الجامعة المستنصرية

كلية الادارة والاقتصاد/قسم الاحصاء

Survival Analysis

محاضرات الدراسة العليا(دبلوم عالي/احصاء حيوي)

اعداد

المدرس الدكتور

رواء صالح محمد الصفار

٢٠١٨

Survival Analysis:

- typically focuses on time to event data. In the most general sense, it consists of techniques for positive valued random variables

- time to death
- • time to onset (or relapse) of a disease
- • length of stay in a hospital
- • duration of a strike
- • money paid by health insurance
- • viral load measurements
- • time to finishing a doctoral dissertation

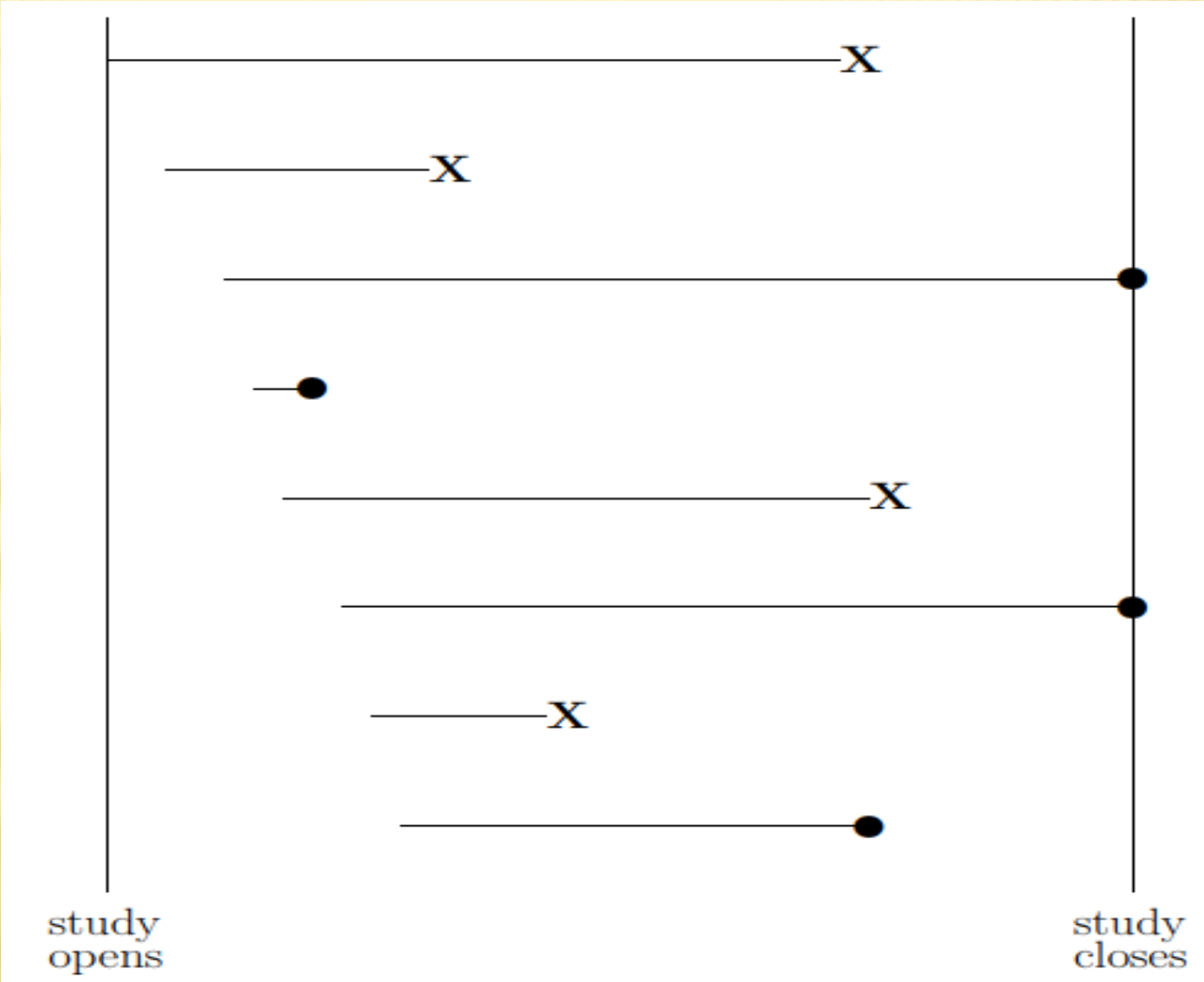
SOME DEFINITIONS AND NOTATION

- Failure time random variables are always non-negative. That is, if we denote the failure time by T, then $T \geq 0$. T can either be discrete (taking a finite set of values, e.g. a1, a2, . . . , an) or continuous (defined on $(0, \infty)$). A random variable X is called a censored failure time random variable if $X = \min(T, U)$, where U is a nonnegative censoring variable.

SOME DEFINITIONS AND NOTATION

- In order to define a failure time random variable, we need: (1) an unambiguous time origin (e.g. randomization to clinical trial, purchase of car) (2) a time scale (e.g. real time (days, years), mileage of a car) (3) definition of the event (e.g. death, need a new car transmission)

ILLUSTRATION OF SURVIVAL DATA



THE ILLUSTRATION OF SURVIVAL DATA ON THE PREVIOUS PAGE SHOWS SEVERAL FEATURES WHICH ARE TYPICALLY

- The illustration of survival data on the previous page shows several features which are typically encountered in analysis of survival data: • individuals do not all enter the study at the same time • when the study ends, some individuals still haven't had the event yet • other individuals drop out or get lost in the middle of the study, and all we know about them is the last time they were still "free" of the event The first feature is referred to as "staggered entry" The last two features relate to "censoring" of the failure time events.

TYPES OF CENSORING:

- • Right-censoring : only the r.v. Xi = min(Ti , Ui) is observed due to – loss to follow-up – drop-out – study termination We call this right-censoring because the true unobserved event is to the right of our censoring time; i.e., all we know is that the event has not happened at the end of follow-up. In addition to observing Xi , we also get to see the failure indicator:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq U_i \\ 0 & \text{if } T_i > U_i \end{cases}$$

Some software packages instead assume we have a censoring indicator:

$$c_i = \begin{cases} 0 & \text{if } T_i \leq U_i \\ 1 & \text{if } T_i > U_i \end{cases}$$

Right-censoring is the most common type of censoring assumption we will deal with in survival analysis

- Left-censoring Can only observe Yi = max(Ti , Ui) and the failure indicators:

$$\delta_i = \begin{cases} 1 & \text{if } U_i \leq T_i \\ 0 & \text{if } U_i > T_i \end{cases}$$

e.g. (Miller) study of age at which African children learn a task. Some already knew (left-censored), some learned during study (exact), some had not yet learned by end of study (right-censored).

Interval-censoring

Observe (Li , Ri) where Ti ∈ (Li , Ri) Ex. 1: Time to prostate cancer, observe l

- Ex. 1: Time to prostate cancer, observe longitudinal PSA measurements
- Ex. 2: Time to undetectable viral load in AIDS studies, based on measurements of viral load taken at each clinic visit
- Ex. 3: Detect recurrence of colon cancer after surgery. Follow patients every 3 months after resection of primary tumor.

INDEPENDENT VS INFORMATIVE CENSORING

- We say censoring is independent (non-informative) if Ui is independent of Ti .
- – Ex. 1 If Ui is the planned end of the study (say, 2 years after the study opens), then it is usually independent of the event times.
- – Ex. 2 If Ui is the time that a patient drops out of the study because he/she got much sicker and/or had to discontinue taking the study treatment, then Ui and Ti are probably not independent.

AN INDIVIDUAL CENSORED

- An individual censored at U should be representative of all subjects who survive to U. This means that censoring at U could depend on prognostic characteristics measured at baseline, but that among all those with the same baseline characteristics, the probability of censoring prior to or at time U should be the same.
- • Censoring is considered informative if the distribution of Ui contains any information about the parameters characterizing the distribution of Ti .

- Suppose we have a sample of observations on n people: (T1, U1),(T2, U2), …,(Tn, Un) There are three main types of (right) censoring times: **Type I**: All the Ui ’s are the same e.g. animal studies, all animals sacrificed after 2 years • **Type II**: Ui = T(r) , the time of the rth failure. e.g. animal studies, stop when 4/6 have tumors • **Type III**: the Ui ’s are random variables, δi ’s are failure indicators:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq U_i \\ 0 & \text{if } T_i > U_i \end{cases}$$

SOME EXAMPLE DATASETS

- Example A. Duration of nursing home stay (Morris et al., Case Studies in Biometry, Ch 12) The National Center for Health Services Research studied 36 for-profit nursing homes to assess the effects of different financial incentives on length of stay. “Treated” nursing homes received higher per diems for Medicaid patients, and bonuses for improving a patient’s health and sending them home. Study included 1601 patients admitted between May 1, 1981 and April 30, 1982.

VARIABLES INCLUDE

- LOS - Length of stay of a resident (in days)
- AGE - Age of a resident
- RX - Nursing home assignment (1:bonuses, 0:no bonuses)
- GENDER - Gender (1:male, 0:female)
- MARRIED - (1: married, 0:not married)
- HEALTH - health status (2:second best, 5:worst)
- CENSOR - Censoring indicator (1:censored, 0:discharged)
- First few lines of data: 37 86 1 0 0 2 0 61 77 1 0 0 4 0

MORE DEFINITIONS AND NOTATION

- There are several equivalent ways to characterize the probability distribution of a survival random variable. Some of these are familiar; others are special to survival analysis. We will focus on the following terms:
- • The density function f(t)
- • The survivor function S(t)
- • The hazard function λ(t)
- • The cumulative hazard function Λ(t)

DENSITY FUNCTION

- Density function (or Probability Mass Function) for discrete r.v.'s Suppose that T takes values in a1, a2, . . . , an.

$$f(t) = Pr(T = t)$$

$$= \begin{cases} f_j & \text{if } t = a_j, j = 1, 2, \ldots, n \\ 0 & \text{if } t \neq a_j, j = 1, 2, \ldots, n \end{cases}$$

  - Density Function for continuous r.v.'s

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t)$$

SURVIVORSHIP FUNCTION: S(T) = P(T ≥ T).

- In other settings, the cumulative distribution function, F(t) = P(T ≤ t), is of interest. In survival analysis, our interest tends to focus on the survival function, S(t)

For a continuous random variable:

$$S(t) = \int_t^{\infty} f(u)du$$

For a discrete random variable:

$$S(t) = \sum_{u \geq t} f(u)$$

$$= \sum_{a_j \geq t} f(a_j)$$

$$= \sum_{a_j \geq t} f_j$$

NOTES:

- • From the definition of S(t) for a continuous variable, S(t) = 1−F(t) as long as F(t) is absolutely continuous w.r.t the Lebesgue measure. [That is, F(t) has a density function.]
- • For a discrete variable, we have to decide what to do if an event occurs exactly at time t; i.e., does that become part of F(t) or S(t)?
- • To get around this problem, several books define S(t) = Pr(T > t), or else define F(t) = Pr(T < t) (eg. Collett)

HAZARD FUNCTION Λ(T)

- Sometimes called an instantaneous failure rate, the force of mortality, or the age-specific failure rate.

– **Continuous random variables:**

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T < t + \Delta t | T \geq t)$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr([t \leq T < t + \Delta t] \cap [T \geq t])}{Pr(T \geq t)}$$

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

$$= \frac{f(t)}{S(t)}$$

HAZARD FUNCTION Λ(T)

– **Discrete random variables:**

$$\begin{aligned}\lambda(a_j) \equiv \lambda_j &= Pr(T = a_j | T \geq a_j) \\ &= \frac{P(T = a_j)}{P(T \geq a_j)} \\ &= \frac{f(a_j)}{S(a_j)} \\ &= \frac{f(t)}{\sum_{k:a_k \geq a_j} f(a_k)}\end{aligned}$$

HAZARD FUNCTION Λ(T)

- **Cumulative Hazard Function $\Lambda(t)$**
  - **Continuous random variables:**

$$\Lambda(t) = \int_0^t \lambda(u)du$$

  - **Discrete random variables:**

$$\Lambda(t) = \sum_{k:a_k<t} \lambda_k$$

RELATIONSHIP BETWEEN S(T) AND Λ(T)

We've already shown that, for a continuous r.v.

$$\lambda(t) = \frac{f(t)}{S(t)}$$

For a left-continuous survivor function $S(t)$, we can show:

$$f(t) = -S'(t) \quad \text{or} \quad S'(t) = -f(t)$$

We can use this relationship to show that:

$$\begin{aligned} -\frac{d}{dt}[\log S(t)] &= -\left(\frac{1}{S(t)}\right) S'(t) \\ &= -\frac{-f(t)}{S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

So another way to write $\lambda(t)$ is as follows:

$$\lambda(t) = -\frac{d}{dt}[\log S(t)]$$

RELATIONSHIP BETWEEN S(T) AND Λ(T)

- **Continuous case:**

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u)du \\
&= \int_0^t \frac{f(u)}{S(u)}du \\
&= \int_0^t -\frac{d}{du}\log S(u)du \\
&= -\log S(t) + \log S(0) \\
&\Rightarrow S(t) = e^{-\Lambda(t)}
\end{aligned}$$

- **Discrete case:**

Suppose that $a_j < t \leq a_{j+1}$. Then

$$\begin{aligned}
S(t) &= P(T \geq a_1, T \geq a_2, \ldots, T \geq a_{j+1}) \\
&= P(T \geq a_1)P(T \geq a_2|T \geq a_1)\cdots P(T \geq a_{j+1}|T \geq a_j) \\
&= (1-\lambda_1)\times\cdots\times(1-\lambda_j) \\
&= \prod_{k:a_k<t}(1-\lambda_k)
\end{aligned}$$

MEASURING CENTRAL TENDENCY IN SURVIVAL

- **Mean survival** - call this μ

$$\mu = \int_0^\infty u f(u) du \quad \text{for continuous } T$$

$$= \sum_{j=1}^{n} a_j f_j \quad \text{for discrete } T$$

- **Median survival** - call this τ, is defined by

$$S(\tau) = 0.5$$

Similarly, any other percentile could be defined.

In practice, we don't usually hit the median survival at exactly one of the failure times. In this case, the estimated median survival is the *smallest* time τ such that

$$\hat{S}(\tau) \leq 0.5$$

SOME HAZARD SHAPES SEEN IN APPLICATIONS

- : • increasing e.g. aging after 65
- • decreasing e.g. survival after surgery
- • bathtub e.g. age-specific mortality
- • constant e.g. survival of patients with advanced chronic diseas

ESTIMATING THE SURVIVAL OR HAZARD FUNCTION

- We can estimate the survival (or hazard) function in two ways:
- • by specifying a parametric model for λ(t) based on a particular density function f(t)
- • by developing an empirical estimate of the survival function (i.e., non-parametric estimation) If no censoring: The empirical estimate of the survival function, S˜(t), is the proportion of individuals with event times greater than t. With censoring:
- If there are censored observations, then S˜(t) is not a good estimate of the true S(t), so other non-parametric methods must be used to account for censoring (life-table methods, Kaplan-Meier estimator)

SOME PARAMETRIC SURVIVAL DISTRIBUTIONS

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t} \text{ for } t \geq 0 \\
S(t) &= \int_t^\infty f(u)du \\
&= e^{-\lambda t} \\
\lambda(t) &= \frac{f(t)}{S(t)} \\
&= \lambda \qquad \text{constant hazard!} \\
\Lambda(t) &= \int_0^t \lambda(u)\, du \\
&= \int_0^t \lambda\, du \\
&= \lambda t
\end{aligned}
$$

Check: Does $S(t) = e^{-\Lambda(t)}$?

median: solve $0.5 = S(\tau) = e^{-\lambda\tau}$:

$$\Rightarrow \tau = \frac{-\log(0.5)}{\lambda}$$

mean:

$$\int_0^\infty u\lambda e^{-\lambda u} du = \frac{1}{\lambda}$$

THE WEIBULL DISTRIBUTION (2 PARAMETERS)

Generalizes exponential:

$$S(t) = e^{-\lambda t^{\kappa}}$$

$$f(t) = \frac{-d}{dt} S(t) = \kappa \lambda t^{\kappa-1} e^{-\lambda t^{\kappa}}$$

$$\lambda(t) = \kappa \lambda t^{\kappa-1}$$

$$\Lambda(t) = \int_0^t \lambda(u) du = \lambda t^{\kappa}$$

λ - the *scale* parameter
κ - the *shape* parameter

The Weibull distribution is convenient because of its simple form. It includes several hazard shapes:

$\kappa = 1 \rightarrow$ constant hazard

$0 < \kappa < 1 \rightarrow$ decreasing hazard

$\kappa > 1 \rightarrow$ increasing hazard

- **Rayleigh** distribution

  Another 2-parameter generalization of exponential:

$$\lambda(t) = \lambda_0 + \lambda_1 t$$

- **compound exponential**

  $T \sim \exp(\lambda)$, $\lambda \sim g$

$$f(t) = \int_0^\infty \lambda e^{-\lambda t} g(\lambda) d\lambda$$

- **log-normal**, **log-logistic**:

  Possible distributions for T obtained by specifying for $\log T$ any convenient family of distributions, e.g.

  $\log T \sim$ normal (non-monotone hazard)

  $\log T \sim$ logistic

WHY USE ONE VERSUS ANOTHER?

- • technical convenience for estimation and inference
- • explicit simple forms for f(t), S(t), and λ(t).
- • qualitative shape of hazard function One can usually distinguish between a one-parameter model (like the exponential) and two-parameter (like Weibull or log-normal) in terms of the adequacy of fit to a dataset. Without a lot of data, it may be hard to distinguish between the fits of various 2-parameter models (i.e., Weibull vs lognormal)

PREVIEW OF COMING ATTRACTIONS

- Next we will discuss the most famous non-parametric approach for estimating the survival distribution, called the Kaplan-Meier estimator. To motivate the derivation of this estimator, we will first consider a set of survival times where there is no censoring. The following are times to relapse (weeks) for 21 leukemia patients receiving control treatment (Table 1.1 of Cox & Oakes): 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23 How would we estimate S(10), the probability that an individual survives to time 10 or later?

What about $\tilde{S}(8)$? Is it $\frac{12}{21}$ or $\frac{8}{21}$?

Values of t	$\hat{S}(t)$
$t \leq 1$	21/21=1.000
$1 < t \leq 2$	19/21=0.905
$2 < t \leq 3$	17/21=0.809
$3 < t \leq 4$	
$4 < t \leq 5$	
$5 < t \leq 8$	
$8 < t \leq 11$	
$11 < t \leq 12$	
$12 < t \leq 15$	
$15 < t \leq 17$	
$17 < t \leq 22$	
$22 < t \leq 23$	

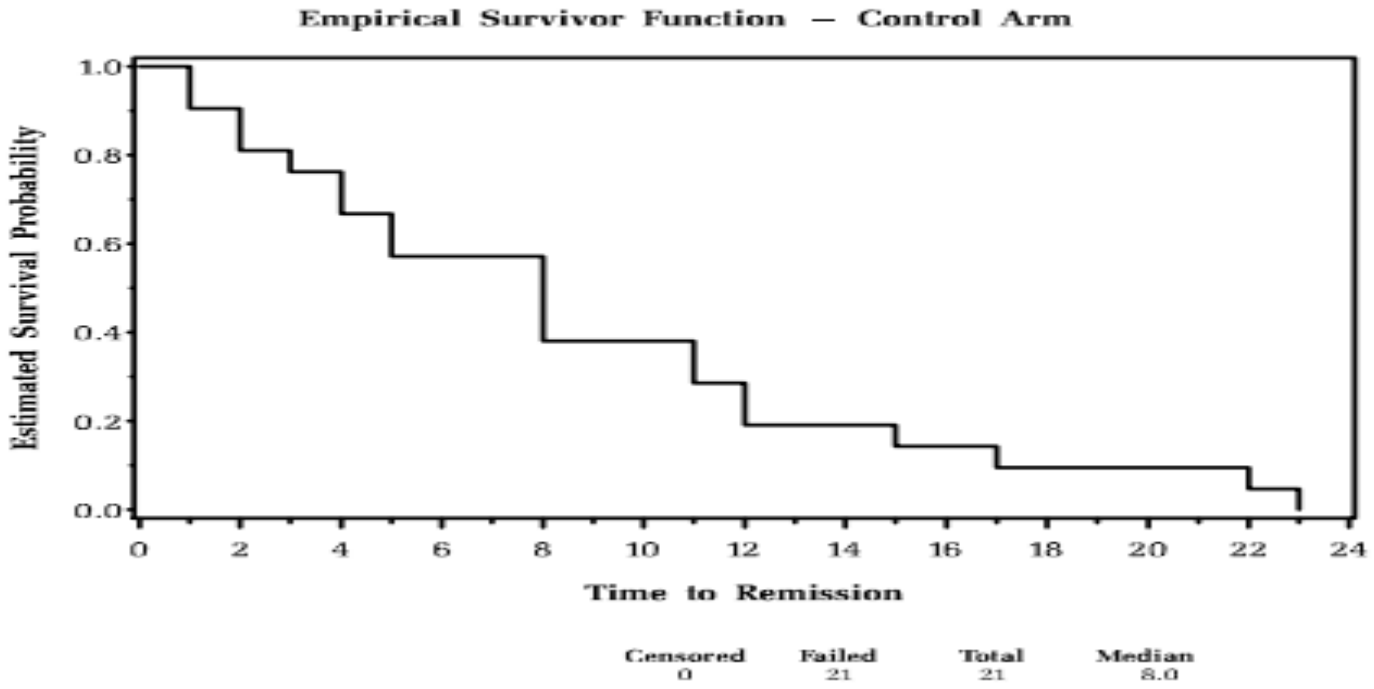
Empirical Survival Function:

When there is no censoring, the general formula is:

$$\tilde{S}(t) = \frac{\#\ individuals\ with\ T \geq t}{total\ sample\ size}$$

- In most software packages, the survival function is evaluated just after time t, i.e., at t +. In this case, we only count the individuals with T > t.

Example for leukemia data (control arm):



Stata Commands for Survival Estimation

```
.use leukem

.stset remiss status if trt==0        (to keep only untreated patients)
(21 observations deleted)

. sts list
         failure _d:  status
   analysis time _t:  remiss
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	21	2	0	0.9048	0.0641	0.6700	0.9753
2	19	2	0	0.8095	0.0857	0.5689	0.9239
3	17	1	0	0.7619	0.0929	0.5194	0.8933
4	16	2	0	0.6667	0.1029	0.4254	0.8250
5	14	2	0	0.5714	0.1080	0.3380	0.7492
8	12	4	0	0.3810	0.1060	0.1831	0.5778
11	8	2	0	0.2857	0.0986	0.1166	0.4818
12	6	2	0	0.1905	0.0857	0.0595	0.3774
15	4	1	0	0.1429	0.0764	0.0357	0.3212
17	3	1	0	0.0952	0.0641	0.0163	0.2612
22	2	1	0	0.0476	0.0465	0.0033	0.1970
23	1	1	0	0.0000	.	.	.

```
.sts graph
```

SAS Commands for Survival Estimation

```
data leuk;
 input t;
cards;
1
1
2
2
3
4
4
5
5
8
8
8
8
11
11
12
12
15
17
22
23
;

proc lifetest data=leuk;
 time t;
run;
```

SAS Output for Survival Estimation

The LIFETEST Procedure

Product-Limit Survival Estimates

t	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	21
1.0000	.	.	.	1	20
1.0000	0.9048	0.0952	0.0641	2	19
2.0000	.	.	.	3	18
2.0000	0.8095	0.1905	0.0857	4	17
3.0000	0.7619	0.2381	0.0929	5	16
4.0000	.	.	.	6	15
4.0000	0.6667	0.3333	0.1029	7	14
5.0000	.	.	.	8	13
5.0000	0.5714	0.4286	0.1080	9	12
8.0000	.	.	.	10	11
8.0000	.	.	.	11	10
8.0000	.	.	.	12	9
8.0000	0.3810	0.6190	0.1060	13	8
11.0000	.	.	.	14	7
11.0000	0.2857	0.7143	0.0986	15	6
12.0000	.	.	.	16	5
12.0000	0.1905	0.8095	0.0857	17	4
15.0000	0.1429	0.8571	0.0764	18	3
17.0000	0.0952	0.9048	0.0641	19	2
22.0000	0.0476	0.9524	0.0465	20	1
23.0000	0	1.0000	0	21	0

SAS Output for Survival Estimation (cont'd)

Summary Statistics for Time Variable t

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	12.0000	8.0000	17.0000
50	8.0000	4.0000	11.0000
25	4.0000	2.0000	8.0000

Mean	Standard Error
8.6667	1.4114

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
21	21	0	0.00

Does anyone have a guess regarding how to calculate the standard error of the estimated survival?

$$\hat{S}(8^+) = P(T > 8) = \frac{8}{21} = 0.381$$

(at $t = 8^+$, we count the 4 events at time=8 as already having failed)

$$se[\hat{S}(8^+)] = 0.106$$

ESTIMATING THE SURVIVAL FUNCTION

- One-sample nonparametric methods:
- We will consider three methods for estimating a survivorship function
  - $S(t) = Pr(T \geq t)$
- without resorting to parametric methods:
- (1) Kaplan-Meier
- (2) Life-table (Actuarial Estimator)
- (3) via the Cumulative hazard estimator

(1) THE KAPLAN-MEIER ESTIMATOR

- The Kaplan-Meier (or KM) estimator is probably the most popular approach. It can be justified from several perspectives:
- • product limit estimator
- • likelihood justification
- • redistribute to the right estimator We will start with an intuitive motivation based on conditional probabilities, then review some of the other justifications.

MOTIVATION:

- First, consider an example where there is no censoring. The following are times of remission (weeks) for 21 leukemia patients receiving control treatment (Table 1.1 of Cox & Oakes): 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23 How would we estimate S(10), the probability that an individual survives to time 10 or later?

What about $\tilde{S}(8)$? Is it $\frac{12}{21}$ or $\frac{8}{21}$?

Let's construct a table of $\tilde{S}(t)$:

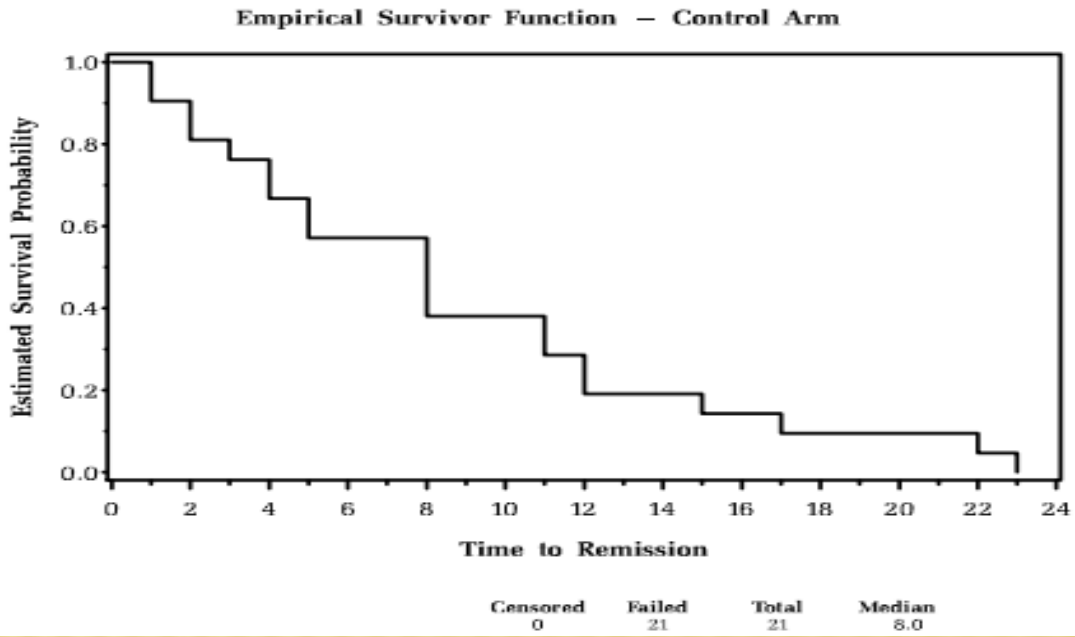
Values of t	$\hat{S}(t)$
$t \leq 1$	21/21=1.000
$1 < t \leq 2$	19/21=0.905
$2 < t \leq 3$	17/21=0.809
$3 < t \leq 4$	
$4 < t \leq 5$	
$5 < t \leq 8$	
$8 < t \leq 11$	
$11 < t \leq 12$	
$12 < t \leq 15$	
$15 < t \leq 17$	
$17 < t \leq 22$	
$22 < t \leq 23$	

EMPIRICAL SURVIVAL FUNCTION:

- When there is no censoring, the general formula is:

$$\tilde{S}(t) = \frac{\#\ individuals\ with\ T \geq t}{total\ sample\ size}$$

Example for leukemia data (control arm):



WHAT IF THERE IS CENSORING?

- Consider the treated group from Table 1.1 of Cox and Oakes:
- 6 + , 6, 6, 6, 7, 9 + , 10+ , 10, 11+ , 13, 16, 17+ 19+ , 20+ , 22, 23, 25+ , 32+ , 32+ , 34+ , 35+ [Note: times with + are right censored]
- We know S(6)= 21/21,
- because everyone survived at least until time 6 or greater. But, we can't say
- S(7) = 17/21,
- because we don't know the status of the person who was censored at time 6. In a 1958 paper in the Journal of the American Statistical Association, Kaplan and Meier proposed a way to nonparametrically estimate S(t), even in the presence of censoring. The method is based on the ideas of conditional probability.

A QUICK REVIEW OF CONDITIONAL PROBABILITY:

- Conditional Probability: Suppose A and B are two events. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Multiplication law of probability: can be obtained from the above relationship, by multiplying both sides by P(B):

$$P(A \cap B) = P(A|B)\,P(B)$$

EXTENSION TO MORE THAN 2 EVENTS

- : Suppose A1, A2…Ak are k different events. Then, the probability of all k events happening together can be written as a product of conditional probabilities:

$$
\begin{aligned}
P(A_1 \cap A_2 \ldots \cap A_k) = \; & P(A_k | A_{k-1} \cap \ldots \cap A_1) \times \\
& \times P(A_{k-1} | A_{k-2} \cap \ldots \cap A_1) \\
& \ldots \\
& \times P(A_2 | A_1) \\
& \times P(A_1)
\end{aligned}
$$

NOW, LET'S APPLY THESE IDEAS TO ESTIMATE S(T):

Suppose $a_k < t \leq a_{k+1}$. Then

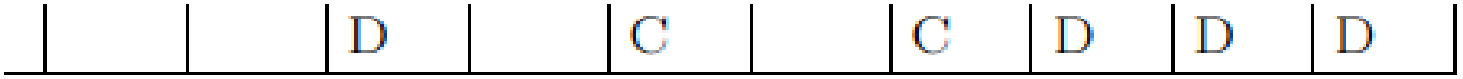
$$
\begin{aligned}
S(t) &= P(T \geq a_{k+1}) \\
&= P(T \geq a_1, T \geq a_2, \ldots, T \geq a_{k+1}) \\
&= P(T \geq a_1) \times \prod_{j=1}^{k} P(T \geq a_{j+1} | T \geq a_j) \\
&= \prod_{j=1}^{k} [1 - P(T = a_j | T \geq a_j)] \\
&= \prod_{j=1}^{k} [1 - \lambda_j] \\
\text{so} \quad \hat{S}(t) &\cong \prod_{j=1}^{k} \left(1 - \frac{d_j}{r_j}\right) \\
&= \prod_{j:a_j<t} \left(1 - \frac{d_j}{r_j}\right)
\end{aligned}
$$

d_j is the number of deaths at a_j

r_j is the number at risk at a_j

INTUITION BEHIND THE KAPLAN-MEIER ESTIMATOR

- Think of dividing the observed timespan of the study into a series of fine intervals so that there is a separate interval for each time of death or censoring:



Using the law of conditional probability,

$$Pr(T \geq t) = \prod_j Pr(\text{survive } j\text{-th interval } I_j \mid \text{survived to start of } I_j)$$

where the product is taken over all the intervals including or preceding time t.

4 POSSIBILITIES FOR EACH INTERVAL:

- (1) No events (death or censoring) - conditional probability of surviving the interval is 1
- (2) Censoring - assume they survive to the end of the interval, so that the conditional probability of surviving the interval is 1
- (3) Death, but no censoring - conditional probability of not surviving the interval is # deaths (d) divided by # 'at risk' (r) at the beginning of the interval. So the conditional probability of surviving the interval is $1-(d/r)$.
- (4) Tied deaths and censoring - assume censorings last to the end of the interval, so that conditional probability of surviving the interval is still $1 - (d/r)$

GENERAL FORMULA FOR JTH INTERVAL:

- It turns out we can write a general formula for the conditional probability of surviving the j-th interval that holds for all 4 cases:

$$1 - \frac{d_j}{r_j}$$

- We could use the same approach by grouping the event times into intervals (say, one interval for each month), and then counting up the number of deaths (events) in each to estimate the probability of surviving the interval (this is called the lifetable estimate).

- However, the assumption that those censored last until the end of the interval wouldn't be quite accurate, so we would end up with a cruder approximation. As the intervals get finer and finer, the approximations made in estimating the probabilities of getting through each interval become smaller and smaller, so that the estimator converges to the true S(t). This intuition clarifies why an alternative name for the KM is the product limit estimator.

The Kaplan-Meier estimator of the survivorship function (or survival probability) $S(t) = Pr(T \geq t)$ is:

$$\hat{S}(t) = \Pi_{j:\tau_j<t} \frac{r_j - d_j}{r_j}$$

$$= \Pi_{j:\tau_j<t} \left(1 - \frac{d_j}{r_j}\right)$$

where

- $\tau_1, ...\tau_K$ is the set of K distinct death times observed in the sample
- d_j is the number of deaths at τ_j
- r_j is the number of individuals "at risk" right before the j-th death time (everyone dead or censored at or after that time).
- c_j is the number of censored observations between the j-th and $(j+1)$-st death times. Censorings tied at τ_j are included in c_j

Note: two useful formulas are:

(1) $r_j = r_{j-1} - d_{j-1} - c_{j-1}$

(2) $r_j = \sum_{l \geq j} (c_l + d_l)$

CALCULATING THE KM - COX AND OAKES EXAMPLE

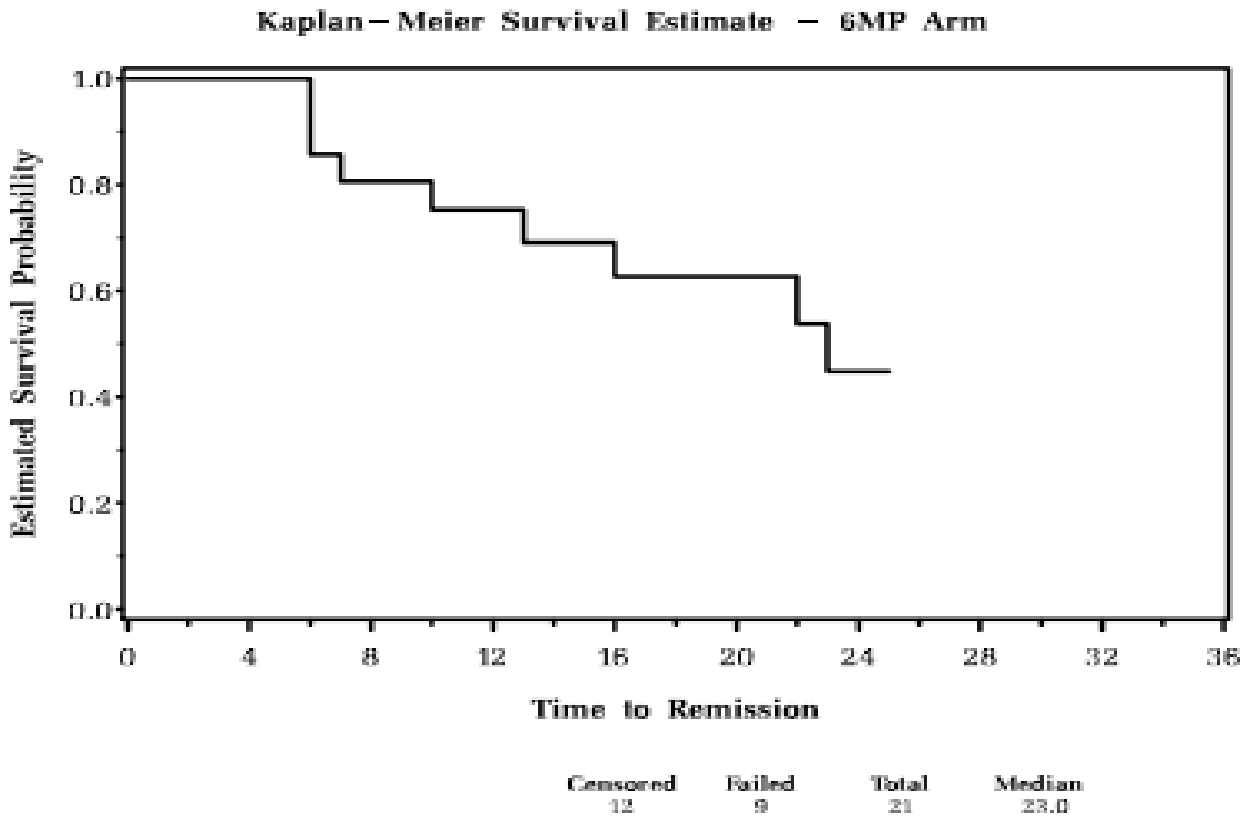
Make a table with a row for every death or censoring time:

τ_j	d_j	c_j	r_j	$1-(d_j/r_j)$	$\hat{S}(\tau_j^+)$
6	3	1	21	$\frac{18}{21}=0.857$	
7	1	0	17		
9	0	1	16		
10					
11					
13					
16					
17					
19					
20					
22					
23					

Note that:

- $\hat{S}(t^+)$ only changes at death (failure) times
- $\hat{S}(t^+)$ is 1 up to the first death time
- $\hat{S}(t^+)$ only goes to 0 if the last event is a death

KM plot for treated leukemia patients



Note: most statistical software packages summarize the KM survival function at τ_j^+, i.e., *just after* the time of the j-th failure.

In other words, they provide $\hat{S}(\tau_j^+)$.

When there is no censoring, the empirical survival estimate would then be:

$$\tilde{S}(t^+) = \frac{\#\ individuals\ with\ T > t}{total\ sample\ size}$$

Output from STATA KM Estimator:

```
  failure time:  weeks
failure/censor:  remiss
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
6	21	3	1	0.8571	0.0764	0.6197	0.9516
7	17	1	0	0.8067	0.0869	0.5631	0.9228
9	16	0	1	0.8067	0.0869	0.5631	0.9228
10	15	1	1	0.7529	0.0963	0.5032	0.8894
11	13	0	1	0.7529	0.0963	0.5032	0.8894
13	12	1	0	0.6902	0.1068	0.4316	0.8491
16	11	1	0	0.6275	0.1141	0.3675	0.8049
17	10	0	1	0.6275	0.1141	0.3675	0.8049
19	9	0	1	0.6275	0.1141	0.3675	0.8049
20	8	0	1	0.6275	0.1141	0.3675	0.8049
22	7	1	0	0.5378	0.1282	0.2678	0.7468
23	6	1	0	0.4482	0.1346	0.1881	0.6801
25	5	0	1	0.4482	0.1346	0.1881	0.6801
32	4	0	2	0.4482	0.1346	0.1881	0.6801
34	2	0	1	0.4482	0.1346	0.1881	0.6801
35	1	0	1	0.4482	0.1346	0.1881	0.6801

TWO OTHER JUSTIFICATIONS FOR KM ESTIMATOR

- I. Likelihood-based derivation (Cox and Oakes) For a discrete failure time variable, define:
- dj number of failures at aj rj number of individuals at risk at aj (including those censored at aj).
- λj Pr(death) in j-th interval (conditional on survival to start of interval) The likelihood is that of g independent binomials:

The likelihood is that of g independent binomials:

$$L(\boldsymbol{\lambda}) = \prod_{j=1}^{g} \lambda_j^{d_j}(1-\lambda_j)^{r_j-d_j}$$

Therefore, the **maximum likelihood estimator** of λ_j is:

$$\hat{\lambda}_j = d_j / r_j$$

Now we plug in the MLE's of λ to estimate S(t):

$$\hat{S}(t) = \prod_{j:a_j<t} (1 - \hat{\lambda}_j)$$

$$= \prod_{j:a_j<t} \left(1 - \frac{d_j}{r_j}\right)$$

II. REDISTRIBUTE TO THE RIGHT JUSTIFICATION

- In the absence of censoring, $\hat{S}(t)$ is just the proportion of individuals with $T \geq t$. The idea behind Efron's approach is to spread the contributions of censored observations out over all the possible times to their right.
- Algorithm:
- • Step (1): arrange the n observed times (deaths or censorings) in increasing order. If there are ties, put censored after deaths.
- • Step (2): Assign weight (1/n) to each time.
- • Step (3): Moving from left to right, each time you encounter a censored observation, distribute its mass to all times to its right. •
- Step (4): Calculate $\hat{S}\ j$ by subtracting the final weight for time j from $\hat{S}\ j-1$

PROPERTIES OF THE KM ESTIMATOR

In the case of no censoring:

$$\hat{S}(t) = \tilde{S}(t) = \frac{\#\ deaths\ at\ t\ or\ greater}{n}$$

where n is the number of individuals in the study.

This is just like an estimated probability from a binomial distribution, so we have:

$$\hat{S}(t) \simeq \mathcal{N}\left(S(t),\ S(t)[1 - S(t)]/n\right)$$

How does censoring affect this?

- $\hat{S}(t)$ is still approximately normal
- The mean of $\hat{S}(t)$ converges to the true S(t)
- The variance is a bit more complicated (since the denominator n includes some censored observations).

GREENWOOD'S FORMULA (COLLETT 2.1.3)

We can think of the KM estimator as

$$\hat{S}(t) = \prod_{j:\tau_j<t} (1 - \hat{\lambda}_j)$$

where $\hat{\lambda}_j = d_j/r_j$.

Since the $\hat{\lambda}_j$'s are just binomial proportions, we can apply standard likelihood theory to show that each $\hat{\lambda}_j$ is approximately normal, with mean the true λ_j, and

$$var(\hat{\lambda}_j) \approx \frac{\hat{\lambda}_j(1 - \hat{\lambda}_j)}{r_j}$$

Also, the $\hat{\lambda}_j$'s are independent in large enough samples.

Since $\hat{S}(t)$ is a function of the λ_j's, we can estimate its variance using the **delta method**:

> **Delta method**: If Y is normal with mean μ and variance σ^2, then $g(Y)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g'(\mu)]^2\sigma^2$.

Two specific examples of the delta method:

(A) $Z = log(Y)$

$$\text{then} \quad Z \sim N\left[log(\mu), \left(\frac{1}{\mu}\right)^2 \sigma^2\right]$$

(B) $Z = exp(Y)$

$$\text{then} \quad Z \sim N\left[e^{\mu}, [e^{\mu}]^2 \sigma^2\right]$$

The examples above use the following results from calculus:

$$\frac{d}{dx} \log u = \frac{1}{u} \left(\frac{du}{dx}\right)$$

$$\frac{d}{dx} e^u = e^u \left(\frac{du}{dx}\right)$$

GREENWOOD'S FORMULA (CONTINUED)

Instead of dealing with $\hat{S}(t)$ directly, we will look at its log:

$$\log[\hat{S}(t)] = \sum_{j:\tau_j<t} \log(1-\hat{\lambda}_j)$$

Thus, by approximate independence of the $\hat{\lambda}_j$'s,

$$\begin{aligned} var(\log[\hat{S}(t)]) &= \sum_{j:\tau_j<t} var[\log(1-\hat{\lambda}_j)] \\ \text{by (A)} \qquad &= \sum_{j:\tau_j<t} \left(\frac{1}{1-\hat{\lambda}_j}\right)^2 var(\hat{\lambda}_j) \\ &= \sum_{j:\tau_j<t} \left(\frac{1}{1-\hat{\lambda}_j}\right)^2 \hat{\lambda}_j(1-\hat{\lambda}_j)/r_j \\ &= \sum_{j:\tau_j<t} \frac{\hat{\lambda}_j}{(1-\hat{\lambda}_j)r_j} \\ &= \sum_{j:\tau_j<t} \frac{d_j}{(r_j-d_j)r_j} \end{aligned}$$

Now, $\hat{S}(t) = \exp[\log[\hat{S}(t)]]$. Thus by (B),

$$var(\hat{S}(t)) = [\hat{S}(t)]^2 var\left[\log[\hat{S}(t)]\right]$$

Greenwood's Formula:

$$var(\hat{S}(t)) = [\hat{S}(t)]^2 \Sigma_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j}$$

- A is $1.96\, se(\hat{L}(t))$
- To calculate this, we need to calculate

$$var(\hat{L}(t)) = var\left[\log(-\log(\hat{S}(t)))\right]$$

- From our previous calculations, we know

$$var(\log[\hat{S}(t)]) = \sum_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j}$$

- Applying the delta method as in example (A), we get:

$$\begin{aligned} var(\hat{L}(t)) &= var(\log(-\log[\hat{S}(t)])) \\ &= \frac{1}{[\log \hat{S}(t)]^2} \sum_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j} \end{aligned}$$

- We take the square root of the above to get $se(\hat{L}(t))$, and then form the confidence intervals as:

$$\hat{S}(t)^{e^{\pm 1.96\, se(\hat{L}(t))}}$$

- This is the approach that Stata uses. Splus gives an option to calculate these bounds (use `conf.type=''log-log''` in `surv.fit`).

THE (2) LIFETABLE ESTIMATOR OF SURVIVAL:

- We said that we would consider the following three methods for estimating a survivorship function
- $S(t) = Pr(T \geq t)$
- without resorting to parametric methods:
- (1) $\surd$ Kaplan-Meier
- (2) $\Longrightarrow$ Life-table (Actuarial Estimator)
- (3) $\Longrightarrow$ Cumulative hazard estimator

(2) THE LIFETABLE OR ACTUARIAL ESTIMATOR

- • one of the oldest techniques around
- • used by actuaries, demographers, etc.
- • applies when the data are grouped Our goal is still to estimate the survival function, hazard, and density function, but this is complicated by the fact that we don’t know exactly when during each time interval an event occurs

- Clinical Life tables - applies to grouped survival data from studies in patients with specific diseases. Because patients can enter the study at different times, or be lost to follow-up, censoring must be allowed

Notation

- the j-th time interval is $[t_{j-1}, t_j)$
- c_j - the number of censorings in the j-th interval
- d_j - the number of failures in the j-th interval
- r_j is the number entering the interval

Example: 2418 Males with Angina Pectoris (Lee, p.91)

Year after Diagnosis	j	d_j	c_j	r_j	$r'_j = r_j - c_j/2$	
$[0, 1)$	1	456	0	2418	2418.0	
$[1, 2)$	2	226	39	1962	1942.5	$(1962 - \frac{39}{2})$
$[2, 3)$	3	152	22	1697	1686.0	
$[3, 4)$	4	171	23	1523	1511.5	
$[4, 5)$	5	135	24	1329	1317.0	
$[5, 6)$	6	125	107	1170	1116.5	
$[6, 7)$	7	83	133	938	871.5	
etc..						

ESTIMATING THE SURVIVORSHIP FUNCTION

- We could apply the K-M formula directly to the numbers in the table on the previous page, estimating S(t) as

$$\hat{S}(t) = \prod_{j:\tau_j<t} \left(1 - \frac{d_j}{r_j}\right)$$

- However, this approach is unsatisfactory for grouped data.... it treats the problem as though it were in discrete time, with events happening only at 1 yr, 2 yr, etc. In fact, what we are trying to calculate here is the conditional probability of dying within the interval, given survival to the beginning of it.

What should we do with the censored people?

We can assume that censorings occur:

- at the beginning of each interval: $r'_j = r_j - c_j$
- at the end of each interval: $r'_j = r_j$
- **on average halfway through** the interval:

$$r'_j = r_j - c_j/2$$

The last assumption yields the Actuarial Estimator. It is appropriate if censorings occur uniformly throughout the interval.

CONSTRUCTING THE LIFETABLE

- First, some additional notation for the j-th interval, [tj−1,tj):
- • Midpoint (tmj) - useful for plotting the density and the hazard function
- • Width (bj = tj−tj−1) needed for calculating the hazard in the j-th interval

QUANTITIES ESTIMATED:

Quantities estimated:

- Conditional probability of dying

$$\hat{q}_j = d_j / r'_j$$

- Conditional probability of surviving

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

- Cumulative probability of surviving at t_j:

$$\begin{aligned} \hat{S}(t_j) &= \prod_{\ell \leq j} \hat{p}_\ell \\ &= \prod_{\ell \leq j} \left(1 - \frac{d_\ell}{r_\ell}\right) \end{aligned}$$

SOME IMPORTANT POINTS TO NOTE:

- Because the intervals are defined as $[t_{j-1}, t_j)$, the first interval typically starts with $t_0 = 0$.
- Stata estimates the survival function at the right-hand endpoint of each interval, i.e., $S(t_j)$
- However, SAS estimates the survival function at the left-hand endpoint, $S(t_{j-1})$.
- The implication in SAS is that $\hat{S}(t_0) = 1$ and $\hat{S}(t_1) = p_1$

CONSTRUCTING THE LIFETABLE USING STATA

- Uses the ltable command. If the raw data are already grouped, then the freq statement must be used when reading the data.

```
. infile years status count using angina.dat
(32 observations read)

. ltable years status [freq=count]
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	1	2418	456	0	0.8114	0.0080	0.7952	0.8264
1	2	1962	226	39	0.7170	0.0092	0.6986	0.7346
2	3	1697	152	22	0.6524	0.0097	0.6329	0.6711
3	4	1523	171	23	0.5786	0.0101	0.5584	0.5981
4	5	1329	135	24	0.5193	0.0103	0.4989	0.5392
5	6	1170	125	107	0.4611	0.0104	0.4407	0.4813
6	7	938	83	133	0.4172	0.0105	0.3967	0.4376
7	8	722	74	102	0.3712	0.0106	0.3505	0.3919
8	9	546	51	68	0.3342	0.0107	0.3133	0.3553
9	10	427	42	64	0.2987	0.0109	0.2775	0.3201
10	11	321	43	45	0.2557	0.0111	0.2341	0.2777
11	12	233	34	53	0.2136	0.0114	0.1917	0.2363
12	13	146	18	33	0.1839	0.0118	0.1614	0.2075
13	14	95	9	27	0.1636	0.0123	0.1404	0.1884
14	15	59	6	23	0.1429	0.0133	0.1180	0.1701
15	16	30	0	30	0.1429	0.0133	0.1180	0.1701

```
. ltable years status [freq=count], hazard
```

Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf	Int]
0	1	2418	0.1886	0.0080	0.2082	0.0097	0.1892	0.2272
1	2	1962	0.2830	0.0092	0.1235	0.0082	0.1075	0.1396
2	3	1697	0.3476	0.0097	0.0944	0.0076	0.0794	0.1094
3	4	1523	0.4214	0.0101	0.1199	0.0092	0.1020	0.1379
4	5	1329	0.4807	0.0103	0.1080	0.0093	0.0898	0.1262
5	6	1170	0.5389	0.0104	0.1186	0.0106	0.0978	0.1393
6	7	938	0.5828	0.0105	0.1000	0.0110	0.0785	0.1215
7	8	722	0.6288	0.0106	0.1167	0.0135	0.0902	0.1433
8	9	546	0.6658	0.0107	0.1048	0.0147	0.0761	0.1336
9	10	427	0.7013	0.0109	0.1123	0.0173	0.0784	0.1462
10	11	321	0.7443	0.0111	0.1552	0.0236	0.1090	0.2015
11	12	233	0.7864	0.0114	0.1794	0.0306	0.1194	0.2395
12	13	146	0.8161	0.0118	0.1494	0.0351	0.0806	0.2182
13	14	95	0.8364	0.0123	0.1169	0.0389	0.0407	0.1931
14	15	59	0.8571	0.0133	0.1348	0.0549	0.0272	0.2425
15	16	30	0.8571	0.0133	0.0000	.	.	.

There is also a "**failure**" option which gives the number of failures (like the default), and also provides a 95% confidence interval on the cumulative failure probability.