# PH 716 Applied Survival Analysis

## Part 2: Kaplan-Meier and Nelson-Aalen Estimators

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2026/02/17 10:59:27

---

### A motivating real-world study: gastric cancer clinical trial

We begin with data from a Phase II clinical trial evaluating XELOX chemotherapy given before surgery to patients with advanced gastric cancer and paraaortic lymph node metastasis (Wang et al., 2014).

The primary outcome is **progression-free survival**, defined as the time from study entry to disease progression or death, whichever occurs first. As is common in clinical trials, not all patients experienced progression during follow-up, resulting in **right-censored observations**.

The dataset contains follow-up times and the event indicator.

```
head(asaur::gastricXelox)
```

```
##   timeWeeks delta
## 1         4     1
## 2         8     1
## 3         8     1
## 4         8     1
## 5         9     1
## 6        11     1
```

From this study, researchers and clinicians naturally ask questions such as:

- What proportion of patients survive beyond a given time since the study entry?
- How does survival change over time in the presence of censoring?

### Notations

- $i$ : subject index, $i = 1, \ldots, n$
- $T_i$ : (authentic) survival time for subject $i$
- $C_i$ : censoring time for subject $i$
- $\widetilde{T}_i = \min(T_i, C_i)$: observed survival time for subject $i$
- $\Delta_i$: event indicator for subject $i$; $= 1$ if $\widetilde{T}_i = T_i$; $= 0$ if $\widetilde{T}_i = C_i$

### Assumptions

- $T_i$ is iid across $i$, i.e., $T_i \sim T$ for all $i$
- $T_i$ is independent of $C_i$

### Kaplan-Meier (KM) estimator

- To estimate $S_T(t)$ $(= S_{T_i}(t)$ for all $i)$ using no covariates

1

- Observed distinct authentic survival times: $t_1 < t_2 < \cdots < t_{n_D}$
  - $n_D$: # of distinct time points at which events are observed
- Recall for discrete survival time
  - $S_T(t) = \prod_{j:t_j \leq t}\{1 - \lambda_T(t_j)\}$
- KM estimator
  - $\widehat{S}_{T,KM}(t) = \prod_{j:t_j \leq t}\{1 - \hat{\lambda}_T(t_j)\}$
    * $\hat{\lambda}_T(t_j) = d_j/r_j$: an estimate of the (conditional) probability for an individual who survives up to time $t_j$ experiences the event at $t_i$, i.e., $\Pr(\text{event occurs in } [t_j, t_{j+1}) \mid T \geq t_j)$
      · $d_j$: # of events that happened exactly at time $t_j$
      · $r_j$: # of individuals at risk prior to $t_j$ (have not yet had an event or been censored prior to $t_j$)

---

- Ex. 2.1: Find the KM estimator for the data below, where the + sign denotes a right-censored subject:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\widetilde{T}_i$ | 2 | 5+ | 8 | 12+ | 15 | 21+ | 25 | 29 | 30+ | 34 |

- Risk table

| $j$ | $t_j$ | $r_j$ | $d_j$ | $d_j/r_j$ | $\widehat{S}_{KM}(t_j)$ |
|---|---|---|---|---|---|
| $-$ | 0 | 10 | 0 | 0 | $1 - 0 = 0$ |
| 1 | 2 | 10 | 1 | .1 | $1 \times (1 - .1) = .9$ |
| 2 | 8 | 8 | 1 | .125 | $.9 \times (1 - .125) = .787$ |
| 3 | 15 | 6 | 1 | .167 | $.787 \times (1 - .167) = .656$ |
| 4 | 25 | 4 | 1 | .25 | $.656 \times (1 - .25) = .492$ |
| 5 | 29 | 3 | 1 | .33 | $.492 \times (1 - .33) = .328$ |
| 6 | 34 | 1 | 1 | 1 | $.328 \times (1 - 1) = 0$ |

```r
library(survival)
ex21 = data.frame(
  time=c(2, 5, 8, 12, 15, 21, 25, 29, 30, 34),
  delta=c(1, 0, 1, 0, 1, 0, 1, 1, 0, 1)
)
km.ex21 = survfit(
  formula = Surv(time, delta)~1,
  data = ex21,
  conf.type = "logit") # type of confidence interval
summary(km.ex21)
```

## Pointwise confidence interval (CI) of survival probability

- Recall the 95% Wald CI of $\theta$:
$$\hat{\theta} \pm 1.96 \times \text{sd}(\hat{\theta})$$

- `conf.type="plain"`
$$\widehat{S}_{T,KM}(t) \pm 1.96 \times \text{sd}\{\widehat{S}_{T,KM}(t)\}$$
  - $\text{var}\{\widehat{S}_{T,KM}(t)\} \approx \{\widehat{S}_{T,KM}(t)\}^2 \sum_{j:t_j \leq t} d_j/\{r_j(r_j - d_j)\}$

- `conf.type="log"`
$$\ln \widehat{S}_{T,KM}(t) \pm 1.96 \times \text{sd}\{\ln \widehat{S}_{T,KM}(t)\}$$

  - $\left[\exp(\ln \widehat{S}_{T,KM}(t) - 1.96 \times \text{sd}\{\ln \widehat{S}_{T,KM}(t)\}), \exp(\ln \widehat{S}_{T,KM}(t) + 1.96 \times \text{sd}\{\ln \widehat{S}_{T,KM}(t)\})\right]$ is a 95% CI for $S_T(t)$
  - $\text{var}\{\ln \widehat{S}_{T,KM}(t)\} \approx \sum_{j:t_j \leq t} d_j/\{r_j(r_j - d_j)\}$
- `conf.type="log-log"`
$$\ln\{-\ln \widehat{S}_{T,KM}(t)\} \pm 1.96 \times \text{sd}[\ln\{-\ln \widehat{S}_{T,KM}(t)\}]$$

  - $\left[\exp\{-\exp(\ln\{-\ln \widehat{S}_{T,KM}(t)\} - 1.96 \times \text{sd}[\ln\{-\ln \widehat{S}_{T,KM}(t)\}])\}, \exp\{-\exp(\ln\{-\ln \widehat{S}_{T,KM}(t)\} + 1.96 \times \text{sd}[\ln\{-\ln \widehat{S}_{T,KM}(t)\}])\}\right]$ is a 95% CI for $S_T(t)$
  - $\text{var}[\ln\{-\ln \widehat{S}_{T,KM}(t)\}] \approx \{\widehat{S}_{T,KM}(t)\}^{-2} \sum_{j:t_j \leq t} d_j/\{r_j(r_j - d_j)\}$
- `conf.type="logit"`
$$\text{logit}\widehat{S}_{T,KM}(t) \pm 1.96 \times \text{sd}\{\text{logit}\widehat{S}_{T,KM}(t)\}$$

  - $\left[\text{logit}^{-1}(\text{logit}\widehat{S}_{T,KM}(t) - 1.96 \times \text{sd}\{\text{logit}\widehat{S}_{T,KM}(t)\}), \text{logit}^{-1}(\text{logit}\widehat{S}_{T,KM}(t) + 1.96 \times \text{sd}\{\text{logit}\widehat{S}_{T,KM}(t)\})\right]$ is a 95% CI for $S_T(t)$
- `conf.type="arcsin"`
$$\text{arcsin}\,\widehat{S}_{T,KM}(t) \pm 1.96 \times \text{sd}\{\text{arcsin}\,\widehat{S}_{T,KM}(t)\}$$

  - $\left[\sin(\text{arcsin}\,\widehat{S}_{T,KM}(t) - 1.96 \times \text{sd}\{\text{arcsin}\,\widehat{S}_{T,KM}(t)\}), \sin(\text{arcsin}\,\widehat{S}_{T,KM}(t) + 1.96 \times \text{sd}\{\text{arcsin}\,\widehat{S}_{T,KM}(t)\})\right]$ is a 95% CI for $S_T(t)$
- `log-log`, `logit`, `arcsin` leading to the confidence interval guaranteed to be inside $[0, 1]$

---

- Visualization of KM estimator

```r
# A plain way
plot(km.ex21)
# A more fancy way
survminer::ggsurvplot(
  km.ex21,
  xlab="Time",
  xlim=c(0,40),
  conf.int = T,
  conf.int.style="step",
  censor=T,
  legend.labs = c("Entire Cohort"),
  risk.table = F,
  cumevents = F,
  tables.height = 0.15
)
```

---

- Based on the KM estimator for Ex. 2.1, clinicians may ask:
  - What is the estimated survival probability at time 10?
  - What is the estimated survival probability at time 15?
  - What is the estimated survival probability at time 20?
  - What is the estimated median event-free survival?

## Properties of KM estimator

- $\widehat{S}_{T,KM}(t)$ is a right-continuous step function, approximating the (likely smooth) $S_T(t)$

- $\widehat{S}_{T,KM}(t)$ is a consistent (but typically biased) estimator of $S_T(t)$

  - As $n$ increases, $\widehat{S}_{T,KM}(t)$ becomes less jagged
  - The bias vanishes when there is no censoring, stemming from the possibility that the last survivor becomes censored.

- Note that $\widehat{S}_{T,KM}(t)$ has $n_D$ jumps

  - One jump at each distinct failure time
  - No jump at the censored times (why?)

- $\widehat{S}_{T,KM}(t)$ is well-defined (it can be specified) up to the last observed time $\max\{\widetilde{T}_1,\ldots,\widetilde{T}_n\}$

  - One cannot estimate $S_T(t)$ for times $\max\{\widetilde{T}_1,\ldots,\widetilde{T}_n\}$ using the KM procedure
  - Because no data available in the sample beyond time $\max\{\widetilde{T}_1,\ldots,\widetilde{T}_n\}$

- If last survivor is censored, KM estimator will NOT drop down to 0

- KM-estimated survival changes over time because of events, but the precision of survival estimates changes because of censoring, because

  - Censoring does NOT
    * Cause the survival curve to drop
    * Directly change survival probability
  - Censoring DOES
    * Reduce the number at risk $r_{ij}$ at later times
    * Increase variability (wider confidence intervals)
    * Make tail estimates less reliable

- Wide confidence intervals at later time points reflect fewer patients remaining at risk

  - Since $\mathrm{var}(d_j/r_j) \approx d_j/\{r_j(r_j - d_j)\}$

## Nelson-Aalen(-Altschuler-Fleming-Harrington) estimator

- To estimate $S_T(t)$ $(= S_{T_i}(t)$ for all $i)$ using no covariates

- Observed distinct authentic survival times: $t_1 < t_2 < \cdots < t_{n_D}$

  - $n_D$: # of distinct time points at which events are observed

- Estimating the cumulative hazard instead

  - Recall for discrete times, $\Lambda_T(t) = \sum_{j:t_j \le t} \lambda_T(t)$
  - $\widehat{\Lambda}_{T,NA}(t) = \sum_{j:t_j \le t} \hat{\lambda}_T(t_j) = \sum_{j:t_j \le t} d_j/r_j$

- Further estimating the survival function

  - Recall for continuous times, $S_T(t) = \exp\{-\Lambda_T(t)\}$
  - $\widehat{S}_{T,NA}(t) = \exp\{-\widehat{\Lambda}_{T,NA}(t)\} = \exp(-\sum_{j:t_j \le t} d_j/n_j)$

- Asymptotically equivalent to KM

  - KM and NA give the same estimator as $n \to \infty$

---

- Revisit Ex. 2.1: Find the NA estimator for the data below, where the + sign denotes a right-censored subject:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\widetilde{T}_i$ | 2 | 5+ | 8 | 12+ | 15 | 21+ | 25 | 29 | 30+ | 34 |

```r
ex21 = data.frame(
  time=c(2, 5, 8, 12, 15, 21, 25, 29, 30, 34),
  delta=c(1, 0, 1, 0, 1, 0, 1, 1, 0, 1)
)
na.ex21 = survival::survfit(
  formula=survival::Surv(time, delta)~1,
  data=ex21,
  conf.type="log-log",
  type = 'fh') # NA estimator
summary(na.ex21)
plot(na.ex21)
```

---

- Based on the NA estimator for Ex. 2.1, clinicians may ask:
  - What is the estimated survival probability at time 10?
  - What is the estimated survival probability at time 15?
  - What is the estimated survival probability at time 20?
  - What is the estimated median event-free survival?

## Revisiting the motivating real-world study: gastric cancer clinical trial

We now return to the Phase II clinical trial evaluating XELOX chemotherapy in patients with advanced gastric cancer. The primary endpoint was progression-free survival, defined as time from study entry to disease progression or death.

Earlier, we posed several scientific questions such as:

- What proportion of patients survive beyond a given time since the study entry?
- How does survival change over time in the presence of censoring?

We now use the KM/NA estimator to answer these questions.

```r
library(survival)
library(asaur)
library(survminer)
data_gastric <- asaur::gastricXelox
km_gastric <- survfit(
  formula = Surv(timeWeeks, delta) ~ 1
  ,data = data_gastric,
  ,conf.type = "log-log"
)
na_gastric <- survfit(
  formula = Surv(timeWeeks, delta) ~ 1
  ,data = data_gastric,
  ,conf.type = "log-log"
  ,type = 'fh'
)
summary(km_gastric)
summary(na_gastric)
survminer::ggsurvplot(
  fit = list("KM" = km_gastric, "NA" = na_gastric),
  xlab = "Time",
  xlim = c(0, 100),
  conf.int = TRUE,
  conf.int.style = "step",
```

```
  censor = TRUE,
  legend.labs = c("Kaplan-Meier", "Fleming-Harrington"),
  risk.table = FALSE,
  cumevents = FALSE,
  tables.height = 0.15,
  combine = TRUE # two curves in one plot
)
```

- From the risk table/estimated survival curve, we can:
  - Estimate the probability of remaining progression-free at specific time points (e.g., 6 or 12 months).
  - Identify the median progression-free survival
  - See survival appear to decline rapidly or gradually over time
    * Early drops in the survival curve indicate periods of higher progression risk
    * A gradual decline suggests more prolonged disease control