Chapter 3
Nonparametric Descriptive Methods

In this chapter we discuss nonparametric estimation methods that can be used to describe the characteristics of the process under study. Because these methods do not make any assumptions about the distribution of the process, they are particularly suited for first exploratory data analyses. Stata contains procedures to calculate life tables and Kaplan-Meier (or product limit) estimates. Both of these methods are helpful for graphical presentations of the survivor function (and their transformations) as well as the transition rate. The life table method is the more traditional procedure and has been used in the case of large data sets because it needs less computing time and space. However, compared to the Kaplan-Meier estimator, the life table method has the disadvantage that the researcher has to define discrete time intervals, as is shown later. Given the capabilities of modern computers, there seems to be no reason anymore to prefer the life table method on the basis of computer time or storage space. We therefore give only a few examples for the life table method and discuss the Kaplan-Meier estimator in more detail.

3.1 Life Table Method

The life table method enables the calculation of nonparametric estimates of the survivor function, the density function, and the transition rate for durations given in a set of episodes.[1] There are two drawbacks to this method. First, it is necessary to group the durations into fixed intervals. The results therefore depend more or less on these arbitrarily defined time intervals. Second, it is only sensible to use this method if there is a relatively large number of episodes, so that estimates conditional for each interval are reliable. However, if this second requirement is fulfilled, the method gives good approximations that can be easily calculated.

Time intervals are defined by split points on the time axis

$$0 \leq \tau_1 < \tau_2 < \tau_3 < \ldots < \tau_L$$

with the convention that $\tau_{L+1} = \infty$, there are L time intervals, each including the left limit, but not the right one.

$$I_l = \{t \mid \tau_l \leq t < \tau_{l+1}\} \qquad l = 1, \ldots, L$$

Given these time intervals, the calculation of life tables by Stata is always done using episode durations. In the following description we therefore assume that all episodes have starting time zero. In addition, we assume that the time intervals start at zero (i.e., $\tau_1 = 0$).

The calculation depends somewhat on the type of input data. The following possibilities are recognized by Stata.[2] (1) If the input data are split into groups, a separate life table is calculated for each of the groups. (2) If there is more than one origin state, the life table calculation is done separately for each subset of episodes having the same origin state. Consequently, the life table calculation is always conditional on a given origin state. (3) If, for a given origin state, there is only a single destination state, an ordinary life table is calculated. If there are two or more destination states, a so-called multiple-decrement life table is produced.

To explain the formulas used for the life table calculation, we proceed in two steps. We first consider the case of a single transition (i.e., only a single origin and a single destination state), then we take into account the possibility of competing risks (i.e., two or more destination states). In both cases, to simplify notation, we assume a sample of N episodes all having the same origin state.

Single Transitions

All formulas used in the calculation of single transition life tables are based on the following quantities, defined for each interval I_l, $l = 1, \ldots, L$.

E_l = the number of episodes with events in I_l

Z_l = the number of censored episodes ending in I_l

The next important point is the definition of a *risk set*, $\mathcal{R}_l$, for each of the time intervals, that is, the set of units (episodes) that are at risk of having an event during the lth interval.[3] Two steps are required to take into account episodes that are censored during the interval. First the number of episodes, N_l, that enter the lth interval, is defined recursively by

$$N_1 = N, \quad N_l = N_{l-1} - E_{l-1} - Z_{l-1}$$

[1] An extensive discussion of the life table method has been given by Namboodiri and Suchindran (1987).

[2] In Stata, there are four kinds of weights: frequency weights (`fweights`), analytic weights (`aweights`), sampling weights (`pweights`), and importance weights (`iweights`). To learn more about these weighting types, enter the command `help weights` at the command line. The Stata command for life tables allows for `fweights`. For a discussion of using weights in longitudinal data analyses, see Hoem (1985, 1989).

[3] We generally denote the risk set by the symbol $\mathcal{R}$, and the number of units contained in the risk set by the symbol R.

Second, one has to decide how many of the episodes that are censored during an interval should be contained in the risk set for that interval. A standard assumption is that one half of their number should be contained, but, clearly, this is a somewhat arbitrary assumption.[4] To provide the possibility of changing this assumption, we assume a constant ω $(0 \leq \omega \leq 1)$ for the definition of the fraction of censored episodes that should be contained in the risk set. The number of elements in the risk set is defined, then, by

$$R_l = N_l - \omega\, Z_l$$

Using these basic quantities, it is easy to define all other concepts used in the life table setup. First, the conditional probabilities for having an event in the lth interval, q_l, and for surviving the interval, p_l, are

$$q_l = \frac{E_l}{R_l} \quad \text{and} \quad p_l = 1 - q_l$$

As an implication, one gets the following estimator for the survivor function:

$$G_1 = 1, \quad G_l = p_{l-1} G_{l-1}$$

Note, however, that in the output of Stata's life table procedure values of the survivor function are given for end points of time intervals.

Having estimates of the survivor function, the density function is evaluated approximately at the midpoints of the intervals as the first derivative

$$f_l = \frac{G_l - G_{l+1}}{\tau_{l+1} - \tau_l} \qquad l = 1, \ldots, q-1$$

Of course, if the last interval is open on the right side, it is not possible to calculate the survivor function for this interval. Also, estimates of the transition rate, r_l, are calculated at the midpoints of the intervals. They are defined by

$$r_l = \frac{f_l}{\bar{G}_l} \qquad \text{where} \qquad \bar{G}_l = \frac{G_l + G_{l+1}}{2}$$

and this can also be written as

$$r_l = \frac{1}{\tau_{l+1} - \tau_l} \frac{q_l}{1 - q_l/2} = \frac{1}{\tau_{l+1} - \tau_l} \frac{E_l}{R_l - E_l/2}$$

Finally, it is possible to calculate approximate standard errors for the estimates of the survivor and density function, and for the transition rates, by

[4] See the discussion given by Namboodiri and Suchindran (1987: 58ff).

Box 3.1.1 Do-file `ehc1.do` (life table estimation)

```
version 9
capture log close
set more off
log using ehc1.log, replace

use rrdat1, clear

gen des = tfin ~= ti          /*destination state*/
gen tf  = tfin - tstart + 1   /*ending time*/

ltable tf des, intervals(30) su h f   /*command for life table estimation*/

log close
```

the formulas

$$\mathrm{SE}(G_l) = G_l \left[\sum_{i=1}^{l-1} \frac{q_i}{p_i R_i} \right]^{1/2}$$

$$\mathrm{SE}(f_l) = \frac{q_l G_l}{\tau_{l+1} - \tau_l} \left[\frac{p_l}{q_l R_l} + \sum_{i=1}^{l-1} \frac{q_i}{p_i R_i} \right]^{1/2}$$

$$\mathrm{SE}(r_l) = \frac{r_l}{\sqrt{q_l R_l}} \left[1 - \left[\frac{r_l (\tau_{l+1} - \tau_l)}{2} \right]^2 \right]^{1/2}$$

Given large samples, it may be assumed that the values of the survivor, density, and rate functions, divided by their standard errors, are approximately standard normally distributed. In these cases it is then possible to calculate confidence intervals.

As an example of life table estimation with Stata, we examine the length of durations until job exit. This means that there is only one type of event: "job exit" from the origin state "being in a job" to the destination state "having left the job." For didactical purposes, we start with a simple example and assume in this application that all job episodes in the data file `rrdat1` (see Boxes 2.2.1 and 2.2.2) can be considered as independent from each other (single episode case) and that there is no important heterogeneity among the individuals.[5] Thus we are going to estimate "average" survivor and transition rate functions across all the job spells and individuals.

In Box 3.1.1 the do-file (`ehc1.do`) for the life table estimation with Stata is shown. The upper part of this file is identical to do-file `ehb2.do`, shown in Box 2.2.4, which was used to define single episode data. In order to

[5] Of course, this is an unrealistic assumption because the individuals are represented in the data file (`rrdat1`) with varying numbers of job spells. Thus there are dependencies between the episodes of each individual.

Box 3.1.2 Result of using do-file `ehc1.do` (Box 3.1.1)

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	30	600	223	28	0.6195	0.0201	0.5788	0.6574
30	60	349	113	23	0.4121	0.0208	0.3712	0.4524
60	90	213	51	15	0.3098	0.0199	0.2711	0.3492
90	120	147	25	16	0.2541	0.0192	0.2172	0.2924
120	150	106	24	15	0.1922	0.0182	0.1578	0.2291
150	180	67	9	5	0.1654	0.0177	0.1323	0.2016
180	210	53	4	9	0.1517	0.0175	0.1193	0.1878
210	240	40	3	5	0.1396	0.0175	0.1075	0.1758
240	270	32	0	5	0.1396	0.0175	0.1075	0.1758
270	300	27	2	7	0.1277	0.0179	0.0952	0.1651
300	330	18	2	5	0.1112	0.0190	0.0775	0.1517
330	360	11	2	1	0.0900	0.0205	0.0552	0.1352
360	390	8	0	3	0.0900	0.0205	0.0552	0.1352
390	420	5	0	4	0.0900	0.0205	0.0552	0.1352
420	450	1	0	1	0.0900	0.0205	0.0552	0.1352

Interval		Beg. Total	Deaths	Lost	Cum. Failure	Std. Error	[95% Conf.	Int.]
0	30	600	223	28	0.3805	0.0201	0.3426	0.4212
30	60	349	113	23	0.5879	0.0208	0.5476	0.6288
60	90	213	51	15	0.6902	0.0199	0.6508	0.7289
90	120	147	25	16	0.7459	0.0192	0.7076	0.7828
120	150	106	24	15	0.8078	0.0182	0.7709	0.8422
150	180	67	9	5	0.8346	0.0177	0.7984	0.8677
180	210	53	4	9	0.8483	0.0175	0.8122	0.8807
210	240	40	3	5	0.8604	0.0175	0.8242	0.8925
240	270	32	0	5	0.8604	0.0175	0.8242	0.8925
270	300	27	2	7	0.8723	0.0179	0.8349	0.9048
300	330	18	2	5	0.8888	0.0190	0.8483	0.9225
330	360	11	2	1	0.9100	0.0205	0.8648	0.9448
360	390	8	0	3	0.9100	0.0205	0.8648	0.9448
390	420	5	0	4	0.9100	0.0205	0.8648	0.9448
420	450	1	0	1	0.9100	0.0205	0.8648	0.9448

Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf.	Int.]
0	30	600	0.3805	0.0201	0.0157	0.0010	0.0137	0.0177
30	60	349	0.5879	0.0208	0.0134	0.0012	0.0110	0.0158
60	90	213	0.6902	0.0199	0.0094	0.0013	0.0069	0.0120
90	120	147	0.7459	0.0192	0.0066	0.0013	0.0040	0.0092
120	150	106	0.8078	0.0182	0.0092	0.0019	0.0056	0.0129
150	180	67	0.8346	0.0177	0.0050	0.0017	0.0017	0.0083
180	210	53	0.8483	0.0175	0.0029	0.0014	0.0001	0.0057
210	240	40	0.8604	0.0175	0.0028	0.0016	0.0000	0.0059
240	270	32	0.8604	0.0175	0.0000	.	.	.
270	300	27	0.8723	0.0179	0.0030	0.0021	0.0000	0.0071
300	330	18	0.8888	0.0190	0.0046	0.0032	0.0000	0.0110
330	360	11	0.9100	0.0205	0.0070	0.0049	0.0000	0.0167
360	390	8	0.9100	0.0205	0.0000	.	.	.
390	420	5	0.9100	0.0205	0.0000	.	.	.
420	450	1	0.9100	0.0205	0.0000	.	.	.

request life table estimation we added the `ltable` command. One needs a specification of time intervals. This is done with the option `intervals`:

```
ltable tf des,intervals(30)
```

This defines time intervals, each having a duration of 30 months, and beginning at time zero: $0 \leq t < 30$, $30 \leq t < 60$, and so on. With the `intervals` option you can also define cutoff points by specifying more than one number as the argument.

As a default, `ltable` displays the estimates of the survivor function. Specifying the option `failure` one gets the cumulative failure table (1 - survival). To obtain estimates of the transition rate, we type `hazard`.

Executing do-file `ehc1.do` with Stata you will get the survival and failure table as well as the estimates of the transition rate aggregated in 30-month intervals. The results of do-file `ehc1.do` are shown in Box 3.1.2.

We want to give a short example of how the numbers in the life table of our example are related to the formulas developed earlier. In the third column of the life table in Box 3.1.2, the number of episodes entering into the successive intervals is given. In the first interval, all 600 episodes entered: $N_1 = N = 600$. The numbers of the following intervals $l = 2, 3, \ldots$ are calculated as

$$N_l = N_{l-1} - E_{l-1} - Z_{l-1}$$

where E_l, the number of events in the lth interval, is printed in column 4, and Z_l, the number of censored episodes in the lth interval, is printed in column 5. In our example:

$$\begin{aligned} N_1 &= 600 \\ N_2 &= 600 - 223 - 28 = 349 \\ N_3 &= 349 - 113 - 23 = 213 \end{aligned}$$

Under the assumption that censored episodes are equally distributed within each interval ($w = 0.5$), one is able to estimate the number of episodes at risk in each interval. For the lth interval:

$$\hat{R}_l = N_l - 0.5\, Z_l$$

In our example:

$$\begin{aligned} \hat{R}_1 &= 600 - 0.5 \cdot 28 = 586.0 \\ \hat{R}_2 &= 349 - 0.5 \cdot 23 = 337.5 \end{aligned}$$

The conditional probability of having an event in the lth interval is given as

$$\hat{q}_l = \frac{E_l}{\hat{R}_l}$$

In our example:

$$\hat{q}_1 = \frac{223}{586.0} = 0.38055$$
$$\hat{q}_2 = \frac{113}{337.5} = 0.33481$$

The conditional probability of experiencing no event in the lth interval is then

$$\hat{p}_l = 1 - \hat{q}_l$$

In our example:

$$\hat{p}_1 = 1 - 0.38055 = 0.61945$$
$$\hat{p}_2 = 1 - 0.33481 = 0.66519$$

Based on these estimates, one can compute estimates of the survivor function (column 6 of the upper panel of the life table):

$$\hat{G}_1 = 1$$
$$\hat{G}_l = \hat{p}_{l-1} \cdot \hat{p}_{l-2} \cdots \hat{p}_1$$

In our example:

$$\hat{G}_1 = 1$$
$$\hat{G}_2 = 0.61945 \cdot 1 = 0.61945$$
$$\hat{G}_3 = 0.66519 \cdot 0.61945 \cdot 1 = 0.41205$$

Finally, we also have to consider the length of the intervals, $\tau_{l+1} - \tau_l$ (for the lth interval). The duration density function is given as

$$\hat{f}_l = \frac{\hat{G}_l - \hat{G}_{l-1}}{\tau_{l+1} - \tau_l}$$

In our example:

$$\hat{f}_1 = \frac{1.00000 - 0.61945}{30 - 0} = 0.01268$$
$$\hat{f}_2 = \frac{0.61945 - 0.41205}{60 - 30} = 0.00691$$

The "average" transition rate, evaluated at the midpoint of each interval, is printed in column 6 of the bottom panel in the life table:

$$\hat{r}_l = \frac{1}{\tau_{l+1} - \tau_l} \frac{E_l}{\hat{R}_l - 0.5\, E_l}$$

Box 3.1.3 Do-file `ehc2.do` to plot a survivor function

```
version 9
set scheme sj
capture log close
set more off
log using ehc2.log, replace

use rrdat1, clear

gen des = tfin ~= ti          /*destination state*/
gen tf  = tfin - tstart + 1   /*ending time*/

ltable tf des, intervals(30) gr title("Life Table Survivor Function")  ///
ysc(r(1)) ylabel(0(0.2)1) xtitle("analysis time") ///
saving("Figure 3_1_1",replace)

log close
```

In our example:

$$\hat{r}_1 = \frac{1}{30 - 0} \frac{223}{586.0 - 0.5 \cdot 223} = 0.01567$$
$$\hat{r}_2 = \frac{1}{60 - 30} \frac{113}{337.5 - 0.5 \cdot 113} = 0.01340$$

The standard errors for the survivor function, cumulative failure table, and the rate function are printed in column 7 of each table, and the 95% confidence interval is printed in columns 8 and 9.

Life tables, as shown in Box 3.1.2, are very complex and are not easily interpreted. It is therefore better to plot the interval-related information of the survival or failure function. An example of a Stata do-file, `ehc2.do`, which can be used to generate a plot of the survivor function, is shown in Box 3.1.3.[6] To draw a graph of the survivor function you simply add the `graph` option to the `ltable` command.[7] The resulting plot is shown in Figure 3.1.1. A similar do-file can be used to generate plots for the failure function.

The plot of the survivor function shows estimates of the proportions of respondents who have not yet changed their jobs up to a specific duration. For example, after 10 years (or 120 months) about 25 % of respondents are still in their jobs, while about 75 % have already left.

[6] Because there may be a considerable number of intervals during which no events take place, and hence the hazard estimate is zero, a graph of the rate function is best created by using a kernel smooth. To plot the results of the hazard table, you employ the command `sts graph`. However, this command is not illustrated here.

[7] The appearance of a Stata graph in the Stata Graph Window is specified by a graph scheme. To ensure that you obtain the same results as in this textbook, type `set scheme sj` and change to the Stata Journal scheme. We use various options to specify the title, scale, and labels of the graph. To learn more about the options possible with `ltable`, type `help ltable` and `help twoway options`.

Box 3.1.4 Do-file `ehc3.do` (life table estimation)

```
version 9
set scheme sj
capture log close
set more off
log using ehc3.log, replace

use rrdat1, clear

gen des = tfin ~= ti          /*destination state*/
gen tf  = tfin - tstart + 1   /*ending time*/

label define sex 1 "Men" 2 "Women"
label value sex sex

ltable tf des, intervals(30) su h f by(sex) /*command for life table estimation*/

ltable tf des, intervals(30) by(sex) gr overlay ///
ysc(r(1)) ylabel(0(0.2)1) xlabel(0(100)500) xtitle("analysis time") ///
saving("Figure 3_1_2",replace)

log close
```

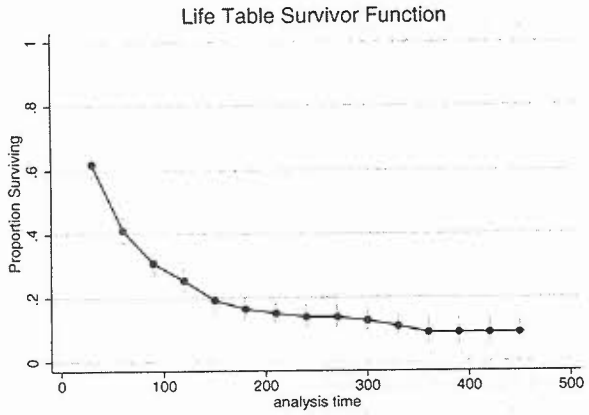

Figure 3.1.1 Plot of the survivor function generated with do-file `ehc2.do`.

Life Table Estimation for Different Groups

Life tables are particularly useful for comparisons of the behavior of subgroups. We therefore extend the example in Box 3.1.1 and demonstrate how separate life tables for men and women can be estimated with Stata. This can be achieved by adding the `by(varname)` option. Box 3.1.4 shows the do-file `ehc3.do`, which is just a modification of do-file `ehc2.do`. The variable `sex` is the indicator variable for men and women.[8] The command to request life table estimation is basically the same. We only added the option `by(sex)` in order to separate life tables for both groups.[9]

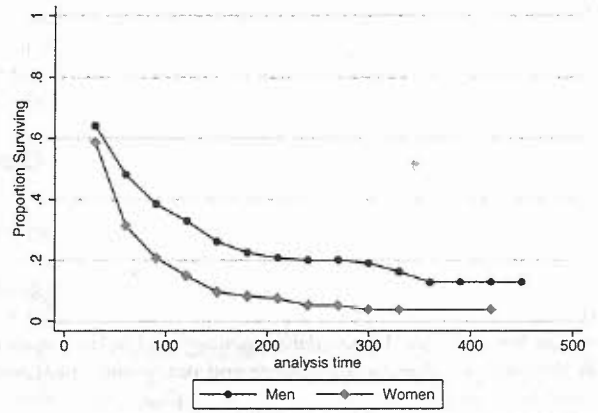

Figure 3.1.2 Plot of survivor functions for men and women, generated with do-file `ehc3.do`.

As a result of using do-file `ehc3.do`, separate tables for men and women are presented. To save space, these tables are not shown here. Instead, Figure 3.1.2 shows a plot of the survivor functions. It is easy to see that at the beginning the process for men and women is quite similar. But after a duration of about three years the survivor function of women decreases more steeply than the survivor function for men. Thus women tend to leave their jobs sooner than men do. After 20 years, about 20 % of men, but only about 5 % of women are still in their jobs. The median job durations are about 57 months for men and about 40 months for women.

Examples of the Application of Survivor Functions in Social Research

In modern social research, survivor functions have been increasingly used to study social change. They are particularly suited to the analysis of how changing historical conditions affect life course transitions of successive birth

[8] In Stata, linking value labels and variables is a two-step process. First, one has to define a value label by using the command `label define`, then one specifies `label value` to attach its contents to a variable.

[9] By default, Stata draws a separate graph within the same image for each group defined with the `by()` option. Sometimes it is useful to overlay these plots on the same graph. To do so, we specify the option `overlay`.

cohorts.[10] They enable the description of the age-graded character of roles and behaviors and, by documenting exits in the timing when individuals enter or leave specific social institutions or positions, the changes of life phases (Blossfeld and Nuthmann 1990; Becker and Blossfeld 1991).

Example 1: "Vulnerable" Phases in the Educational Careers of German Students

An example of the application of survivor functions in educational research is given in Blossfeld (1990, 1992). He studied the process of entry into vocational training in Germany and showed that the particular organizational structure of the German educational system creates what he calls a "vulnerable" life phase for students. *Vulnerability* means that the time span between having left the general educational system and entry into vocational training is limited to a short period of about two or three years, during which prevailing historical events, economic conditions, and demographic constellations strongly determine the opportunities of each generation to acquire vocational training.

The survivor functions in Figure 3.1.3 demonstrate this "vulnerable" phase for three different birth cohorts, and for men and women. They show the proportions of school leavers who did not yet enter the German vocational training system for every point in time after leaving the general educational system. The curves are very different for the three birth cohorts and for men and women. In particular, the economic and political breakdown in the immediate postwar period (1945–50) had a strong negative effect on enrollment in vocational training for the 1929–31 cohort. Confronted with the existing historical conditions, school leavers of this birth cohort did not rank entering vocational training highly because they had more urgent problems to deal with (e.g., making a living), and it would have been very difficult to find trainee positions at all (Mayer 1987, 1988, 1991). Compared to this immediate postwar period, the later social and economic development until the mid-1970s led to a constant rise in the standard of living, a decrease in unemployment, and a substantial growth in the number of trainee positions offered in the private sector.

The upper part of Figure 3.1.3 shows that about 50 % of men in the 1929–31 birth cohort started vocational training immediately after leaving the general educational system. An additional 27 % of these men undertook vocational training within three years of leaving the general educational system. But about 23 % never entered vocational training. In comparison, about 71 % and 79 % of the men in the 1939–41 and 1949–51 birth cohorts, respectively, began vocational training immediately, and an additional 14 %

[10] See Mayer and Schwarz 1989; Hogan 1978, 1981; Marini 1978, 1984, 1985; Elder 1975, 1978, 1987.

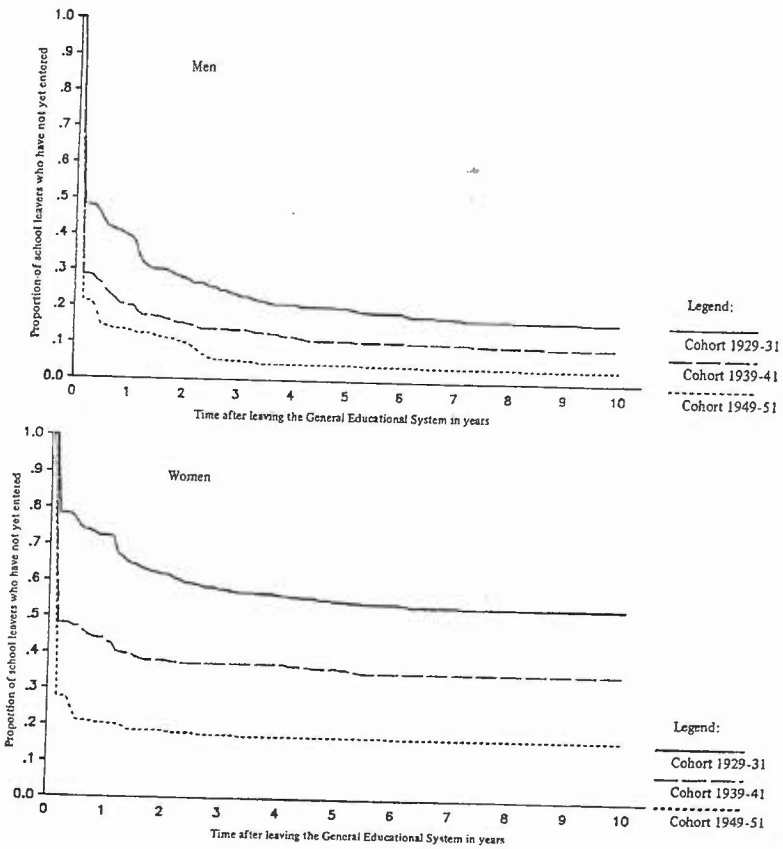

Figure 3.1.3 Process of entering vocational training (survivor functions).

of the men in both cohorts started within three years. Thus only about 15 % of the men in the 1939–41 cohort and about 7 % of the men in the 1949–51 cohort never entered vocational training.

For women (lower part of Figure 3.1.3), these differences were even more pronounced. Within three years of leaving the general educational system, only 40 % of the women in the 1929–31 cohort, but 60 % of the women in the 1939–41 cohort, and 80 % of the women in the 1949–51 cohort undertook vocational training. In other words, with regard to educational opportunities, women in particular carried the burden of the immediate postwar social and economic crises, but they also profited more than men from the rapid

economic recovery in the 1960s and early 1970s.

In summary, this description based on survivor functions reveals that in the German vocational training system entries are hard to postpone beyond a "vulnerable" life phase of about two or three years after leaving the general educational system. Individuals' careers are directed to vocational training relatively early and are hard to change later on. The tremendous increase in the proportion of trainee positions until the early 1970s had therefore almost no effect on the process of entering vocational training for the 1929–31 cohort (Figure 3.1.3). It was not possible for the members of this cohort to "stop" their life course and to "resume" their educational careers when the trainee positions finally became available. In the German educational system a temporary lack of trainee positions therefore not only is a short-term problem but also leads to long-term life-course effects for many people. There is a long-term disadvantage, because it is difficult to acquire vocational degrees in later life stages when one is more removed from the institutions of vocational training and has additional social commitments (such as maintaining one's own home, marriage, and children), making entrance into the institutions of vocational training more and more unlikely. Hence, in terms of educational opportunities, there are disadvantaged generations in Germany, such as the cohorts born around 1930, who completed their training in the immediate postwar period, or the large birth cohorts that crowded into vocational training at the beginning of the 1980s.

Example 2: Changes in Women's Ages at Family Formation

Another illustration of the utility of survivor functions for describing social change is given by Blossfeld and Jaenichen (1992). They discussed the changes in the process of women's entry into marriage and motherhood across successive birth cohorts in Germany. Tables 3.1.1 and 3.1.2 show the percentages of women who have not yet entered first marriage or first birth for each birth cohort and specific ages. These percentages are based on life table estimates of survivor functions for each cohort for the events of entry into marriage and first birth.

As shown in Table 3.1.1, age at first marriage fell sharply from the 1919–23 cohort to the 1944–48 cohort, and has since been rising again until the youngest birth cohort. The greatest movements occurred among women aged 20–24, where the unmarried proportion dropped from 46 % to 15 % and subsequently increased again to 40 %. The result is that as far as the youngest cohorts, 1964–68 and 1959–63, can be followed, they have more or less the same age pattern at entry into marriage as we find for the oldest cohorts, 1924–28 and 1919–23.

Looking at ages at first birth in Table 3.1.2, we observe a similar trend. Again, it is the 1944–48 cohort that entered motherhood at the youngest

Table 3.1.1 Changes in the timing of entry into marriage, as measured by proportions unmarried at specific ages (percentages).

Birth cohort	Proportion of unmarried women at age							
	20	24	28	32	36	40	44	48
1964–68	89	–	–	–	–	–	–	–
1959–63	78	40	–	–	–	–	–	–
1954–58	73	32	19	–	–	–	–	–
1949–53	65	24	11	7	–	–	–	–
1944–48	65	15	7	4	3	–	–	–
1939–43	80	21	8	5	3	3	–	–
1934–38	76	23	9	6	5	5	4	–
1929–33	86	32	13	7	6	5	5	4
1924–28	90	40	16	11	8	6	5	5
1919–23	90	46	20	13	10	9	7	7

Table 3.1.2 Changes in the timing of entry into motherhood, as measured by proportions childless at specific ages (percentages).

Birth cohort	Proportion of childless women at age							
	20	24	28	32	36	40	44	48
1964–68	92	–	–	–	–	–	–	–
1959–63	90	57	–	–	–	–	–	–
1954–58	84	55	30	–	–	–	–	–
1949–53	78	47	21	15	–	–	–	–
1944–48	77	32	16	11	9	–	–	–
1939–43	87	41	18	13	10	10	–	–
1934–38	83	45	18	13	11	11	11	–
1929–33	92	54	28	19	16	16	16	16
1924–28	87	56	26	19	15	14	14	14
1919–23	91	57	30	19	15	15	15	15

ages. For this cohort, not only marriages but entries into motherhood were highly concentrated. And again, we find more or less the same time pattern of entry into motherhood for the youngest cohorts, 1964–68 and 1959–63, and the oldest cohorts, 1924–28 and 1919–23, at least as far as the youngest cohorts can be followed.

Both tables show that in Germany the delay of entry into marriage and motherhood seems to be less dramatic than has been shown for other countries, such as the Scandinavian ones, especially Sweden (see Blossfeld 1995; Hoem 1986, 1991; Hoem and Rennermalm 1985). In Germany, more or less the same entrance pattern of ages into marriage and motherhood is observed as was established 50 years ago. However, it is also clear that in

Germany the earlier movement toward younger and universal marriage and motherhood had come to a halt at the end of the 1960s and the beginning of the 1970s. But this reversal of the timing of marriage and motherhood is not in line with the monotonic trend in women's educational attainment across cohorts (Blossfeld and Shavit 1993). It is therefore questionable whether changes in marriage and motherhood can be attributed mainly to women's growing economic independence (see Blossfeld 1995), as argued for example by Becker (1981).

3.2 Product-Limit Estimation

Another method for the nonparametric estimation of the survivor function and its derivatives is the product-limit, also called the Kaplan-Meier (1958), method. One of the advantages of this approach, compared with the life table method, is that it is not necessary to group the episode durations according to arbitrarily defined time intervals. Instead, the product-limit method is based on the calculation of a risk set at every point in time where at least one event occurred. In this way, the information contained in a set of episodes is optimally used. The only drawback of this method results from the fact that all episodes must be sorted according to their ending (and starting) times, but with efficient sorting algorithms the method can be employed with fairly large sets of episodes.

This section describes the product-limit estimation method and its implementation in Stata.[11] The options, depending on the type of input data, are essentially the same as with the life table method: (1) If the input data are split into groups, separate product-limit estimates are calculated for each of the groups. (2) If there is more than a single origin state, one or more product-limit estimates are calculated for each subset of episodes having the same origin state. (3) If there is more than a single destination state, separate product-limit estimates are calculated for each transition found in the input data.

The following description proceeds in two steps. First, we consider the case of a single transition, then the case of two or more destination states.

Single Transitions

We assume a sample of N episodes, all having the same origin state and either having the same destination state or being right censored. If groups are defined, it is assumed that all episodes belong to the same group. For the moment we also assume that all episodes have the starting time zero.

[11] You must declare your data to be event history data before you can use the Stata command for product-limit estimation. If you work with weighted data, `fweights`, `iweights`, and `pweights` may be specified at this point.

The first step is to consider the points in time where at least one of the episodes ends with an event. There are, say, q such points in time.

$$\tau_1 < \tau_2 < \tau_3 < \ldots < \tau_q$$

The second step is to define three basic quantities, all defined for $l = 1, \ldots, q$, with the convention that $\tau_0 = 0$.

E_l = the number of episodes with events at τ_l

Z_l = the number of censored episodes ending in $[\tau_{l-1}, \tau_l)$

R_l = the number of episodes in the risk set at τ_l, denoted $\mathcal{R}_l$, that is, the number of episodes with starting time less than τ_l and ending time $\geq \tau_l$

Note that the implied definition of the risk set allows the handling of episodes with starting times greater than zero. Also note that the risk set at τ_l includes episodes that are censored at this point in time. It is assumed that a censored episode contains the information that there was no event up to *and including* the observed ending time of the episode. As sometimes stated, censoring takes place an infinitesimal amount to the right of the observed ending time.

Given these quantities, the product-limit estimator of the survivor function is defined as

$$\hat{G}(t) = \prod_{l:\tau_l<t} \left(1 - \frac{E_l}{R_l}\right)$$

This is a step function with steps at the points in time, τ_l. The commonly used formula to calculate estimates of standard errors for the survivor function is

$$\mathrm{SE}(\hat{G}(t)) = \hat{G}(t) \left[\sum_{l:\tau_l<t} \frac{E_l}{R_l\,(R_l - E_l)}\right]^{1/2}$$

In addition to survivor function estimates, the product-limit method gives a simple estimate of the cumulated transition rate.

$$\hat{H}(t) = -\log\left(\hat{G}(t)\right)$$

This is again a step function. It is especially useful for simple graphical checks of distributional assumptions about the underlying durations (see chapter 8).

Unfortunately, unlike the life table estimation, the product-limit method does not provide direct estimates of transition rates. Of course, it is possible

Box 3.2.1 Do-file ehc5.do (Kaplan-Meier estimation)

```
version 9
capture log close
set more off
log using ehc5.log, replace

use rrdat1, clear

gen des = tfin ~= ti           /*destination state*/
gen tf  = tfin - tstart + 1    /*ending time*/

stset tf, f(des)   /*define single episode data*/

sts list /*list survivor function*/

log close
```

Box 3.2.2 Result of do-file ehc5.do (Box 3.2.1)

```
      failure _d:  des
analysis time _t:  tf
        Beg.              Net    Survivor      Std.
Time    Total   Fail      Lost    Function     Error  [95% Conf. Int.]
-------------------------------------------------------------------
   2      600      2         1      0.9967    0.0024   0.9867   0.9992
   3      597      5         2      0.9883    0.0044   0.9757   0.9944
   4      590      9         0      0.9732    0.0066   0.9567   0.9835
   5      581      3         1      0.9682    0.0072   0.9506   0.9796
   6      577     10         0      0.9514    0.0088   0.9309   0.9660
   7      567      9         1      0.9363    0.0100   0.9136   0.9533
   8      557      6         3      0.9262    0.0107   0.9022   0.9446
   9      548      7         1      0.9144    0.0115   0.8889   0.9343
  10      540      8         4      0.9009    0.0123   0.8739   0.9223
                          output omitted
  42      273      1         0      0.5040    0.0209   0.4623   0.5442
  43      272      2         1      0.5003    0.0209   0.4586   0.5405
  44      269      5         1      0.4910    0.0210   0.4493   0.5313
  45      263      1         0      0.4891    0.0210   0.4474   0.5295
                          output omitted
 275       26      1         0      0.1345    0.0175   0.1025   0.1709
                          output omitted
 293       20      1         0      0.1278    0.0179   0.0953   0.1652
                          output omitted
 312       16      1         1      0.1198    0.0185   0.0866   0.1588
 326       14      1         0      0.1112    0.0190   0.0775   0.1518
                          output omitted
 332       11      1         0      0.1011    0.0198   0.0666   0.1440
                          output omitted
 350        9      1         0      0.0899    0.0205   0.0550   0.1353
                          output omitted
 428        1      0         1      0.0899    0.0205   0.0550   0.1353
```

to get estimates by numerical differentiation of $\hat{H}(t)$, but this requires that one first applies a smoothing procedure to the cumulative rate.

In illustrating the application of the product-limit estimator with Stata in Box 3.2.1, we again apply the job-exit example in which we assumed that there are only single episodes and two states ("being in a job" and "having left the job"). To estimate the Kaplan-Meier survivor function, type `sts list`. The default is to list the function at all the unique time values in the data. To reduce output, you can also choose the points in time at which the estimated survivor function is to be listed. This is achieved by using the option `at()`. In our example, the following command will produce the output shown in Box 3.2.2: `sts list, at(0 2/10 42/45 275 293 312 326 332 350 428)`.[12]

The column labeled `Time` shows the points in time where at least one event takes place. The risk set is given in the second column. For example,

Box 3.2.3 Do-file ehc6.do to plot a survivor function

```
version 9
set scheme sj
capture log close
set more off
log using ehc6.log, replace

use rrdat1, clear

gen des = tfin ~= ti           /*destination state*/
gen tf  = tfin - tstart + 1    /*ending time*/

stset tf, f(des)    /*define single episode data*/
sts graph, title("Product-Limit Survivor Function") saving("Figure 3_2_1", replace)

log close
```

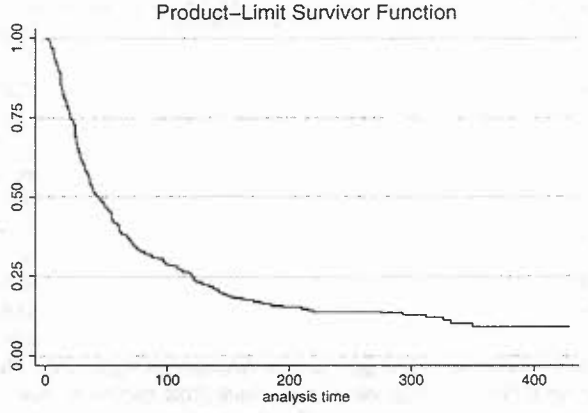

Figure 3.2.1 Plot of survivor function (product-limit estimation) generated with do-file ehc6.do.

[12] If you are not yet familiar with number lists in Stata, type `help numlist`.

at the job duration of two months, 600 episodes are still at risk. The number of events is given in column 3. For example, at a job duration of 6 months we observe 10 job moves, and at a job duration of 42 months we observe 1 event. The last four columns show estimates of the survivor function, its standard errors, and the 95% confidence intervals. For example, after about 4 years (or 45 months) a proportion of 0.4891 of workers are still in the same job. The survivor function of the product-limit estimator is only defined up to the highest event time. The highest event time in Box 3.2.2 is 350 months. Eight additional censored cases with longer job durations follow. Under these circumstances, the estimated survivor function can no longer approach zero and can only be interpreted up until 350 months.

Again, the survivor function in Box 3.2.2 is not very comprehensible. The shape of the function can be evaluated more easily, if it is plotted against job duration. An example do-file, `ehc6.do`, in order to plot the product-limit estimator of the survivor function with Stata is shown in Box 3.2.3. When you type `sts graph`, or simply `sts`, you are shown a graph of the Kaplan-Meier survival estimate. The resulting plot is given in Figure 3.2.1. A comparison with the life table estimate of the survivor function, shown in Figure 3.1.1, shows fairly identical results. This is in accordance with our experience in practical research applications that show that the difference between life table and product-limit estimations is normally very small.

3.3 Comparing Survivor Functions

In analyzing episode data, one often has to compare survivor functions and test if there are significant differences. Basically, two different methods are available. The first relies on the calculation of confidence intervals for each of the survivor functions and then checks if they overlap or not. This is possible with both the life table and the product-limit methods. Both methods provide estimates of standard errors for the survivor function. Another possibility is to calculate specific test statistics to compare two or more survivor functions. This section describes both possibilities.

Defining Groups of Episodes

To make any comparisons, there must be two or more groups of episodes. This is easily done using indicator variables that define membership in a group. In Stata, the syntax is

```
sts list, by(varlist)
```

In our example we specify a single `by()` variable, `by(sex)`, but it is also possible to select up to five variables. The set of episodes given in the current data matrix is then split into groups, and separate calculations are pursued for each group identified by equal values of the variables in `varlist`.

Box 3.3.1 Do-file `ehc7.do` (comparing survivor functions)

```
version 9
set scheme sj
capture log close
set more off
log using ehc7.log, replace

use rrdat1, clear

gen des = tfin ~= ti          /*destination state*/
gen tf  = tfin - tstart + 1   /*ending time*/

label define sex 1 "Men" 2 "Women"
label value sex sex

stset tf, f(des)    /*define single episode data*/

sts list, by(sex)

sts graph, by(sex) gwood saving("Figure 3_3_1", replace)

sts test sex             /*Log-rank test for equality of survivor functions*/
sts test sex, wilcoxon   /*Wilcoxon (Breslow)*/
sts test sex, tware      /*Wilcoxon (Taron-Ware)*/
sts test sex, peto       /*Wilcoxon (Prentice)*/

log close
```

To illustrate grouping in the case of product-limit estimation, we extend the example in Box 3.2.1. The new do-file, `ehc7.do`, is a small modification of do-file `ehc5.do` already shown in Box 3.2.1. The additional commands are shown in Box 3.3.1.

Using this modified do-file, a product-limit estimation is done separately for men and women. The results window contains two tables, one with estimates for men, another one for women. Figure 3.3.1 shows a plot of these two survivor functions. In Stata, you can obtain this plot with the following command: `sts graph, by(sex)`. The option `gwood` is used to show the pointwise Greenwood confidence bands around the survivor function.[13] After about 3 years, the confidence bands of the survivor functions of men and women no longer intersect. Thus there are statistically significant differences in the job-exit behavior of men and women for greater durations.

Construction of Test Statistics

Many different test statistics have been proposed to compare two or more survivor functions. We describe four that can be calculated with Stata. All of them are based on product-limit estimates of survivor functions.

It is assumed that m groups have been defined that do not intersect.

[13] If you specify `gwood` to draw pointwise confidence bands, the curves are automatically placed on separate graphs.

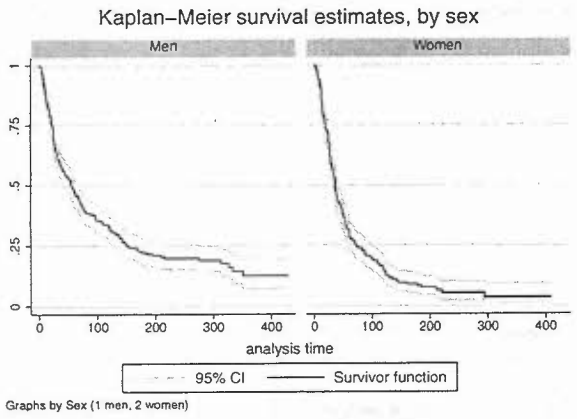

Figure 3.3.1 Plots of survivor functions (product-limit estimation) for men and women with 95 % confidence intervals (grey-scaled). The plot has been generated with do-file `ehc7.do`.

The whole sample is implicitly defined as the set of all episodes that are contained in one of these groups. Then, in exactly the same way as explained in connection with the product-limit method, all calculations are done for each transition in the whole sample separately. Therefore we only consider a sample of episodes that have the same origin state and are censored or have the same destination state.

In general, a sample defined this way consists of m groups, and the following table can be calculated.

$$\begin{array}{llllllll} \tau_1 & R_{11} & E_{11} & R_{12} & E_{12} & \dots & R_{1m} & E_{1m} \\ \tau_2 & R_{21} & E_{21} & R_{22} & E_{22} & \dots & R_{2m} & E_{2m} \\ & & & & \vdots & & & \\ \tau_q & R_{n1} & E_{q1} & R_{q2} & E_{q2} & \dots & R_{qm} & E_{qm} \end{array}$$

These are the basic quantities for the product-limit estimation, for the whole sample, and for each group separately. $\tau_1 < \tau_2 < \ldots < \tau_q$ are the points in time where at least one episode contained in the sample has an event. E_{lg} is the number of episodes contained in group g and having an event at τ_l; R_{lg} is defined as the number of elements in the risk set at τ_l for the episodes contained in group g (i.e., all episodes belonging to group g that have starting times less than τ_l and ending times equal to or greater than τ_l). All together, these quantities are sufficient for a product-limit estimation in each of the m groups.

Given this, the four test statistics can be defined, and they are denoted S_ν ($\nu = 1, \ldots, 4$). Because the calculations only differ in different weights, we give their definitions first. The weights are denoted $W_l^{(\nu)}$, and they are defined for $l = 1, \ldots, q$ by

$$\begin{aligned} W_l^{(1)} &= 1 \\ W_l^{(2)} &= R_l \\ W_l^{(3)} &= \sqrt{R_l} \\ W_l^{(4)} &= \prod_{i=1}^{l} \frac{R_i - E_i + 1}{R_i + 1} \end{aligned} \tag{3.1}$$

The next step is to construct for each of the four test statistics one (m)-vector $U^{(\nu)}$ and one (m,m)-matrix $V^{(\nu)}$. The definitions are[14]

$$U_g^{(\nu)} = \sum_{l=1}^{q} W_l^{(\nu)} \left(E_{lg} - R_{lg} \frac{E_{l0}}{R_{l0}}\right)$$

$$V_{g_1 g_2}^{(\nu)} = \sum_{i=1}^{n} W_l^{(\nu)2} \frac{E_{l0}\,(R_{l0} - E_{l0})}{R_{l0} - 1} \frac{R_{lg_1}}{R_{l0}} \left(\delta_{g_1 g_2} - \frac{R_{lg_2}}{R_{l0}}\right)$$

Finally, the test statistics are defined by

$$S_\nu = U^{(\nu)\prime}\, V^{(\nu)-1}\, U^{(\nu)} \tag{3.2}$$

All of them follow a χ^2-distribution with $m-1$ degrees of freedom given the null hypothesis that there are no significant differences. Note that, accordingly, the rank of $V^{(\nu)}$ is only $m-1$. Therefore in the calculation of (3.2), one can use a generalized inverse or omit the last dimension without loss of generality. Stata follows the latter of these two possibilities.

Unfortunately, there is no uniform convention to name the different test statistics, so we state the names used by Stata and give some remarks about other naming conventions. In the order given by (3.1), we have

1. *Log-Rank.* Other names are *Generalized Savage Test* (Andreß 1985, p. 158; 1992). The same name is used by BMDP, with *Mantel-Cox* added. SAS calculates this test statistic under the name *Logrank*.
2. *Wilcoxon-Breslow-Gehan.* BMDP gives the name *Generalized Wilcoxon (Breslow)*. SAS uses only the label *Wilcoxon*.
3. *Tarone-Ware.* This test statistic was proposed by Tarone and Ware (1977) and is named accordingly. It is also calculated by BMDP, using the label *Tarone-Ware*.

[14] δ_{ij} is the Kronecker symbol, which is one if $i = j$ and zero otherwise.

Box 3.3.2 Results of do-file ehc7.do (Box 3.3.1)

```
. sts test sex
        failure _d:  des
  analysis time _t:  tf

Log-rank test for equality of survivor functions
      |   Events         Events
sex   |  observed       expected
------+-------------------------
Men   |       245         290.89
Women |       213         167.11
------+-------------------------
Total |       458         458.00
                                    chi2(1) =      20.60
                                    Pr>chi2 =     0.0000

. sts test sex, wilcoxon
        failure _d:  des
  analysis time _t:  tf

Wilcoxon (Breslow) test for equality of survivor functions
      |   Events         Events        Sum of
sex   |  observed       expected        ranks
------+---------------------------------------
Men   |       245         290.89       -11883
Women |       213         167.11        11883
------+---------------------------------------
Total |       458         458.00            0
                                    chi2(1) =       9.36
                                    Pr>chi2 =     0.0022

. sts test sex, tware
        failure _d:  des
  analysis time _t:  tf

Tarone-Ware test for equality of survivor functions
      |   Events         Events        Sum of
sex   |  observed       expected        ranks
------+---------------------------------------
Men   |       245         290.89    -717.3311
Women |       213         167.11     717.3311
------+---------------------------------------
Total |       458         458.00            0
                                    chi2(1) =      14.33
                                    Pr>chi2 =     0.0002

. sts test sex, peto
        failure _d:  des
  analysis time _t:  tf

Peto-Peto test for equality of survivor functions
      |   Events         Events        Sum of
sex   |  observed       expected        ranks
------+---------------------------------------
Men   |       245         290.89   -21.622585
Women |       213         167.11    21.622585
------+---------------------------------------
Total |       458         458.00            0
                                    chi2(1) =      10.70
                                    Pr>chi2 =     0.0011
```


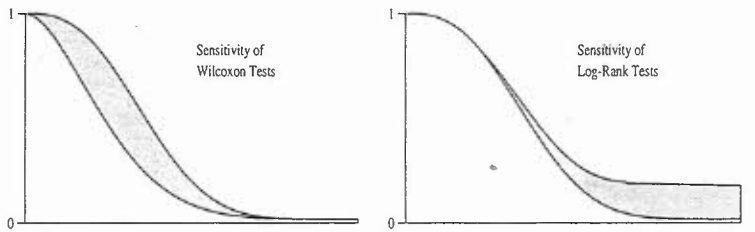

Figure 3.3.2 Regions of sensitivity for Wilcoxon and Log-Rank tests.

4. *Peto-Peto-Prentice.* Finally, there is a test statistic explained by Lawless (1982, p. 423) with reference to R. L. Prentice.

The Stata command to request calculation of test statistics to compare survivor functions is `sts test` as shown in Box 3.3.1. The `sts test` command, by default, performs the log-rank test. To compute the other test statistics you have to select one of the following options: `wilcoxon` to specify the Wilxocon-Breslow-Gehan test, `tware` to conduct the Taron-Ware test, or `peto` to run the Peto-Peto-Prentice test.[15] The results of the do-file `ehc7.do` is shown in Box 3.3.2. All test statistics are based on the null hypothesis that the survivor functions of men and women do not differ. They are χ^2-distributed with $m-1$ degrees of freedom (in the example we have two groups, men and women: $m = 2$). In our example, all four test statistics are significant. In other words, the null hypothesis that survivor functions of men and women do not differ must be rejected. However, it is easy to see that there is a great difference between the Log-Rank (or Savage) test statistic and the other test statistics. The reason for this is that the Wilcoxon tests stress differences of the survivor functions at the beginning of the duration, whereas the Log-Rank (or Savage) test statistic stresses increasing differences at the end of the process time (see Figure 3.3.2).

Multiple Destination States

We now turn to the case of multiple transitions. Here we have a situation of competing risks conditional on a given origin state. There are different concepts to describe such a situation. The simplest generalization of the single transition case leads to product-limit estimates for pseudosurvivor functions.[16]

[15] You may also obtain the test statistic for the Cox test and the Fleming-Harrington test, which are not discussed here. For more details you are referred to the Stata manual.

[16] This generalization is commonly used with the product-limit method. See, for instance, the discussion in Lawless (1982, p. 486f) and Tuma and Hannan (1984, p. 69f).

Box 3.3.3 Do-file ehc9.do

```
version 9
set scheme sj
capture log close
set more off
log using ehc9.log, replace

use rrdat1, clear

gen     des = 2                          /*destination state*/
replace des = 1 if (presn/pres -1)>0.2
replace des = 3 if (presn/pres -1)<0.0
replace des = 0 if presn==-1

gen tf = tfin - tstart + 1               /*ending time*/

stset tf, f(des=1) /* upward moves */
sts list if pres <= 65, at(0 2 4 6/8 160 170 326 428)
sts gen surv1 = s

stset tf, f(des=2) /* lateral moves */
sts list, at(0 3 4/7 184 194 209 350 428)
sts gen surv2 = s

stset tf, f(des=3) /* downward moves */
sts list, at(0 2/6 275 293 312 332 428)
sts gen surv3 = s

for any surv1 surv2 surv3 \ any "upward" "lateral" "downward": label var X "Y"

graph twoway line surv1 surv2 surv3 _t, sort  ///
ysc(r(0 1)) ylabel(0(0.2)1) xtitle("analysis time")legend(row(1))

log close
```

The method is analogous to the single transition case. One starts with N episodes having the same origin state. Then, for each possible destination state k, one looks at the points in time, $\tau_{k,l}$, where at least one transition to destination state k takes place. There are, say, $l = 1, \ldots, q_k$ such points in time.

Let $E_{k,l}$ denote the number of events at $\tau_{k,l}$, and let $\mathcal{R}_l$ denote the risk set at the same point in time. Note that the risk set does not depend on the destination state, but is defined as in the single transition case as the set of all episodes with a starting time less than $\tau_{k,l}$, and with an ending time equal to or greater than $\tau_{k,l}$. The product-limit estimate of the pseudosurvivor functions may then be formally defined by

$$\tilde{G}_k(t) = \prod_{l:\tau_{k,l}<t} \left(1 - \frac{E_{k,l}}{R_l}\right)$$

Obviously, a calculation of this estimate can use the same algorithm as in the single transition case. In the calculation for a specific destination state,

Box 3.3.4a Part 1 (upward moves)

```
      failure _d:  des == 1
analysis time _t:  tf
          Beg.                    Survivor      Std.
Time     Total      Fail          Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------
   0         0         0            1.0000          .          .          .
   2       591         1            0.9983     0.0017*     0.9880     0.9998
   4       582         2            0.9949     0.0030     0.9842     0.9983
   6       569         3            0.9896     0.0042     0.9771     0.9953
   7       559         1            0.9879     0.0046     0.9747     0.9942
   8       549         1            0.9861     0.0049     0.9723     0.9930
 160        61        74            0.7301     0.0316     0.6625     0.7864
 170        58         1            0.7175     0.0334     0.6459     0.7772
 326        14         1            0.6663     0.0583     0.5382     0.7664
 428         1         0            0.6663     0.0583     0.5382     0.7664
-------------------------------------------------------------------------
Note:  Survivor function is calculated over full data and evaluated at
indicated times; it is not calculated from aggregates shown at left.
```

Box 3.3.4b Part 2 (lateral moves)

```
      failure _d:  des == 2
analysis time _t:  tf

          Beg.                    Survivor      Std.
Time     Total      Fail          Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------
   0         0         0            1.0000          .          .          .
   3       597         2            0.9966     0.0024     0.9867     0.9992
   4       590         4            0.9899     0.0041     0.9776     0.9954
   5       581         2            0.9865     0.0047     0.9732     0.9932
   6       577         4            0.9796     0.0058     0.9644     0.9884
   7       567         3            0.9745     0.0065     0.9580     0.9845
 184        53       201            0.4181     0.0316     0.3557     0.4791
 194        50         1            0.4097     0.0321     0.3466     0.4717
 209        41         1            0.3997     0.0328     0.3353     0.4632
 350         9         1            0.3553     0.0510     0.2571     0.4546
 428         1         0            0.3553     0.0510     0.2571     0.4546
-------------------------------------------------------------------------
Note:  Survivor function is calculated over full data and evaluated at
indicated times; it is not calculated from aggregates shown at left.
```

one only has to treat *all* episodes that do not end in this destination as if they were censored.

As an illustration, we use (in do-file ehc9.do in Box 3.3.3) our example data set, rrdat1, and construct a new variable, des, to distinguish three different destination states.[17] Variable des takes the values 1, 2, and 3 for

[17] We defined upward shifts as job mobility leading to an increase in the prestige score of 20 % or more, downward shifts as job mobility connected with a decrease in the prestige score, and lateral shifts as having no effect or experiencing an increase in the prestige score of up to 20 %. It is important to note here that the prestige score in our example data varies between 18 and 78. Those already in good positions with a prestige score greater than 65 are no more at risk to experience an upward career move. These records

Box 3.3.4c Part 3 (downward moves)

```
      failure _d:  des == 3
analysis time _t:  tf

        Beg.                    Survivor      Std.
Time    Total      Fail         Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------
   0        0         0           1.0000          .          .          .
   2      600         1           0.9983     0.0017     0.9882     0.9998
   3      597         3           0.9933     0.0033     0.9823     0.9975
   4      590         3           0.9883     0.0044     0.9755     0.9944
   5      581         0           0.9883     0.0044     0.9755     0.9944
   6      577         3           0.9831     0.0053     0.9689     0.9909
 275       26        84           0.6468     0.0445     0.5522     0.7263
 293       20         1           0.6144     0.0528     0.5025     0.7084
 312       16         1           0.5760     0.0619     0.4458     0.6862
 332       11         0           0.5760     0.0619     0.4458     0.6862
 428        1         0           0.5760     0.0619     0.4458     0.6862
-------------------------------------------------------------------------
Note:  Survivor function is calculated over full data and evaluated at
indicated times; it is not calculated from aggregates shown at left.
```

upward, lateral, and downward moves, respectively, or the value 0 for right censored episodes.

Moreover, we change the content of this new variable `des` to zero for subjects who have left their last job before the interview date because these subjects are no more at risk. In our example, we simply add the line `replace des = 0 if presn==-1` because the prestige score of the consecutive job episode, `presn`, is -1 for both censored observations and all records with ending times before the date of the interview. Executing the do-file `ehc9.do`, shown in Box 3.3.3, with Stata generates three tables with product-limit estimates for the three destination states at specified times. We can also plot the results. First, you need to generate a new variable containing the estimated survivor function. The syntax is: `sts generate newvar = s`. Next we use the graph command as shown in Box 3.3.3.

Figure 3.3.3 shows the survivor functions for these three directional moves. One observes that workers move down faster than they move up. After a duration of approximately 120 months (or 10 years) in a job, only about 22 % have experienced an upward move, while about 24 % moved down.[18]

are therefore excluded from the analysis.

[18]These results differ from the results in Blossfeld and Rohwer (2002) because we have taken into account here the ceiling effect for upward moves (only people who can experience a 20 % or more increase in the prestige score from job n to job $n+1$ are at risk to move upward), the bottom effect (only people who are not already at the bottom of the prestige score are at risk to move downward), and we have corrected for all job exits without a job destination (in this case `presn` is coded -1).

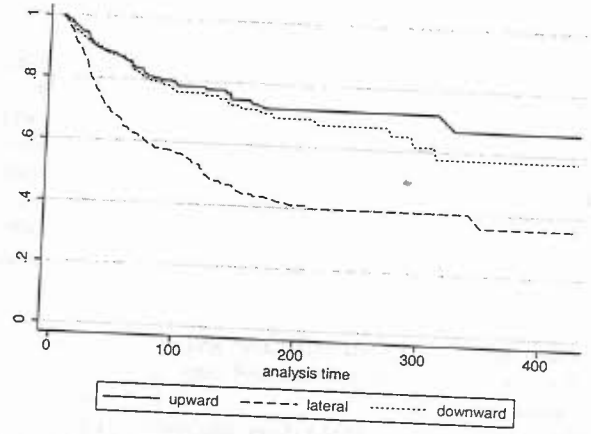

Figure 3.3.3 Plot of survivor functions (product-limit estimation) for upward, lateral, and downward moves. The plot was generated with do-file `ehc9.do`.

Multiepisode Data (or Repeating Events)

The presentation of the life table method and the product-limit estimation in this book has been limited to the special case of single episodes (with competing risks). For didactical reasons we assumed in our practical examples that the episodes in the data set `rrdat1` are statistically independent from each other and that the population is homogeneous. Of course, this is not the case because, in the GLHS, complete job histories of men and women up to the time of the interview are recorded. Thus individuals with a higher number of jobs are represented in the file `rrdat1` more often than individuals with a lower number. One solution to this problem is that one only looks at episodes of a certain job number. For example, one could study job behavior in only the first job, second job, and so on. Another solution would be that one compares survivor functions for several—say the first four—jobs (using the variable "serial number of job" as a group variable) and tests whether the survivor functions are equal. If they do not significantly differ, one could pool them and continue as in the single episode case.

Generally, great differences in the number of episodes between observation units suggests that the sample is quite heterogeneous. However, even in the case of single episodes, homogeneous populations are assumed. Neglected (or unobserved) heterogeneity between observation units can lead to apparent time-dependence (see chapter 10) and wrong substantive conclusions. One should therefore be careful and only estimate survivor functions

for actually homogeneous groups by disaggregating the sample according to theoretically important variables. Unfortunately, in practical applications, this approach is normally of limited use due to the huge sample size that would be necessary for studying a great number of various subpopulations separately. Thus comparisons of survivor functions between subgroups and the possibility to detect time-dependence based on transformations of survivor functions (see chapter 8) usually possess only a heuristic character. In many cases it is easier to include population heterogeneity and differences in the event histories of individuals into transition rate models, as is shown in the following chapters.

Chapter 4

Exponential Transition Rate Models

In practical research, the analysis of event history data with nonparametric estimation methods is associated with several disadvantages. First, as discussed in the previous chapter, with an increasing number of subgroups normally a point is rapidly reached, at which it is no longer sensible to estimate and compare survivor functions due to the small number of cases left in the various subgroups. Second, even in the case where it is feasible to estimate a rising number of survivor functions for important subgroups, comparisons of these functions quickly become complex and interpretation difficult. Third, in the case of quantitative characteristics (e.g., income, age, etc.), variables must be grouped (e.g., "high income group" vs. "low income group," etc.), with a loss of information, to be able to estimate and compare survivor functions. Finally, multiepisode processes can hardly be analyzed with nonparametric methods. Over the last 20 years, transition rate models have therefore increasingly been used in practical research for the analysis of event history data instead of nonparametric methods.

Transition rate models are a general statistical technique through which one can analyze how the transition rate is dependent on a set of covariates. As discussed in section 1.2, viewing the transition rate as a function of change in covariates is naturally linked with a causal approach to the study of social processes. In general, this modeling approach requires that covariates be measured on an interval or ratio scale, but nominal and ordinal covariates can be incorporated into the models through the use of "dummies" (i.e., by substituting the original variables by a set of 0-1 variables). If permitted by measurement, there is also the interesting possibility of controlling for various factors by introducing their metric versions as proxies in the analysis.[1] Well-known examples are the inclusion of social inequality via metric prestige scores (Treiman 1977; Handl, Mayer, and Müller 1977; Wegener 1985; Shavit and Blossfeld 1993) or the approximation of qualification levels by the average number of school years necessary to obtain a specific level of educational attainment (Blossfeld 1985; Shavit and Blossfeld 1993). The previous history of the process can also be easily taken into account in transition rate models. For example, in the job duration example, the history of the process might be incorporated through a variable "general la-

[1] In this case it is assumed that qualitative states reflect points (or intervals) on an underlying metric scale. If the states are ordered, one might argue that the sequence of states corresponds to segments of an underlying continuous variable.