BIOS 600: Principles of Statistical Inference

Survival Analysis

Fall 2016

Reading

- Pagano and Gauvreau, Chapter 21
- For more information on survival analysis, take BIOS 680!

Goal

Survival analysis is a complex topic, and you are strongly encouraged to take a survival analysis course if you plan to analyze data of this type. The goal of our coverage is to give you the skills you need to understand results of simple descriptive statistics in this setting, and we will not have time to discuss more complex modeling of survival data in this course.

Survival Data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest, e.g.

- Time from transplant surgery until new organ failure

Typically, the event of interest is called a *failure* (even if it is a good thing). The time interval between a starting point and the failure is known as the *survival time* and is often represented by t.

Survival Data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest, e.g.

- Time from transplant surgery until new organ failure
- Time to death in a pancreatic cancer trial

Typically, the event of interest is called a *failure* (even if it is a good thing). The time interval between a starting point and the failure is known as the *survival time* and is often represented by t.

Survival Data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest, e.g.

- Time from transplant surgery until new organ failure
- Time to death in a pancreatic cancer trial
- Time to first sex

Typically, the event of interest is called a *failure* (even if it is a good thing). The time interval between a starting point and the failure is known as the *survival time* and is often represented by t.

Survival Data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest, e.g.

- Time from transplant surgery until new organ failure
- Time to death in a pancreatic cancer trial
- Time to first sex
- Time to menopause

Typically, the event of interest is called a *failure* (even if it is a good thing). The time interval between a starting point and the failure is known as the *survival time* and is often represented by t.

Survival Data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest, e.g.

- Time from transplant surgery until new organ failure
- Time to death in a pancreatic cancer trial
- Time to first sex
- Time to menopause
- Time to divorce

Typically, the event of interest is called a *failure* (even if it is a good thing). The time interval between a starting point and the failure is known as the *survival time* and is often represented by t.

Survival Data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest, e.g.

- Time from transplant surgery until new organ failure
- Time to death in a pancreatic cancer trial
- Time to first sex
- Time to menopause
- Time to divorce
- Time to receipt of PhD

Typically, the event of interest is called a *failure* (even if it is a good thing). The time interval between a starting point and the failure is known as the *survival time* and is often represented by t.

Survival Data

Certain aspects of survival data make data analysis particularly challenging.

- Typically, not all the individuals are observed until their times of failure

Survival Data

Certain aspects of survival data make data analysis particularly challenging.

- Typically, not all the individuals are observed until their times of failure
    - An organ transplant recipient may die in an automobile accident before the new organ fails

Survival Data

Certain aspects of survival data make data analysis particularly challenging.

- Typically, not all the individuals are observed until their times of failure
  - An organ transplant recipient may die in an automobile accident before the new organ fails
  - A PhD student may withdraw from the program to start a multi-billion dollar health company

Survival Data

Certain aspects of survival data make data analysis particularly challenging.

- Typically, not all the individuals are observed until their times of failure
    - An organ transplant recipient may die in an automobile accident before the new organ fails
    - A PhD student may withdraw from the program to start a multi-billion dollar health company
    - Not everyone gets divorced

Survival Data

Certain aspects of survival data make data analysis particularly challenging.

- Typically, not all the individuals are observed until their times of failure
  - An organ transplant recipient may die in an automobile accident before the new organ fails
  - A PhD student may withdraw from the program to start a multi-billion dollar health company
  - Not everyone gets divorced
  - A pancreatic cancer patient may move to Aitutaki instead of undergoing further treatment

Survival Data

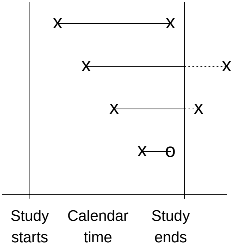
Certain aspects of survival data make data analysis particularly challenging.

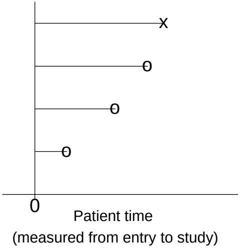
- Typically, not all the individuals are observed until their times of failure
  - An organ transplant recipient may die in an automobile accident before the new organ fails
  - A PhD student may withdraw from the program to start a multi-billion dollar health company
  - Not everyone gets divorced
  - A pancreatic cancer patient may move to Aitutaki instead of undergoing further treatment
- In this case, an observation is said to be *censored* at the last point of contact with the patient.

Aitutaki

I hope you do visit Aitutaki someday! The Cook Islands are really nice.

Study Time and Patient Time

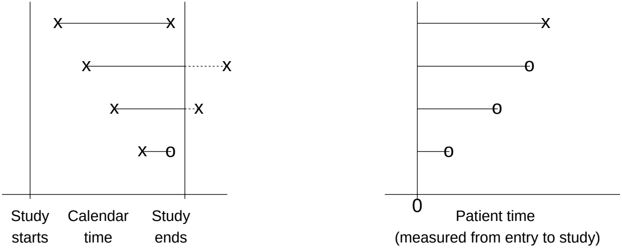




It is important to distinguish between study time and patient time.

- A study may start enrolling patients in September and continue until all 500 patients have been enrolled

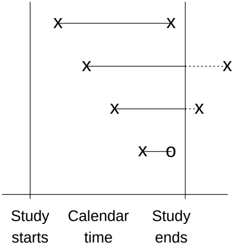
Study Time and Patient Time

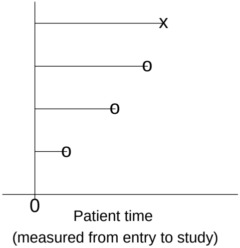


It is important to distinguish between study time and patient time.

- A study may start enrolling patients in September and continue until all 500 patients have been enrolled
- This is likely to take months or years

Study Time and Patient Time





It is important to distinguish between study time and patient time.

- A study may start enrolling patients in September and continue until all 500 patients have been enrolled
- This is likely to take months or years
- Time is typically converted to patient time (time between enrollment and failure or censoring) before analysis

Survival Function

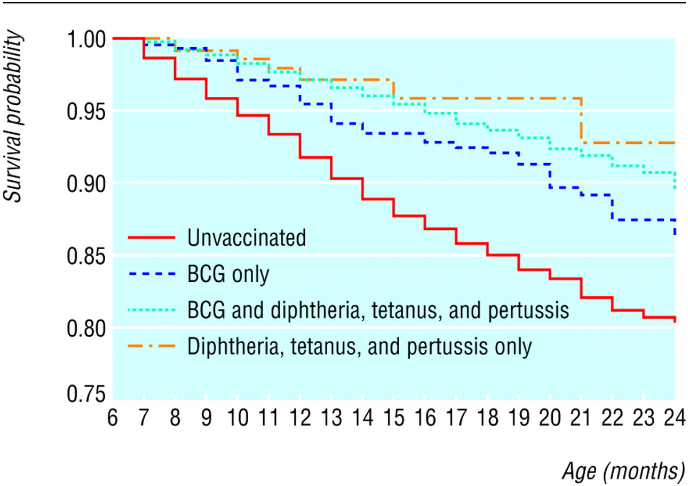
The distribution of survival times is characterized by the *survival function*, represented by $S(t)$. For a continuous random variable T,

$$S(t) = Pr(T > t),$$

and $S(t)$ represents the proportion of individuals who have not yet failed.

The graph of $S(t)$ versus t is called a survival curve. The survival curve shows the proportion of survivors at any given time.

Survival of Children in Burkina Faso by Vaccination Status



Simple Example

A small study enrolls 10 patients, whose outcomes are below.

Patient	Event time (x)	Event type
1	4.5	Death
2	7.5	Death
3	8.5	Censored
4	11.5	Death
5	13.5	Censored
6	15.5	Death
7	16.5	Death
8	17.5	Censored
9	19.5	Death
10	21.5	Censored

How do we estimate the survival curve for these data?

Kaplan-Meier Estimate

Perhaps the most popular estimate of a survival curve is the *Kaplan-Meier* or *product-limit* estimate. This method is actually fairly intuitive.

First, define the following quantities.

- l_t: # at risk of failure at time t (i.e., those who did not fail before t and those who were not censored before t)

Kaplan-Meier Estimate

Perhaps the most popular estimate of a survival curve is the *Kaplan-Meier* or *product-limit* estimate. This method is actually fairly intuitive.

First, define the following quantities.

- l_t: # at risk of failure at time t (i.e., those who did not fail before t and those who were not censored before t)
- d_t: # who fail at time t

Kaplan-Meier Estimate

Perhaps the most popular estimate of a survival curve is the *Kaplan-Meier* or *product-limit* estimate. This method is actually fairly intuitive.

First, define the following quantities.

- l_t: # at risk of failure at time t (i.e., those who did not fail before t and those who were not censored before t)
- d_t: # who fail at time t
- $q_t = \frac{d_t}{l_t}$: estimated probability of failing at time t

Kaplan-Meier Estimate

Perhaps the most popular estimate of a survival curve is the *Kaplan-Meier* or *product-limit* estimate. This method is actually fairly intuitive.

First, define the following quantities.

- l_t: # at risk of failure at time t (i.e., those who did not fail before t and those who were not censored before t)
- d_t: # who fail at time t
- $q_t = \frac{d_t}{l_t}$: estimated probability of failing at time t
- $S(t)$: cumulative probability of surviving beyond time t, estimated as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right).$$

Kaplan-Meier Estimate

Perhaps the most popular estimate of a survival curve is the *Kaplan-Meier* or *product-limit* estimate. This method is actually fairly intuitive.

First, define the following quantities.

- l_t: $\#$ at risk of failure at time t (i.e., those who did not fail before t and those who were not censored before t)
- d_t: $\#$ who fail at time t
- $q_t = \frac{d_t}{l_t}$: estimated probability of failing at time t
- $S(t)$: cumulative probability of surviving beyond time t, estimated as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right).$$

- The $\prod$ symbol is for multiplication, e.g. $\prod_{i=1}^{3} x_i = x_1 x_2 x_3$ and $\prod_{i=1}^{5} i = 1 \times 2 \times 3 \times 4 \times 5$.

How is that intuitive?

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

At each time t, the probability of surviving is just $1 - Pr(failing)$. Before there are any failures in the data, our estimated $\hat{S}(t) = 1$. At the time of the first failure, this probability falls below 1 and is simply one minus the probability of failing at that time, or $1 - \frac{\#\ failures}{\#\ at\ risk\ of\ failing}$.

After the first failure, things get more complicated. At the time of the second failure, you can calcuate $1 - \frac{\#\ failures}{\#\ at\ risk\ of\ failing}$, but this doesn't provide the whole picture, as someone else has already died. In fact, this is the conditional probability of surviving now that you've made it past the time of the first failure.

How is that intuitive?

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

How do you then calculate the total (unconditional) probability of survival? That is just the product of the probability of surviving past the first failure times the conditional probability of surviving beyond the second failure given that you made it past the first, or

$$Pr(\text{survived past first and second times})$$

$$= Pr(\text{survive past first time}) Pr(\text{survive past second time} \mid \text{survived past first time})$$

$$= \left(1 - \frac{\#\ failures\ at\ failure\ time\ 1}{\#\ at\ risk\ of\ failing\ at\ failure\ time\ 1}\right)\left(1 - \frac{\#\ of\ failures\ at\ failure\ time\ 2}{\#\ at\ risk\ of\ failing\ at\ failure\ time\ 2}\right)$$

If someone is censored, they are no longer at risk of failing at the next failure time and are taken out of the calculation

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0		
4.5	1	0		
7.5	1	0		
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0		
7.5	1	0		
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	$1 - \frac{1}{10} = 0.9$
7.5	1	0		
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	$0.9\times(1-\frac{1}{9})=0.8$
8.5	0	1		
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	$0.8 \times (1 - \frac{0}{8})$=0.8
11.5	1	0		
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	$0.8 \times (1 - \frac{1}{7}) = 0.69$
13.5	0	1		
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0		
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	$0.69 \times (1 - \frac{1}{5}) = 0.552$
16.5	1	0		
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	0.552
16.5	1	0	3	$0.552 \times (1 - \frac{1}{4}) = 0.414$
17.5	0	1		
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	0.552
16.5	1	0	3	0.414
17.5	0	1	2	0.414
19.5	1	0		
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	0.552
16.5	1	0	3	0.414
17.5	0	1	2	0.414
19.5	1	0	1	$0.414\times(1-\frac{1}{2})$=0.207
21.5	0	1		

What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

Kaplan-Meier (KM) Estimate

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{l_{t_i}}\right)$$

t	# Failed (d_t)	# Censored	#Left (l_{t+1})	$\hat{S}(t)$
0.0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	0.552
16.5	1	0	3	0.414
17.5	0	1	2	0.414
19.5	1	0	1	0.207
21.5	0	1	0	0.207

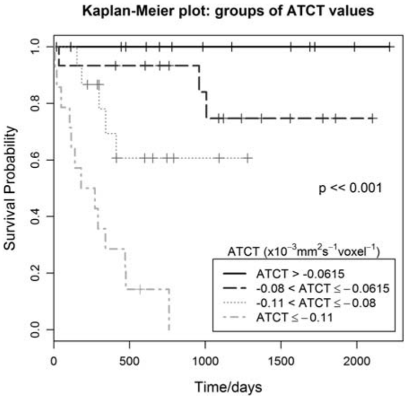
What would $\hat{S}(21.5)$ be if the last observation were a failure instead of censored?

KM Estimate

In between failure times, the KM estimate does not change but is constant. This gives the estimated survival function its step-like appearance (we call this type of function a *step function*).

Tumors in Children, 2012 *Neuro-Oncology*

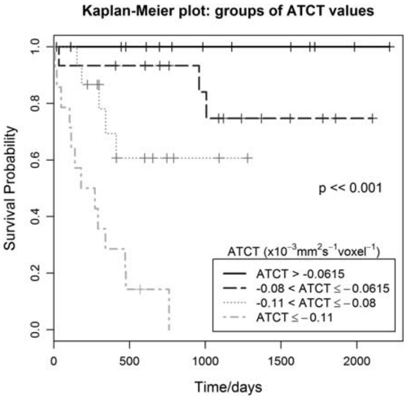
ATCT is an imaging-based biomarker of tumor prognosis.



- Which biomarker values are associated with the best survival?

Tumors in Children, 2012 *Neuro-Oncology*

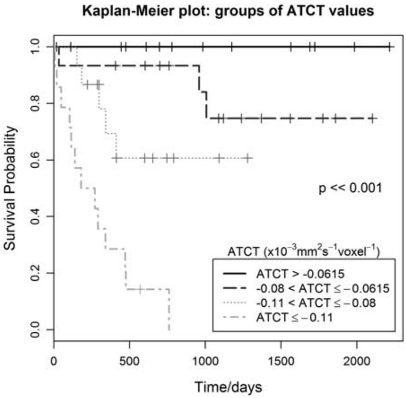
ATCT is an imaging-based biomarker of tumor prognosis.



- Which biomarker values are associated with the best survival?
- Which values are associated with the worst survival?

Tumors in Children, 2012 *Neuro-Oncology*

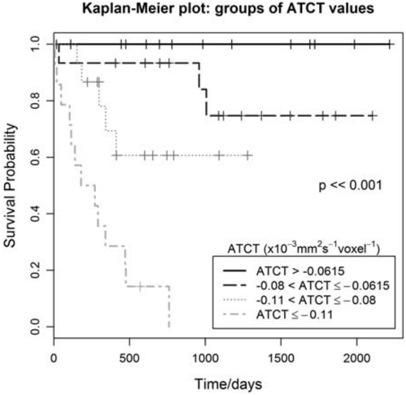
ATCT is an imaging-based biomarker of tumor prognosis.



- Which biomarker values are associated with the best survival?
- Which values are associated with the worst survival?
- What is the median survival time in the group with the smallest ATCT values?

Tumors in Children, 2012 *Neuro-Oncology*

ATCT is an imaging-based biomarker of tumor prognosis.



- Which biomarker values are associated with the best survival?
- Which values are associated with the worst survival?
- What is the median survival time in the group with the smallest ATCT values?
- If a child is in the group with the largest ATCT values, what is his/her estimated 5-year survival probability?

Revascularization in GRACE

The Global Registry of Acute Coronary Events (GRACE) registry now contains data on over 60,000 subjects with an acute coronary syndrome. We consider the outcome of days until death as a function of whether a revascularization procedure was performed or not. First, we generate KM survival estimates for the revascularization group and for the non-revascularization group in a sample of 1000 patients from the registry.

The Stata survival commands are a bit more complex than those for linear and logistic regression. First, we need to tell Stata the name of the time to event variable (here, days) and the name of a variable that takes value 1 for failures and 0 for censored observations (here, death) by typing stset days, failure(death). Then the data are prepared for other survival analysis commands. To generate the KM estimates, we type sts graph, by(revasc) where in these data the variable REVASC is the group indicator.

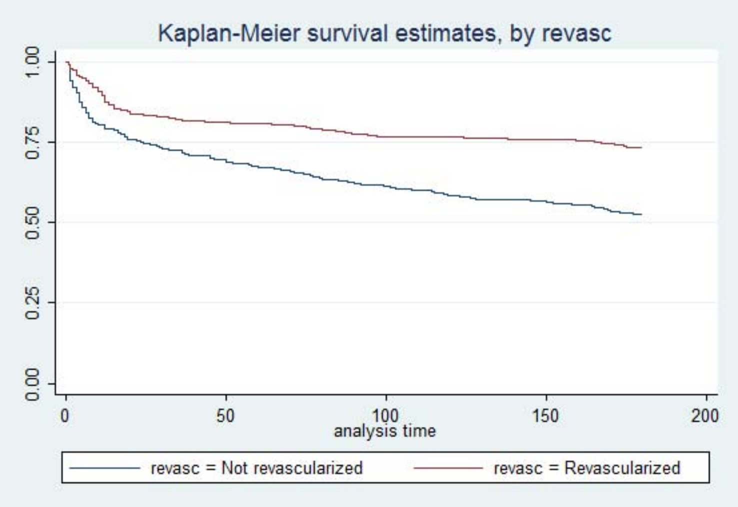
R Code for Kaplan-Meier Plot

The Kaplan-Meier plot can be as fancy as we want to make it. Here is some basic code, which you can make prettier (you'll want to do this if you end up making a lot of plots for work). Note that instead of working with a variable that takes value 1 for deaths and 0 for censoring, R wants the opposite (a variable that takes value 1 for censored and 0 for dead).

```
library(survival)
censor=1-death
res1=survfit(Surv(days,censor)~strata(revasc))
plot(res1)
#fancier plot below
plot(res1,lty=2:3,col=2:3,xlab="Days",ylab="Survival Probability")
```

The legend command can be used to add a legend to the plot as well.

KM Curve for GRACE Data



Is the risk of death lower among those who had revascularization?

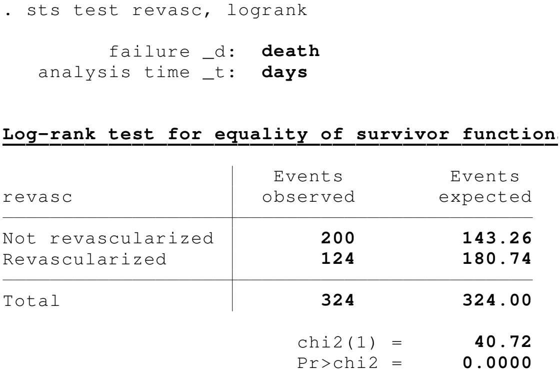
Log-Rank Test

How do we determine whether the difference in survival curves is statistically significant?

The log-rank test is quite intuitive. The idea behind it is to construct a 2 $\times$ 2 contingency table by group (revascularization or not for the GRACE data) at each time t at which a failure occurs. Then, the Mantel-Haenszel test statistic is used to test for differences between the two groups. For this test, the null hypothesis is that the survival curves in the two groups are the same, e.g.

$$H_0 : S_{revascularized}(t) = S_{not\ \ revascularized}(t).$$

Log-Rank Test for GRACE Data



```
. sts test revasc, logrank

         failure _d:  death
   analysis time _t:  days

Log-rank test for equality of survivor function

                     |   Events         Events
revasc               |  observed       expected
---------------------+--------------------------
Not revascularized   |       200         143.26
Revascularized       |       124         180.74
---------------------+--------------------------
Total                |       324         324.00

                          chi2(1) =      40.72
                          Pr>chi2 =     0.0000
```

Log-Rank Test in R

```
survdiff(Surv(days,censor)~revasc)
```