

Survival Analysis: Kaplan-Meier Method

CRP 245 Tutorial

Duke University Clinical Research Training Program

2026-01-11

Table of contents

Introduction	2
0.1 Study Context and Sample Description	2
0.2 Data Dictionary	2
1 Setup and Data Loading	3
1.1 Loading the Data	3
2 Exploration: Survival Distributions	4
3 Kaplan-Meier Survival Modeling	6
3.1 Question 1: Generating Survival Estimates	6
3.2 Question 2: The Kaplan-Meier Plot	7
4 Interpretation of Clinical Outcomes	8
4.1 Clinical Thresholds	8
4.2 Median Survival Time	9
Practice Checkpoint	10

Introduction

Learning Objectives

After completing this tutorial, you will be able to:

- **Define** censoring and understand its impact on survival analysis.
- **Calculate** Kaplan-Meier estimates for survival probability.
- **Visualize** survival distributions using professional KM curves with risk tables.
- **Interpret** 95% confidence intervals for survival estimates.
- **Explain** and determine median survival time.

0.1 Study Context and Sample Description

This analysis uses data from a randomized clinical trial comparing two treatments for advanced ovarian carcinoma (stages IIIB and IV). Understanding survival time is critical in oncology to evaluate treatment efficacy and counsel patients on prognosis.

Key Points About the Study Sample:

- **Population:** Patients with advanced ovarian cancer.
- **Intervention:** Comparing cyclophosphamide alone vs. cyclophosphamide plus adriamycin.
- **Outcome:** Time from randomization to death.

💡 Why Study Survival?

In many clinical trials, we don't observe the event of interest (e.g., death) for every patient by the time the study ends. Standard measures like "mean survival" cannot be calculated using simple averages because that would ignore patients who are still alive. The **Kaplan-Meier method** allows us to include data from every patient, whether they reached the event or not.

0.2 Data Dictionary

Table 1: Ovarian Cancer Study Variables

Variable	Description
futime	Follow-up time (days) - Primary Outcome
fustat	Survival status (1=Died, 0=Censored/Alive)

Variable	Description
age	Patient age (years)
resid.ds	Residual disease (1=No, 2=Yes)
rx	Treatment arm (0=Control, 1=Experimental)
ecog.ps	Performance status (1=Better, 2=Worse)

1 Setup and Data Loading

We begin by loading the specialized R packages required for survival analysis.

```
# Check for and install required packages
if (!requireNamespace("survival", quietly = TRUE)) install.packages("survival")
if (!requireNamespace("survminer", quietly = TRUE)) install.packages("survminer")
if (!requireNamespace("graphics", quietly = TRUE)) install.packages("graphics")

library(survival)
library(survminer)
library(graphics)
```

1.1 Loading the Data

We load the ovarian cancer trial dataset into a dataframe called `ovarian2`.

```
# Load the trial data
load(url("https://www.duke.edu/~sgrambow/crp241data/ovarian2.RData"))

# Examine sample characteristics
summary(ovarian2)
```

futime		fustat		age		resid.ds	
Min.	: 59.0	Min.	:0.0000	Min.	:38.89	Min.	:1.000
1st Qu.	: 368.0	1st Qu.	:0.0000	1st Qu.	:50.17	1st Qu.	:1.000
Median	: 476.0	Median	:0.0000	Median	:56.85	Median	:2.000
Mean	: 599.5	Mean	:0.4615	Mean	:56.17	Mean	:1.577
3rd Qu.	: 794.8	3rd Qu.	:1.0000	3rd Qu.	:62.38	3rd Qu.	:2.000
Max.	:1227.0	Max.	:1.0000	Max.	:74.50	Max.	:2.000

	rx	ecog.ps
Min.	:0.0	Min. :1.000
1st Qu.:	0.0	1st Qu.:1.000
Median	:0.5	Median :1.000
Mean	:0.5	Mean :1.462
3rd Qu.:	1.0	3rd Qu.:2.000
Max.	:1.0	Max. :2.000

Statistical Interpretation

1. **Sample Size:** $n = 26$ patients.
2. **Events:** Out of 26 patients, 12 died (`fustat = 1`) and 14 were censored (`fustat = 0`).
3. **Median Age:** 56.5 years.

Clinical Context: Censoring

A patient is “censored” if they are still alive at the end of the study or if they were lost to follow-up. We know they survived *at least* until their last contact, but we don’t know when they finally died. Survival analysis explicitly accounts for this “partial” information.

2 Exploration: Survival Distributions

Before modeling, we look at the raw distribution of follow-up times for both those who died and those who were censored.

```
# Distribution for Deaths
hist(ovarian2$futime[ovarian2$fustat == 1],
     main = "Patients who Died",
     xlab = "Days",
     col = "darkred",
     breaks = seq(0, 1400, by = 200))
```

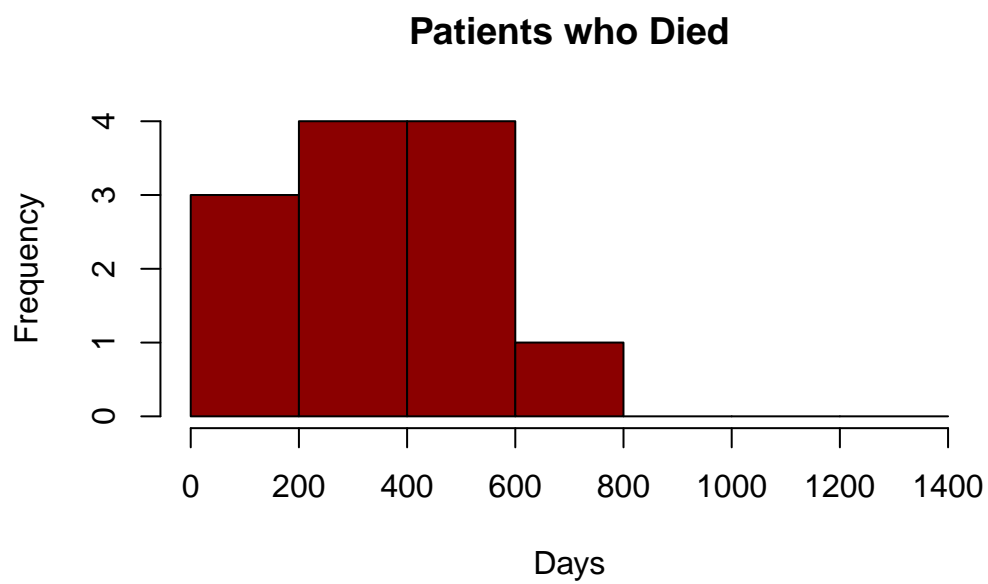


Figure 1: Distribution of follow-up times

```
# Distribution for Censored  
hist(ovarian2$futime[ovarian2$fustat == 0],  
      main = "Censored Patients",  
      xlab = "Days",  
      col = "darkblue",  
      breaks = seq(0, 1400, by = 200))
```

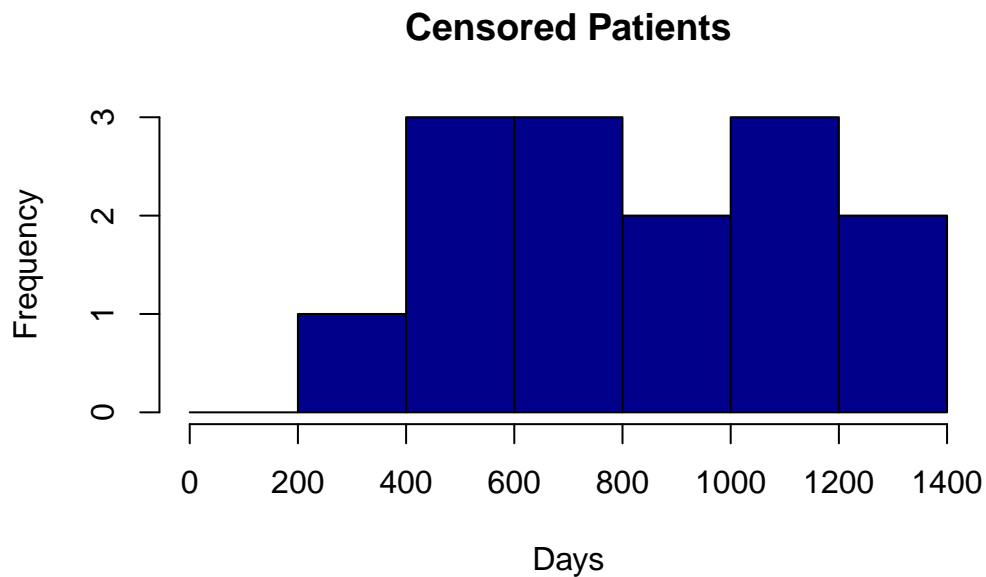


Figure 2: Distribution of follow-up times

! Clinical Insight

Notice that censoring often occurs later in the follow-up period. This suggests that patients who remained in the study longer were successfully tracked even if they didn't experience the event.

3 Kaplan-Meier Survival Modeling

3.1 Question 1: Generating Survival Estimates

How do we represent the probability of staying alive over time? The `survfit` function calculates these probabilities at every event time.

```
# Create the KM survival object
# ~1 indicates we are looking at the cohort as a whole
fit.km <- survfit(Surv(futime, fustat) ~ 1, data = ovarian2)
```

```
# Show survival estimates at intervals
summary(fit.km, times = c(0, 365, 730, 1095))
```

Call: `survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian2)`

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	26	0	1.000	0.000	1.000	1.000
365	20	7	0.731	0.087	0.579	0.923
730	10	5	0.497	0.105	0.328	0.752
1095	4	0	0.497	0.105	0.328	0.752

i Statistical Interpretation

- **At 1 Year (365 days):** The estimate is **73.1%**.
- **At 2 Years (730 days):** The estimate is **49.7%**.
- The **95% Confidence Intervals** represent our uncertainty about the true population survival rate based on this sample.

3.2 Question 2: The Kaplan-Meier Plot

Clinical Tool: The KM plot is the most common way to visualize prognosis across time.

```
ggsurvplot(fit.km, data = ovarian2,
            risk.table = TRUE,
            main = "Survival Probability: Advanced Ovarian Cancer",
            xlab = "Days Since Randomization",
            ylab = "Survival Probability",
            palette = "blue")
```

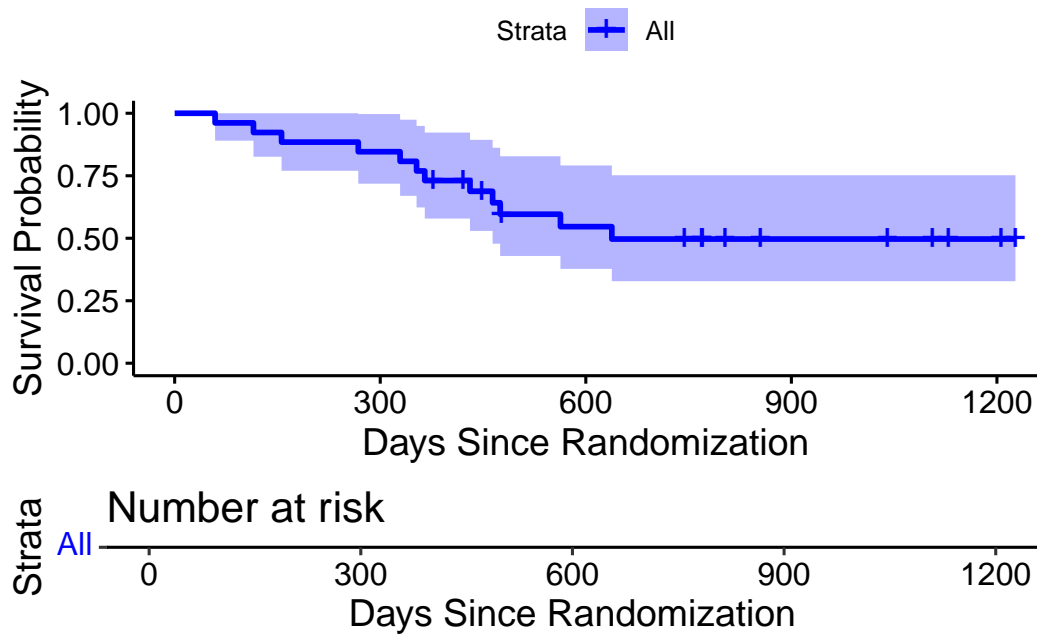


Figure 3: Standard Kaplan-Meier Survival Curve with Risk Table

! Critical Concept: The Risk Table

Always look at the **Number at Risk** table beneath the plot. As time passes, the “step-downs” in the curve are based on fewer and fewer patients. By 1000 days, only 2 patients remain at risk. This means the estimate at the far right of the plot is much less reliable than the estimate at the start.

4 Interpretation of Clinical Outcomes

4.1 Clinical Thresholds

What is the risk of early mortality, and what are the chances of long-term survival?

```
# Probability of surviving at least 2 years
prob_2yr <- summary(fit.km, times = 730)$surv
```

```
# Probability of dying before 6 months (180 days)
mort_6mo <- 1 - summary(fit.km, times = 180)$surv
```

- **2-Year Survival Probability:** 49.7%
- **6-Month Mortality Risk:** 11.5%

💡 Clinical Perspective

Knowing that approximately half of the patients survive to 2 years is a vital statistic for managing patient expectations and comparing this population to other cancer types or newer treatments.

4.2 Median Survival Time

Definition: The point in time where 50% of the population is expected to have experienced the event (died).

```
# Get median value
summary(fit.km)$table["median"]
```

```
median
638
```

```
# Plot with median indicator
ggsurvplot(fit.km, data = ovarian2,
            risk.table = TRUE,
            surv.median.line = "hv",
            palette = "darkgreen")
```

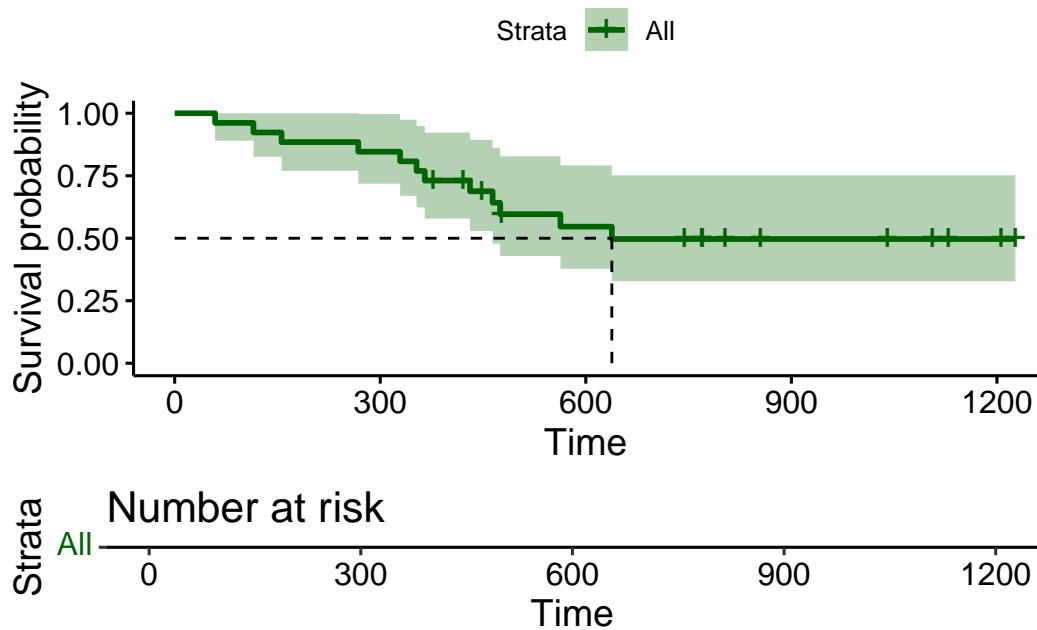


Figure 4: KM Curve highlighting Median Survival Time

i Statistical Analysis

- **Median survival:** 638 days (approx 1.75 years).
- **95% CI:** [464, NA].
- The upper bound of the confidence interval is **NA (Not Available)** because the survival curve in this sample did not drop low enough to determine the upper limit of the “50% mark” with 95% confidence.

Practice Checkpoint

! Key Teaching Points

- ☐ **Censoring** allows us to include “partial” survivors in our analysis.
- ☐ **KM Curves** always “step down” at events but stay flat at censoring times.
- ☐ **Risk Tables** tell you how many patients contribute to the estimate at any given time.

□ **Median Survival** is the standard “average person” outcome in cancer research.

This tutorial was developed for CRP 245 at Duke University.