sage.