Estimation with Event-Time Data

Dave Harrington

May 14 - 18, 2018

The Kaplan-Meier estimator

Estimating standard errors

The cumulative hazard estimator

Derivations

The Kaplan-Meier estimator

APPROACHES TO ESTIMATING $S(t)$

- Parametric models and maximum likelihood
- *The non-parametric Kaplan-Meier (KM) estimate*
  - KM also called the product limit estimator because of original derivation

THE KAPLAN-MEIER ESTIMATOR: GENERAL IDEA

The Kaplan-Meier estimator is probably the most popular approach.

When there is no censoring, the general formula is:

$$\widehat{S}(t) = \frac{\#\ \textit{individuals with}\ T > t}{\textit{total sample size}}$$

AN EXAMPLE: COX AND OAKES, NO CENSORING

Time to relapse (weeks) for 21 leukemia patients receiving control treatment[1]:

- 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

What is $\widehat{S}(10) = \widehat{P}(T > 10)$, the probability that an individual survives more than 10 weeks?

- This is $8/21 = 0.38$ since 8 people survive more than 10 weeks.

What about $\widehat{S}(8)$?

- $\widehat{S}(8) = \widehat{P}(T > 8) = 8/21 = 0.38$
- The four events at $t = 8$ are counted as having already failed.

---

[1] Table 1.1 of Cox & Oakes, 1984

EMPIRICAL SURVIVAL FUNCTION

When there is no censoring, the general formula is:

$$\widehat{S}(t) = \frac{\# \text{ individuals with } T > t}{\text{total sample size}}$$

What is the standard error of $\widehat{S}(t)$?

- When there is no censoring, the estimated survival function is a proportion $\hat{p}$ with the standard error:

$$\text{s.e.}[\widehat{S}(t)] = \sqrt{p(1-p)/n}$$

Example: $\text{s.e.}[\widehat{S}(8)] = \sqrt{(0.38)(0.62)/21} = 0.106$

A TABLE OF $\widehat{S}(t)$

Values of t	# individuals with $T > t$	$\widehat{S}(t)$
$0 \leq t < 1$	21	21/21=1.000
$1 \leq t < 2$	19	19/21=0.905
$2 \leq t < 3$	17	17/21=0.809
$3 \leq t < 4$	16	16/21=0.762
$4 \leq t < 5$	14	14/21=0.667
$5 \leq t < 8$	12	12/21=0.571
$8 \leq t < 11$	8	8/21=0.381
$11 \leq t < 12$	6	6/21=0.286
$12 \leq t < 15$	5	4/21=0.191
$15 \leq t < 17$	3	3/21=0.143
$17 \leq t < 22$	2	2/21=0.095
$22 \leq t < 23$	1	1/21=0.048

WHAT ABOUT CENSORING?

Consider time to relapse (weeks) for leukemia patients in the treatment group.[2] Times with $^+$ are right censored:

$$6^+, 6, 6, 6, 7, 9^+, 10^+, 10, 11^+, 13, 16, 17^+$$

$$19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$$

Naturally, $\widehat{S}(6-) = 21/21$

- because everyone survived until at least time 6 or greater

Not right to claim $\widehat{S}(6) = 17/21$

- due to unknown status of person censored at time 6

---

[2]Table 1.1 of Cox and Oakes

Censoring with the Kaplan-Meier

In a 1958 paper in the *Journal of the American Statistical Association*, Kaplan and Meier proposed a way to nonparametrically estimate $S(t)$, in the presence of censoring.

The method is based on the ideas of *conditional probability*.

Censoring and the KM estimator

$S(t)$ in the discrete case:

To estimate $S(t)$ for time t within the interval t_k and t_{k+1}, e.g. $t_k \leq t < t_{k+1}$, consider the intervals defined by the ordered k failure times,

$$[t_0, t_1), [t_1, t_2), \ldots, [t_{k-1}, t_k), [t_k, \infty)$$

The KM estimate is constructed based on events within each interval $[t_j, t_{j+1})$

- d_j is the number of deaths in the interval $[t_j, t_{j+1})$

- r_j is the number of individuals at risk in the interval $[t_j, t_{j+1})$

Initial assumptions: $t_0 = 0$, $P(T > t_0) = 1$.

Censoring and the KM: discrete case

Then,

$$
\begin{aligned}
S(t) = P(T > t) &= P(T > t_k) \\
&= P(T > t_1, T > t_2, \ldots, T > t_k) \\
&= P(T > t_1) \times \prod_{j=2}^{k} P(T > t_j | T > t_{j-1}) \\
&\overset{(*)}{=} \prod_{j=1}^{k} [1 - P(T = t_j | T > t_{j-1})] = \prod_{j=1}^{k} [1 - \lambda_j] \\
\text{so} \quad \widehat{S}(t) &\cong \prod_{j=1}^{k} \left(1 - \frac{d_j}{r_j}\right) = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)
\end{aligned}
$$

(*) Initial assumptions: $t_0 = 0$, $P(T > t_0) = 1$.

Censoring and the KM: continuous case

For continuous data, the Kaplan-Meier estimator of the survivorship function $S(t) = P(T > t)$ is

$$\widehat{S}(t) = \prod_{j:\tau_j \leq t} \frac{r_j - d_j}{r_j} = \prod_{j:\tau_j \leq t} \left(1 - \frac{d_j}{r_j}\right), \text{ where}$$

- $\tau_1, \ldots, \tau_K$ are the K distinct death times observed
- d_j is the number of deaths at τ_j
- r_j is the number of individuals "at risk" right before the j-th death time (everyone dead or censored *at or after* that time).
    - $r_j = r_{j-1} - d_{j-1} - c_{j-1}$
    - Alternatively, $r_j = \sum_{l \geq j}(c_l + d_l)$
- c_j is the number of censored observations between the j-th and $(j+1)$-th death times.
    - Censorings tied at τ_j are included in c_j

Most widely used software packages (SAS, Stata, R) have modules for survival analysis.

We will focus on R since it is free and has very good survival routines written by Terry Therneau.

Fitting a Kaplan-Meier in R

```r
library(survival)
library(eventtimedata)
data("cox.oakes.leukemia")
leukemia.remission <- survfit(Surv(time, relapse) ~ group,
                              data = cox.oakes.leukemia)
plot(leukemia.remission, lty = 2:3, mark.time = TRUE, xlab = "Weeks",
     ylab = "Survival Function" )
```
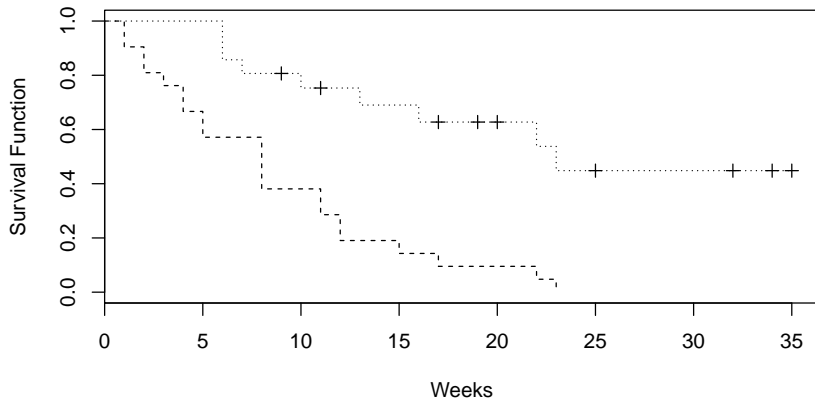
Numerical output

```r
library(survival)
library(eventtimedata)
print(leukemia.remission)
```

```
## Call: survfit(formula = Surv(time, relapse) ~ group, data = cox.oake
##
##          n events median 0.95LCL 0.95UCL
## group=0 21     21      8       4      12
## group=1 21      9     23      16      NA
```

```r
leukemia.group.0 = subset.data.frame(cox.oakes.leukemia, group == 0)
km.group.0 = survfit(Surv(time, relapse) ~ 1, data = leukemia.group.0)

summary(km.group.0)
```

```
## Call: survfit(formula = Surv(time, relapse) ~ 1, data = leukemia.group.0)
##
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    1     21       2   0.9048  0.0641      0.78754        1.000
##    2     19       2   0.8095  0.0857      0.65785        0.996
##    3     17       1   0.7619  0.0929      0.59988        0.968
##    4     16       2   0.6667  0.1029      0.49268        0.902
##    5     14       2   0.5714  0.1080      0.39455        0.828
##    8     12       4   0.3810  0.1060      0.22085        0.657
##   11      8       2   0.2857  0.0986      0.14529        0.562
##   12      6       2   0.1905  0.0857      0.07887        0.460
##   15      4       1   0.1429  0.0764      0.05011        0.407
##   17      3       1   0.0952  0.0641      0.02549        0.356
##   22      2       1   0.0476  0.0465      0.00703        0.322
##   23      1       1   0.0000     NaN           NA           NA
```

KM NUMERICAL ESTIMATES, GROUP $== 1$

```
leukemia.group.1 = subset.data.frame(cox.oakes.leukemia, group == 1)
km.group.1 = survfit(Surv(time, relapse) ~ 1, data = leukemia.group.1)

summary(km.group.1)

## Call: survfit(formula = Surv(time, relapse) ~ 1, data = leukemia.group.1)
##
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    6     21       3    0.857  0.0764        0.720        1.000
##    7     17       1    0.807  0.0869        0.653        0.996
##   10     15       1    0.753  0.0963        0.586        0.968
##   13     12       1    0.690  0.1068        0.510        0.935
##   16     11       1    0.627  0.1141        0.439        0.896
##   22      7       1    0.538  0.1282        0.337        0.858
##   23      6       1    0.448  0.1346        0.249        0.807
```

Subsets used here only to fit output on slides.

summary(leukemia.remission) prints values for both groups.

Estimating standard errors

Pointwise confidence intervals for the KM

Why *pointwise*?

- Since the KM is a function of time, there is an estimate of the standard error (or the variance) at each time.

Greenwood's formula is the most commonly used estimate of the KM standard error.

$$\widehat{\text{var}}(\widehat{S}(t)) = [\widehat{S}(t)]^2 \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j - d_j)r_j}$$

Derivation given later in the slides.

Confidence intervals for the KM

A 95% confidence interval could be based on

$$\widehat{S}(t) \pm z_{1-\alpha/2} \times \text{s.e.}[\widehat{S}(t)],$$

with s.e.$[\widehat{S}(t)]$ estimated using Greenwood's formula.

- However, this approach can yield values > 1 or < 0.

The better approach is to use the *log-log* transformation and base
intervals around

$$L(t) = \log[-\log[S(t)]]$$

In R, use the option `conf.type = "log-log"`. The default transformation in R
is $L(t) = -\log[S(t)]$.

Confidence intervals . . .

To transform back, use $S(t) = \exp[-\exp[L(t)]]$.

Since. . .

- $0 \leq S(t) \leq 1$,

- $0 \leq -\log[S(t)] < \infty$, and

- $-\infty < \log[-\log[S(t)]] < \infty$,

the confidence interval will be in the proper range when transformed back.

LOG-LOG APPROACH FOR CONFIDENCE INTERVALS:

1. Define $L(t) = \log[-\log[S(t)]]$.

2. Form a 95% confidence interval for $L(t)$, $(\widehat{L}(t) - A, \widehat{L}(t) + A)$, with $A = 1.96 \times$ s.e.$[\widehat{L}(t)]$.

3. Apply $S(t) = \exp[-\exp[L(t)]]$ to obtain the confidence bounds for the 95% CI on $S(t)$,

$$\left(\exp[-e^{(\widehat{L}(t)+A)}], \exp[-e^{(\widehat{L}(t)-A)}] \right)$$

4. Substituting $\widehat{L}(t) = \log[-\log[\widehat{S}(t)]]$ back into the above bounds yields confidence bounds of

$$\left([\widehat{S}(t)]^{e^A}, [\widehat{S}(t)]^{e^{-A}} \right)$$

Confidence intervals for median survival

The median is usually defined as

$$q_{0.5} = \min\{t_j : \widehat{S}(t) \geq 0.5\}.$$

Other quantiles are defined similarly.

Confidence limits for median survival are based on confidence intervals for $S(t)$.

R uses the method due to Brookmeyer and Crowley (Biometrics 1982, 38, 29–41).
- SAS and other packages use this as well.

The formulas are complex and not shown here.

The cumulative hazard estimator

ESTIMATING $S(t)$ VIA THE NELSON-AALEN CUMULATIVE HAZARD

The cumulative hazard $\Lambda(t)$ can be approximated by a sum over j intervals,

$$\Lambda(t) \approx \sum_j \lambda_j \Delta$$

where

- λ_j is the value of the hazard in the jth interval
- Δ is the width of each interval

Since $\widehat{\lambda}_j \Delta$ is approximately the probability of having an event in an interval j, conditional on having survived until the beginning of the interval, $\Lambda(t)$ can be approximated further as

$$\Lambda(t) \approx \sum_j \lambda_j \Delta \approx \sum_j \frac{d_j}{r_j}$$

Thus, the *Nelson-Aalen estimator* can be written as

$$\widehat{\Lambda}_{NA}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j}$$

From $\widehat{\Lambda}_{NA}(t)$, an alternative to the KM estimator of $S(t)$ can be calculated:

$$\widehat{S}_{FH}(t) = \exp[-\widehat{\Lambda}_{NA}(t)]$$

The *Fleming-Harrington estimator* is generally very close to $\widehat{S}_{KM}(t)$.

EXAMPLE: TIME TO RECIDIVISM, ROSSI (1980)

Recidivism is the event of rearrest and reincarceration after release from prison.

A randomized study[3] with 52 weeks of follow-up after randomization collected information on the following variables:

- `fin`: Financial support vs no financial support after release

- `week`: Time in weeks to either re-arrest or censoring

- `arrest`: $1 = $ arrest during the follow-up, $0 = $ no arrest

---

[3] rossi dataset in `eventtimedata` package.

```r
library(survival)
library(eventtimedata)
data("rossi")
rossi.recidivism.km <- survfit(Surv(week, arrest) ~ fin,
                               data = rossi)
plot(rossi.recidivism.km, lty = 2:3, col = 3:4, mark.time = TRUE,
     xlab = "Weeks",
     ylab = "Probability of Recidivism",
     axes = FALSE,
     conf.times = c(10,30,50),
     main = "KM of Recidivism Probability, with Conf. Int.",
     cex = 0.6, cex.main = 0.8)
axis(1)
axis(2)
legend(10, .5, c("Financial Support", "No Financial Support"),
       lty = 2:3, col = 3:4, cex = 0.6)
```

KM of Recidivism Probability, with Conf. Int.

Probability of Recidivism

- - - Financial Support
······ No Financial Support
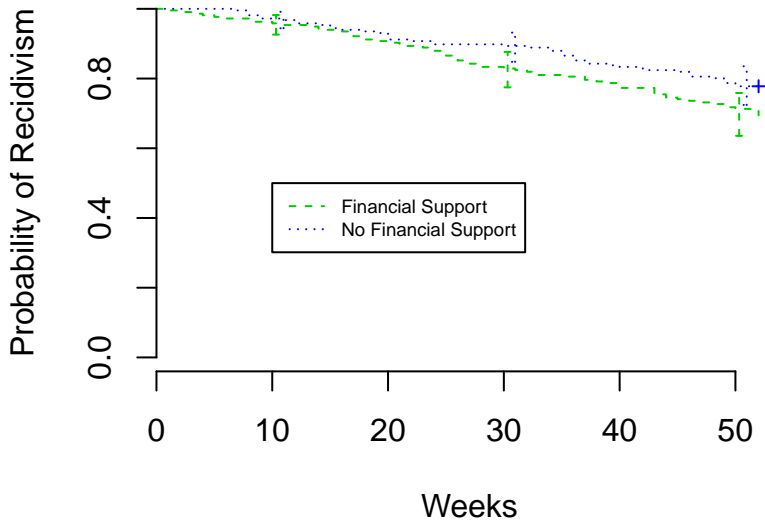
Weeks

Cumulative hazard (risk) of recidivism, w/CIs

```r
library(survival)
library(eventtimedata)
data("rossi")
rossi.recidivism.ch <- survfit(Surv(week, arrest) ~ fin,
                               data = rossi)
plot(rossi.recidivism.km, lty = 2:3, col = 3:4, mark.time = TRUE,
     fun = "cumhaz",
     xlab = "Weeks",
     ylab = "Cumulative Hazard of Recidivism",
     axes = FALSE,
     conf.times = c(10,30,50),
     main = "Cumulative Risk of Recidivism Probability,
     with Conf. Int.",
     cex = 0.6, cex.main = 0.8)
axis(1)
axis(2)
legend("topleft", inset = c(0.1, 0.3),
       c("Financial Support", "No Financial Support"),
       lty = 2:3, col = 3:4, cex = 0.6)
```

Cumulative Risk of Recidivism Probability, with Conf. Int.

Financial Support
No Financial Support

Weeks

Cumulative Hazard of Recidivism
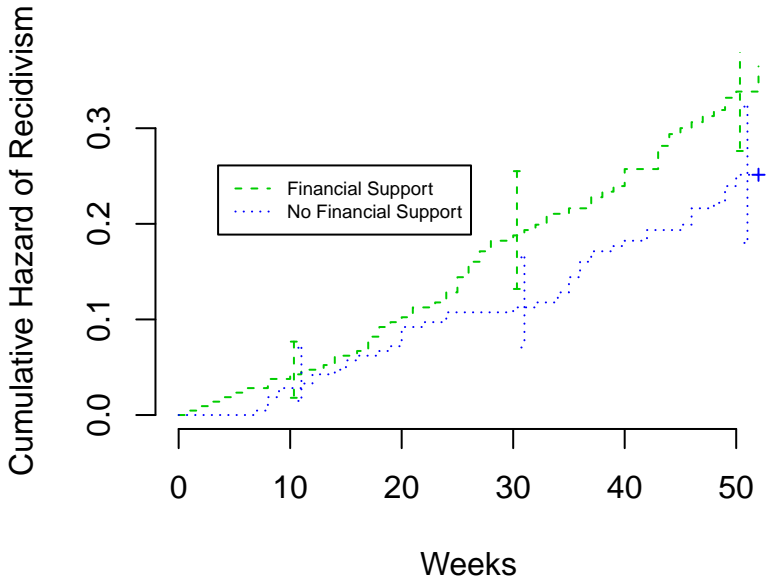
Confidence intervals vs confidence bands

Examining many confidence intervals may cause the same problem as simultaneous hypothesis tests.

- Overall coverage probability for the curve is not right.

Hall and Wellner (*Biometrika*, 1980) solved that problem by deriving confidence bands:

- 95% bands have probability 0.95 of covering the entire survival curve.

- These bands will be wider than pointwise intervals.

- Formulas complex, not shown here.

```r
library(survival)
library(eventtimedata)
data("rossi")
rossi.recidivism.km <- survfit(Surv(week, arrest) ~ fin,
                                data = rossi)
plot(rossi.recidivism.km, lty = 2:3, col = 3:4, mark.time = TRUE,
     xlab = "Weeks",
     ylab = "Probability of Re-arrest",
     axes = FALSE,
     conf.int = TRUE,
     main = "KM of Recidivism Probability, with Conf. Bands",
     cex = 0.6, cex.main = 0.8)
axis(1)
axis(2)
legend(10, .5, c("Financial Support", "No Financial Support"),
       lty = 2:3, col = 3:4, cex = 0.6)
```

KM of Recidivism Probability, with Conf. Bands

Probability of Re−arrest

- - - Financial Support
- · · · No Financial Support

Weeks
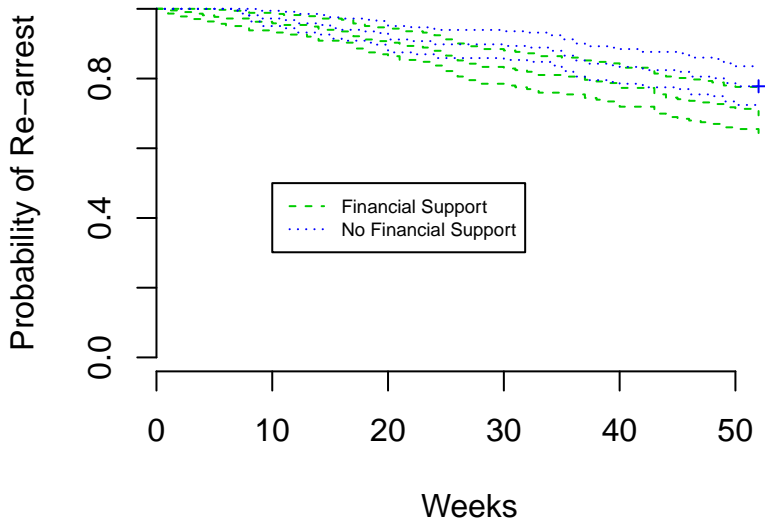
# Example: Application to FDA (7 March 2018)

On 7 March 2018, Amgen asked for FDA approval of the drug blinatumumab in patients with a sub-type of acute lymphoblastic leukemia (ALL).

- The drug would be given to patients who experienced a clinical complete remission, but had evidence of minimal residual disease (MRD).

Figure on the next slide from the FDA analysis of the data shows

- Relapse free survival by MRD status
- Shows confidence bands (Hall and Wellner)

Figure 1: Study 20120148 - Kaplan-Meier Plot of Hematological RFS of Patients by Baseline MRD Level
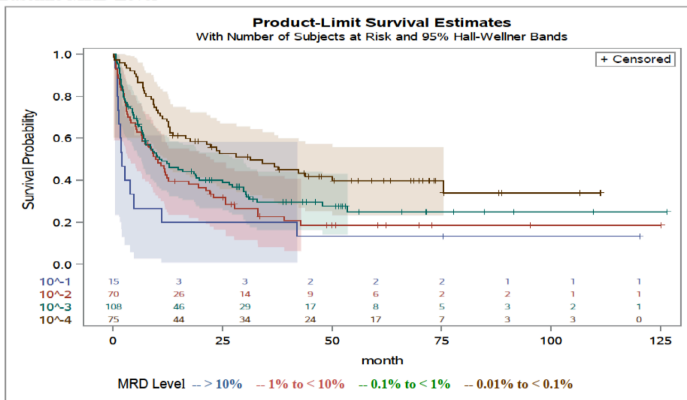
Source: FDA analysis

Figure 1: FDA presentation, 7 March 2018

Derivations

Conditional Probability: Suppose A and B are two events. Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplication Law: Multiply both sides of the above by $P(B)$.

$$P(A \cap B) = P(A|B)P(B)$$

Extension to more than 2 events: Suppose $A_1, A_2, \ldots, A_k$ are k different events. Then, the probability of all k events occurring can be written as a product of conditional probabilities.

$$\begin{aligned}
P(A_1 \cap A_2 \ldots \cap A_k) = {} & P(A_k | A_{k-1} \cap \ldots \cap A_1) \\
& \times P(A_{k-1} | A_{k-2} \cap \ldots \cap A_1) \\
& \times \ldots \\
& \times P(A_2 | A_1) \\
& \times P(A_1)
\end{aligned}$$

Think of dividing the observed time-span of the study into a series of small intervals so that there is a separate interval for each time of death or censoring (with possible ties):

Using the law of conditional probability,

$$P(T > t) = \prod_j P(\text{survive } j\text{-th interval } I_j \mid \text{survived to start of } I_j),$$

over all intervals preceding time t.

KM estimator derivation, continuous case . . .

Four possibilities for each interval:

1. No event: conditional probability of surviving the interval is 1.

2. Censoring: assume individual survives to end of the interval, so that the conditional probability of surviving the interval is 1.

3. Death, but no censoring: conditional probability of *not* surviving the interval is # deaths (d) divided by # "at risk" (r) at the beginning of the interval. Thus, the conditional probability of surviving the interval is $1 - \frac{d}{r}$.

4. Tied deaths and censoring: assume censorings survive to end of the interval, so that conditional probability of surviving the interval is still $1 - \frac{d}{r}$.

Thus, the general formula for the conditional probability of surviving the j-th interval that holds for all 4 cases is $1 - \frac{d_j}{r_j}$.

As the intervals become smaller,

- The approximations made in estimating the probabilities of surviving each interval become smaller.
- The estimator converges to the true $S(t)$ as the sample size increases.

This argument clarifies why an alternative name for the KM is the *product limit estimator*.

Result stated earlier

For continuous data, the Kaplan-Meier estimator of the survivorship function $S(t) = P(T > t)$ is

$$\widehat{S}(t) = \prod_{j:\tau_j \leq t} \frac{r_j - d_j}{r_j} = \prod_{j:\tau_j \leq t} \left(1 - \frac{d_j}{r_j}\right), \text{ where}$$

- $\tau_1, \ldots, \tau_K$ are the K distinct death times observed
- d_j is the number of deaths at τ_j
- r_j is the number of individuals "at risk" right before the j-th death time (everyone dead or censored *at or after* that time).
    - $r_j = r_{j-1} - d_{j-1} - c_{j-1}$
    - Alternatively, $r_j = \sum_{l \geq j}(c_l + d_l)$
- c_j is the number of censored observations between the j-th and $(j+1)$-th death times.
    - Censorings tied at τ_j are included in c_j

Derivation of Greenwood's formula

KM estimator can be thought of as

$$\widehat{S}(t) = \prod_{j:\tau_j \leq t} (1 - \widehat{\lambda}_j), \text{ where } \widehat{\lambda}_j = \frac{d_j}{r_j}.$$

Since the $\widehat{\lambda}_j$'s are (conditionally) binomial proportions, standard likelihood theory can be used to to show each $\widehat{\lambda}_j$ is approximately normally distributed, with mean λ_j, and variance[4]

$$\text{var}(\widehat{\lambda}_j) = \frac{\lambda_j(1 - \lambda_j)}{r_j}$$

The $\widehat{\lambda}_j$'s are independent in large enough samples.

---

[4] The estimated variance is $\widehat{\text{var}}(\widehat{\lambda}_j) = \frac{\widehat{\lambda}_j(1-\widehat{\lambda}_j)}{r_j}$.

DERIVATION OF GREENWOOD'S FORMULA ...

Since $\widehat{S}(t)$ is a function of the λ_j's, its variance can be estimated using the *delta method*,

- an approach for calculating the variance of non-linear functions.

Delta method: If Y is normal with mean μ and variance σ^2, then $g(Y)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g'(\mu)]^2\sigma^2$.

DIGRESSION: THE DELTA METHOD

Two specific examples that will be used in the derivation:

- Ex. 1: $Z = g(Y) = \log(Y)$, then $g'(y) = (1/y)$:

$$Z \sim N\left(\log(\mu), \left(\frac{1}{\mu}\right)^2 \sigma^2\right)$$

- Ex. 2: $Z = g(Y) = \exp(Y)$, then $g'(y) = e^y$:

$$Z \sim N\left(e^\mu, [e^\mu]^2 \sigma^2\right)$$

DERIVATION OF GREENWOOD'S FORMULA ...

Instead of dealing with $\widehat{S}(t)$ directly, use $\log[\widehat{S}(t)]$ since calculating variance of a sum is easier than calculating variance of a product,

$$\log[\widehat{S}(t)] = \sum_{j:\tau_j \leq t} \log(1 - \widehat{\lambda}_j)$$

By approximate independence of the $\widehat{\lambda}_j$'s,

$$\mathrm{var}(\log[\widehat{S}(t)]) = \sum_{j:\tau_j \leq t} \mathrm{var}[\log(1 - \widehat{\lambda}_j)].$$

Apply the delta method (Ex. 1), where $\mu = 1 - \lambda_j$ and $\sigma^2 = \frac{\lambda_j(1-\lambda_j)}{r_j}$.

$$\widehat{\mathrm{var}}(\log[\widehat{S}(t)]) = \sum_{j:\tau_j \leq t} \left(\frac{1}{1-\widehat{\lambda}_j}\right)^2 \left(\frac{\widehat{\lambda}_j(1-\widehat{\lambda}_j)}{r_j}\right)$$

$$= \sum_{j:\tau_j \leq t} \frac{\widehat{\lambda}_j}{(1-\widehat{\lambda}_j)r_j} = \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j - d_j)r_j}$$

GREENWOOD'S FORMULA

To obtain $\widehat{\text{var}}(\widehat{S}(t))$, apply the delta method again (Ex. 2), using the relationship $\widehat{S}(t) = \exp[\log[\widehat{S}(t)]]$,

$$\widehat{\text{var}}(\widehat{S}(t)) = [\widehat{S}(t)]^2 \; \widehat{\text{var}}\left[\log[\widehat{S}(t)]\right]$$

Substitute the previous result for $\widehat{\text{var}}\left[\log[\widehat{S}(t)]\right]$ to obtain Greenwood's Formula,

$$\widehat{\text{var}}(\widehat{S}(t)) = [\widehat{S}(t)]^2 \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j - d_j)r_j}$$