Nonparametric methods: empirical survival function and the Kaplan-Meier estimator

Shannon Pileggi

STAT 417

OUTLINE

Survival estimators

Kaplan-Meier estimator

KM in Minitab/R

Nonparametric methods

- We have covered several parametric models to describe time-to-event data; however, it might not be a simple matter to assign a specific probability distribution (e.g. consider the age at first drink of alcohol).
- In practice, when the probability distribution is not known, we will need to use *nonparametric estimators* of the survival function, hazard function, and cumulative hazard functions that do not assume any form of the probability model for T.
- In addition, descriptive quantities such as the mean and median will need to be estimated.
- The main issue faced with time-to-event data is that some times may be *incomplete* (censored or truncated). We'll discuss methods and techniques that accommodate *right censored* observations.
- Why do we need to accommodate censoring?

Example: descriptive statistics with censored values

Consider the following extreme example: A statistics exam was administered to 10 students. At the end of the 50 minute class period, one student had finished (in 45 minutes), and the others had not finished it.

1. Describe the time-to-event variable T.
2. What are the observed values of T?
3. What distribution do you think T follows?

Example: descriptive statistics with censored values, cont.

4. Treat the right censored times as complete. Estimate the average time to complete the exam.

5. Throw out the right censored times. Estimate the average time to complete the exam.

6. What issues do you see with either of the strategies for dealing with censored data?

Estimating survival probabilities

The following times compose a *subsample* of 7 of the 57 motorist reaction times in seconds:

$$1.41 \quad 1.41+ \quad 2.76+ \quad 3.56 \quad 4.18 \quad 4.71+ \quad 13.18$$

Estimate the probability that a motorist takes longer than 1.41 seconds to react aggressively under each of the following schemes:

1. Treat the censored times as complete.
2. Throw out the censored times.

Estimators of the survival function

- Recall that the primary function used to describe the time-to-event variable T is the survival function, $S(t) = P(T > t)$.
- When we treat all times as complete (# 1 in previous example), this illustrates one strategy or method for estimating $S(t)$ known as the **empirical survival function (ESF)**, denoted $\widehat{S}_E(t)$.
- $\widehat{S}_E(t)$ is typically calculated for each observed time t_i in the sample, and then plotted as a *step function*.
- Note: this method is unbiased for $S(t)$ **if**...

Empirical survival function

Let $t_1, t_2, \ldots, t_n$ represent the observed event times for n individuals (includes censored and complete times). The ESF, $\widehat{S}_E(t)$, is calculated as:

Empirical survival function, motorist reaction times example

1.41 1.41+ 2.76+ 3.56 4.18 4.71+ 13.18

1. Find $\widehat{S}_E(t)$ for each value of t_i.

Empirical survival function, motorist reaction times example, cont.

2. Sketch the graph of $\widehat{S}_E(t)$ for all t.
3. Using the graph, estimate $S(2)$, $S(6)$, and $S(10)$.

Summary: ignoring censored event times

- By treating the censored times as complete times, we are assuming that the event times are shorter than what they actually are, thereby underestimating the true quantities, e.g. mean survival time and probabilities of survival.
- By disregarding the censored times, we lose information about the event times (consider the statistics exam times).
- Treating the censored observations as complete or ignoring them will *bias* any estimates based on the remaining complete times.

Next we'll consider a way to estimate the survival probabilities with a method that take into account the censoring status of subjects.

Survival estimators

Kaplan-Meier estimator

KM in Minitab/R

Kaplan-Meier estimator

- The **Kaplan-Meier estimator**, denoted $\widehat{S}(t)$, adjusts $\widehat{S}_E(t)$ to reflect the presence of right censored observations.
- When constructing $\widehat{S}(t)$, denote the n observed event times as $t_1, t_2, \ldots, t_n$, and the number of distinct *complete* event is m, where $m \leq n$.

$$1.41 \quad 1.41+ \quad 2.76+ \quad 3.56 \quad 4.18 \quad 4.71+ \quad 13.18$$

In the motorist reaction time subsample, $m =$______ and $n =$______.

1. $m = 6$, $n = 7$
2. $m = 7$, $n = 6$
3. $m = 4$, $n = 7$
4. $m = 5$, $n = 7$

Kaplan-Meier estimator, cont.

- Then the *ordered* complete event times from smallest to largest are denoted as:

  with the convention that $t_{(0)} = 0$.
- Using the complete times, $t_{(1)}$ through $t_{(m)}$, and beginning with time $t_{(0)} = 0$, create a series of time intervals:

  with some modifications to the last interval depending on whether the largest observed time is complete or censored.

Kaplan-Meier estimator, cont.

- By convention, the 0^{th} time interval is $[0, t_{(1)})$
- If the largest observed time is complete (denoted $t_{(m)}$), then the last (m^{th}) interval is $[t_{(m)}, t_{(m)}]$.
- If the largest observed time is censored (denoted $t_{\text{max}+}$), then the last interval is $[t_{(m)}, t_{\text{max}+})$.

Kaplan-Meier estimator example

$$1.41 \quad 1.41+ \quad 2.76+ \quad 3.56 \quad 4.18 \quad 4.71+ \quad 13.18$$

1. List the values corresponding to $t_{(1)}$ through $t_{(m)}$.

2. Write out the time intervals.

Kaplan-Meier estimator, cont.

Once the complete event times have been identified, and the m time intervals have been constructed, then for $i = 0, \ldots, m$, define:

- n_i

- d_i

- $n_i - d_i$

Kaplan-Meier estimator example, cont.

i	Interval	$t_{(i)}$	n_i	d_i	$n_i - d_i$	$\hat{p}_i$	$\widehat{S}(t_{(i)})$
0	$[0, 1.41)$	0					
1	$[1.41, 3.56)$	1.41					
2	$[3.56, 4.18)$	3.56					
3	$[4.18, 13.18)$	4.18					
4	$[13.18, 13.18]$	13.18					

Kaplan-Meier estimator, cont.

How would we evaluate $S(t)$ at the complete event times: $t_{(1)}, \ldots, t_{(m)}$?

- Recall the definition of the survival function:

$$S(t) = P(T > t)$$

- Then $S(t_{(1)}), S(t_{(2)}), \ldots, S(t_{(m)})$ can each be written as a product of conditional probabilities using applications of the **multiplication rule**:

- Note: by definition: $S(t_{(0)}) = P(T > 0) = 1$.

Kaplan-Meier estimator, cont.

Then using the multiplication rule we can write $S(t_{(1)}) = P(T > t_{(1)})$ as:

Kaplan-Meier estimator, cont.

Furthermore:

- $S(t_{(2)}) = P(T > t_{(2)}) =$

- Finally we have $S(t_{(i)}) = P(T > t_{(i)}) =$

We'll need estimates of:

$$P(T > t_{(i)} | T > t_{(i-1)})$$

for $i = 1, \ldots, m$.

$\widehat{S}(t)$: motorist reaction times

Using the information in the table, estimate the probability that a motorist will not react aggressively after 1.41 seconds, given that the motorist hasn't reacted aggressively after 0 seconds.

Kaplan-Meier estimator, cont.

The estimate of $P(T > t_{(i)}|T > t_{(i-1)})$, denoted by $\hat{p}_i$ is given by:

This is the estimated conditional probability of an individual surviving past time $t_{(i)}$ given that the individual has survived past time $t_{(i-1)}$.

Example: Calculate the values of $\hat{p}_i$ and fill in the table.

Kaplan-Meier estimator, cont.

Hence, the Kaplan-Meier estimator of the survival function at the complete event time $t_{(i)}$ is given by:

or for general time t, the Kaplan-Meier estimator is given by:

Kaplan-Meier example, cont.

Compute the values of $\widehat{S}(t_{(i)})$ and fill in the last column in the table.

Kaplan-Meier notes

1. $\widehat{S}(t) = 1$ for $t < t_{(1)}$.

2. $\widehat{S}(t)$ remains constant for $t_{(i)} \leq t < t_{(i+1)}$.

3. The Kaplan-Meier estimates of survival probability are always greater than or equal to estimates based on the empirical survival function.

4. If no censoring is present in the data, then the Kaplan-Meier and empirical survival functions are identical. What do you think happens when all the times are (right) censored?

Group Exercise

Will $\widehat{S}(t)$ reach 0?

1. yes
2. no
3. it depends

Comparing ESF and KM estimators

Use both the ESF and KM estimators to estimate $S(2)$, $S(6)$, and $S(10)$. How do the estimates compare?

	$\widehat{S}_E(t)$ (ESF)	$\widehat{S}(t)$ (KM)
$\hat{S}(2)$		
$\hat{S}(6)$		
$\hat{S}(10)$		

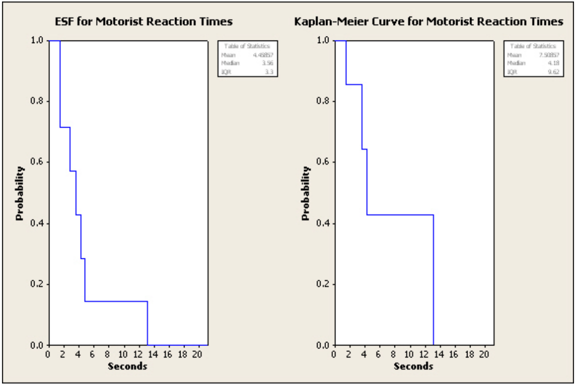
Survival estimators

Kaplan-Meier estimator

KM in Minitab/R

Graphs of the survival estimators

The **Kaplan-Meier curve** is a graphical display of $\widehat{S}(t)$, and is a (right continuous) step function which steps down only at *complete* event times.



Minitab: Kaplan-Meier estimates

Minitab output for the Nonparametric Distribution Analysis of the subsample of motorist reaction times:

```
Distribution Analysis: Time
Variable: Time

Censoring Information  Count
Uncensored value           4
Right censored value       3

Censoring value: Censor 1 = 0
```

Minitab: Kaplan-Meier estimates

```
Nonparametric Estimates

Characteristics of Variable

            Standard   95.0% Normal CI
Mean(MTTF)     Error     Lower     Upper
   7.50857   2.54935   2.51194   12.5052

Median = 4.18
IQR = 9.62  Q1 = 3.56  Q3 = 13.18
```

Minitab: Kaplan-Meier estimates

```
Kaplan-Meier Estimates

         Number  Number     Survival  Standard   95.0% Normal CI
 Time   at Risk  Failed  Probability     Error     Lower     Upper
 1.41         7       1     0.857143  0.132260  0.597918   1.00000
 3.56         4       1     0.642857  0.210424  0.230433   1.00000
 4.18         3       1     0.428571  0.224258  0.000000   0.86811
13.18         1       1     0.000000  0.000000  0.000000   0.00000
```

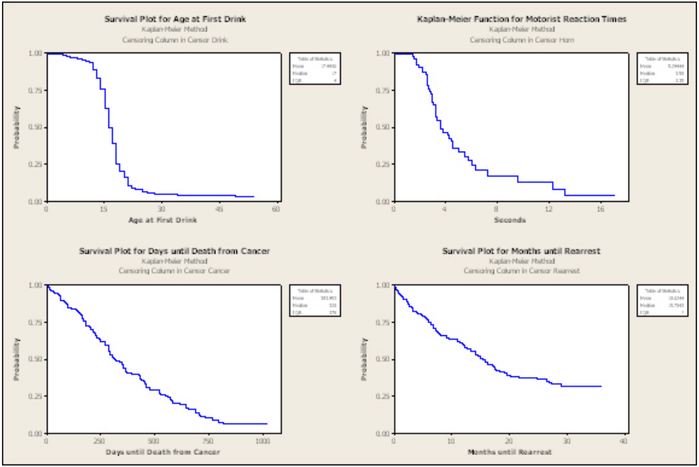
Describing survival experiences

The following slide (from left-right, top-bottom) provides $\widehat{S}(t)$ for the four example data sets:

1. Age at first drink of alcohol
2. Seconds until driver reacts aggressively
3. Days until death from lung cancer (North Central Cancer Treatment Group Study)
4. Months until rearrest

Examine the Kaplan-Meier curves for each data set. Comment on some notable features of the "survival" experiences of a couple of the examples (in particular the rearrest data).

Describing survival experiences



Describing survival experiences

Kaplan-Meier estimates using R

Motorist reaction times example: R code

```
> library(survival)
> time <- c(1.41,1.41,2.76,3.56,4.18,4.71,13.18)
> censor <- c(1,0,0,1,1,0,1)
> motorist.surv <- Surv(time,censor)
> KM.obj <- survfit(motorist.surv˜1, conf.type="none")
> summary(KM.obj)
```

R output:

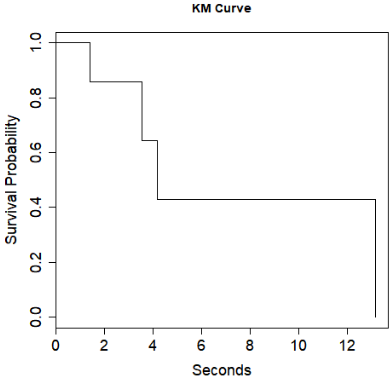
```
Call: survfit(formula = motorist.surv ~ 1, conf.type = "none")

  time n.risk n.event survival std.err
  1.41      7       1    0.857   0.132
  3.56      4       1    0.643   0.210
  4.18      3       1    0.429   0.224
 13.18      1       1    0.000     NaN
```

Kaplan-Meier estimates using R

Motorist reaction times example: R code

```
> plot(KM.obj, xlab="Seconds", ylab="Survival Probability",
    main="KM Curve")
```


KM Curve
Survival Probability
Seconds