Biost 517 / Biost 514
Applied Biostatistics I / Biostatistics I

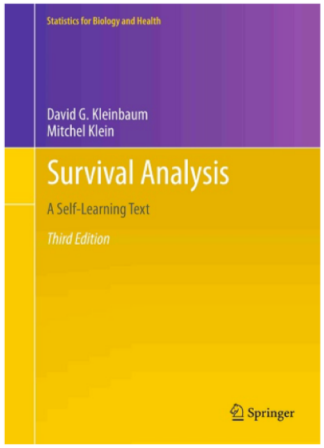
Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

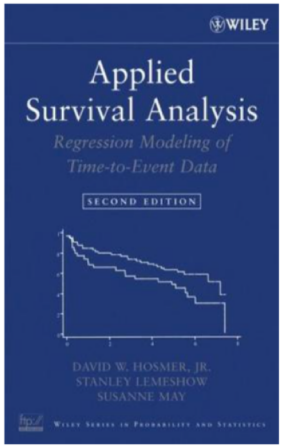
Lecture 20

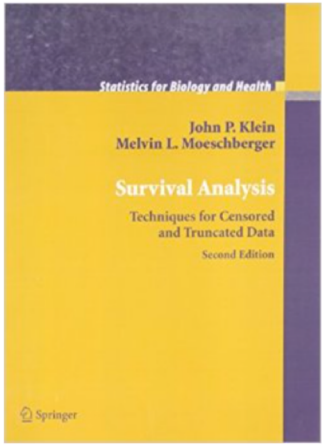
Survival Analysis: Introduction; Nonparametric Analysis; Kaplan-Meier Estimates of Survival; Logrank Test of Equal Survival

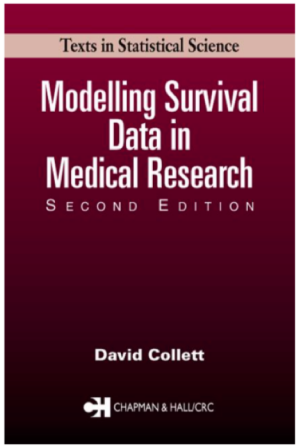
Survival Analysis

- BIOST 514/517's material on survival analysis is not intended to be comprehensive!
- Regression approaches for survival analysis are seen, briefly, in Biost 515/518
- For a book-length look at this topic, the following are recommended:









- Biostat 537 course on survival analysis is high recommended for a more comprehensive treatment of this material

Introduction to Survival Data

- In many biomedical studies the primary endpoint of interest is time to a certain event. Examples are
  - time to death
  - time it takes for a patient to respond to a therapy;
  - time from response until disease relapse (i.e., disease returns);
  - etc.
- For proper analysis of survival data, we must take into account the time until the event occurs, or until the end of follow-up if the event has not yet occurred during our observation period

Example: Leukemia Randomized Control Study

- To illustrate common characteristics of survival data, consider a randomized control study of 6-mercaptopurine (6-MP) as maintenance therapy for children in remission from acute lymphoblastic leukemia (Freireich et al. 1963.)
- 42 patients were in remission and were randomized in equal numbers to 6-MP or placebo add followed:

Weeks in remission among leukemia patients

Placebo	**1**	**1**	**2**	**2**	**3**	**…**	**12**	**15**	**17**	**22**	**23**
6-MP	**6**	**6**	**6**	**6+**	**3**	**4**	**…**	**19+**	**20+**	**32**	**34+**	**35+**

A "+" indicates that the exact time to relapse was not observed.

Example: Leukemia Randomized Control Study

- All 21 patients in the placebo group relapses, and only 9 of 21 patients in the 6-MP group relapsed.
- However, a crucial characteristic is that for the 12 patients in the 6-MP group, the exact time to relapse was unobserved. We only know that if a relapse occurred, the time was greater than the follow-up time.
- Should we treat these incompletely observed times as missing data? Should we simply throw them out?
- Very bad idea!
  - These observations contain valuable information:
    20+ = the participant did not experience a relapse before 20 weeks; only know that the actual relapse time is somewhere in the interval $(20,\infty)$

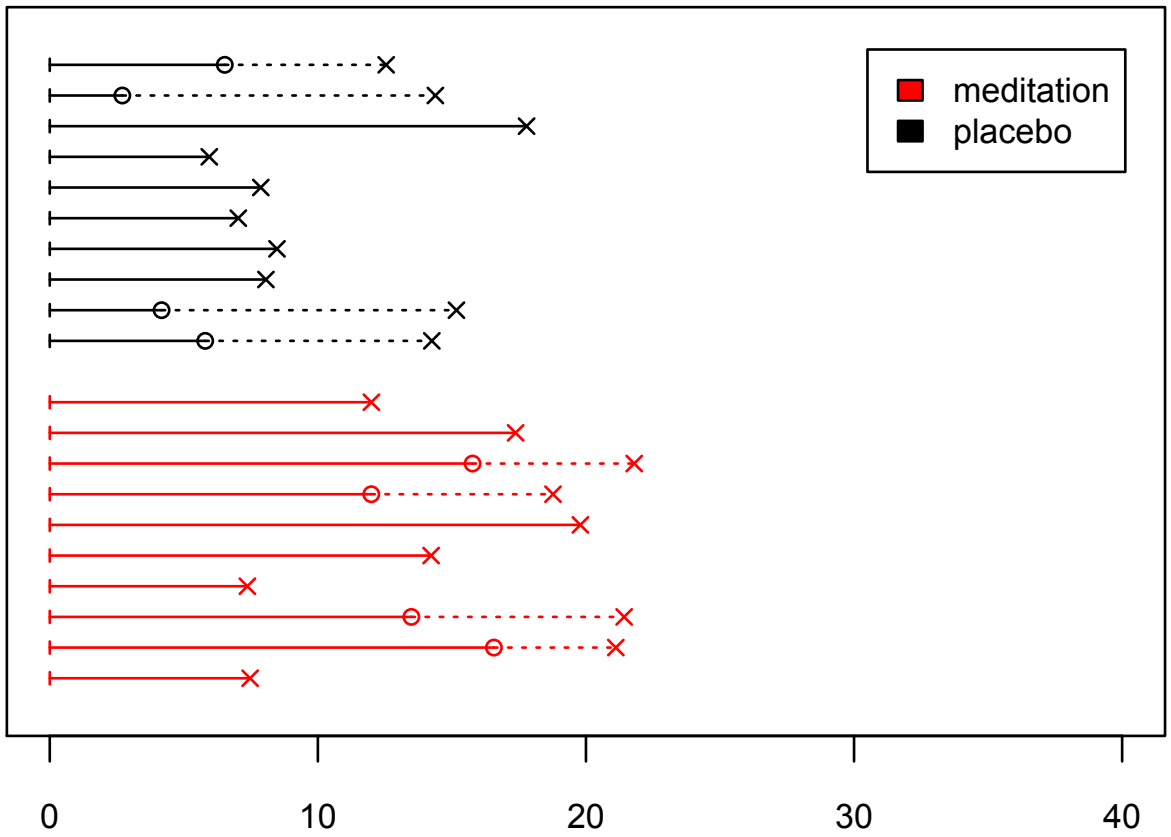
Example: Leukemia Randomized Control Study

- The censored participants are not representative of the whole study population.
- Which participants are more likely to be censored: those with smaller or larger times?
- Systematically excluding censored times would lead to biased estimates. For the leukemia data, would this lead to an overestimation or underestimation of the mean time until relapse?

Right Censored Data

- It is common to not have been able to exactly observe the time to event for all study participants.
- Why so? Most common reasons are:
  - the study ended (say, after 30 weeks) and some participants had not yet had an event (e.g., relapse); (administrative censoring)
  - the participant left the study before having a relapse; (loss to follow-up)
- These all lead to **right-censored data**.
- A special type of missing data: the exact value is not always known
  - Some measurements are known exactly
  - Some measurements are only known to exceed some specified value (perhaps different for each subject)

Right Censored Data



- Right-censored data: on a graph, the unobserved event time would lie somewhere to the **right** of the censoring time

Survival Analysis Data

- What is survival analysis?
  - it is the branch of statistics concerned with the analysis of time-to-event data;
- often, the goal of a survival analysis is to:
  - to describe the distribution of a time-to-event;
  - to compare the time-to-event distribution in different subpopulations; investigate the relationship between explanatory variables and the time-to-event distribution;
  - in epidemiology, we may be interested in estimating the distribution of lifetime, disease duration, age at onset and time until infection, for example;

Survival Analysis Key Terminology Concepts

- Suppose T is a continuous time-to-event random variable.
- Let f be the **density function** for T:

$$f(t) := \lim_{\Delta t \to 0} \frac{P\left(t \le T < t + \Delta t\right)}{\Delta t}$$

$f(t)\Delta t$ approximates the probability that T has a value in $[t, t + \Delta t)$;

- The **survival function** S is defined as:

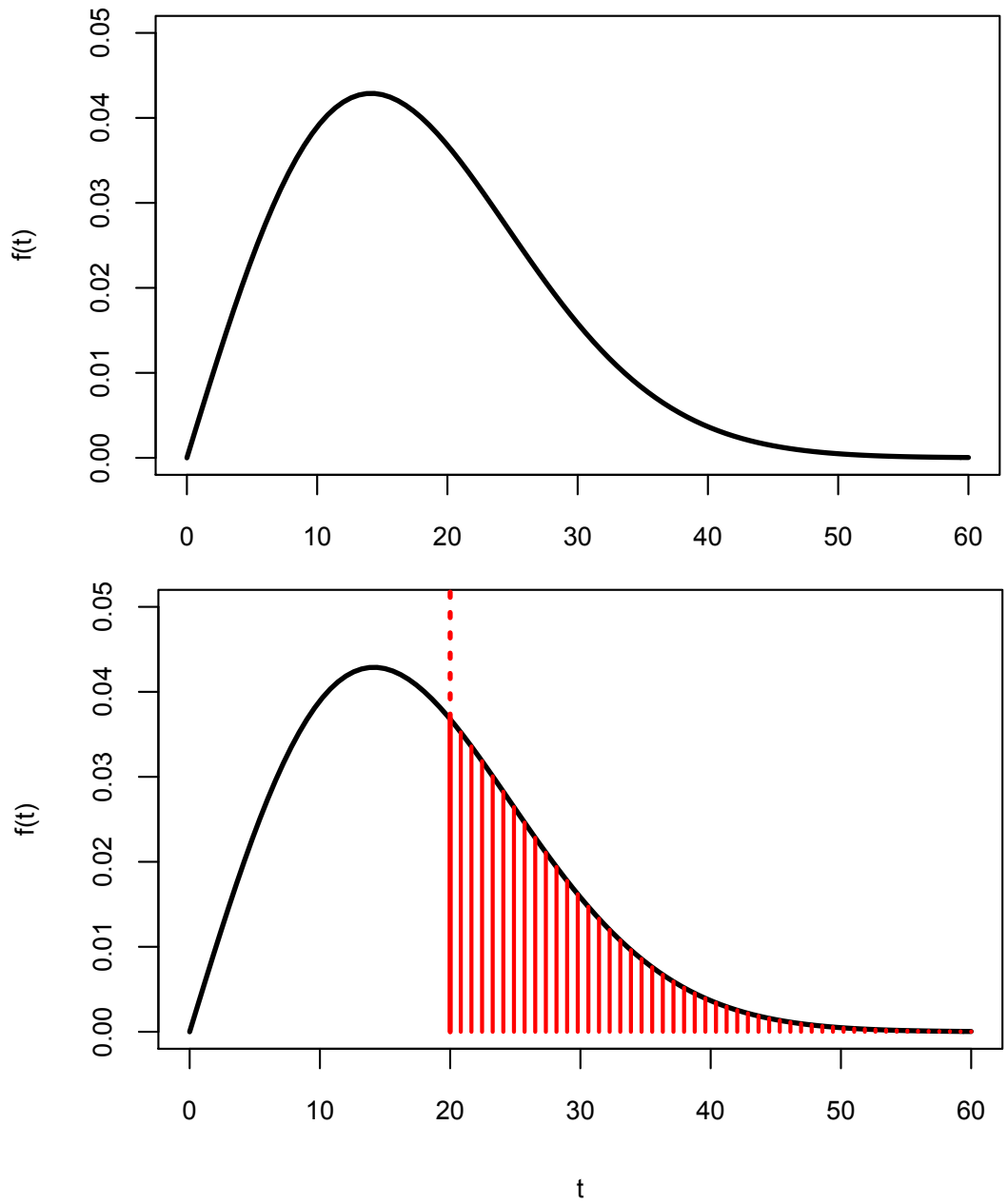
$$S(t) := P(T > t) = \int_t^\infty f(u)du$$

$S(t)$ is the proportion of the population with a time-to-event greater than t

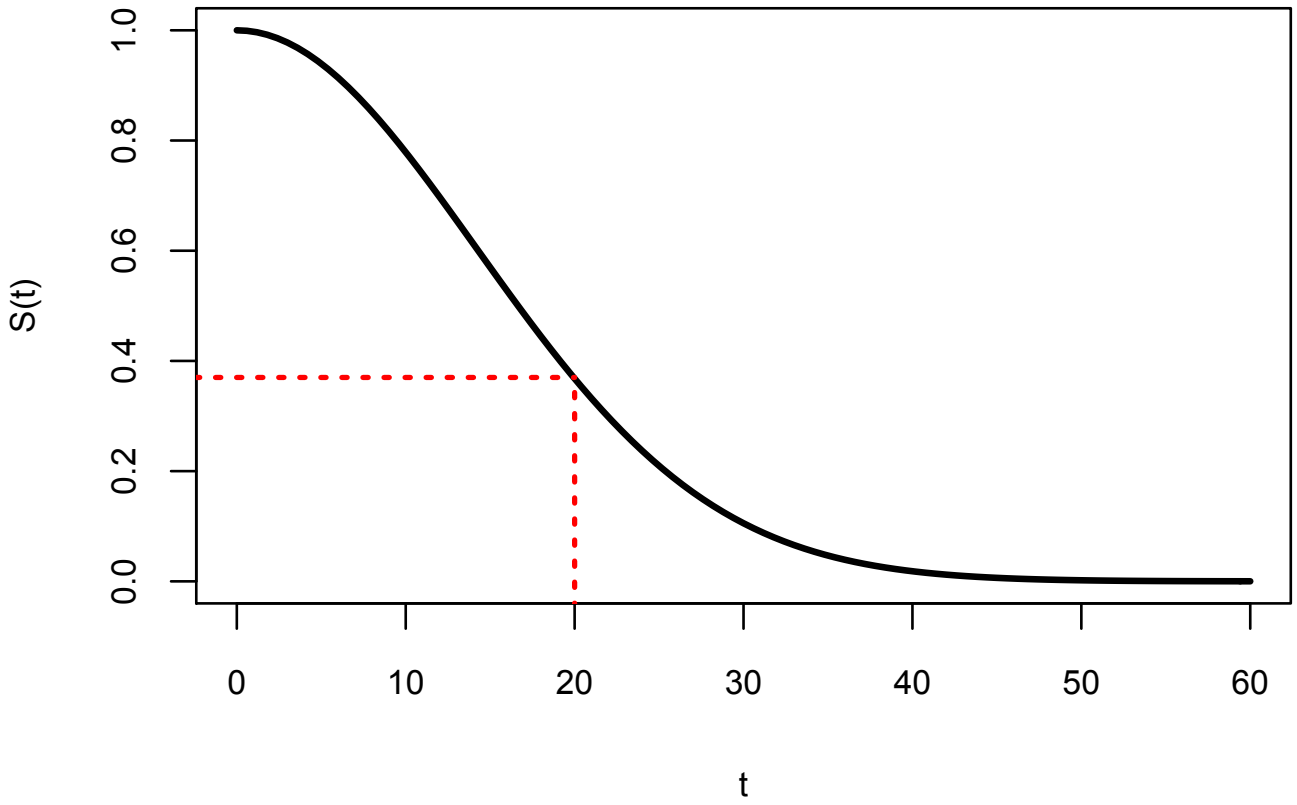
S is non-increasing. It starts at 1 (i.e., $S(0) = 1$) and ends at 0 (i.e., $S(\infty) = 0$)

- In the leukemia example, if S(20)=0.37, it follows that 37% of children with leukemia will not have a relapse within the first 20 weeks.

Example of a Time-to-Event Density Funtion



Survival Function Example



Survival Analysis Key Terminology

- The **hazard function** (also hazard rate, or failure rate) h is defined as:

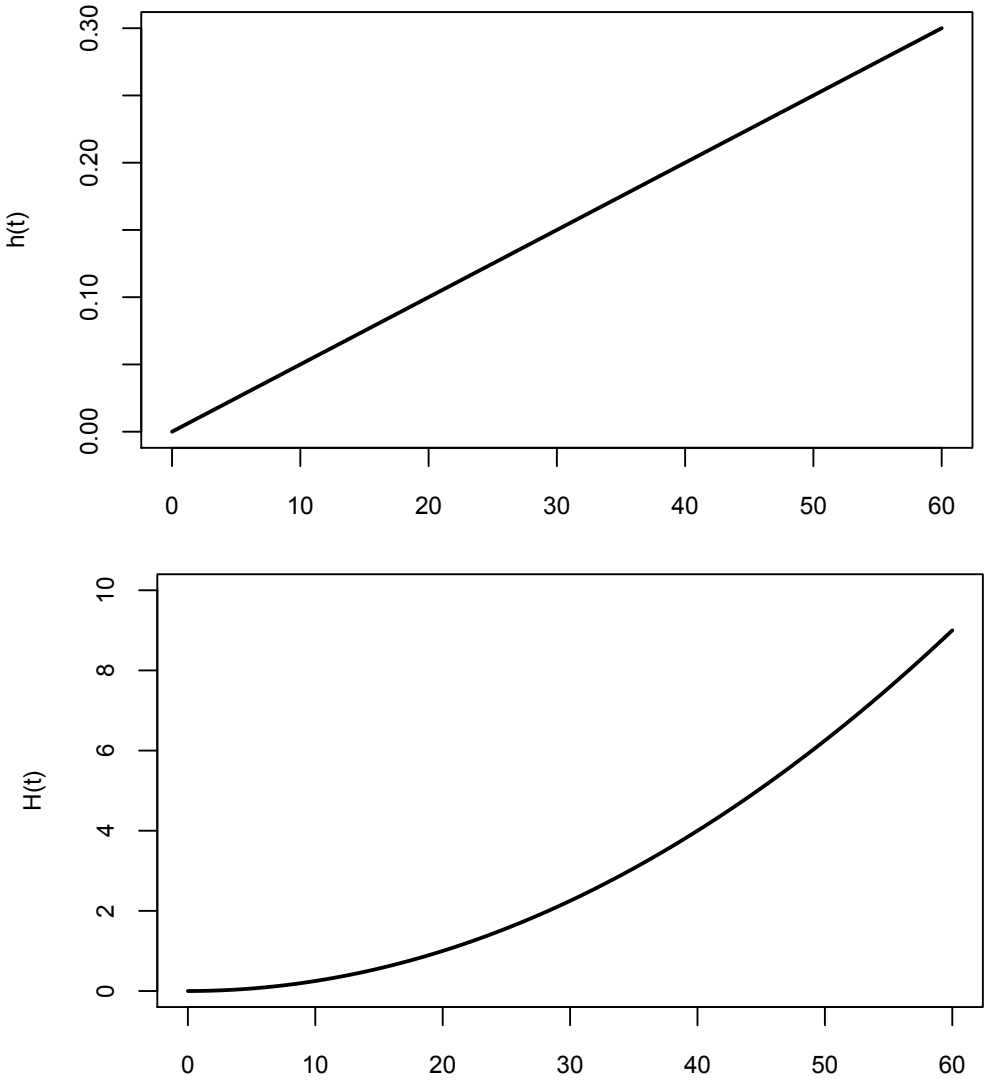
$$h(t) := \lim_{\Delta t \to 0} \frac{P\left(t \le T < t + \Delta t \mid T \ge t\right)}{\Delta t}$$

$h(t)$ is the instantaneous failure rate at t given survival until t.

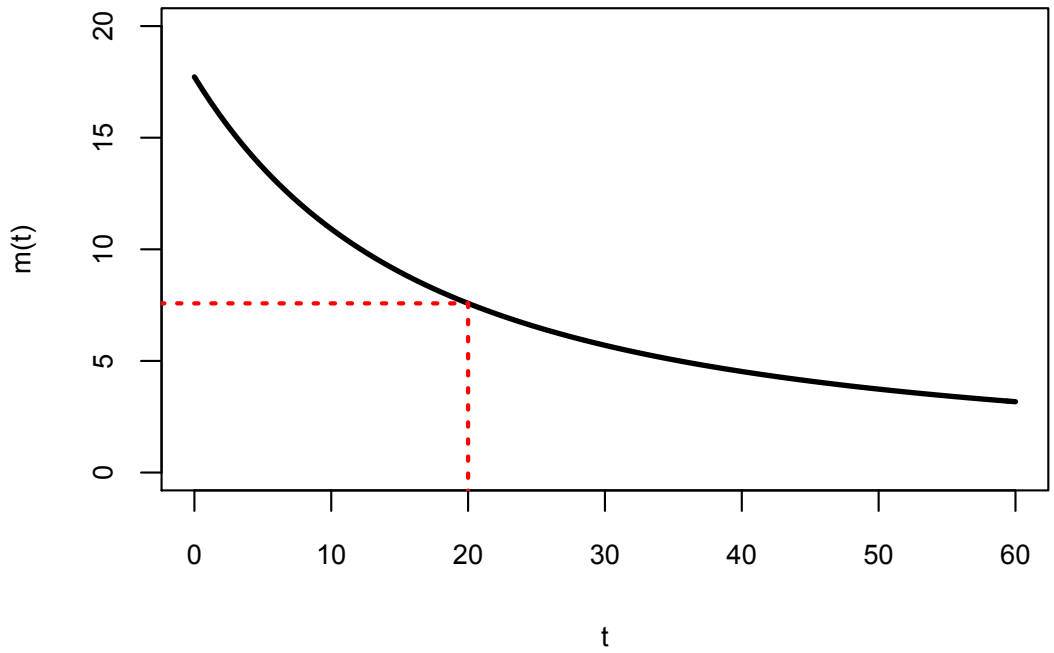
$h(t)\Delta t$ approximates the probability that T has a value in $[t, t + \Delta t)$ given $T \ge t$;

- The **cumulative hazard function** H is defined as: $H(t) := \int_0^t h(u)du$
- The **mean residual time** m is defined as: $m(t) := E(T - t \mid T \ge t)$
  - $m(t)$ is the average remaining time until the terminating event given the event has not occurred by time t.
  - In the leukemia example, if m(20)=7.2, then a person having not experienced a relapse in 20 weeks will experience it on average in another 7.2 weeks.

Examples of Hazard and Cumulative Hazard Functions



Example of Mean Residual Function



Survival Analysis Data

- The distribution of T can entirely be calculated using any of the functions $T, f, S, h, H,$ and m
- Most methods in survival analysis focus on modeling and/or estimating either the survival function or the hazard rate.
- Why the hazard rate?
  - its conditional interpretation may be useful in many epidemiological applications;
  - it can be easily estimated using right-censored data.

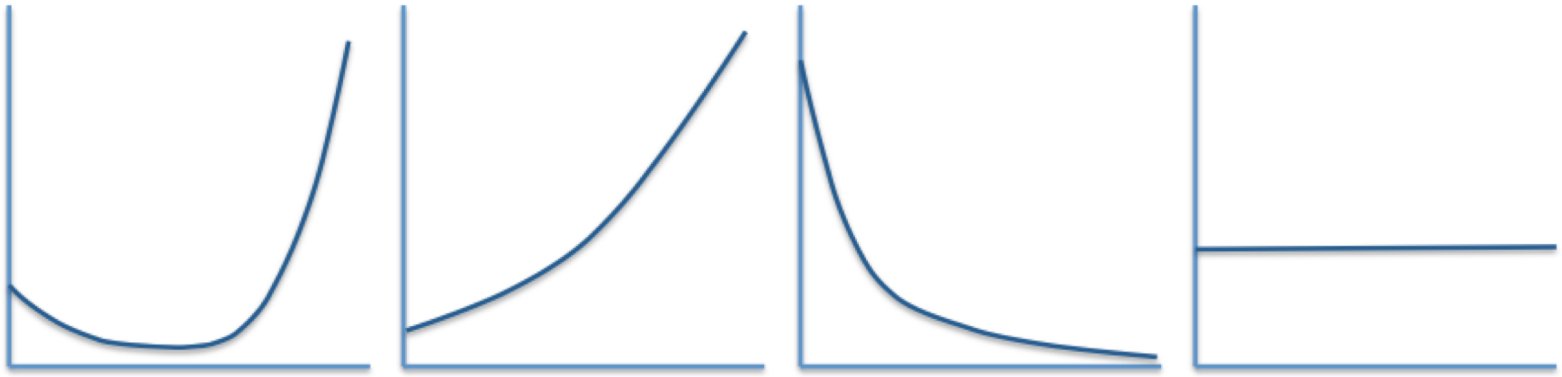
What shape do we expect the hazard function to have? Well, it depends...

(a) survival from surgery until death due to complications;

(b) survival from onset of a progressive disease;

(c) the time before a radioactive atom disintegrates.

(d) an individual's total lifetime;



Noninformative Censoring

- When estimating survivor functions using censored data, censoring must not be informative
  - Censored subjects neither more nor less likely to have an event in the immediate future
- Censored individuals must be a random sample of those at risk at time of censoring: missing at random (MAR) based on time of censoring
  - Missingness depends on time last observed
  - But random among all subjects at that time
- Methods are available to allow a random sample from all subjects at risk having similar modeled covariates:
- Missingness depends on time last observed and some other measured and modeled covariates

Informative Censoring Examples

- Subjects in a RCT are withdrawn due to treatment failure
  - (likely they would die sooner than those remaining)

- Subjects in a RCT in a fatal condition are lost to follow up when they go on vacation
  - (likely they are healthier than those remaining)

Nonparametric Approach to Survival Analysis

NONPARAMETRIC ESTIMATION FROM INCOMPLETE OBSERVATIONS*

E. L. KAPLAN
University of California Radiation Laboratory
AND
PAUL MEIER
University of Chicago

In lifetesting, medical follow-up, and other fields the observation of the time of occurrence of the event of interest (called a *death*) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a *loss*). Losses may be either accidental or controlled, the latter resulting from a decision to terminate certain observations. In either case it is usually assumed in this paper that the lifetime (age at death) is independent of the potential loss time; in practice this assumption deserves careful scrutiny. Despite the resulting incompleteness of the data, it is desired to estimate the proportion $P(t)$ of items in the population whose lifetimes would exceed t (in the absence of such losses), without making any assumption about the form of the function $P(t)$. The observation for each item of a suitable initial event, marking the beginning of its lifetime, is presupposed.

For random samples of size N the product-limit (PL) estimate can be defined as follows: List and label the N observed lifetimes (whether to death or loss) in order of increasing magnitude, so that one has $0 \leq t_1' \leq t_2' \leq \cdots \leq t_N'$. Then $\hat{P}(t) = \prod_r [(N-r)/(N-r+1)]$, where r assumes those values for which $t_r' \leq t$ and for which t_r' measures the time to death. This estimate is the distribution, unrestricted as to form, which maximizes the likelihood of the observations.

Other estimates that are discussed are the actuarial estimates (which are also products, but with the number of factors usually reduced by grouping); and reduced-sample (RS) estimates, which require that losses not be accidental, so that the limits of observation (potential loss times) are known even for those items whose deaths are observed. When no losses occur at ages less than t, the estimate of $P(t)$ in all cases reduces to the usual binomial estimate, namely, the observed proportion of survivors.

Notation for Kaplan-Meier Estimator of a Survivor Function

Unobserved :

True times to event : $\left\{T_1^0, T_2^0, \ldots, T_n^0\right\}$

Censoring Times : $\left\{C_1, C_2, \ldots, \mathrm{C}_n\right\}$

Observed data :

Observation Times : $T_i = \min\left(T_i^0, C_i\right)$

Event indicators : $\mathrm{D}_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$

Nonparametric Estimation of a Survivor Function: Kaplan-Meier

- Definition of intervals, number at risk, failures

$$\text{Ordered distinct observation times}:$$

$$t_1 \le t_2 \le \cdots \le t_k$$

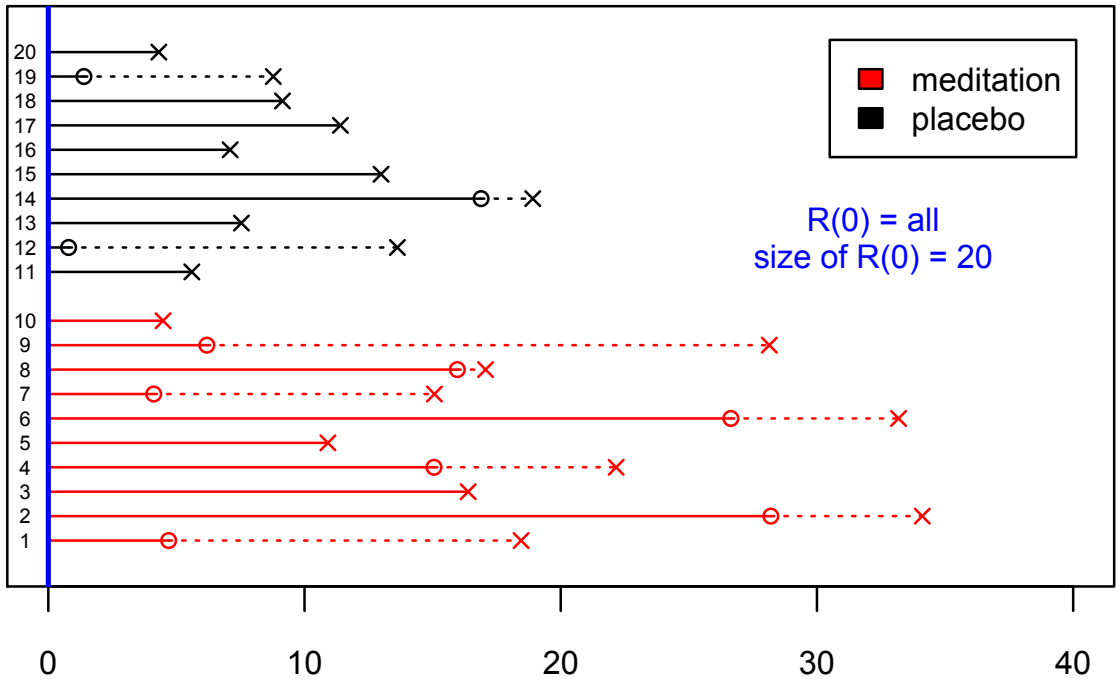
$$\text{Time interval}: \qquad (t_{j-1}, t_j]$$

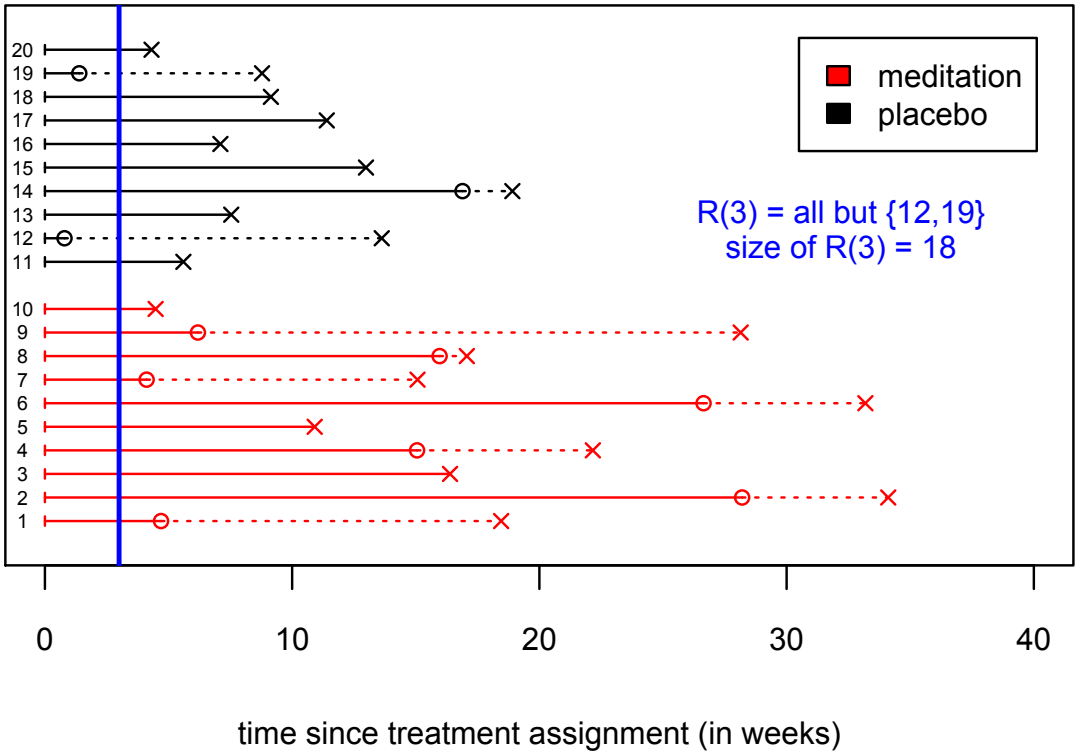
$$\text{Number at risk at } t_j: \qquad N_j$$

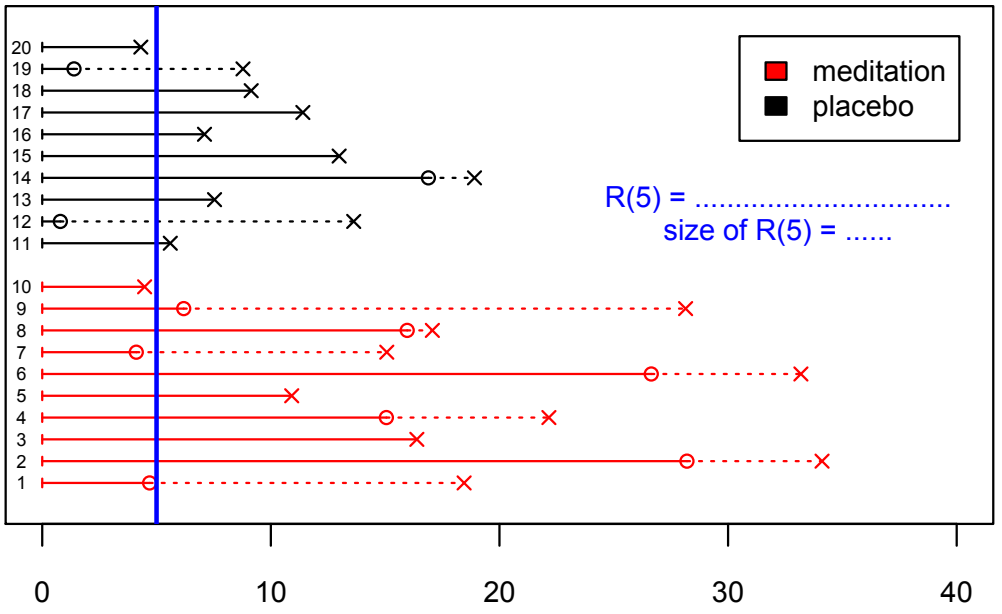
$$\text{Number of events at } t_j: \qquad D_j$$

- **Note:**
  - **Number at risk at t_j corresponds to the number who were at risk in the interval**
  - **Number of events at t_j corresponds to the number of events that occurred in the interval.**

- Let R(t) be the set of individuals who are at risk at time t




20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
meditation
placebo
R(3) = all but {12,19}
size of R(3) = 18
0
10
20
30
40
time since treatment assignment (in weeks)


20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
meditation
placebo
R(5) = ...
size of R(5) = ...
0
10
20
30
40
time since treatment assignment (in weeks)

Nonparametric Survivor Analysis: Kaplan-Meier Hazard Estimates

- Simple random sampling and non-informative censoring are assumed; event times for those censored, beyond their censoring time, are no different from others who survive to that time.
- Computation of hazard and conditional probability of survival in interval based on
  - number at risk at the beginning of the interval, and
  - number having an event during the interval

$$\text{Hazard for event in interval:} \qquad \frac{D_j}{N_j}$$

$$\text{Conditional probability of survival in interval:}$$

$$\Pr\left(T^0 \geq t_j \mid T^0 \geq t_{j-1}\right) = 1 - \frac{D_j}{N_j}$$

Kaplan-Meier Survival Estimate

- Estimating survival probability with noninformative censoring

$$S(t) = Pr\,(T^0 > t)$$

Cumulative probability of survival:

$$\Pr\left(T^0 > t_j\right) = \Pr\left(T^0 > t_j \mid T^0 > t_{j-1}\right)\Pr\left(T^0 > t_{j-1}\right)$$

$$\hat{S}\left(t_j\right) = \left(1 - \frac{D_j}{N_j}\right) \times \left(1 - \frac{D_{j-1}}{N_{j-1}}\right) \times \cdots \times \left(1 - \frac{D_1}{N_1}\right)$$

$$= \prod_{i=1}^{j}\left(1 - \frac{D_i}{N_i}\right)$$

Kaplan-Meier Survival Estimate

$S(t)$ is a probability; we can estimate the Std Error of $\hat{S}(t)$ by

$$\widehat{\text{StdErr}}[\,\hat{S}(t)\,] \;=\; \hat{S}(t)\sqrt{\sum_{j:t_j<t}\frac{D_j}{N_j-D_j}}, \quad \left(\begin{array}{c}\text{Greenwood's}\\ \text{Formula}\end{array}\right)$$

where D_j= #deaths at t_j, and N_j= #surviving to t_j. But this gives CIs that can go beyond [0,1], so the default instead uses

$$\widehat{\text{StdErr}}\left[\log\left(-\log(\hat{S}(t))\right)\right] \;=\; \frac{\sqrt{\sum_{j:t_j<t}\frac{D_j}{N_j-D_j}}}{\sum_{j:t_j<t}\log\left(\frac{N_j-D_j}{D_j}\right)}.$$

Writing $LLS = \log\left(-\log(\hat{S}(t))\right)$, i.e. $\hat{S}(t) = e^{-e^{LLS}}$, note that

$$\begin{aligned}
\text{95\% CI for } \log(-\log(S(t))) &= (LLS - 1.96\hat{se}, LLS + 1.96\hat{se})\\
\text{95\% CI for } -\log(S(t)) &= (e^{LLS-1.96\hat{se}}, e^{LLS+1.96\hat{se}})\\
\text{95\% CI for } S(t) &= (e^{-e^{LLS+1.96\hat{se}}}, e^{-e^{LLS-1.96\hat{se}}})\\
&= (\hat{S}(t)^{e^{1.96\hat{se}}}, \hat{S}(t)^{-e^{1.96\hat{se}}}),\ \text{within [0,1]}
\end{aligned}$$

Survival Analysis in R

- For survival analysis in R, we will let R know that we use the **"survival"** package
- A survival object must first be created with the ***Surv()*** function
- It takes two variables to specify a censored time-to event data.
- The following is required for the ***Surv()*** function for right censored data:
  - A specified "time" variable: this is the follow up time
  - A specified "event" variable: this is an indicator for the event being observed during the follow up time, where a 0 corresponds to a right censored event and a 1 corresponds to an event (at the observed time).

Survival Package in R

- Below are commands to create a survival object with the **Surv()** function for the leukemia randomized control study data

```
> leukdata=read.csv("./DataSets/leukdata.csv",header=T)
> leukdata$survobj <- with(leukdata,Surv(time = time, event = relapse))
> head(leukdata,15)
   time relapse group survobj
1     6       1  6-MP       6
2     6       1  6-MP       6
3     6       1  6-MP       6
4     7       1  6-MP       7
5    10       1  6-MP      10
6    13       1  6-MP      13
7    16       1  6-MP      16
8    22       1  6-MP      22
9    23       1  6-MP      23
10    6       0  6-MP      6+
11    9       0  6-MP      9+
12   10       0  6-MP     10+
13   11       0  6-MP     11+
14   17       0  6-MP     17+
15   19       0  6-MP     19+
```

Q. What does "+" indicate about the observations for the survobj?

Survival Package in R

- Kaplan-Meier estimates of the distribution of time to relapse for the leukemia data (available on Canvas) can be with the ***survfit()*** function.
- Using ***summary()*** of an object created using the ***survfit()*** function provides point estimates and inference for survival probabilities:
  - point estimates of survival probabilities
  - complementary log-log scale used to obtain standard errors of estimated survival probabilities and confidence intervals in R

Kaplan-Meier Estimates and CI

- Kaplan-Meier estimates and standard errors of the survival probabilities for time to relapse with the leukemia data (available on Canvas) can be obtained with the ***survfit()*** function
- The 1 in the formula `(survobj ~ 1)` indicates we are estimating the survival function for one group

```
> km1 <- survfit(survobj~1, data=leukdata)
> summary(km1)
Call: survfit(formula = survobj ~ 1, data = leukdata)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     42       2    0.952  0.0329        0.8901        1.000
    2     40       2    0.905  0.0453        0.8202        0.998
    3     38       1    0.881  0.0500        0.7883        0.985
    4     37       2    0.833  0.0575        0.7279        0.954
    5     35       2    0.786  0.0633        0.6709        0.920
    6     33       3    0.714  0.0697        0.5899        0.865
    7     29       1    0.690  0.0715        0.5628        0.845
    8     28       4    0.591  0.0764        0.4588        0.762
   10     23       1    0.565  0.0773        0.4325        0.739
   11     21       2    0.512  0.0788        0.3783        0.692
   12     18       2    0.455  0.0796        0.3227        0.641
   13     16       1    0.426  0.0795        0.2958        0.615
   15     15       1    0.398  0.0791        0.2694        0.588
   16     14       1    0.369  0.0784        0.2437        0.560
   17     13       1    0.341  0.0774        0.2186        0.532
   22      9       2    0.265  0.0765        0.1507        0.467
   23      7       2    0.189  0.0710        0.0909        0.395
```

Visualizing Survival Data

- Scatterplots of censored data are not scientifically meaningful

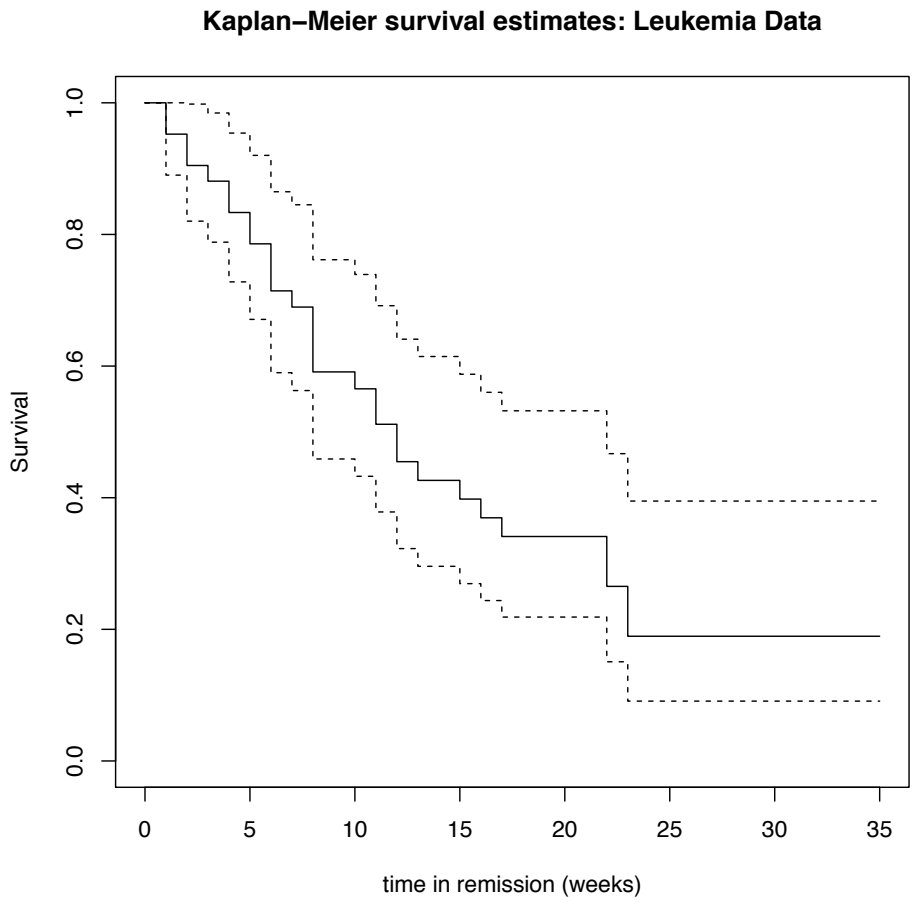
- It is thus better not to generate them unless you do something to indicate the censored data
  - We can label censored data, but we have to remember the true value may be anywhere larger than that

- Instead we look at Kaplan-Meier curves

K-M plots: Plotting the K-M curve

- Use `plot()` on the estimated survival function from `survfit()`:

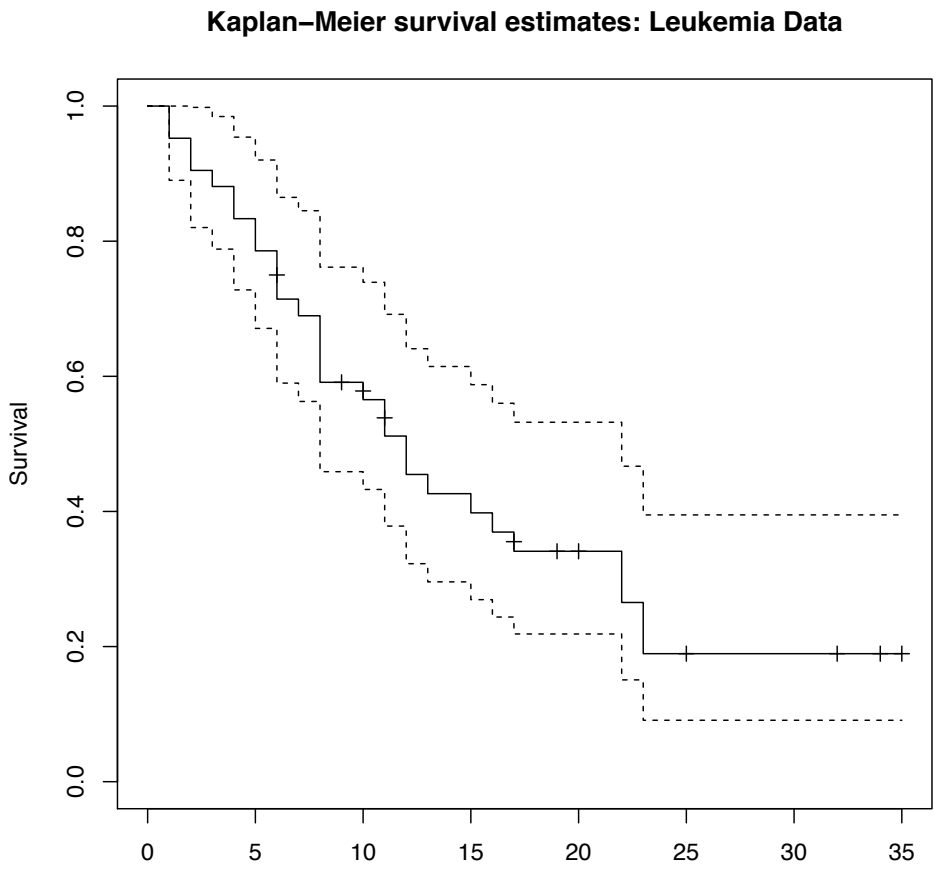
```
plot(km1,xlab="time in remission (weeks)", ylab="Survival",
main="Kaplan-Meier survival estimates: Leukemia Data")
```



K-M plots: Plotting the K-M curve

- To include censoring times, use the option “mark.time=TRUE” with `plot()`

```
plot(km1,xlab="time in remission (weeks)", ylab="Survival",
main="Kaplan-Meier survival estimates: Leukemia
Data",mark.time=TRUE)
```



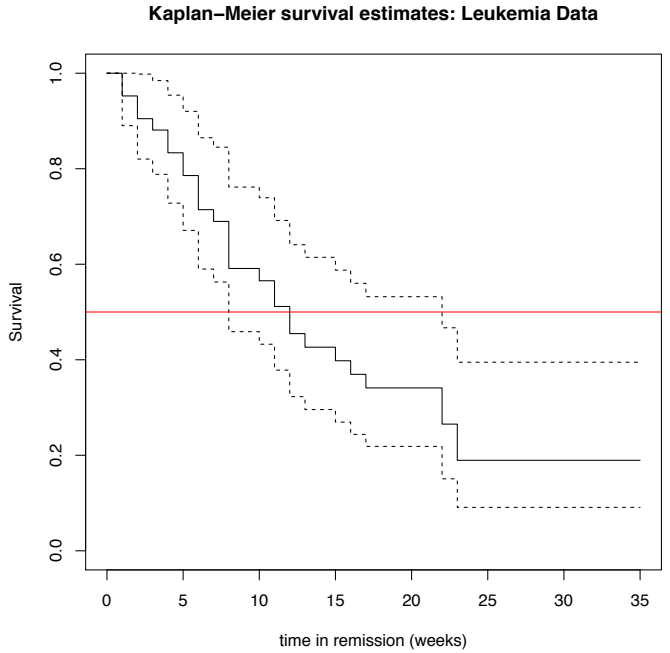
K-M plots: Plotting the K-M curve

- For inference on a quantile of the survival times, use e.g. quantile(km1, prob=0.5) – CIs obtained by inversion:

```
> quantile(km1,prob=.5)
$quantile
50
12

$lower
50
 8

$upper
50
22
```



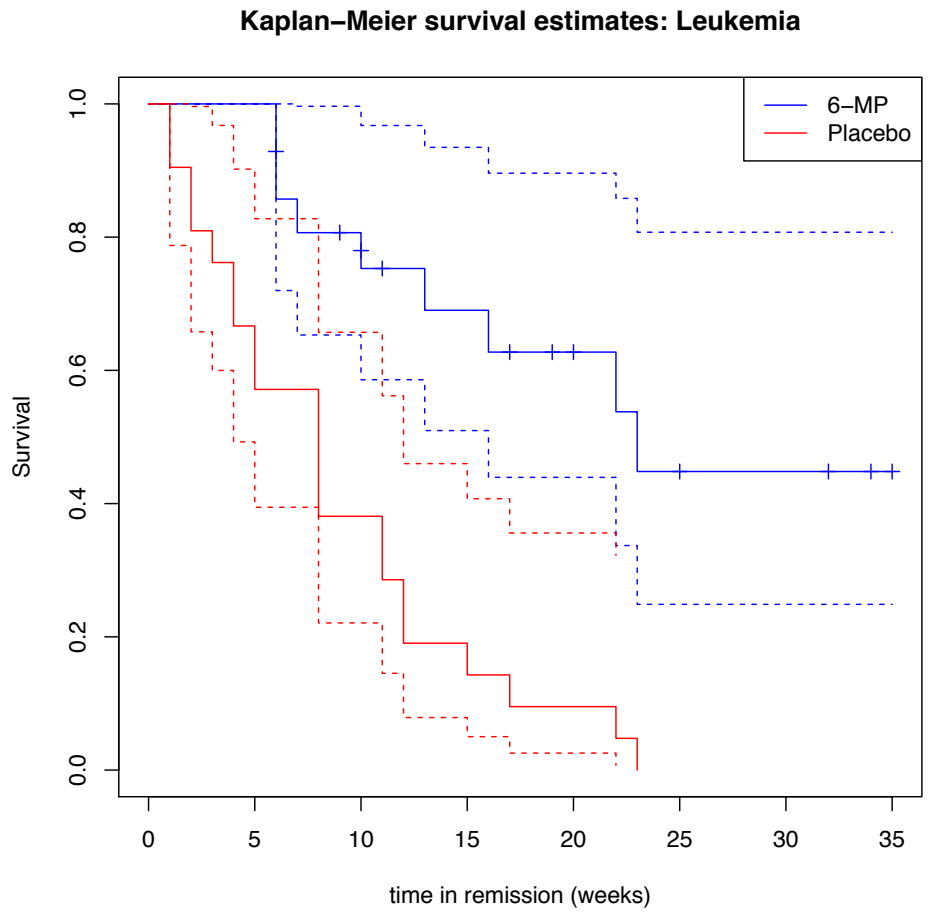
- Giving Median=12 (8,22). (But prob is 1-S(t) – beware!)
- CIs around the KM curve are not a statement about where the whole KM curve S lies – just statements about each S(t)

Comparing Survival Functions of Two or More Groups:

- It is often of interest to test if there is any association between survival across groups defined by a predictor of interest
- Under the null hypothesis, the survival function *S(t)* is the same across groups.
- Can test for an association by comparing **observed** survival probabilities to **expected** survival probabilities under the null hypothesis (O-E)

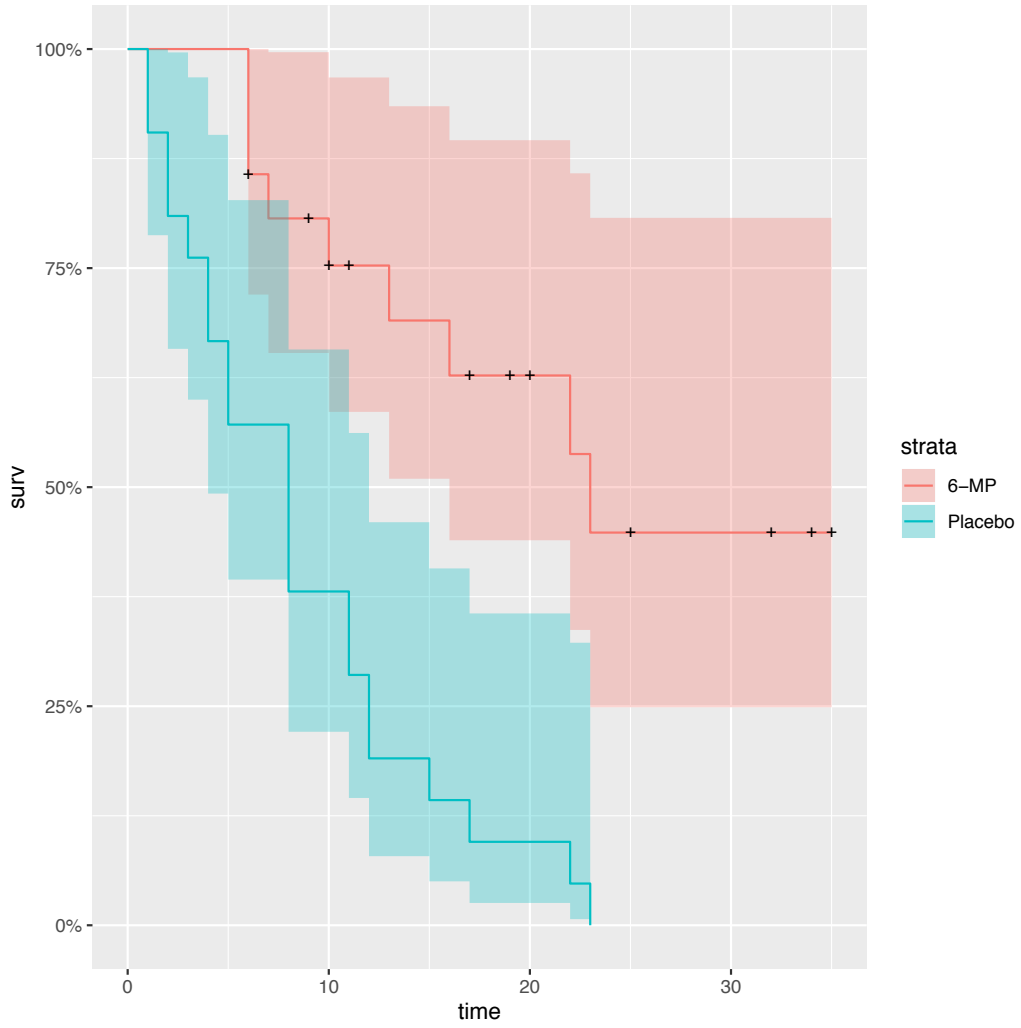
Plotting K-M Survival Functions for two or more groups: Leukemia Data

```
km2 <- survfit(survobj~group, data=leukdata)
plot(km2, mark.time=TRUE,col=c("blue", "red"), xlab="time in remission (weeks)",
ylab="Survival",  main="Kaplan-Meier survival estimates: Leukemia")
legend( "topright", lty=1, col=c("blue", "red"), legend=levels(leukdata$group))
```



Plotting K-M Survival Functions for Leukemia Data: ggfortify package

```
library(ggfortify)
km2 <- survfit(survobj~group, data=leukdata)
autoplot(km2)
```



Tabulated Survival Estimates

- Obtain a table of the Kaplan-Meier estimates for each group

```
> summary(km2)
Call: survfit(formula = survobj ~ group, data = leukdata)

                group=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807

                group=Placebo
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
   12      6       2   0.1905  0.0857      0.07887        0.460
   15      4       1   0.1429  0.0764      0.05011        0.407
   17      3       1   0.0952  0.0641      0.02549        0.356
   22      2       1   0.0476  0.0465      0.00703        0.322
   23      1       1   0.0000     NaN           NA           NA
```

Test of Equal Survival: Logrank Test

One test statistic T that measures extremity considers a 2×2 table at each of the event-times t_j, allowing for censoring;

	Group 1	Group 2	Total
#Deaths	D_{1j}	D_{2j}	D_j
#Survive to t_{j+1}	$N_{1j} - D_{1j}$	$N_{2j} - D_{2j}$	$N_j - D_j$
#At risk	N_{1j}	N_{2j}	N_j

If there were equal survival in the groups (and non-informative censoring) the expected counts are;

	Group 1	Group 2	Total
#Deaths	$N_{1j}\frac{D_j}{N_j}$	$N_{2j}\frac{D_j}{N_j}$	D_j
#Survive to t_{j+1}	$N_{1j}\frac{N_j - D_j}{N_j}$	$N_{2j}\frac{N_j - D_j}{N_j}$	$N_j - D_j$
#At risk	N_{1j}	N_{2j}	N_j

The approach is very similar to Pearson's χ^2 test.

Test of Equal Survival: Logrank Test

- The logrank test combines measures of observed minus expected death counts, quite similar to those we saw for Pearson's χ^2 test, across time intervals:

$$O - E = \sum_j D_{1j} - N_{1j} \frac{D_j}{N_j}$$

  - i.e. total deaths beyond expected, for group 1
  - The same calculation gives deaths below expected for group 2, so no need to also consider that discrepancy from the null

- An estimate of the variance of *O*-*E* is

$$\widehat{Var}\left(O - E\right) = \sum_j N_j \frac{N_j}{N_j - 1} \frac{N_{1j}}{N_j} \frac{N_{2j}}{N_j} \frac{D_j}{N_j} \frac{N_j - D_j}{N_j}$$

Logrank Test of equal survival for two or more groups:

- The test statistic for the logrank test is:

$$T = \frac{(O-E)^2}{\widehat{Var}(O-E)}$$

- Under the null hypothesis of equal survival functions for the two groups:

$$T \text{ approximately follows a } \chi^2_1 \text{ distribution}$$

- P-values are calculated using the null distribution.
- Testing equality of survival functions for more than 2 groups uses the same ideas, but the formula are too complex to give here – (see page 82 of the K&K book)

R: Logrank Test of equal survival

- Use the *survdiff()* function in R to perform a logrank test of differences of Kaplan-Meier estimates of survival for two or more groups In this data's placebo group, we observed 10.3 more deaths than expected under the null

```
> survdiff( surv ~ group, data=leukdata );
Call:
survdiff(formula = surv ~ group, data = leukdata)

                N Observed Expected (O-E)^2/E (O-E)^2/V
group=6-MP     21        9     19.3      5.46      16.8
group=Placebo  21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

R: Logrank Test of equal survival

```
> survdiff( surv ~ group, data=leukdata );
Call:
survdiff(formula = surv ~ group, data = leukdata)

                N Observed Expected (O-E)^2/E (O-E)^2/V
group=6-MP     21        9     19.3      5.46      16.8
group=Placebo  21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

- In this data's placebo group, we observed 10.3 more deaths than expected under the null
- ... and 10.3 fewer deaths on treatment, than expected under H0 (Reporting just one of these is recommended!)
- Data at least this extreme would only be expected in proportion 0.00004 (i.e. 0.004%) of similar studies, under H0, the null hypothesis of equal survival in the two groups
- p-value is from a large-n approximation, we are assuming simple random sampling and non-informative censoring

logrank test

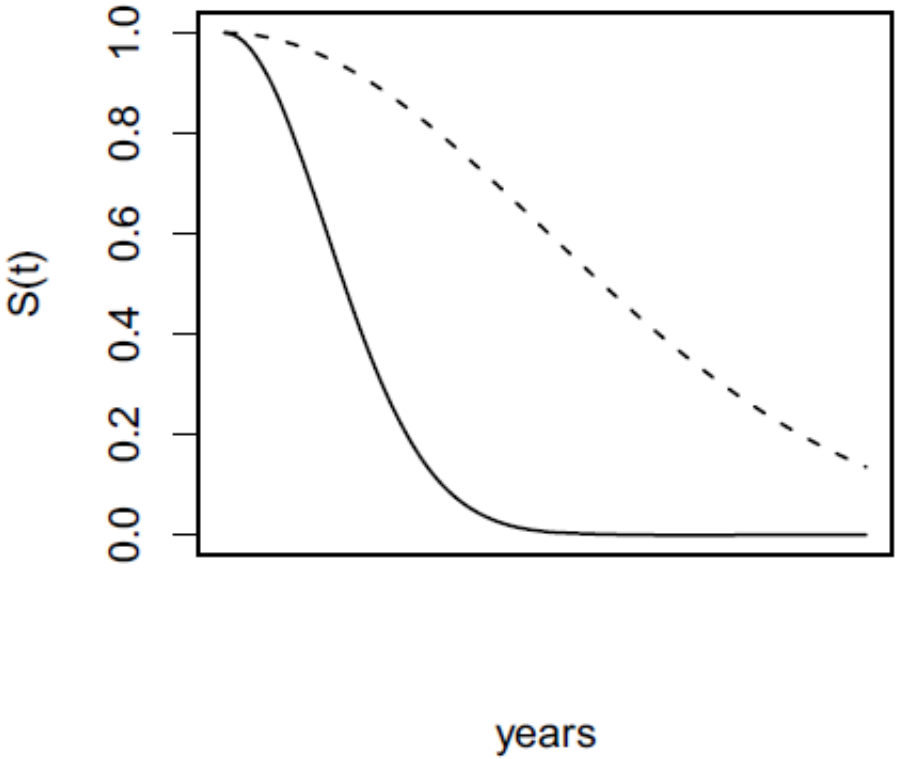
- The p-value for the logrank test is interpretable as a test of the null hypothesis that the survival distribution is the same for two groups

  H_0: $S_1(t)=S_2(t)$ for all t, or $h_1(t)=h_2(t)$ for all t

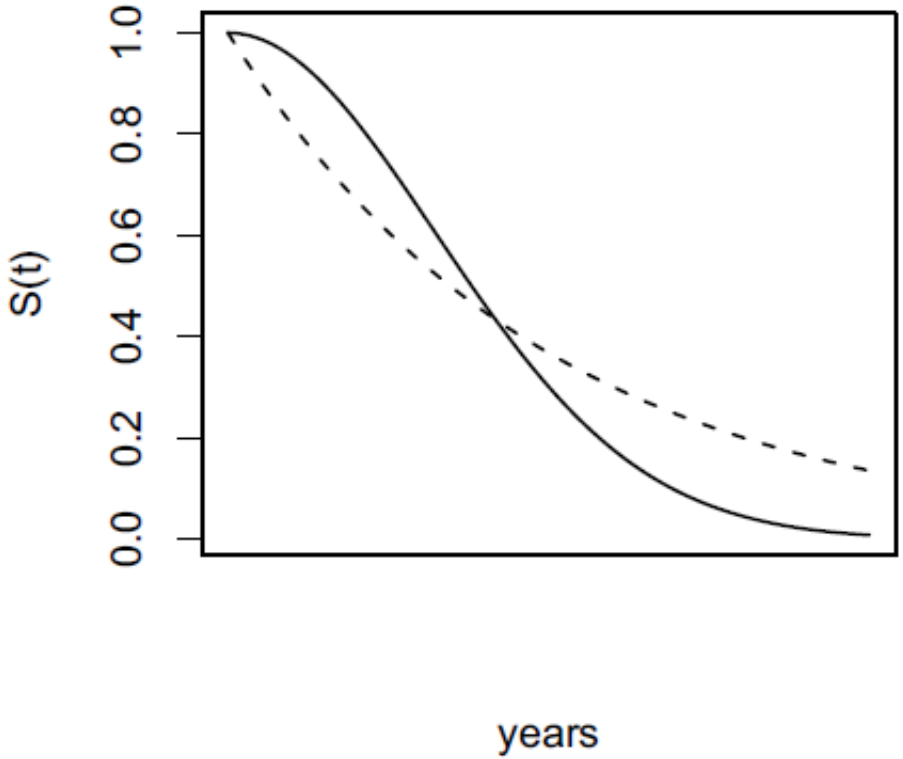
- Notice from the way the logrank test works, the ordering (ranking) of the data matters, but not the values. For this reason it is considered a rank based test.
- The logrank test is most powerful when the true alternative is “proportional hazards”

logrank test

Can detect:



but not:



Example 2: Prostate cancer (PSA dataset)

- Prognostic value of nadir PSA on time in remission from prostate cancer
- PSA data set: 50 men who received hormonal treatment for advanced prostate cancer
- Followed at least 24 months for clinical progression, but exact time of follow-up varies
- Nadir PSA: lowest level of serum prostate specific antigen achieved post-treatment

Example 2: Prostate cancer (PSA dataset)

- Prognostic value of nadir PSA on time in remission from prostate cancer
- PSA data set: 50 men who received hormonal treatment for advanced prostate cancer
- Followed at least 24 months for clinical progression, but exact time of follow-up varies
- Nadir PSA: lowest level of serum prostate specific antigen achieved post-treatment

Stratified Kaplan-Meier plots in R: setup

We start by loading in the 'survival' package and creating the event indicator:

```
## Load the required package, "survival"
library("survival")

# The event indicator is 'inrem': 1 = no, 2 = yes.
# Recode event as 1 if patient is not in remission (relapse) at observed time.
psa$event <- ifelse(psa$inrem == "no", 1, 0)
```

- event = 1 if the event (discontinuation of remission) is observed, event = 0 if the event is not observed (still in remission).
- End of remission (event = 1) is equivalent to when "inrem" is "no"

K-M plots: Survival object

`Surv()` creates the survival object from the observed times and event indicator; this will be our outcome variable in our Kaplan-Meier estimator and proportional hazards model.

```
> survobj <- with(psa, Surv(obstime, event))
> survobj
 [1] 42+ 48+ 40+ 75+ 30+ 24+ 58+ 36+ 40  60+ 60  48  30  60+ 48+ 35+ 42+
[18] 45  43  42  40  39  36  26  26  22  21  20  18  17  16  16  15  12
[35]  8   9   7   6   3   3   3   6   6   1  31  10  14  12  32  42+
```

K-M plots: Fitting the K-M curve

`survfit()` computes the Kaplan-Meier estimate from the survival object created earlier.

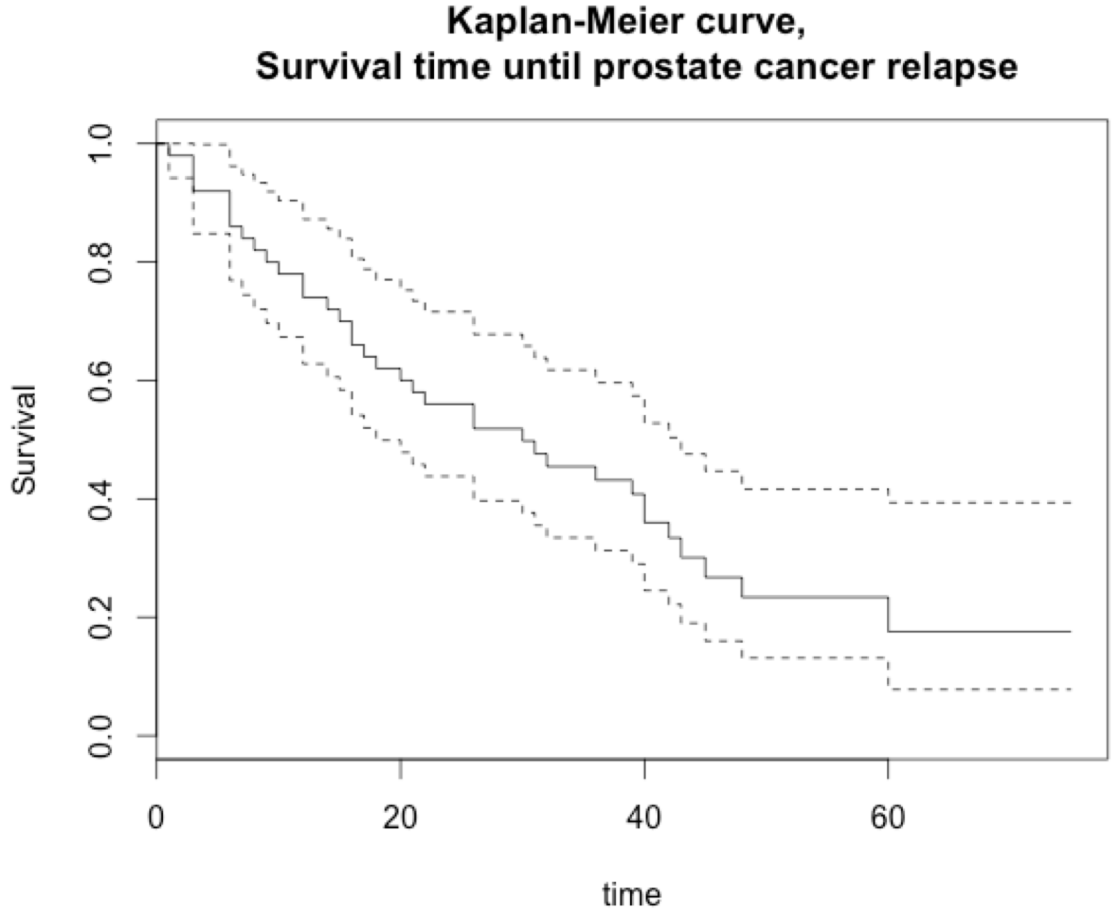
- The 1 in the formula `(survobj ~ 1)` indicates we are estimating the survival function for one group

```
> eS <- survfit( survobj ~ 1 )
> summary(eS)
Call: survfit(formula = survobj ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     50       1    0.980  0.0198        0.9420        1.000
    3     49       3    0.920  0.0384        0.8478        0.998
    6     46       3    0.860  0.0491        0.7690        0.962
    7     43       1    0.840  0.0518        0.7443        0.948
    8     42       1    0.820  0.0543        0.7201        0.934
    9     41       1    0.800  0.0566        0.6965        0.919
   10     40       1    0.780  0.0586        0.6732        0.904
   12     39       2    0.740  0.0620        0.6279        0.872
   14     37       1    0.720  0.0635        0.6057        0.856
```

K-M plots: Plotting the K-M curve

```
> # 95% CI for estimated survival function, no censoring marks
> plot(eS, main="Kaplan-Meier curve, \n Survival time until prostate cancer relapse",
+      xlab="time", ylab="Survival", conf.int = TRUE, mark.time = FALSE);
```



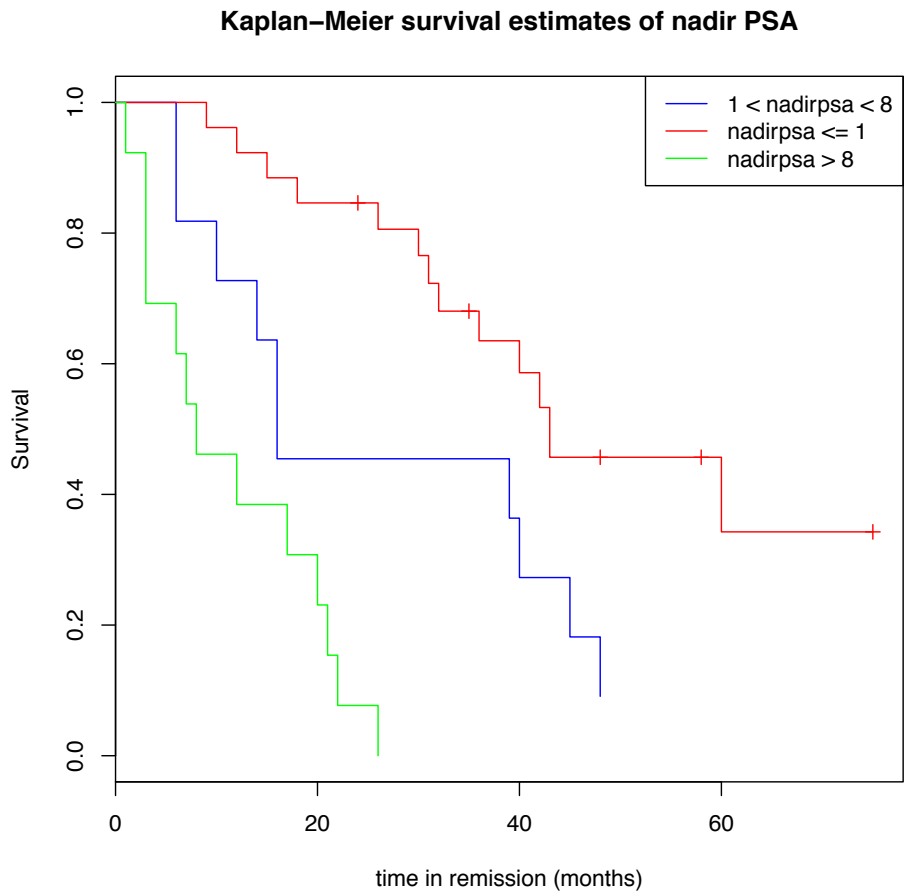
Stratified Kaplan-Meier Plots

- Create a categorical nadir PSA group variable with three levels:
  - Nadir PSA <=1
  - 1 < Nadir PSA <= 8
  - Nadir PSA >8

```
## Create three nadir PSA groups
psa$nadirpsaCTG<-NA
psa$nadirpsaCTG[psa$nadirpsa<=1]="nadirpsa <= 1"
psa$nadirpsaCTG[psa$nadirpsa>1 & psa$nadirpsa<=8]="1 < nadirpsa < 8"
psa$nadirpsaCTG[psa$nadirpsa>8]="nadirpsa > 8"
```

Stratified Kaplan-Meier Plots

```
pdf("Kaplan_Meier_nadir_PSA.pdf")
kms <- survfit( survobj ~ nadirpsaCTG, data=psa);
plot(kms, mark.time=TRUE,col=c("blue", "red","green"), xlab="time in remission (months)", ylab="Survival",
main="Kaplan-Meier survival estimates of nadir PSA");
legend( "topright", lty=1, col=c("blue", "red","green"), legend=levels(psa$nadirpsaCTG))
dev.off()
```



Tabulated Survival Estimates

```
> summary(kms)
Call: survfit(formula = survobj ~ nadirpsaCTG, data = psa)

                nadirpsaCTG=1 < nadirpsa < 8
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     11       2   0.8182  0.1163       0.6192        1.000
   10      9       1   0.7273  0.1343       0.5064        1.000
   14      8       1   0.6364  0.1450       0.4071        0.995
   16      7       2   0.4545  0.1501       0.2379        0.868
   39      5       1   0.3636  0.1450       0.1664        0.795
   40      4       1   0.2727  0.1343       0.1039        0.716
   45      3       1   0.1818  0.1163       0.0519        0.637
   48      2       1   0.0909  0.0867       0.0140        0.589

                nadirpsaCTG=nadirpsa <= 1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    9     26       1    0.962  0.0377        0.890        1.000
   12     25       1    0.923  0.0523        0.826        1.000
   15     24       1    0.885  0.0627        0.770        1.000
   18     23       1    0.846  0.0708        0.718        0.997
   26     21       1    0.806  0.0780        0.667        0.974
   30     20       1    0.766  0.0839        0.618        0.949
   31     18       1    0.723  0.0894        0.567        0.921
   32     17       1    0.681  0.0937        0.520        0.891
   36     15       1    0.635  0.0978        0.470        0.859
   40     13       1    0.586  0.1018        0.417        0.824
   42     11       1    0.533  0.1055        0.362        0.786
   43      7       1    0.457  0.1147        0.279        0.747
   60      4       1    0.343  0.1311        0.162        0.725

                nadirpsaCTG=nadirpsa > 8
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     13       1   0.9231  0.0739       0.7890        1.000
```

Log rank test of differences in K-M survival estimates

- Perform a log rank test of differences of the Kaplan-Meier estimates of survival for three nadir PSA groups

```
> survdiff( survobj ~ nadirpsaCTG, data=psa)
Call:
survdiff(formula = survobj ~ nadirpsaCTG, data = psa)

                              N Observed Expected (O-E)^2/E (O-E)^2/V
nadirpsaCTG=1 < nadirpsa < 8 11       10     7.62     0.744     0.975
nadirpsaCTG=nadirpsa <= 1    26       13    24.63     5.495    18.550
nadirpsaCTG=nadirpsa > 8     13       13     3.75    22.853    28.364

 Chisq= 32.4  on 2 degrees of freedom, p= 9e-08
```