LECTURE NOTES FOR SURV 616

Statistical Methods II

Statistical Methods in Epidemiology

CLASS #14

OVERVIEW: In this class we will discuss the non-parametric analysis of *life time distributions.* We will begin with a brief review of cumulative distribution functions, and introduce the concept of a survivor function and hazard rate function. Then we will discuss the estimation of the intervalized survivor function by life-table methods, and then discuss the *Product-Limit, or Kaplan-Meier Estimator.* Finally we will discuss the Cox Proportional hazards regression model for introducing covariates into the analysis.

References

Lawless, J.F. (1982). **Statistical Models and Methods for Lifetime Data.** John Wiley & Sons, New York.

1. Distributions and Survivor Functions.

We will be interested in analyzing strictly nonnegative random variables $T \geq 0$ which will represent survival times for individuals. For the moment we will consider the random variable T to be continuous. In general the distribution of T can be characterized by its density function

$$(1.1) \qquad f_T(t) \geq 0 \qquad for \qquad t \geq 0$$

and its cumulative distribution function (CDF)

$$(1.2) \qquad F_T(t) = P\{T \leq t\} = \int_0^t f_T(u)\,du$$

Usually in mathematical statistics we assume a parametric form of the CDF, say $F_T(t;\theta)$, and then estimated the parameter vector θ by the method of maximum likelihood. The estimated CDF is then $F_T(t;\hat{\theta})$ where $\hat{\theta}$ is the maximum likelihood estimator of θ.

There are many parametric distributions which are commonly used in the analysis of life distributions, with the most common being

(1.3) The Exponential Distribution $f_T(t) = \lambda e^{-\lambda t}$

The Weibull Distribution $f_T(t) = \lambda\beta(\lambda t)^{\beta-1} e^{-(\lambda t)^\beta}$

The Gamma Distribution $f_T(t) = \dfrac{\lambda(\lambda t)^{k-1} e^{-\lambda t}}{\Gamma(k)}$

The Log-Normal Distribution $f_T(t) = \dfrac{1}{\sqrt{2\pi}\sigma t} \exp\left\{ -\dfrac{1}{2}\left(\dfrac{\log(t)-\mu}{\sigma} \right)^2 \right\}$

Often we are interested in estimating the distribution function $F_T(t)$ in the absence of any particular parametric form. In this case we ay that we want a nonparametric estimate of the distribution function $F_T(t)$. To make things specific, we will assume that we have an independent random sample $T_1, T_2, ..., T_n$ from the distribution function $F_T(t)$. Then it can be shown that the *nonparametric maximum likelihood estimator* of $F_T(t)$ is given by

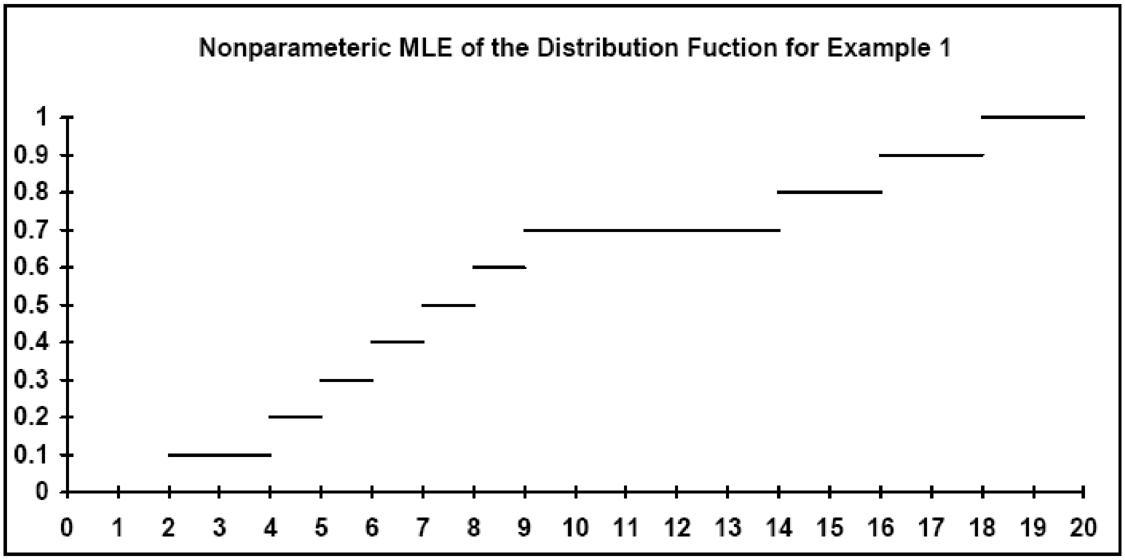
$$(1.4) \qquad \hat{F}_T(t) = n^{-1} \sum_{i=1}^{n} 1\{T_i \le t\}$$

where the indicator function is defined by

$$(1.5) \qquad \begin{aligned} 1\{T_i \le t\} &= 1 \quad \text{when} \quad T_i \le t \\ &= 0 \quad \text{when} \quad T_i > t \end{aligned}$$

The non-parametric maximum likelihood estimator puts mass n^{-1} at each of the observed data points. As an illustration we consider the following example

Example 1. Let 5,6,8,18,16,14,7,2,4,9 be an observed sample of size 10 from a common distribution function F_T. Then graphically the maximum likelihood estimator of F_T looks like



Often the graph above will have an open circle on the right side of the line segment, an a closed circle on the left side of the line segment, which would indicate that the distribution function is right continuous. ■

The random variable T will represent for us the time of death for an individual, and the distribution function $F_T(t)$ represents the probability that an individual dies before time t. A more optimistic, as well as heavily used function is the *survival function* defined as

$$S_T(t) = P\{T > t\} \tag{1.6}$$

which is the probability that a person lives beyond time t. Clearly, the survival function is related to the cumulative distribution function by

$$S_T(t) = 1 - F_T(t) \tag{1.7}$$

The survival function, just as the cumulative distribution function completely defines the distribution of the random variable T.

Another characteristic, called the *hazard rate* or *hazard function*, of the distribution of the random variable T is of interest. The hazard function can be motivated by the following situation. Assume that you have lived to time t, what is the probability you will die in the next period of time Δt? Mathematically this says, what is the probability

$$(1.8) \qquad P\left\{t \le T < t + \Delta t \middle| T \ge t\right\}$$

As the time period Δt goes to zero the probability in (1.8) goes to zero, so we speak of a *limiting,* or *instantaneous* hazard rate

$$(1.9) \qquad h_T\left(t\right) = \lim_{\Delta t \to 0} \frac{P\left\{t \le T < t + \Delta t \middle| T \ge t\right\}}{\Delta t}$$

which can be shown to equal

$$(1.10) \qquad h_T\left(t\right) = \frac{f_T\left(t\right)}{S_T\left(t\right)} = \frac{f_T(t)}{1 - F_T(t)}$$

The hazard rate gives the instantaneous probability of death at time t given that the person has survived up until time t.

It can be shown that the hazard rate also completely characterizes the distribution of the random variable T. For a continuous distribution it can be shown that the following relationships hold

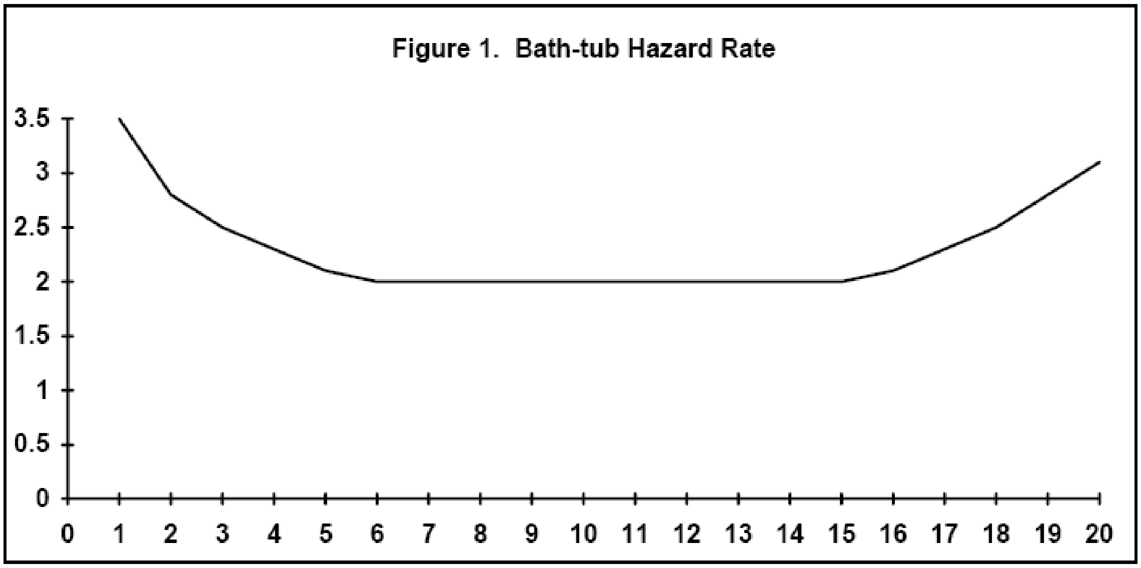
$$(1.11) \qquad h_T\left(t\right) = -\frac{d}{dt}\log\left(S_T\left(t\right)\right)$$

$$(1.12) \qquad S_T\left(t\right) = \exp\left\{-\int_0^t h_T\left(x\right)dx\right\}$$

$$(1.13) \qquad f_T\left(t\right) = h_T\left(t\right)\exp\left\{-\int_0^t h_T\left(x\right)dx\right\} = h_T(t)S_T(t)$$

(1.14) $$h_T(t) \geq 0 \qquad \int_0^\infty h_T(t)dt = \infty$$

The hazard rate carries a great deal of information about the underlying distribution, that is particularly suited to life distributions. Several shapes of the hazard rate have taken on particular names such as the *bath-tub hazard rate* pictured in Figure 1 below



The characteristic feature of the bath-tub hazard rate is that there is relatively high rate of death early, then levels out, and finally the hazard rate increases again for longer time periods. This is indicative of human populations where the death rate for infants is relatively high until about age 1, then the hazard rate is relatively flat until about age thirty when the hazard rate increases again.

2. Life Tables and Estimation of the Survivor Function

In this section we look at estimating the survivor function discussed in the previous section. We begin by looking at the method of life tables which historically is the oldest of the methods of estimating the survivor function. We then look at the Product limit estimator, also called the Kaplan-Meier estimator.

Life Table Methods

To introduce the method of life tables we begin with the following *typical* set of data

Table 1. Typical data for Life Table Analysis

Time Interval (months)	Number Under Observation at Start of Interval	Number that Died During Interval
$[0,2)$	300	195
$[2,4)$	105 (= 300-195)	27
$[4,6)$	78	15
$[6,8)$	63	12
$[8,10)$	51	9
$[10,12)$	42	12
$[12,14)$	30	

A key aspect of Table 1 is the *time* is broken up into discrete intervals, in this case units of two months. For each time period we are given how many people are alive at the beginning of the interval, and how many people died during the interval. In order to make further progress we need to introduce the following notation

(2.1) x time at beginning of interval
O_x number under observation at exact time x
n length of interval
${}_nd_x$ number dying in interval x to $x+n$
${}_np_x$ probability of surviving from time x to time $x+n$
${}_nq_x$ probability of dying during time x to time $x+n$

Finally we will denote with a capital P the probability

(2.2) ${}_nP_x$ probability of surviving from time x to time $x+n$

when we are interested in periods longer than a single interval. As an example of the notation ${}_{12}P_0$ would denote the probability of surviving one year (i.e. twelve months). IN terms of the notation above we can write

(2.3) $${}_{12}P_0 = \prod_{i=0}^{5} {}_{2}p_{2i} = ({}_{2}p_0)({}_{2}p_2)({}_{2}p_4)({}_{2}p_6)({}_{2}p_8)({}_{2}p_{10})$$

which says that the probability of living twelve months is the same as the probability of surviving the first two months, times the probability of living the next two moths, and so on. The method of life table analysis is based on the basic formula in (2.3), and the primary concern is on estimating the individual probabilities ${}_{n}p_x$.

The estimate we use for the individual probability ${}_{n}p_x$ is just

(2.4) $${}_{n}\hat{p}_x = 1 - \frac{{}_{n}d_x}{O_x}$$

which is just one minus the proportion of those that died in the interval as a proportion of those who were alive at the start of the interval. Note that we can also write

(2.5) $${}_{n}\hat{q}_x = \frac{{}_{n}d_x}{O_x}$$

Using this notation we can write Table 2.

Table 2. Data from Table 1 in terms of Life Table Notation

			Probability		
			Dying in interval	Surviving given alive at start of interval	Surviving to end of interval given alive and t=0
x	O_x	${}_2d_x$	${}_2\hat{q}_x$	${}_2\hat{p}_x$	${}_{x+2}\hat{P}_0$
0	300	195	0.65	0.35	0.35
2	105	27	0.26	0.74	0.26
4	78	15	0.19	0.81	0.21
6	63	12	0.19	0.81	0.17
8	51	9	0.18	0.82	0.14
10	42	12	0.29	0.71	0.10
12	30				

Interestingly we could also have written

$$(2.6)\qquad {}_{12}P_0 = \frac{O_{12}}{O_0} = \frac{30}{300} = 0.10$$

since

$$(2.7)\qquad O_{x+2} = O_x - {}_2d_x$$

so

$$(2.8)\qquad {}_{12}P_0 = \left(1-\frac{{}_2d_0}{O_0}\right)\left(1-\frac{{}_2d_2}{O_2}\right)\left(1-\frac{{}_2d_4}{O_4}\right)\left(1-\frac{{}_2d_6}{O_6}\right)\left(1-\frac{{}_2d_8}{O_8}\right)\left(1-\frac{{}_2d_{10}}{O_{10}}\right)$$

$$= \left(\frac{O_0-{}_2d_0}{O_0}\right)\left(\frac{O_2-{}_2d_2}{O_2}\right)\left(\frac{O_4-{}_2d_4}{O_4}\right)\left(\frac{O_6-{}_2d_6}{O_6}\right)\left(\frac{O_8-{}_2d_8}{O_8}\right)\left(\frac{O_{10}-{}_2d_{10}}{O_{10}}\right)$$

$$= \left(\frac{O_2}{O_0}\right)\left(\frac{O_4}{O_2}\right)\left(\frac{O_6}{O_4}\right)\left(\frac{O_8}{O_6}\right)\left(\frac{O_{10}}{O_8}\right)\left(\frac{O_{12}}{O_{10}}\right) = \frac{O_{12}}{O_0}$$

which looks like a fairly obvious estimator, and less complicated than the product form in (2.3). The real reason that the product form in (2.3) is important is because of *withdrawals. Withdrawals* are individuals which leave the data set (during an interval) without dying before they leave. Withdrawals are a real applied aspect of survival data.

They can occur for a number of reasons. For example we could be following a group of individuals over time, and some people may move or otherwise are unable to continue with the study. Another example is the following. Say we started a study of survival from breast cancer surgery, beginning on January 1, 1980, and we would end the study on January 1, 1990. For each woman who had surgery during the period 1/1/80 to 1/1/90 we would be able to observe whether they died before January 1, 1990. Notice that the form of the study will artificially *force* withdrawals. For example, a woman who had surgery on January 1, 1985 would be able to be followed for a maximum of 5 years.

To handle withdrawals we need to introduce some new notation

(2.9) $\quad {}_n w_x \quad$ number of withdrawals from time x to time $x+n$

The problem with withdrawals in intervals is that we lose information. For example, we do not know when during the interval the individual withdrew. So for example if we knew that all of the withdrawals occurred just before the end of the interval, this tells us that for most of the interval the withdrawals were alive, while if the withdrawals occurred at the beginning then we have no information about the withdrawals during the intervals. Based on these two scenarios we could have two possible estimates of ${}_2q_x$ namely

(2.10)

$${}_2\hat{q}_x \approx \frac{{}_2d_x}{O_x} \qquad \text{All withdrawals at end of interval}$$

$${}_2\hat{q}_x \approx \frac{{}_2d_x}{\underbrace{O_x - {}_2w_x}_{\text{At risk}}} \qquad \text{All withdrawals at beginning of interval}$$

A third way to estimate ${}_2q_x$ is the following

(2.11)

$${}_2\hat{q}_x = \frac{{}_2d_x + {}_2\hat{d}_x}{O_x} = \frac{(\text{deaths among continuing}) + (\text{deaths among withdrawals})}{(\text{at risk at beginning})}$$

where ${}_2\hat{d}_x$ is an estimate of the number of deaths in the interval among the withdrawals. If we assume that withdrawals are uniformly distributed across the interval then a reasonable estimate of ${}_2\hat{d}_x$ is

$$(2.12) \qquad {}_2\hat{d}_x = {}_2w_x\left(\frac{1}{2}\right){}_2\hat{q}_x = (\text{number of withdrawals})\,\Pr(\text{death in } x, x+2)\tfrac{1}{2}$$

Combining this with (2.11) we obtain

$$(2.13) \qquad {}_2\hat{q}_x = \frac{{}_2d_x + {}_2w_x\left(\frac{1}{2}\right){}_2\hat{q}_x}{O_x}$$

which solving for ${}_2\hat{q}_x$ yields

$$(2.14) \qquad {}_2\hat{q}_x = \frac{{}_2d_x}{O_x - {}_2w_x/2}$$

which is called the *standard life-table estimate.* The number at risk is (number at x)-($\frac{1}{2}$ of withdrawals). The thinking is withdrawals are not subjected to the same risks as continuing cases. In addition we call

$$(2.15) \qquad O'_x = O_x - {}_2w_x/2$$

the adjusted O_x, and we write

$$(2.16) \qquad {}_2\hat{q}_x = \frac{{}_2d_x}{O'_x}$$

As an example of these concepts consider the data in Table 3 below

Table 3. Illustration of Standard Life Table Estimates with Withdrawals

					Probability		
x	O_x	${}_2d_x$	${}_2w_x$	O'_x	${}_2\hat{q}_x$	${}_2\hat{p}_x$	${}_{x+2}\hat{P}_0$
0	300	193	6	297	0.65	0.35	0.35
2	101	25	8	97	0.26	0.74	0.26
4	68	12	10	63	0.19	0.81	0.21
6	46	8	10	41	0.20	0.80	0.17
8	28	4	10	23	0.17	0.83	0.14
10	14	3	10	9	0.33	0.67	0.09
12	1						

In the example above we can estimate the variance of ${}_{x+2}\hat{P}_0$ by

(2.17) $$\hat{V} = \left\{ {}_{x+2}\hat{P}_0 \right\} = {}_{x+2}\hat{P}_0^2 \sum_{i=1}^{x/2} \frac{{}_2\hat{q}_x}{O'_i - {}_2d_i}$$

This estimates the variance of the approximate distribution of ${}_{x+2}\hat{P}_0$ in large samples. It can be used to form approximate 95% confidence intervals for the true survivor function.

We have been dealing with the particular examples in Tables 1-3 which had data in two month intervals. A slightly more general setting with slightly more general notation is to define the intervals $I_1, \ldots, I_K$ as

(2.18) $$\begin{aligned} I_1 &= [a_0, a_1) \\ I_2 &= [a_1, a_2) \\ &\vdots \\ I_K &= [a_{K-1}, a_K) \end{aligned}$$

and

(2.19)

O_i number of observations at the start of I_i

d_i number of deaths in interval I_i

w_i number of withdrawals in interval I_i

then

(2.20) $$\hat{p}_i = 1 - \frac{d_i}{O_i - w_i / 2}$$

$$\hat{P}_I = \prod_{i=1}^{I} \hat{p}_i$$

$$\hat{V}\left\{\hat{P}_I\right\} = \hat{P}_I^2 \sum_{i=1}^{I} \frac{\hat{q}_i}{O_i' - d_i}$$

The variance estimator in (2.20) is commonly called Greenwood's formula.

Kaplan-Meier Estimate

The Kaplan-Meier, or product limit, estimator is the nonparametric maximum likelihood estimator of the survivor function when the exact times of death and withdrawal are known. To compute the Kaplan-Meier estimator we begin by ordering the survival time from smallest to largest, including the times of withdrawals. In the case when deaths occur at the same time as withdrawals, list the withdrawals after the deaths in the list. No define the notation

(2.21) $i =$ specific survival time
$O_i =$ number at risk at time i
$d_i =$ number of deaths at time i
${}_nP_0 =$ survival function from time 0 through time n

The number at risk are those who have not withdrawn or died as of time i. Then the Kaplan-Meier estimator of ${}_nP_0$ is given by

(2.22) $${}_n\hat{P}_0 = \prod_{i \le n} \left(1 - \frac{d_i}{O_i} \right)$$

The variance estimator for ${}_n\hat{P}_0$ is given by

(2.23) $$\hat{V}\left\{ {}_n\hat{P}_0 \right\} = {}_n\hat{P}_0^2 \sum_{i=1}^{n} \frac{d_i}{O_i \left(O_i - d_i \right)}$$

where this is an estimator of the variance of the approximate distribution in large samples.

We present the next example to illustrate the Kaplan Meier estimator.

Example 2. Consider the following data from a hypothetical study

Data for Example 2 (* indicates withdrawal)

			Interval Probability		Survival
i	O_i	d_i	$\hat{q}_i$	$\hat{p}_i$	${}_iP_0$
2	10	1	1/10	9/10	0.900
4*	9	0	0	1	0.900
5	8	1	1/8	7/8	0.788
6	7	1	1/7	6/7	0.675
9	6	2	2/6	4/6	0.450
12	4	1	1/4	3/4	0.338
12*	3	0	0	1	0.338
15*	2	0	0	1	0.338
17	1	1	1	0	0.000

Notice that $\hat{q}_i$ takes on distinct values when a death occurs, and survivor function does not change when a withdrawal takes place. Often data is presented with only the times of death reported, which in this case would be

Data for Example 2 Without Withdrawals

			Interval Probability		Survival
i	O_i	d_i	$\hat{q}_i$	$\hat{p}_i$	${}_iP_0$
2	10	1	1/10	9/10	0.900
5	8	1	1/8	7/8	0.788
6	7	1	1/7	6/7	0.675
9	6	2	2/6	4/6	0.450
12	4	1	1/4	3/4	0.338
17	1	1	1	0	0.000

Notice that the variance estimator for the survival function can be used with either form of the table, since in the first form of the table the *zero deaths* would contribute a 0 to the sum, and thus would give the same value when applied to the second table.

In the next example we demonstrate the estimation of two survival functions based on two groups.

Example 3. We take data from a clinical trial reported by Freireich, E.O. et al. (1963, *Blood*, **21**, 699-716) in which the drug 6-mercaptopurine (6-MP) was compared to a placebo with respect to the ability to maintain remission in acute leukemia patients. The table below gives the lengths of remission (in weeks) where the *starred* observations denote withdrawals.

Data for Example 3

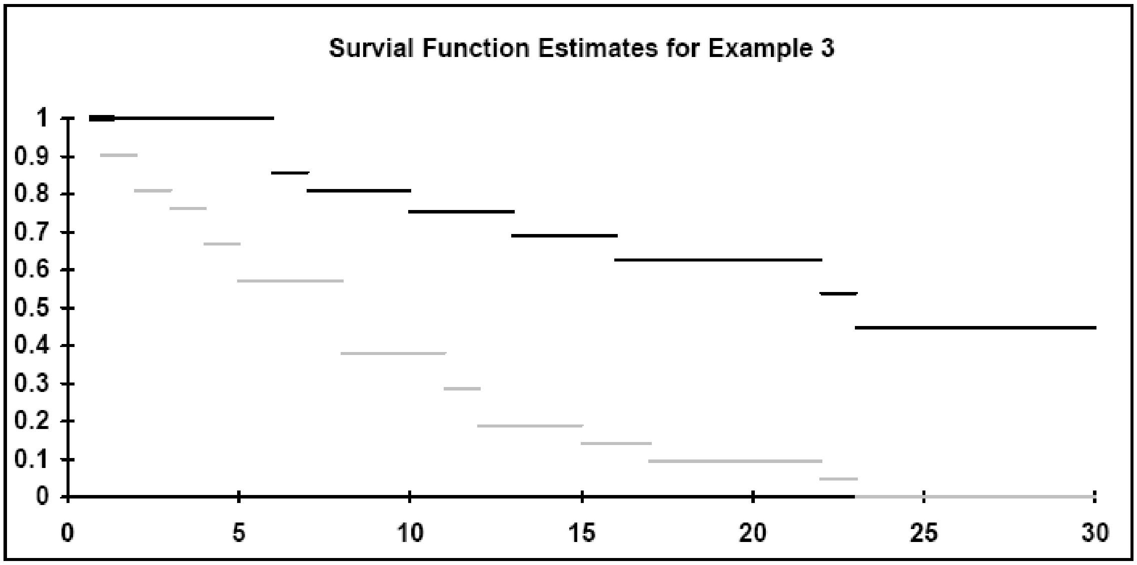
6-MP	6,6,6,6*,7,9*,10,10*,11*,13,16,17*,19* , 20*,22,23,25*,32*,32*,34*,35*
Placebo	1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17 ,22,23

Withdrawals are not considered as deaths. Below we give the estimation of the survival function

Estimation of Survival Function for Example 3

Drug 6-MP				Placebo			
Time	O_i	d_i	${}_i\hat{P}_0$	Time	O_i	d_i	${}_i\hat{P}_0$
6	21	3	0.857	1	21	2	0.905
7	17	1	0.807	2	19	2	0.810
10	15	1	0.753	3	17	1	0.762
13	12	1	0.690	4	16	2	0.667
16	11	1	0.627	5	14	2	0.571
22	7	1	0.538	8	12	4	0.381
23	6	1	0.448	11	8	2	0.286
				12	6	2	0.190
				15	4	1	0.143
				17	3	1	0.095
				22	2	1	0.048
				23	1	1	0.000

We have graphed the two functions below where the survival function of Drug 6-MP is graphed as the dark line and the placebo is graphed as the lighter line.



It is clear from the graph that Drug 6-MP increases the survival rate over the placebo. In the next section we discuss how to compare the survival functions.

We next show how to perform this analysis in SAS. We used the following code below

```
data set1;
  input weeks group censored;
  cards;
    6  1  0
    6  1  0
    6  1  0
    6  1  1
    7  1  0
    9  1  1
   10  1  0
   10  1  1
   11  1  1
   13  1  0
   16  1  0
   17  1  1
   19  1  1
   20  1  1
   22  1  0
   23  1  0
```

```
     25 1 1
     32 1 1
     32 1 1
     34 1 1
     35 1 1
      1 2 0
      1 2 0
      2 2 0
      2 2 0
      3 2 0
      4 2 0

      4 2 0
      5 2 0
      5 2 0
      8 2 0
      8 2 0
      8 2 0
      8 2 0
     11 2 0
     11 2 0
     12 2 0
     12 2 0
     15 2 0
     17 2 0
     22 2 0
     23 2 0
      ;
Proc lifetest plots=(s);
       Time weeks*censored(1);
       Strata group;

 run;
```

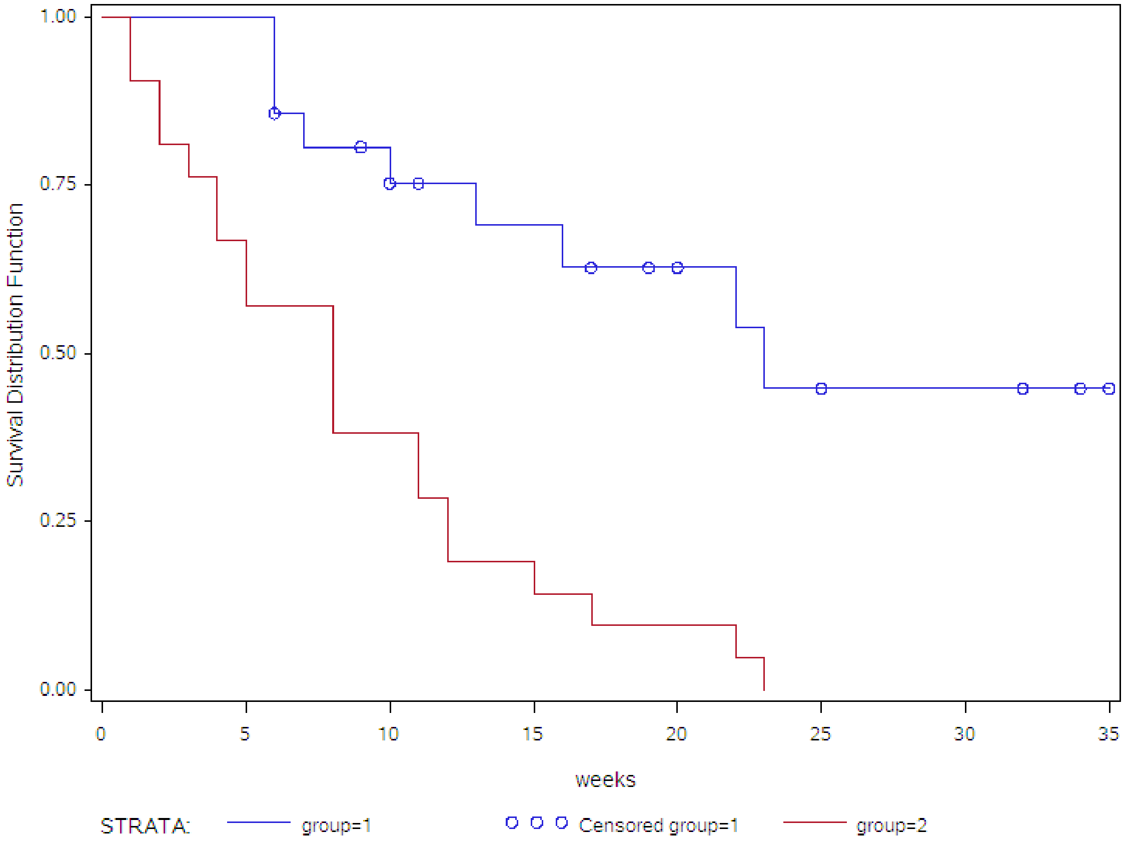
The procedure PROC LIFETEST estimates the survival function with the Kaplan-Meier estimator. The variable CENSORED tells the program which variables have been withdrawn, and in this the TIME statement with CENSORED(1) indicated that those observations with the value CENSORED=1 were withdrawn. The STRATA statement allows us to compute the survival functions for the two groups. The output from the code is given below

```
                               The LIFETEST Procedure

                          Product-Limit Survival Estimates
                                     GROUP = 1

                                                   Survival
                                                   Standard     Number      Number
              weeks     Survival     Failure        Error       Failed       Left
```

```
             0.0000        1.0000           0           0         0          21
             6.0000             .           .           .         1          20
             6.0000             .           .           .         2          19
             6.0000        0.8571      0.1429      0.0764         3          18
             6.0000*            .           .           .         3          17
             7.0000        0.8067      0.1933      0.0869         4          16
             9.0000*            .           .           .         4          15
            10.0000        0.7529      0.2471      0.0963         5          14
            10.0000*            .           .           .         5          13
            11.0000*            .           .           .         5          12
            13.0000        0.6902      0.3098      0.1068         6          11
            16.0000        0.6275      0.3725      0.1141         7          10
            17.0000*            .           .           .         7           9
            19.0000*            .           .           .         7           8
            20.0000*            .           .           .         7           7
            22.0000        0.5378      0.4622      0.1282         8           6
            23.0000        0.4482      0.5518      0.1346         9           5
            25.0000*            .           .           .         9           4
            32.0000*            .           .           .         9           3
            32.0000*            .           .           .         9           2
            34.0000*            .           .           .         9           1
            35.0000*            .           .           .         9           0

                NOTE: The marked survival times are censored observations.


                        Summary Statistics for Time Variable weeks

                                    Quartile Estimates

                              Point         95% Confidence Interval
                 Percent    Estimate    Transform      [Lower      Upper)

                      75        .       LOGLOG        23.0000        .
                      50     23.0000    LOGLOG        13.0000        .
                      25     13.0000    LOGLOG         6.0000     22.0000


                                    Mean    Standard Error

                                 17.9092           1.6474

NOTE: The mean survival time and its standard error were underestimated because the
                                       largest




                                  The LIFETEST Procedure

                                   Stratum 2: group = 2

                             Product-Limit Survival Estimates
```

```
                                    Survival
                                    Standard     Number      Number
  weeks     Survival    Failure      Error       Failed       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           .       21           0

             Summary Statistics for Time Variable weeks

                         Quartile Estimates

                   Point          95% Confidence Interval
      Percent    Estimate    Transform      [Lower      Upper)

           75     12.0000    LOGLOG         8.0000     22.0000
           50      8.0000    LOGLOG         4.0000     11.0000
           25      4.0000    LOGLOG         1.0000      5.0000

                          Mean    Standard Error

                        8.6667            1.4114


      Summary of the Number of Censored and Uncensored Values

                                                            Percent
Stratum           group       Total  Failed    Censored    Censored

      1               1          21       9          12       57.14
      2               2          21      21           0        0.00
-------------------------------------------------------------------
  Total                          42      30          12       28.57
```

```
                    The LIFETEST Procedure

                 Survival Function Estimates
```



There is additional SAS output that we did not include which compares the two survivor functions. We do not have time to discuss those particular statistics.

3. Cox Proportional Hazards Regression

Up to now we have been discussing the estimation of survivor function for a random variable T, and have based inference about the survivor function based on an independent identically distributed sample. Most of the time the individuals in our sample are different in some measurable way, such as gender, age, health status, and we usually want to take that into account in our analysis. In the previous class we introduced the concept of incidence rates, and called the ratio of two incidence rates as the relative risk of one state compared to another. The hazard rate we defined earlier is analogous to an incidence rate, which is continuous in time. It is natural to think of modeling the hazard function as dependent on a set of covariates which we are able to measure. If we collect the covariates in a $p\times 1$ dimensional vector $\mathbf{X}$, then following Cox (1972) we write the hazard function as

$$(3.1) \qquad h(t|\mathbf{X}) = h_0(t)\exp(\mathbf{X}'\beta) \quad \text{or} \quad \ln h(t|\mathbf{X}) = \ln h_0(t) + \mathbf{X}'\beta$$

where $h_0(t)$ is called the *baseline hazard function.* The baseline hazard function serves as a theoretical point of reference, and can be viewed as an overall hazard rate, or overall time dependent incidence rate. The model in (3.1) is called the *Cox Proportional Hazards Model.* We can also write (3.1) as

$$(3.2) \qquad \ln\left(\frac{h(t|\mathbf{X})}{h_0(t)}\right) = \mathbf{X}'\beta$$

where we can view the left-hand side of (3.2) as the log of a *time dependent relative risk.* Therefore the time dependent log relative risk is modeled as a linear regression function of the covariates. Remember that we can write

$$(3.2) \qquad S_0(t) = \exp\left(-\int_0^t h_0(u)\,du\right)$$

$$S(t|\mathbf{X}) = \exp\left(-\int_0^t h(u|\mathbf{X})\,du\right)$$

$$= \exp\left(-\int_0^t h_0(u)\exp(\mathbf{X}'\beta)\,du\right)$$

$$= \left[S_0(t)\right]^{\exp(\mathbf{X}'\beta)}$$

$S(t \mid \mathbf{X})$ is like sum of failure probabilities in many small intervals. $S(t \mid \mathbf{X})$ says that the survival function which depends on covariates, can be written as an exponentiated version on the *baseline survivor function.* Note that in general there is no specific restrictions placed on the baseline survivor function, such as belonging to a particular parametric family of distributions.

In the Cox proportional hazards approach there are usually two objectives. The first is to estimate the regression coefficients β for the purposes of determining the effect of the covariates on the survival function. The second objective is to estimate the baseline survival function. A strict likelihood function can not be written down for a general random sample of observations, but Cox proposed that the following function be maximized as a function of β

$$(3.3) \qquad L(\beta) = \prod_{i=1}^{k} \left(\frac{e^{\mathbf{X}'_{(i)}\beta}}{\sum_{j \in R_i} e^{\mathbf{X}'_j \beta}} \right)$$

where we assume that there are k observed deaths at times $t_1, t_2, ..., t_k$ with associated covariates $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, ..., \mathbf{X}_{(k)}$, and those individuals who have not died or withdrawn by time t_i are place in the *Risk Set* R_i. The heuristic justification of (3.3) is that if it is known that a individual is to die at time t_i, then the probability that it is individual i out all the possible individuals that could die (i.e. the Risk Set R_i) should be

$$(3.4) \qquad \frac{h\left(t \middle| \mathbf{X}_{(i)}\right)}{\sum_{j \in R_i} h\left(t \middle| \mathbf{X}_j\right)} = \frac{e^{\mathbf{X}'_{(i)}\beta}}{\sum_{j \in R_i} e^{\mathbf{X}'_j \beta}}$$

It can be shown that the estimator of β obtained by maximizing the function in (3.3) is consistent for the true regression coefficients and asymptotically normally distributed.

The estimator of the baseline survivor function is more difficult to explain. It turns out that it can be estimated by a nonparametric estimator similar in form to the

Kaplan-Meier estimator. We will not present it here. SAS has a procedure PROC PHREG which estimates the Cox proportional hazards model.