STAT40810 — Stochastic Models

Brendan Murphy

Week 5

Non-Parametric Survival Analysis

Non-Parametric Estimate

- Suppose we wish to estimate the survival function without assuming a particular family of distributions.
- We want to be able to do this when we have (right) censored observations.
- We will look at a method called the Kaplan-Meier estimator.
- It is the most cited statistical paper of all time.

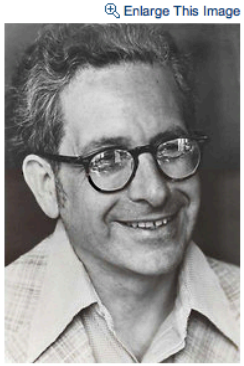
Paul Meier, Statistician Who Revolutionized Medical Trials, Dies at 87

By DENNIS HEVESI
Published: August 12, 2011

Paul Meier, a leading medical statistician who had a major influence on how the federal government assesses and makes decisions about new treatments that can affect the lives of millions, died on Sunday at his home in Manhattan. He was 87.

RECOMMEND
TWITTER
LINKEDIN
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Enlarge This Image



Paul Meier

The cause was complications of a stroke, his daughter Diane Meier said.

As early as the mid-1950s, Dr. Meier was one of the first and most vocal proponents of what is called "randomization."

Under the protocol, researchers randomly assign one group of patients to receive an experimental treatment and another to receive the standard treatment. In that way, the researchers try to avoid unintentionally skewing the results by choosing, for example, the healthier or younger patients to receive the new treatment.

If the number of subjects is large enough, the two groups will be the same in every respect except the treatment they receive. Such randomized

No Censoring

- Suppose we had a data set with no censoring.
- Let the observed values be

$$t_1, t_2, \ldots, t_n.$$

- How would we do this?
- If we want to estimate $S(t)$, for any t we could use

$$\hat{S}(t) = \frac{\text{Number of } t_i > t}{\text{Number of observations}}.$$

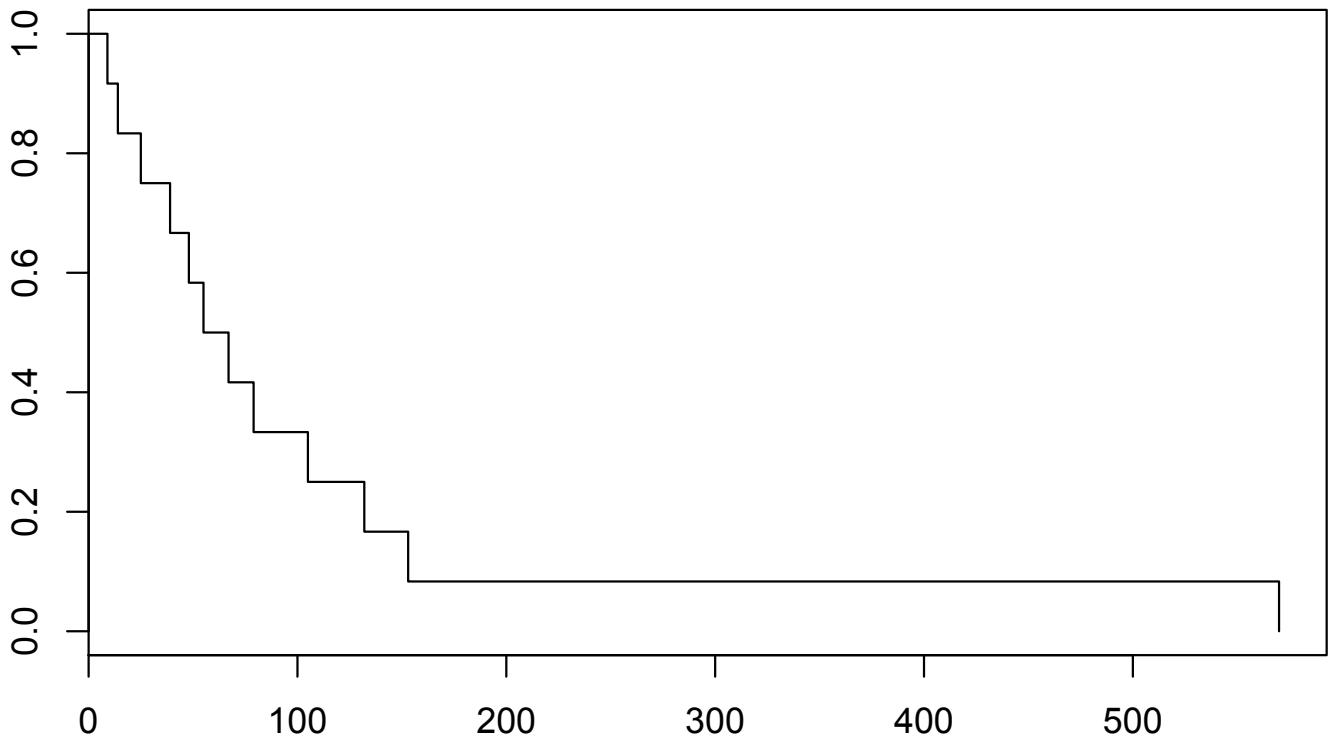
Calculation

- Let's do the calculation manually.

j	t_j	$obs > t_j$	$\hat{S}(t_j)$
1	9.00	11	0.92
2	14.00	10	0.83
3	25.00	9	0.75
4	39.00	8	0.67
5	48.00	7	0.58
6	55.00	6	0.50
7	67.00	5	0.42
8	79.00	4	0.33
9	105.00	3	0.25
10	132.00	2	0.17
11	153.00	1	0.08
12	570.00	0	0.00

Example: Failure Times

- For the failure time data we get...



Alternative Formulation

- Suppose at each time where an event occurs, we record the following:
  - d_j: the number of events that occurred at time t_j
  - n_j: the number of observations in the study at time t_j
  - $\lambda_j = d_j/n_j$: this is called the discrete hazard at time t_j.
- We can estimate $S(t_j)$ as

$$\hat{S}(t_j) = \prod_{k=1}^{j}(1-\lambda_k) = \prod_{k=1}^{j}\left(1-\frac{d_k}{n_k}\right).$$

- This is precisely how the Kaplan-Meier estimate works.

Alternative Calculation

- Let's do the calculation manually.

j	t_j	d_j	n_j	λ_j	$S(t_j)$
1	9.00	1.00	12	0.08	0.92
2	14.00	1.00	11	0.09	0.83
3	25.00	1.00	10	0.10	0.75
4	39.00	1.00	9	0.11	0.67
5	48.00	1.00	8	0.12	0.58
6	55.00	1.00	7	0.14	0.50
7	67.00	1.00	6	0.17	0.42
8	79.00	1.00	5	0.20	0.33
9	105.00	1.00	4	0.25	0.25
10	132.00	1.00	3	0.33	0.17
11	153.00	1.00	2	0.50	0.08
12	570.00	1.00	1	1.00	0.00

R Code

```
# Read in Failure Times
x <- scan()
79 105  14 153  67  25  39   9  55 132  48 570

#Fit Kaplan-Meier Curve
library(survival)
fit <- survfit(Surv(x)~1,se=FALSE)

# Plot the fit
plot(fit)

#Add the exponential model fit
lambda<-1/mean(x)
tvec<-seq(0,1000,length=201)
points(tvec,1-pexp(tvec,lambda),type="l",col="blue",lty=3)
```

Fit

```
Call: survfit(formula = Surv(x) ~ 1, se = FALSE)

 time n.risk n.event survival
    9     12       1   0.9167
   14     11       1   0.8333
   25     10       1   0.7500
   39      9       1   0.6667
   48      8       1   0.5833
   55      7       1   0.5000
   67      6       1   0.4167
   79      5       1   0.3333
  105      4       1   0.2500
  132      3       1   0.1667
  153      2       1   0.0833
  570      1       1   0.0000
```

Fit Comparison

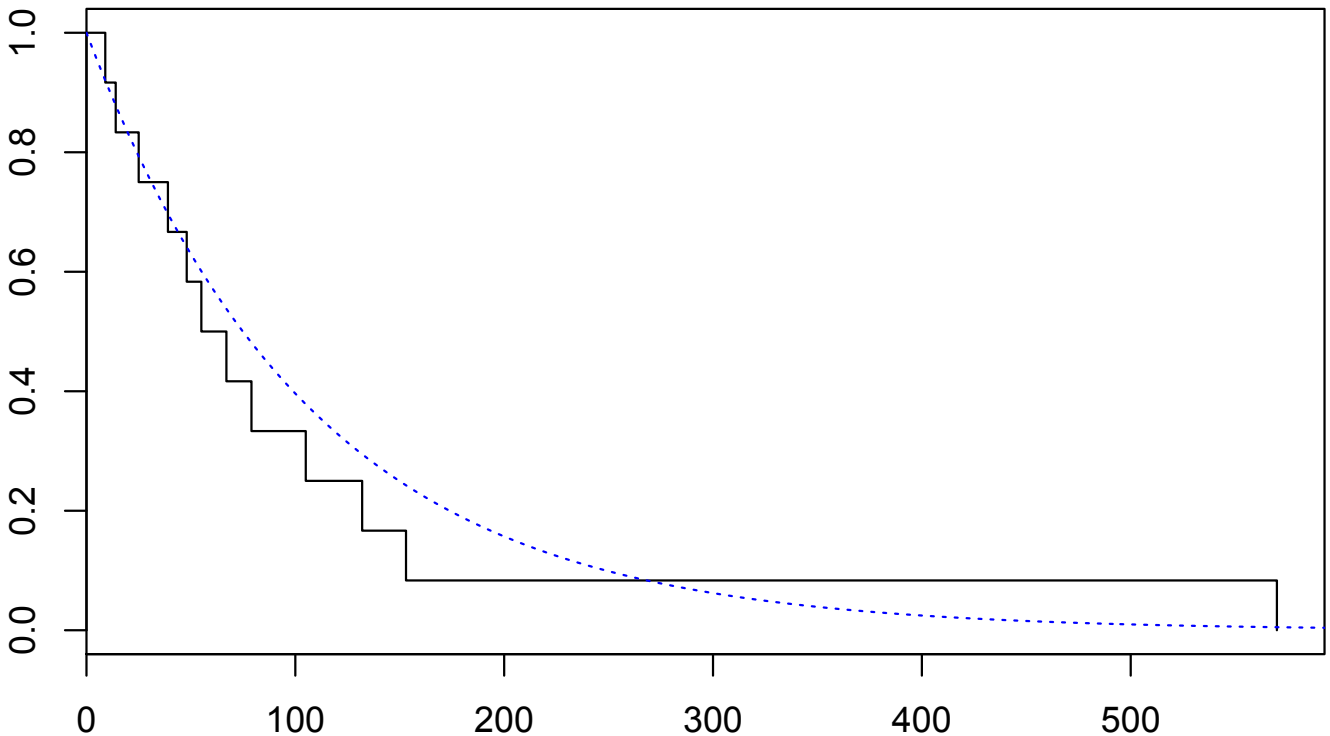


Figure : Survival curves are very similar.

Example: Leukemia

- The survival times of a number of leukemia patients who were on a maintained treatment were recorded.
- The times recorded are:

  9 13 13+ 18 23 28+ 31 34 45+ 48 161+

- The times marked with a $+$ are right censored.
- We want to do estimate the survival function non-parametrically.
- How do we do it?

Kaplan-Meier

- Let's try the same approach...

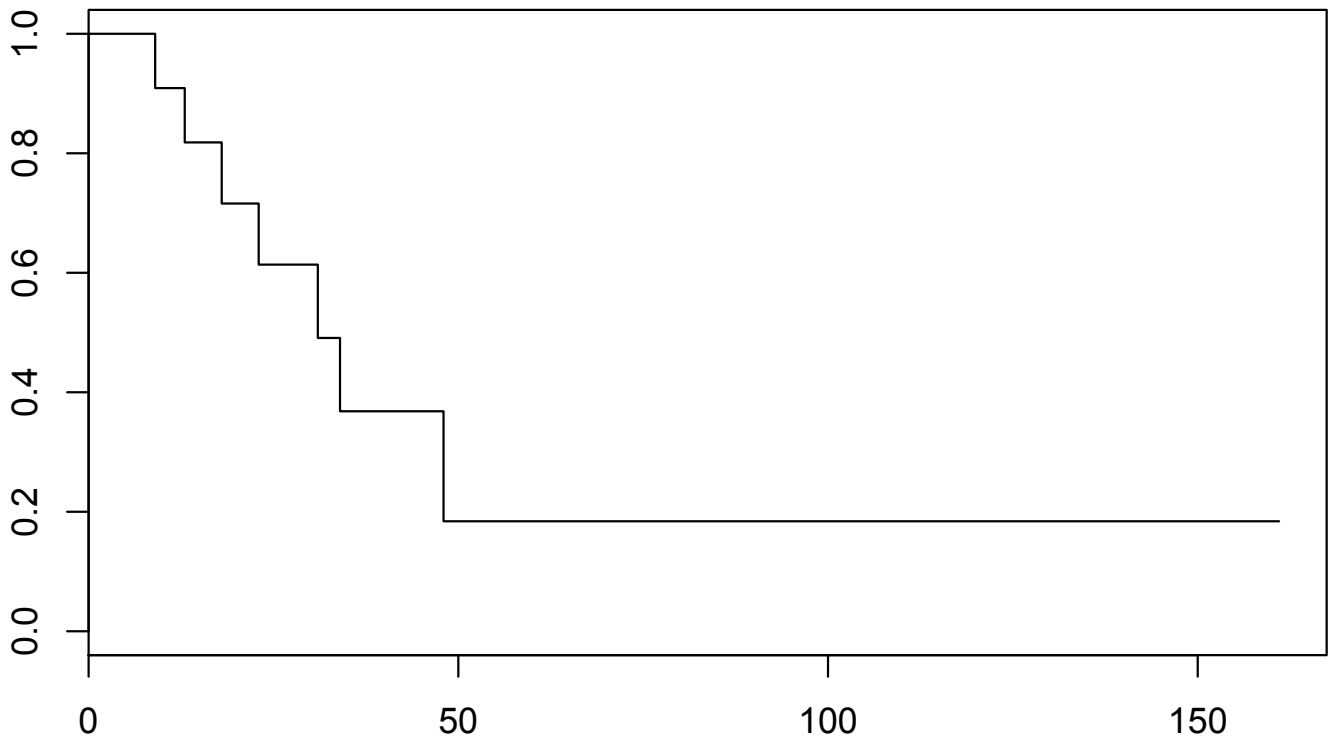
9 13 13+ 18 23 28+ 31 34 45+ 48 161+

j	t_j	d_j	n_j	λ_j	$S(t_j)$
1	9	1.00	11	0.09	0.91
2	13	1.00	10	0.10	0.82
3	28	1.00	8	0.13	0.72
4	23	1.00	7	0.14	0.61
5	31	1.00	5	0.20	0.49
6	34	1.00	4	0.25	0.37
7	48	1.00	2	0.50	0.18

- This is the Kaplan-Meier estimate with censoring.

Example: Leukemia

- We plot the survival estimate to get



Example: Leukemia

```
Call: survfit(formula = survdat ~ 1, se = FALSE)

 time n.risk n.event survival
    9     11       1    0.909
   13     10       1    0.818
   18      8       1    0.716
   23      7       1    0.614
   31      5       1    0.491
   34      4       1    0.368
   48      2       1    0.184
```

R Code

```
#Load survival package
library(survival)

#Load leukemia data
data(leukemia)

#Extract data for Maintained group
dat<-leukemia[leukemia$x=="Maintained",1:2]
dat

#Set up Surv() object
survdat <- Surv(time=dat[,1],event=dat[,2])
survdat

#Fit the KM to the data
fit <- survfit(survdat~1,se=FALSE)
summary(fit)
plot(fit)
```