432 Class 15

https://thomaselove.github.io/432-2025/

2025-03-04

## Today’s Agenda

* Data which describe time to an event
  + The Survival Function, S(t)
  + Kaplan-Meier Estimation of the Survival Function
  + Creating Survival Objects in R
  + Drawing Survival Curves with **survminer**

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(janitor)  
library(conflicted)  
library(gt)  
library(here)  
library(mosaic)  
library(survival) ## new today  
library(survminer) ## new today  
library(easystats)  
library(tidyverse)  
  
theme\_set(theme\_bw())   
conflicts\_prefer(base::max, dplyr::filter)

# Introduction to Time-to-Event Data

## Working with Time to Event Data

In many medical studies, the main outcome variable is the time to the occurrence of a particular event.

* In a randomized controlled trial of cancer, for instance, surgery, radiation, and chemotherapy might be compared with respect to time from randomization and the start of therapy until death.

## Time-to-Event data

* In this case, the event of interest is the death of a patient, but in other situations it might be remission from a disease, relief from symptoms or the recurrence of a particular condition.
* Such observations are generally referred to by the generic term **survival data** even when the endpoint or event being considered is not death but something else.

## A First Example: Survival with AML

The aml data provided in the **survival** package describe survival times (in months) for 23 subjects with acute myelogenous leukemia.

Question: Should the standard course of chemotherapy be extended (maintained) for additional cycles or not?

* The study followed 11 subjects who received this maintenance chemotherapy, and 12 who did not, and tracked how long they survived (in months).
* But the problem was that 18 of the subjects were still alive at the time of their last follow-up.

## Data Ingest

I ingested the aml data from the **survival** package into a new tibble that I called aml\_432, with 23 rows and 4 columns…

aml\_432 <- survival::aml |> tibble() |>  
 rename(maintain = x, censored = status, months = time) |>  
 arrange(desc(months)) |>  
 mutate(subject = as.character(row\_number())) |>  
 relocate(subject)  
  
head(aml\_432)

# A tibble: 6 × 4  
 subject months censored maintain   
 <chr> <dbl> <dbl> <fct>   
1 1 161 0 Maintained   
2 2 48 1 Maintained   
3 3 45 0 Maintained   
4 4 45 1 Nonmaintained  
5 5 43 1 Nonmaintained  
6 6 34 1 Maintained

## The aml\_432 codebook

| aml\_432 | Description | *aml* |
| --- | --- | --- |
| subject | subject identifying code | – |
| months | known months of survival in follow-up period | *time* |
| censored | 1 if subject’s follow-up time was censored, else 0 | *status* |
| maintain | Chemotherapy was Maintained or “Nonmaintained” | *x* |

* If censored is 0, this means we know when the person died.

## What’s in a Time-to-Event Study?

Survival analysis is concerned with prospective studies. We start with a cohort of patients and follow them forwards in time to determine some clinical outcome.

* Follow-up continues until either some event of interest occurs, the study ends, or further observation becomes impossible.

## Time-to-Event (Survival) Outcomes

The outcomes in a survival analysis consist of the patient’s **fate** and **length of follow-up** at the end of the study.

* For some patients, the outcome of interest may not occur during follow-up.
* For such patients, whose follow-up time is *censored*, we know only that this event did not occur while the patient was being followed. We do not know whether or not it will occur at some later time.

## Two subjects from aml\_432

aml\_432 |> filter(subject %in% c(1, 8)) |> gt() |>   
 tab\_options(table.font.size = 24) |>   
 opt\_stylize(style = 2, color = "blue")

| subject | months | censored | maintain |
| --- | --- | --- | --- |
| 1 | 161 | 0 | Maintained |
| 8 | 31 | 1 | Maintained |

* Subject 1 was part of the Maintained group, and survived 161 months, at which point they died.
* Subject 8 was in the Maintained group, and survived at least 31 months, but was then censored[[1]](#footnote-29).

## Two more subjects from aml\_432

aml\_432 |> filter(subject %in% c(15, 22)) |> gt() |>   
 tab\_options(table.font.size = 24) |>   
 opt\_stylize(style = 2, color = "blue")

| subject | months | censored | maintain |
| --- | --- | --- | --- |
| 15 | 16 | 0 | Nonmaintained |
| 22 | 5 | 1 | Nonmaintained |

* Subject 15 was in the Nonmaintained group, and survived 16 months, and then died.
* Subject 22 was in the Nonmaintained group, and survived at least 5 months, but was then censored (so they were not followed further.)

## Problems with Time to Event Data

The primary problems are *censoring* and *non-Normality*…

1. At the completion of the study, some patients may not have reached the endpoint of interest (death, relapse, etc.). Consequently, the exact survival times are not known.
   * All that is known is that the survival times are greater than the amount of time the individual has been in the study.
   * The survival times of these individuals are said to be **censored** (precisely, they are right-censored).

## Problems with Time to Event Data

The primary problems are *censoring* and *non-Normality*…

1. Survival data are not symmetrically distributed. They will often appear positively skewed, with a few people surviving a very long time compared with the majority; so assuming a normal distribution will not be reasonable.

Next, we’ll define some special functions to build models that address these concerns.

## The Survival Function,

The **survival function**, (sometimes called the survivor function) is the probability that the survival time, , is greater than or equal to a particular time, .

* = proportion of people surviving to time or beyond

## If there’s no censoring, the survival function is easy to estimate

When there is no censoring, this function is easily estimated.

but this won’t work if there is censoring.

Even with censoring, the Kaplan-Meier approach essentially estimates the survival function by the number of patients alive at time divided by the total number of study subjects remaining at that time.

## Kaplan-Meier Estimator

The Kaplan-Meier estimator first orders the (unique) survival times from smallest to largest, then estimates the survival function at each unique survival time.

* The survival function at the second death time, is equal to the estimated probability of not dying at time conditional on the individual being still at risk at time .

## Kaplan-Meier Estimator

1. Order the survival times from smallest to largest, where is the th largest unique survival time, so we have…

## Kaplan-Meier Estimator

1. The Kaplan-Meier estimate of the survival function is

where is the number of people at risk just before , including those censored at time , and is the number of people who experience the event at time .

## Creating a Survival Object in R

The Surv() function, part of the survival package in R, will create a **survival object** from two arguments:

1. time = follow-up time
2. event = a status indicator, where
   * event = 1 or TRUE means the event was **observed** (for instance, the patient died)
   * event = 0 or FALSE means the follow-up time was **censored**

## The aml\_432 tibble

We have follow-up time in months, and we know if the event (death) occurred if censored is 0, and that the subject was censored if censored is 1.

| aml\_432 | Description |
| --- | --- |
| months | known months of survival in follow-up period |
| censored | 1 if subject’s follow-up time was censored, else 0 |

A little re-arranging might help us.

## Creating a death variable in aml\_432

aml\_432 <- aml\_432 |> mutate(death = 1 - censored)  
  
aml\_432 |> tabyl(death, censored) |> adorn\_title()

censored   
 death 0 1  
 0 0 18  
 1 5 0

OK, we have what we need.

* 18 subjects were censored so we only know they lived at least X months.
* 5 subjects died during the study, so we know their time to death.

## death and months relationship?

In our aml\_432 tibble,

* months is follow-up time, in months
* death = 1 if subject died, 0 if censored.

favstats(months ~ death, data = aml\_432) |>  
 gt() |> fmt\_number(columns = mean:sd, decimals = 1) |>   
 tab\_options(table.font.size = 24) |>   
 opt\_stylize(style = 2, color = "blue")

| death | min | Q1 | median | Q3 | max | mean | sd | n | missing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 5 | 9.75 | 23 | 32.5 | 48 | 23.1 | 14.1 | 18 | 0 |
| 1 | 13 | 16.00 | 28 | 45.0 | 161 | 52.6 | 61.9 | 5 | 0 |

## Building a Survival Object

surv\_aml <- Surv(time = aml\_432$months, event = aml\_432$death)  
  
head(surv\_aml, 4)

[1] 161 48+ 45 45+

* Subject 1 survived 161 months and then died.
* Subject 2 survived 48 months before being censored.
* Subject 3 survived 45 months and then died.
* Subject 4 survived 45 months before being censored.

Remember that only 5 of the 23 subjects died, the rest were censored at the last time they were seen for follow-up.

## Building a Kaplan-Meier Estimate

Remember that surv\_aml is the survival object we created.

km\_aml <- survfit(surv\_aml ~ 1, conf.int = 0.90)  
  
print(km\_aml, print.rmean = TRUE)

Call: survfit(formula = surv\_aml ~ 1, conf.int = 0.9)  
  
 n events rmean\* se(rmean) median 0.9LCL 0.9UCL  
[1,] 23 5 109 23.3 161 45 NA  
 \* restricted mean with upper limit = 161

* 5 events (deaths) occurred in 23 subjects.
* Restricted mean survival time is 109 months (upper limit 161?)
* Median survival time is 161 (why?) but has a lower bound for the 90% CI.

## Kaplan-Meier Estimate

summary(km\_aml)

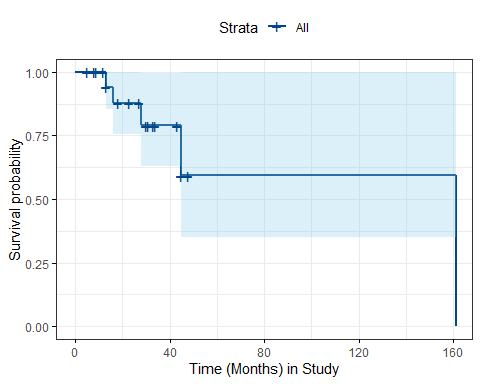
Call: survfit(formula = surv\_aml ~ 1, conf.int = 0.9)  
  
 time n.risk n.event survival std.err lower 90% CI upper 90% CI  
 13 17 1 0.941 0.0571 0.852 1.000  
 16 15 1 0.878 0.0807 0.755 1.000  
 28 10 1 0.791 0.1105 0.628 0.995  
 45 4 1 0.593 0.1902 0.350 1.000  
 161 1 1 0.000 NaN NA NA

* We started with 23 subjects.
* Up to 13 months, no one died, but 6 were censored (so 17 are at risk.) Estimated survival probability = 0.941
* By the next death at 16 months, only 15 people were still at risk. Estimated Pr(survival) now 0.878.

## Kaplan-Meier Plot, via survminer

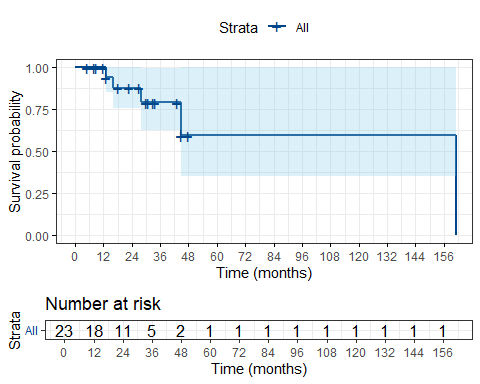
* The solid line indicates survival probability at each time point (in months)
* The crosses indicate time points where censoring has occurred.
* The steps down indicate events (deaths.)
* The shading indicates 95% pointwise confidence intervals.

ggsurvplot(km\_aml, data = aml\_432, conf.int = TRUE,   
 ggtheme = theme\_bw(), palette = "lancet",  
 xlab = "Time (Months) in Study", conf.int.fill = "skyblue")



## Adding a “Number at Risk” Table

ggsurvplot(km\_aml, data = aml\_432,  
 ggtheme = theme\_bw(), palette = "lancet",  
 conf.int = TRUE, # Add confidence interval  
 conf.int.fill = "skyblue",   
 risk.table = TRUE, # Add risk table  
 xlab = "Time (months)", # Adjust X axis label  
 break.time.by = 12 # X ticks every 12 months  
 )



## Save our aml\_432 data

write\_rds(aml\_432, here("c15/data/aml\_432.Rds"))

I use write\_rds() to do this, rather than other options. The main reason I prefer it is that it does not compress the file by default, and space is generally cheaper than time. It’s part of the **readr** package in the core tidyverse.

* Details at <https://readr.tidyverse.org/reference/read_rds.html>
* Posted the aml\_432.Rds file to [our 432-data page](https://github.com/THOMASELOVE/432-data).

## Where We Are So Far

* Created a small (n = 23) tibble, aml\_432.
* Observed 5 deaths, and 18 subjects censored before death.
* Survival object (containing time and fate) called surv\_aml
* Created Kaplan-Meier estimate of survival function, km\_aml.
* Plotted the Kaplan-Meier estimate with ggsurvplot().
* Added a number at risk table to our Kaplan-Meier curve

Now, let’s consider a potential predictor (maintenance status) of our time-to-event outcome.

## Comparing Survival, by Group

Suppose we want to compare the survival functions for subjects classified by their maintenance group…

aml\_432 |>   
 tabyl(maintain, death) |>   
 adorn\_totals(where = "col") |>   
 adorn\_title()

death   
 maintain 0 1 Total  
 Maintained 7 4 11  
 Nonmaintained 11 1 12

* In our sample, 4 of 11 in the “maintained” group and 1 of 12 in the “nonmaintained” group had the event (died).

## Survival Function, by Group

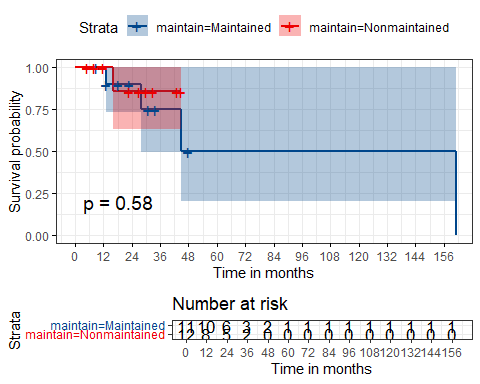
km\_aml\_grp <- survfit(surv\_aml ~ aml\_432$maintain)  
print(km\_aml\_grp, print.rmean = TRUE)

Call: survfit(formula = surv\_aml ~ aml\_432$maintain)  
  
 n events rmean\* se(rmean) median 0.95LCL  
aml\_432$maintain=Maintained 11 4 97.2 28.4 103 28  
aml\_432$maintain=Nonmaintained 12 1 140.3 19.2 NA NA  
 0.95UCL  
aml\_432$maintain=Maintained NA  
aml\_432$maintain=Nonmaintained NA  
 \* restricted mean with upper limit = 161

* 4 of 11 maintained subjects died; estimated restricted mean survival time is 97.2 months.
* 1 of 12 non-maintained subjects died, est. restricted mean survival = 140.3 months.

## Kaplan-Meier Plot, by Group

ggsurvplot(km\_aml\_grp, data = aml\_432,  
 ggtheme = theme\_bw(), palette = "lancet",  
 conf.int = TRUE,  
 xlab = "Time in months",  
 break.time.by = 12,  
 risk.table = TRUE,  
 risk.table.height = 0.25,   
 pval = TRUE)



## Next Time: Time-to-event data

* Testing the difference between two survival curves
  + log rank tests
  + Peto-Peto modification of the Gehan-Wilcoxon test
* Customizing a Kaplan-Meier plot
  + Plotting Cumulative Event Rates
* A larger example
* Classes 22-23: Cox models for regression on time-to-event outcomes.

1. perhaps because the study ended, or they were lost to the investigators [↑](#footnote-ref-29)