Descriptive Techniques in Discrete-Time Survival Analysis

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

GCM, 2010

Descriptive Techniques in Discrete-Time Survival Analysis

1. Introduction
2. The Life Table
3. The Hazard Function
4. The Survivor Function
    - Estimating the Discrete Survivor Function
    - Median Lifetime
5. Developing Intuition: Some Sample Survival Analyses
    - Cocaine Relapse
    - First Sexual Intercourse
    - First Suicide Ideation
    - Careers of Female Congressional Representatives

Introduction

In this module, we examine the basic descriptive techniques used in discrete-time survival analysis.

The Life Table

1. The fundamental tool for summarizing the sample distribution of event occurrences is the *life table*.
2. The file *teachers.csv* contains the basic data for the study tracking the careers of 3941 teachers.
3. These data can be converted readily into a life table in the format shown in Figure 10.1 on page 327 of Singer and Willett.

The Life Table

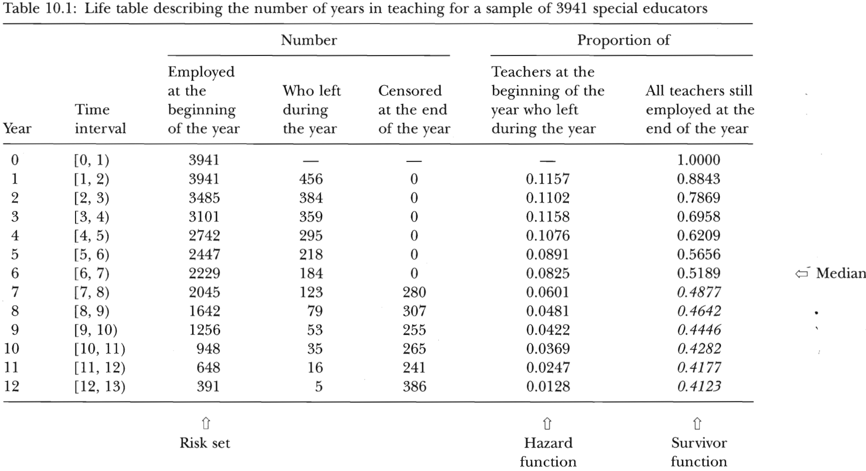


Table 10.1: Life table describing the number of years in teaching for a sample of 3941 special educators

		Number			Proportion of		
Year	Time interval	Employed at the beginning of the year	Who left during the year	Censored at the end of the year	Teachers at the beginning of the year who left during the year	All teachers still employed at the end of the year	
0	[0, 1)	3941	—	—	—	1.0000	
1	[1, 2)	3941	456	0	0.1157	0.8843	
2	[2, 3)	3485	384	0	0.1102	0.7869	
3	[3, 4)	3101	359	0	0.1158	0.6958	
4	[4, 5)	2742	295	0	0.1076	0.6209	
5	[5, 6)	2447	218	0	0.0891	0.5656	
6	[6, 7)	2229	184	0	0.0825	0.5189	⇦ Median
7	[7, 8)	2045	123	280	0.0601	*0.4877*	
8	[8, 9)	1642	79	307	0.0481	*0.4642*	
9	[9, 10)	1256	53	255	0.0422	*0.4446*	
10	[10, 11)	948	35	265	0.0369	*0.4282*	
11	[11, 12)	648	16	241	0.0247	*0.4177*	
12	[12, 13)	391	5	386	0.0128	*0.4123*	
		⇧ Risk set			⇧ Hazard function	⇧ Survivor function	

The Life Table

```
> library(survival)
> teachers<-read.table("teachers.csv", sep=",", header=T)
> ts <- survfit( Surv(t, 1-censor)~ 1, conf.type="none", data=teachers)
> h<-ts$n.event/ts$n.risk
> nlost<-ts$n.risk-ts$n.event- ts$n.risk[-1]
> nlost[12] = ts$n.risk[12]-ts$n.event[12]
> tab10.1<-cbind(time=ts$time, risk=ts$n.risk, left=ts$n.event,
+ censored=nlost, hazard=h, survival=ts$surv)
> tab10.1
```

```
      time risk left censored  hazard survival
 [1,]    1 3941  456        0 0.11571   0.8843
 [2,]    2 3485  384        0 0.11019   0.7869
 [3,]    3 3101  359        0 0.11577   0.6958
 [4,]    4 2742  295        0 0.10759   0.6209
 [5,]    5 2447  218        0 0.08909   0.5656
 [6,]    6 2229  184        0 0.08255   0.5189
 [7,]    7 2045  123      280 0.06015   0.4877
 [8,]    8 1642   79      307 0.04811   0.4642
 [9,]    9 1256   53      255 0.04220   0.4446
[10,]   10  948   35      265 0.03692   0.4282
[11,]   11  648   16      241 0.02469   0.4177
[12,]   12  391    5      386 0.01279   0.4123
```

The Hazard Function – Continuous Case

Let T stand for the time the event of interest occurs.

In the general continuous case the *hazard function*, $h(t)$, is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\left(t \leq T < t + \Delta t \mid T \geq t\right)}{\Delta t} \tag{1}$$

To decipher this expression, look inside the limit. The numerator is the conditional probability that the event occurs in the interval between $T = t$ and $T = t + \Delta t$, given that it did not occur earlier. By dividing this probability by Δt, we *transform it to a rate.* By taking the limit as the width of the interval (i.e., Δt) becomes infinitesimally small, we are calculating the *instantaneous rate* at which the event occurs at a given time. The hazard function in the continuous case is never negative, and can vary from 0 to infinity.

The Discrete-Time Hazard Function

Singer and Willett specialize the hazard function for the discrete case by assuming

1. Each interval is 1 time unit
2. The hazard function $h(t_{ij})$ is the conditional probability of the event of interest occurring to the ith individual in the jth interval, given that it has not occurred previously.

That is,

$$h(t_{ij}) = \Pr\left(T_i = j \,|\, T_i \geq j\right) \tag{2}$$

The Discrete-Time Hazard Function

The maximum likelihood estimator of the Discrete-Time Hazard function is simply the list of values in Column 6 of Table 10.1.

This column of the table gives the proportion of teachers entering each time period who left the profession during that period.

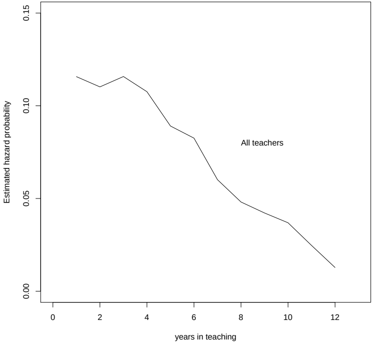
More formally,

$$\hat{h}(t_j) = \frac{n\ events_j}{n\ at\ risk_j} \tag{3}$$

A hazard function defined this way, since it is a probability, varies between 0 and 1.

The Hazard Function

```
> plot(ts$time, h, type="l", ylab="Estimated hazard probability",
+  xlab="years in teaching", ylim=c(0, .15), xlim=c(0, 13))
> text(8, .08, "All teachers", adj=0)
```



The Survivor Function – Continuous Case

The survivor function $S(t)$ is the probability that the individual will survive *beyond* time t. Formally,

$$S(t) = \Pr(T > t) \tag{4}$$

The Survivor Function – Discrete Case

Singer and Willett (p. 334) define the survivor function for the discrete case in essentially the same way. $S(t_{ij})$, the discrete survivor function for individual i at time j, is the probability that individual i will survive *beyond* time j, i.e.,

$$S(t_{ij}) = \Pr(T_i > j) \tag{5}$$

Estimating the Discrete Survivor Function

The survivor function is simple to compute if there is no censoring. It is slightly more complicated once censoring occurs. Singer and Willett (pp. 334–335) refer to the two situations (before and after the first censored case occurs) as the "direct method" and the "indirect method."

The Direct Method

$$\hat{S}(t_j) = \frac{n \text{ still surviving at the end of time period } j}{n \text{ in the data set}} \tag{6}$$

Once censoring starts occurring, the above equation can no longer be counted on to provide a consistent estimate of the survival function. So what can we do?

The answer is, we can draw on our knowledge of conditional probability, and utilize the “keep it alive sequential approach” from Psychology 310.

The Indirect Method

Remember that, in general, $\Pr(A \cap B) = \Pr(A)\Pr(B|A)$.

Surviving past year j can be written as

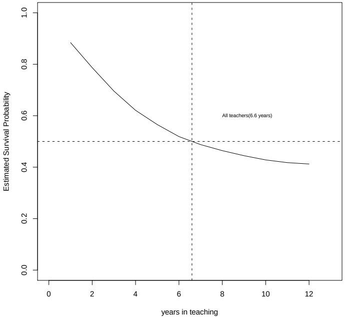
$$\begin{aligned} \text{Surviving past year } j &= \text{Surviving past year } j-1 \\ &\quad \cap \text{ Surviving year j} \\ &= \text{Surviving past year } j-1 \\ &\quad \cap \text{ Not experiencing event in year } j \end{aligned}$$

It immediately follows that a consistent estimate of the probability of surviving past year j is given by

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \times \left[1 - \hat{h}(t_j)\right] \tag{7}$$

A Plot of a Survivor Function

```
> plot(ts$time, ts$surv, type="l", ylab="Estimated Survival Probability",
+ xlab="years in teaching",  ylim=c(0, 1.), xlim=c(0, 13))
> abline(h=c(.5), lty=2)
> abline(v=c(6.6), lty=2)
> text(8, .6, "All teachers(6.6 years)", adj=0, cex=.7)
```



The Product-Limit Formula

Recall that $h(0) = 0$, and so $1 - h(0) = 1$.

Consequently, $\hat{S}(t_1) = (1) \times \left[1 - \hat{h}(t_1)\right] = 1 - \hat{h}(t_1)$.

This implies that $\hat{S}(t_2) = \left[1 - \hat{h}(t_2)\right] \times \left[1 - \hat{h}(t_1)\right]$.

More generally,

$$\hat{S}(t_j) = \prod_{i=1}^{j} 1 - \hat{h}(t_i) \tag{8}$$

This is known as the *product-limit* formula.

Median Lifetime

The estimated median lifetime identifies that value of T for which the value of the estimated survivor function is 0.50.

This is the point in time we estimate that half the population of interest has experienced the target event.

Linear interpolation (which may not make sense when the time is discrete), is frequently used. For example, with Table 10.1, we estimate the median length of stay to be

$$6 + \frac{.5189 - .5}{.5189 - .4877} = 6.61 \tag{9}$$

Cocaine Relapse

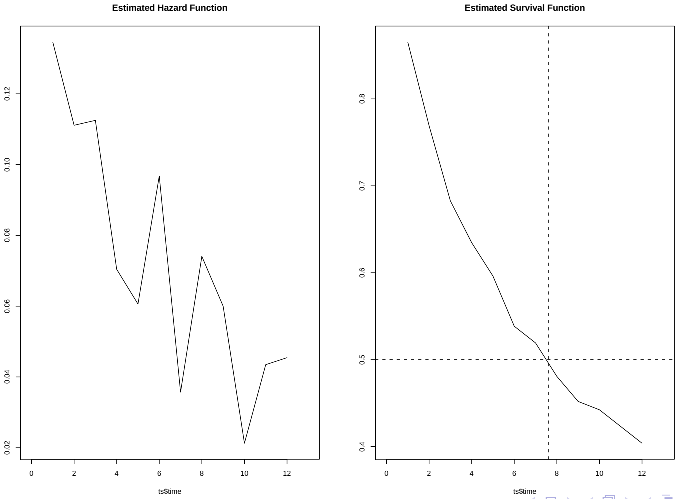
Havassy, Hall, and Wasserman (1991) studied relapse to cocaine use following release from an inpatient treatment program. After 12 weekly follow ups, 62 of 104 former addicts had relapsed.

Cocaine Relapse

Below is code for setting up the survival function and the hazard function.

```
> par(mfrow=c(1, 2), cex=0.7)
> # Panel A
> cocaine<-read.table("cocaine_relapse.csv", sep=",", header=T)
> ts <- survfit( Surv(week, 1-censor)˜ 1, conf.type="none", data=cocaine)
> h<-ts$n.event/ts$n.risk
> plot(ts$time, h, type="l", ylab=" ", main="Estimated Hazard Function",
+   xlim=c(0, 13))
> plot(ts$time, ts$surv, type="l", ylab=" ",
+   main="Estimated Survival Function", xlim=c(0, 13))
> abline(h=c(.5), lty=2)
> abline(v=c(7.6), lty=2)
```

Cocaine Relapse



Interpreting the Hazard Function

Note how the hazard peaks immediately after release, then drifts downward. According to Singer and Willett (p. 341), such a monotonically decreasing hazard function is quite common in areas like substance abuse, mental illness, child abuse, and incarceration.

Interpreting the Survivor Function

Of course, the survivor function mirrors the hazard function.

1. When the hazard function is high, the survivor function drops quickly
2. When the hazard function is low, the survivor function drops slowly
3. When the hazard function is zero, the survivor function is flat

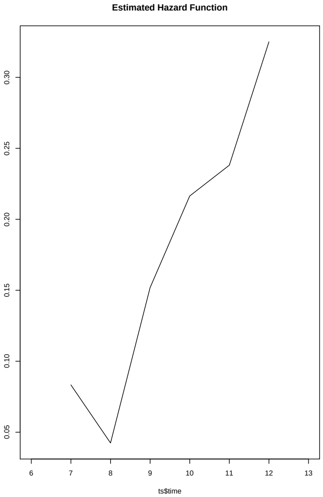
First Sexual Intercourse

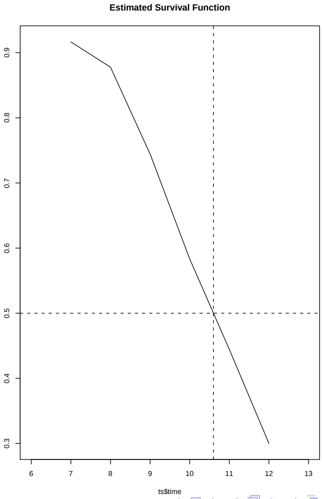
Capaldi, Crosby, and Stoolmiller (1996) studied the grade when a sample of at-risk adolescent males had their first heterosexual intercourse.

First Sexual Intercourse

```
> par(mfrow=c(1, 2), cex=0.7)
> ## firstsex<-read.table("http://www.ats.ucla.edu/stat/examples/alda/firstsex.csv",
> ## sep=",", header=T)
> firstsex<-read.table("firstsex.csv", sep=",", header=T)
> ts <- survfit( Surv(time, 1-censor)~ 1, conf.type="none", data=firstsex)
> h<-ts$n.event/ts$n.risk
> plot(ts$time, h, type="l", ylab=" ", main="Estimated Hazard Function",
+  xlim=c(6, 13))
> plot(ts$time, ts$surv, type="l", ylab=" ", main="Estimated Survival Function",
+  xlim=c(6, 13))
> abline(h=c(.5), lty=2)
> abline(v=c(10.6), lty=2)
```

First Sexual Intercourse





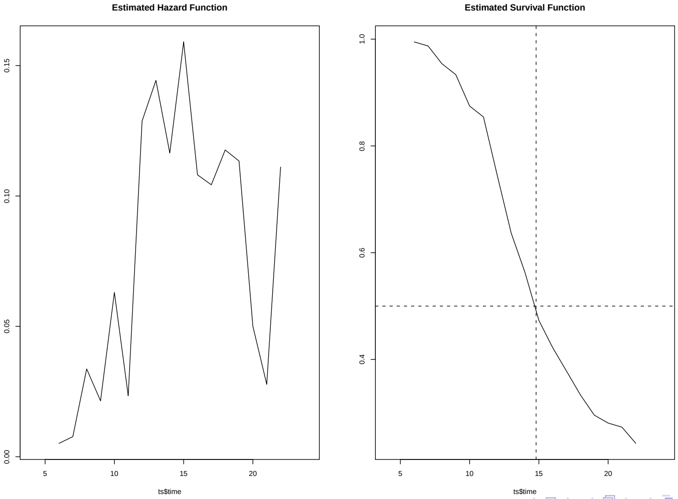
First Suicide Ideation

Bolger, et al. (1989) studied age at first suicide ideation for 391 undergraduates.

First Suicide Ideation

```
> par(mfrow=c(1, 2), cex=0.7)
> rm(list=ls())
> suicide<-read.table("suicide.csv", sep=",", header=T)
> ts <- survfit( Surv(time, 1-censor)˜ 1, conf.type="none", data=suicide)
> h<-ts$n.event/ts$n.risk
> plot(ts$time, h, type="l", ylab=" ", main="Estimated Hazard Function",
+  xlim=c(4, 24))
> plot(ts$time, ts$surv, type="l", ylab=" ", main="Estimated Survival Function",
+  xlim=c(4, 24))
> abline(h=c(.5), lty=2)
> abline(v=c(14.8), lty=2)
```

First Suicide Ideation



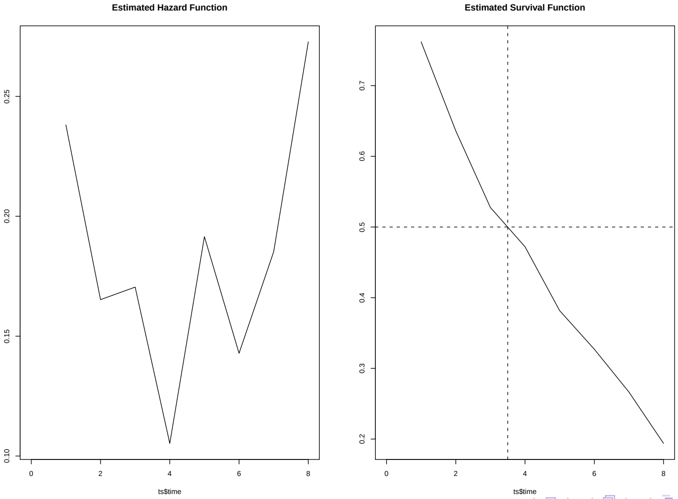
Careers of Female Congressional Representatives

This study tracks the careers of all 168 women elected to the U.S. House of Representatives between 1919 and 1996, for up to 8 terms or until 1998. Because the tracking ended in 1998, 37.5% of the participants' data were censored.

Careers of Female Congressional Representatives

```
> congress<-read.table("congress.csv", sep=",", header=T)
> ts <- survfit( Surv(time, 1-censor)˜ 1, conf.type="none", data=congress)
> h<-ts$n.event/ts$n.risk
> plot(ts$time, h, type="l", ylab=" ", main="Estimated Hazard Function",
+  xlim=c(0,8))
> plot(ts$time, ts$surv, type="l", ylab=" ", main="Estimated Survival Function",
+  xlim=c(0,8))
> abline(h=c(.5), lty=2)
> abline(v=c(3.5), lty=2)
```

Careers of Female Congressional Representatives


Estimated Hazard Function
ts$time
Estimated Survival Function
ts$time