IDS 702

Introduction to Survival Analysis

What is survival analysis?

Survival analysis refers to the analysis of **time-to-event** data.

A key feature of survival analysis is that **not everyone experiences the event of interest during the observation window**

- Does a new feature lead users to sign up for a premium subscription?
- Does income impact policy renewal?
- Does a new cancer treatment increase survival time?

Application	Event
Clinical research	Patient death
clinical research	Hospital discharge
Hardware	Equipment failure
Customer analytics	Churn
Customer analytics	First purchase
Unit economics	Break-even revenue
Product analytics	Adoption
Employee tenure	Resignation
Hiring	Accepted offer
Engineering	Bug ticket closed
Customer care	Support ticket closed
Loan repayment	Loan paid in full
Inventory	Stock depleted

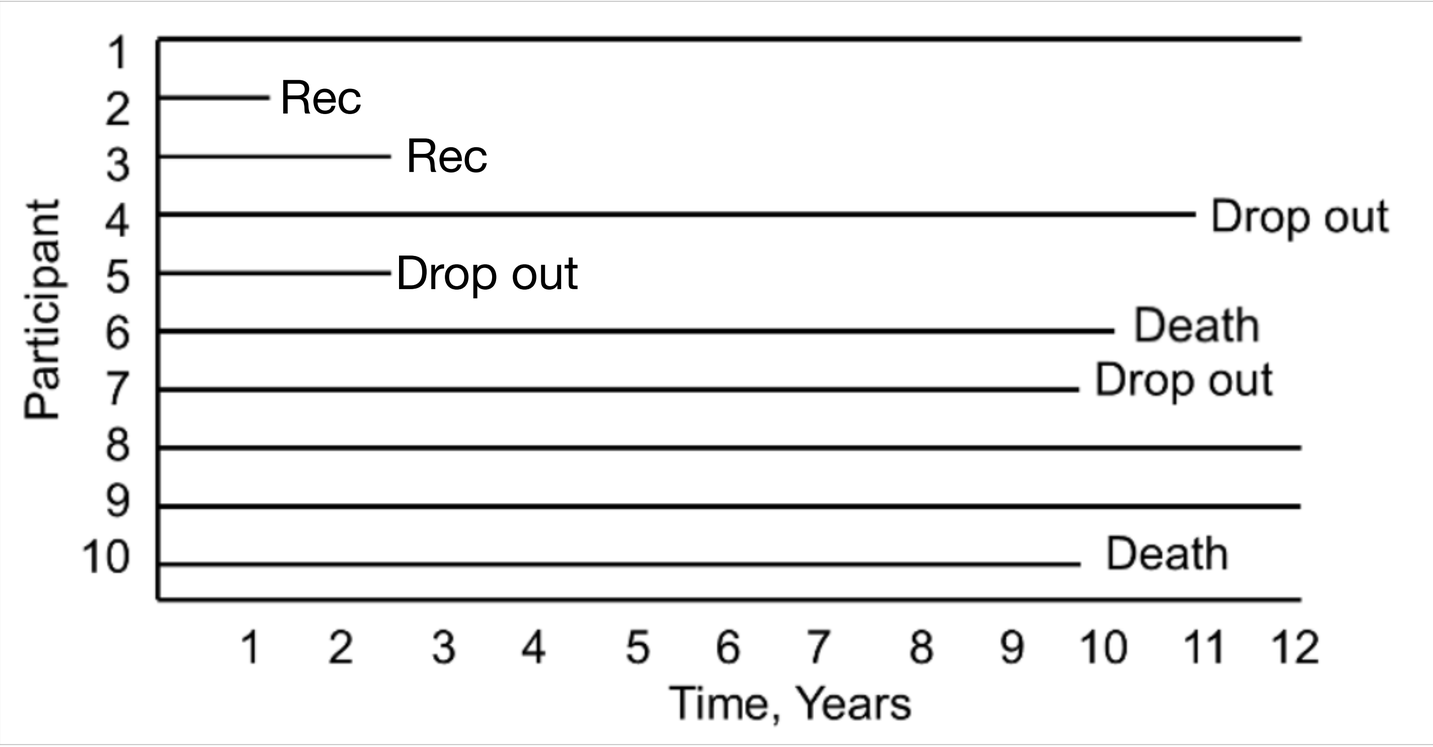
https://odsc.com/blog/when-to-use-survival-analysis-applications-in-industry-data-science/

Breast cancer data

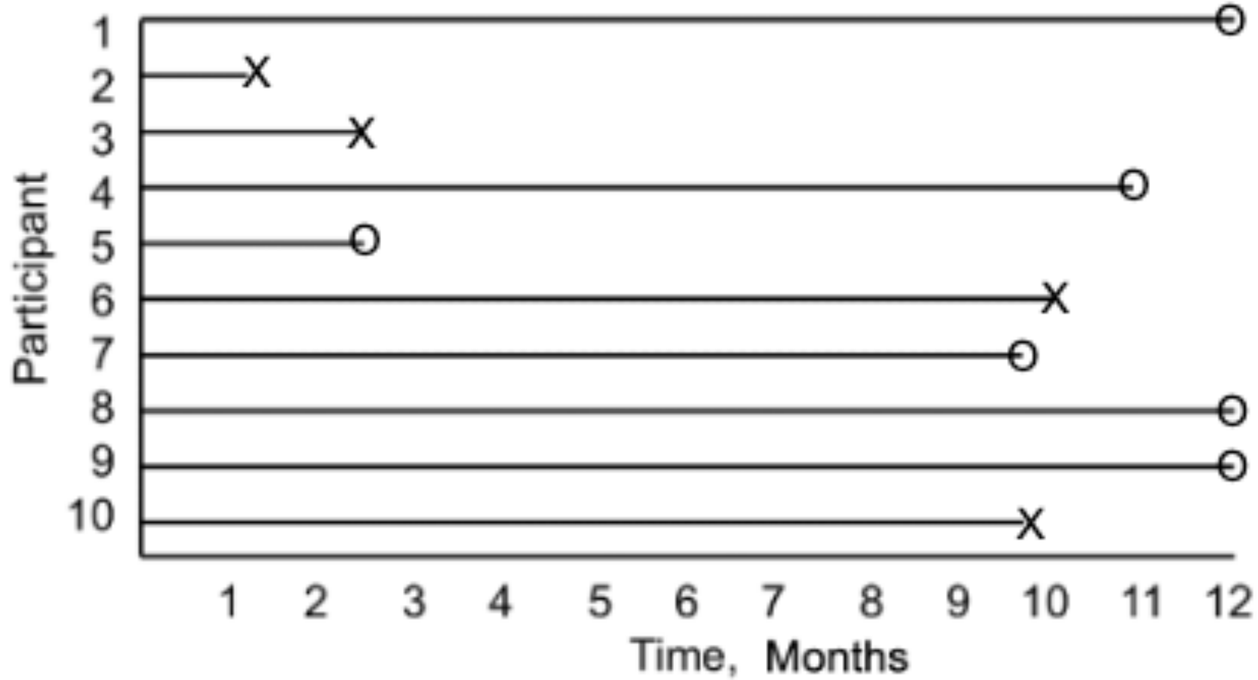
A dataset that contains patient records from a 1984-1989 trial conducted by the German Breast Cancer Study Group (GBSG) of 720 patients with node positive breast cancer. 686 patients have complete data for the prognostic variables.

hormon (1=received hormonal treatment)	rfstime (recurrence free survival time; days to first recurrence, death, or last follow-up)	Status (0=alive without recurrence, 1=recurrence or death)
0	1838	0
0	403	1
1	1855	0
0	842	1

Visualizing the structure of survival data



Visualizing the structure of survival data



Types of censoring

- Right censoring: event occurred after the observation time
- Left censoring: event occurred before the observation time
- Interval censoring: event occurred during an interval of time that was unobserved

Non-informative censoring

A key assumption in survival analysis is **non-informative censoring**: that the reason that the observation is censored is independent of the outcome

Kaplan-Meier Survival Method

Goal: We want to estimate the **survivor function**

The **survivor function** gives the probability that a patient will survive past time t

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

If T represents the time that death occurs, then $S(t)$ gives the probability that the patient survives past time t

Kaplan-Meier Survival Method

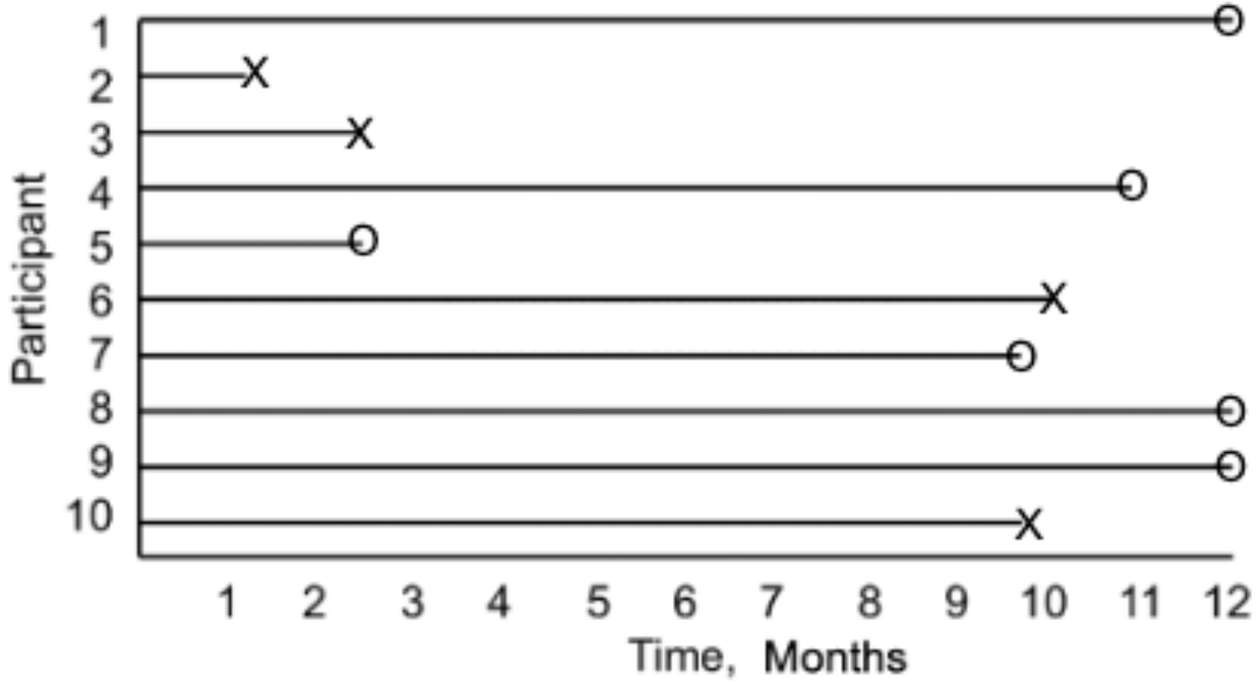
$$\hat{S}(t) = \Pi_{i:t_i \leq t}(1 - \frac{d_i}{n_i})$$

t_i: time when at least one death occurred

d_i: number of deaths at that time

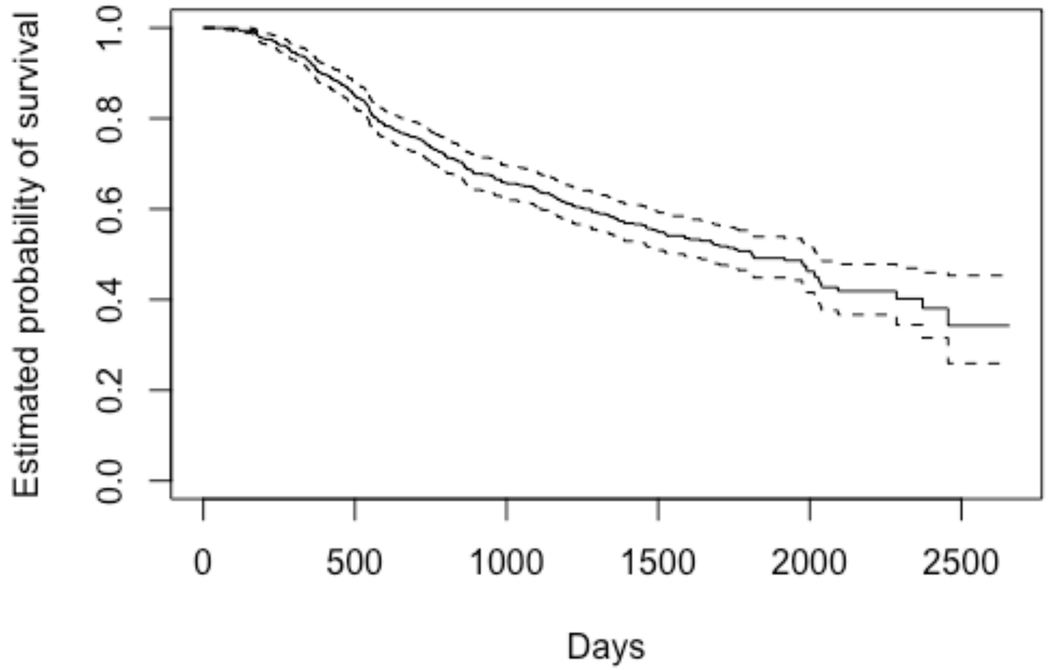
n_i: total number at risk at that time (accounts for censoring)

Kaplan-Meier Survival Method



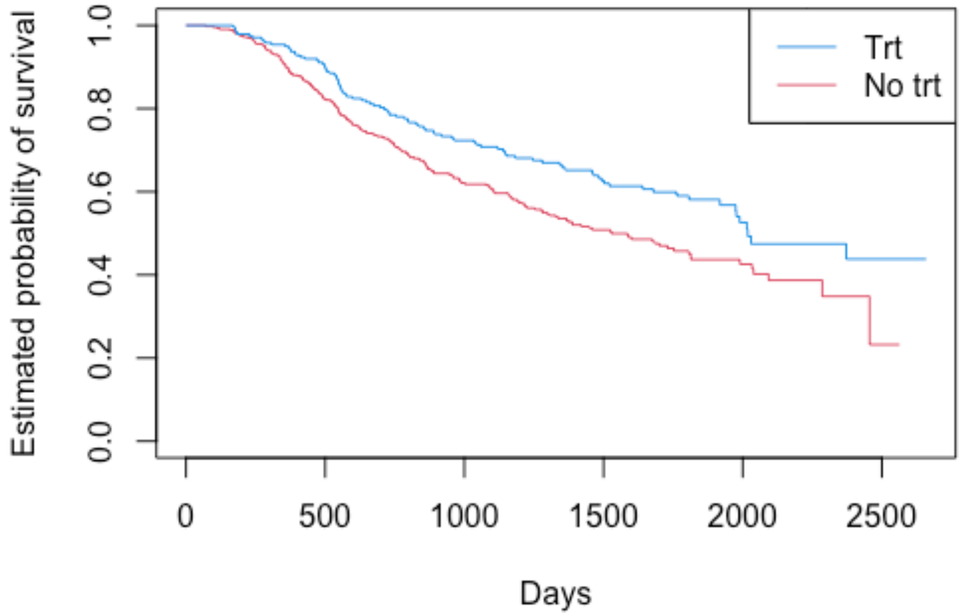
Kaplan-Meier curve in R

```r
{r}
library(survival)
attach(breastcancer)
fit.surv1 <- survfit(Surv(rfstime, status) ~ 1)
plot(fit.surv1, xlab="Days", ylab="Estimated probability of survival")
```



Kaplan-Meier curve in R

```
library(survival)
attach(breastcancer)
fit.surv1 <- survfit(Surv(rfstime, status) ~ factor(hormon))
plot(fit.surv1, xlab="Days", ylab="Estimated probability of survival",
     col=c(2,4))
legend("topright",c("Trt","No trt"),col=c(4,2),lty=1)
```



Log-rank test in R

```
> survdiff(Surv(rfstime, status) ~ factor(hormon))
Call:
survdiff(formula = Surv(rfstime, status) ~ factor(hormon))

                   N Observed Expected (O-E)^2/E (O-E)^2/V
factor(hormon)=0 440      205      180      3.37      8.56
factor(hormon)=1 246       94      119      5.12      8.56

 Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```