Class 2: Kaplan-Meier

Yubo Rasmussen

# Introduction

This notebook demonstrates how to compute and visualize Kaplan-Meier estimates of the survival function using the survival package in R. Each example explores a different dataset and scenario to highlight the flexibility of Kaplan-Meier survival analysis.

## Example 1: Time to Discontinuation of a Medical Device

The data represent an experiment studying the time to discontinuation of a medical device. Observations include both complete and censored data, indicated by Censor.

### Load the Data

library(survival)  
  
# Time to discontinuation (in weeks)  
Time <- c(10,13,18,19,23,30,36,38,54,56,59,75,93,97,104,107,107,107)  
  
# Censoring indicator (1 = event occurred, 0 = censored)  
Censor <- c(1,0,0,1,0,1,1,0,0,0,1,1,1,1,0,1,0,0)  
  
# Create a survival object  
Surv(Time, Censor)

[1] 10 13+ 18+ 19 23+ 30 36 38+ 54+ 56+ 59 75 93 97 104+  
[16] 107 107+ 107+

**Questions and Answers**:

1. **What is the purpose of the Surv() function?** - The Surv() function creates a survival object, which encodes survival time and censoring information. This format is essential for downstream survival analysis methods like Kaplan-Meier or Cox regression.

2. **Why do we need the censoring indicator?** - The censoring indicator differentiates between observed events (1) and censored data (0). This ensures that censored observations are included in the analysis without incorrectly being treated as complete events.

### Kaplan-Meier Estimate

# Fit the Kaplan-Meier survival curve  
KM.Surv <- survfit(Surv(Time, Censor) ~ 1, conf.type = "plain")  
  
# Summary of the survival object  
summary(KM.Surv)

Call: survfit(formula = Surv(Time, Censor) ~ 1, conf.type = "plain")  
  
 time n.risk n.event survival std.err lower 95% CI upper 95% CI  
 10 18 1 0.944 0.0540 0.8386 1.000  
 19 15 1 0.881 0.0790 0.7267 1.000  
 30 13 1 0.814 0.0978 0.6220 1.000  
 36 12 1 0.746 0.1107 0.5290 0.963  
 59 8 1 0.653 0.1303 0.3972 0.908  
 75 7 1 0.559 0.1412 0.2827 0.836  
 93 6 1 0.466 0.1452 0.1816 0.751  
 97 5 1 0.373 0.1430 0.0927 0.653  
 107 3 1 0.249 0.1392 0.0000 0.522

**Questions and Answers**:

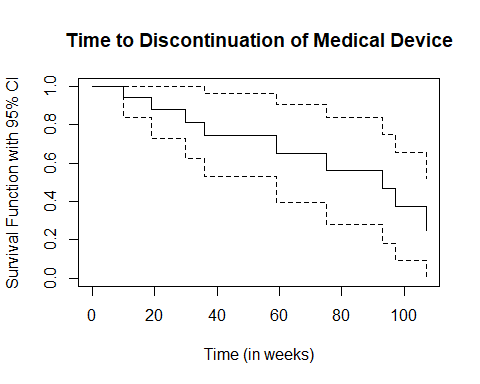
1. **What does ~ 1 mean in this context?** - The ~ 1 specifies that the survival analysis is performed without grouping by any covariates. It assumes all observations belong to a single group.

2. **What does conf.type = "plain" mean in this context?** - It specifies the method for calculating confidence intervals, in this case, the plain (non-logarithmic) method.

3. **What does the Kaplan-Meier estimate represent?** - It provides the probability of survival beyond specific time points, accounting for censoring.

### Plot the Survival Curve

plot(KM.Surv, main = "Time to Discontinuation of Medical Device",  
 xlab = "Time (in weeks)", ylab = "Survival Function with 95% CI")



**Questions and Answers**:

1. **Why is it important to visualize the Kaplan-Meier curve?** - Visualizing the survival curve helps interpret the probability of survival over time and identify trends, such as sudden drops.

2. **What do the confidence intervals indicate on the plot?** - The confidence intervals provide a range of uncertainty around the estimated survival probabilities.

## Example 2: Acute Myelogenous Leukaemia (AML) Data

This example uses AML data where: - 1 indicates a relapse occurred. - 0 indicates the patient left the study or was still in remission.

### Load and Explore the Data

# Load the aml dataset  
library(survival)  
data("aml")

Warning in data("aml"): data set 'aml' not found

attach(aml)  
  
# Create a survival object  
Surv(time, status)

[1] 9 13 13+ 18 23 28+ 31 34 45+ 48 161+ 5 5 8 8   
[16] 12 16+ 23 27 30 33 43 45

**Questions and Answers**:

1. **What does status represent in the AML dataset?** - It indicates whether a patient experienced a relapse (1) or was censored (0).

2. **How does the Kaplan-Meier estimate handle censored patients in this dataset?** - Censored patients contribute to the number at risk until their censoring time but do not affect the probability of survival beyond that point.

### Kaplan-Meier Estimate by Group

# Fit the Kaplan-Meier survival curve by group (x)  
fit <- survfit(Surv(time, status) ~ x, data = aml)  
  
# Summary of the survival object  
summary(fit)

Call: survfit(formula = Surv(time, status) ~ x, data = aml)  
  
 x=Maintained   
 time n.risk n.event survival std.err lower 95% CI upper 95% CI  
 9 11 1 0.909 0.0867 0.7541 1.000  
 13 10 1 0.818 0.1163 0.6192 1.000  
 18 8 1 0.716 0.1397 0.4884 1.000  
 23 7 1 0.614 0.1526 0.3769 0.999  
 31 5 1 0.491 0.1642 0.2549 0.946  
 34 4 1 0.368 0.1627 0.1549 0.875  
 48 2 1 0.184 0.1535 0.0359 0.944  
  
 x=Nonmaintained   
 time n.risk n.event survival std.err lower 95% CI upper 95% CI  
 5 12 2 0.8333 0.1076 0.6470 1.000  
 8 10 2 0.6667 0.1361 0.4468 0.995  
 12 8 1 0.5833 0.1423 0.3616 0.941  
 23 6 1 0.4861 0.1481 0.2675 0.883  
 27 5 1 0.3889 0.1470 0.1854 0.816  
 30 4 1 0.2917 0.1387 0.1148 0.741  
 33 3 1 0.1944 0.1219 0.0569 0.664  
 43 2 1 0.0972 0.0919 0.0153 0.620  
 45 1 1 0.0000 NaN NA NA

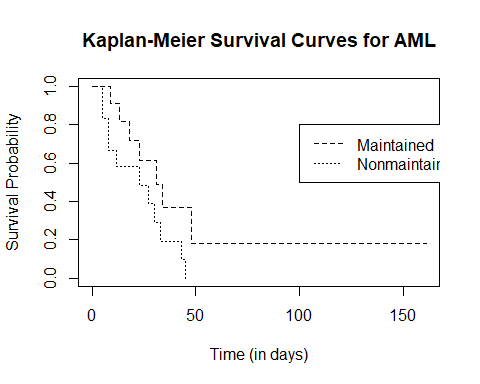
**Questions and Answers**:

1. **What does stratifying by x achieve in this analysis?** - Stratifying by x compares survival probabilities between groups, such as treatment versus control.

2. **What key differences might you expect between groups in this example?** - Differences in survival probabilities or patterns of relapse over time, potentially due to treatment efficacy.

### Plot the Survival Curve

plot(fit, lty = 2:3, main = "Kaplan-Meier Survival Curves for AML",  
 xlab = "Time (in days)", ylab = "Survival Probability")  
  
# Add legend  
legend(100, 0.8, c("Maintained", "Nonmaintained"), lty = 2:3)



**Questions and Answers**:

1. **What does the legend add to this plot?** - It clarifies which line corresponds to which group, aiding in interpretation.

2. **Why are different line types used for groups?** - Differentiating line types ensures clarity, especially in black-and-white visualizations.

## Example 3: Rat Drug Trial

This example involves survival times of rats in a drug trial. Observations include both events and censored data.

### Load the Data

# Time to event (in days)  
Time <- c(3,4,6,11,11,17,21,24,25,26,30,30,30,30,30)  
  
# Censoring indicator (1 = event occurred, 0 = censored)  
Censor <- c(1,1,0,1,1,0,1,0,0,1,0,0,0,0,0)  
  
# Ensure lengths match  
length(Time) == length(Censor)

[1] TRUE

# Create a survival object  
Surv(Time, Censor)

[1] 3 4 6+ 11 11 17+ 21 24+ 25+ 26 30+ 30+ 30+ 30+ 30+

**Questions and Answers**:

1. **Why is it important to check that the lengths of Time and Censor match?** - Mismatched lengths can result in errors or misinterpretations in the analysis.

2. **What assumptions are made about censored data in this example?** - Censoring is assumed to be non-informative, meaning it is unrelated to the survival times.

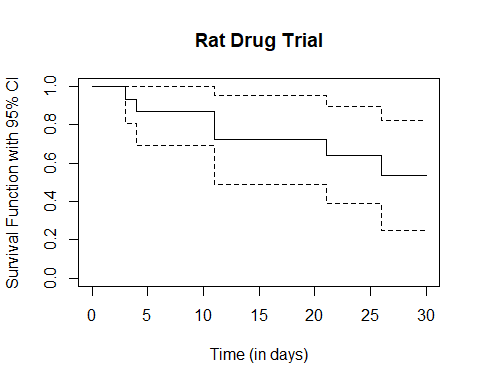
### Kaplan-Meier Estimate

# Fit the Kaplan-Meier survival curve  
KM.Surv <- survfit(Surv(Time, Censor) ~ 1, conf.type = "plain")  
  
# Summary of the survival object  
summary(KM.Surv)

Call: survfit(formula = Surv(Time, Censor) ~ 1, conf.type = "plain")  
  
 time n.risk n.event survival std.err lower 95% CI upper 95% CI  
 3 15 1 0.933 0.0644 0.807 1.000  
 4 14 1 0.867 0.0878 0.695 1.000  
 11 12 2 0.722 0.1185 0.490 0.954  
 21 9 1 0.642 0.1297 0.388 0.896  
 26 6 1 0.535 0.1457 0.249 0.820

### Plot the Survival Curve

plot(KM.Surv, main = "Rat Drug Trial",  
 xlab = "Time (in days)", ylab = "Survival Function with 95% CI")



**Questions and Answers**:

1. **What features of the survival curve might indicate the effect of the drug?** - Sharp declines in survival probability may indicate periods where many events occurred, suggesting the drug’s effectiveness or lack thereof.