STAT 6390: Analysis of Survival Data

Textbook coverage: Chapter 2

Steven Chiou

Department of Mathematical Sciences,
University of Texas at Dallas

Survivor, hazard and cumulative hazard functions

- Suppose the actual (uncensored, untruncated) survival time of an individual is t and can be regarded as the observed value of a variable, T.
- We assume the support of T is non-negative or $(0, \infty)$.
- We call T the *random variable* associated with the survival time, and we define T has a cumulative distribution function given by $F(t) = \mathrm{P}(T \leq t)$.
- The survival function of T is then defined as

$$S(t) = 1 - \mathrm{P}(T \leq t) = 1 - F(t).$$

- Why are we more interested in $S(t)$?

Survivor, hazard and cumulative hazard functions

- The *hazard function* is widely used to survival analysis.
- The hazard function $h(t)$ is defined below

$$h(t) = \lim_{\delta \to 0} \frac{\mathrm{P}(t \leq T < t + \delta | T \geq t)}{\delta}. \tag{1}$$

- $\mathrm{P}(t \leq T < t + \delta | T \geq t)$ is a conditional probability.
- The conditional probability is then expressed as a probability per nit time by dividing by the time interval, δ, to give a *rate*.
- The function $h(t)$ is also referred to as the *hazard rate*, the *instantaneous death rate*, the *intensity rate*, or the *force of mortality*.
- Event rate at time t, conditional on the event not having occurred before t.

Survivor, hazard and cumulative hazard functions

- In terms of probability, if t is measured in days, $h(t)$ is the approximate probability that an individual, who is *at risk* of the event occurring at the start of day t, experiences the event during that day.
  - In this case $\delta = 1$.
  - $\lim_{\delta \to 0}$ can be thought of as changing the unit from days to hours, minutes, seconds, milliseconds...
- If the event of interest is not death, $h(t)$ can also be regarded as the *expected number of events* experienced by an individual in unit time, given that the event has not occurred before then.
  - Think of $\mathrm{E}\{\mathrm{I}(\cdot)\} = \mathrm{P}(\cdot)$.
  - The part "given that the event..." might be ignored if events follow the Poisson process.

Survivor, hazard and cumulative hazard functions

- The definition in (1) leads to some useful relationships between survivor and hazard functions:

$$(1) = \lim_{\delta \to 0} \frac{\mathrm{P}(t \leq T < t + \delta)}{\delta \cdot \mathrm{P}(T < t)} = \lim_{\delta \to 0} \frac{F(t+\delta) - F(t)}{\delta} \cdot \frac{1}{\mathrm{P}(T < t)} = \frac{\mathrm{d}F(t)}{\mathrm{d}t} \cdot \frac{1}{S(t)}.$$

- $h(t)$ is approximately the probability that an individual experiences an event at this instant (t) given that he/she is risk free up to t.
- If T is a continuous random variable, then we have

$$h(t) = \frac{f(t)}{S(t)}. \tag{2}$$

- This shows that from any one of the three functions, $f(t)$, $S(t)$, and $h(t)$, the other two can be determined.

Survivor, hazard and cumulative hazard functions

- Equation (2) also implies

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\{\log S(t)\} \text{ and } S(t) = e^{-H(t)},$$

  where $H(t) = \int_0^t h(u)\,\mathrm{d}u$ is the *cumulative hazard function*.
- Similarly, the cumulative hazard function can also be obtained from

$$H(t) = -\log S(t).$$

- The cumulative hazard function is the cumulative risk of an event occurring by time t.
- If the event is death, then $H(t)$ summarizes the risk of death up to time t, given that death has not occurred by t.
- If the event is not death, $H(t)$ can be interpreted as the expected number of events that occur in the interval $(0, t)$.

Survivor, hazard and cumulative hazard functions

- It is possible for $H(t) > 1$, $h(t) > 1$, or $f(t) > 1$.
- Like $F(t)$, $S(t)$ is bounded in $[0, 1]$.
- $F(t)$ and $H(t)$ is non-decreasing; $S(t)$ is non-increasing.
- $h(t)$ can go up and down.
- For example, suppose $T \sim \exp(\lambda)$, where λ is the rate. Then
  - $S(t) = e^{-\lambda t}$.
  - $h(t) = \lambda$.
  - $H(t) = \lambda t$.

Empirical survivor function

- The $S(t)$ can be estimated non-parametrically with the *product limit* estimator, which is also known as the *Kaplan-Meier* estimator.
- We first assume there is a single sample of survival times, and none of these are censored.
- In this case, the survivor probability at t, $S(t)$, is defined as

$$\hat{S}_e(t) = \frac{\text{\# individuals with survival times} \geq t}{\text{\# individuals in the data set}}. \tag{3}$$

- Equation (3) is called *empirical survivor function*.
- Similar $\hat{F}_e(t) = 1 - \hat{S}_e(t)$ is called the *empirical cumulative distribution function*.

Empirical survivor function

- We illustrate with the first 10 uncensored subjects in the `whas100` data.
- Make sure **`tidyverse`** package and `whas100` are properly loaded*.

```
> whas10 <- whas100 %>% filter(fstat > 0) %>% filter(row_number() <= 10)
> whas10
# A tibble: 10 x 9
      id admitdate  foldate      los lenfol fstat   age gender   bmi
   <int> <fct>      <fct>      <int>  <int> <int> <int>  <int> <dbl>
 1     1 3/13/1995  3/19/1995      4      6     1    65      0  31.4
 2     2 1/14/1995  1/23/1996      5    374     1    88      1  22.7
 3     3 2/17/1995  10/4/2001      5   2421     1    77      0  27.9
 4     4 4/7/1995   7/14/1995      9     98     1    81      1  21.5
 5     5 2/9/1995   5/29/1998      4   1205     1    78      0  30.7
 6     6 1/16/1995  9/11/2000      7   2065     1    82      1  26.5
 7     7 1/17/1995  10/15/1997     3   1002     1    66      1  35.7
 8     8 11/15/1994 11/24/2000    56   2201     1    81      1  28.3
 9     9 8/18/1995  2/23/1996      5    189     1    76      0  27.1
10    12 5/26/1995  9/29/1996     11    492     1    83      0  24.7
```

* see note 1 for details.

Empirical survivor function

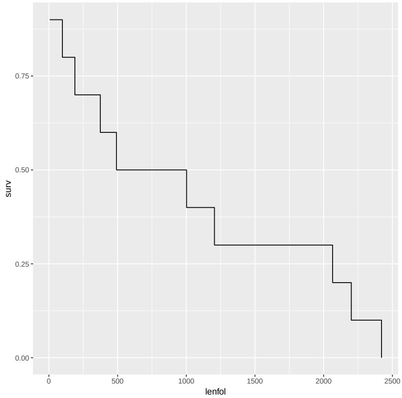
- The empirical estimates can be easily computed with `ecdf`.

```
> whas10 <- whas10 %>% mutate(surv = 1 - ecdf(lenfol)(lenfol))
> whas10
# A tibble: 10 x 10
      id admitdate  foldate      los lenfol fstat   age gender   bmi  surv
   <int> <fct>      <fct>      <int>  <int> <int> <int>  <int> <dbl> <dbl>
 1     1 3/13/1995  3/19/1995      4      6     1    65      0  31.4 0.9
 2     2 1/14/1995  1/23/1996      5    374     1    88      1  22.7 0.6
 3     3 2/17/1995  10/4/2001      5   2421     1    77      0  27.9 0
 4     4 4/7/1995   7/14/1995      9     98     1    81      1  21.5 0.8
 5     5 2/9/1995   5/29/1998      4   1205     1    78      0  30.7 0.3
 6     6 1/16/1995  9/11/2000      7   2065     1    82      1  26.5 0.200
 7     7 1/17/1995  10/15/1997     3   1002     1    66      1  35.7 0.4
 8     8 11/15/1994 11/24/2000    56   2201     1    81      1  28.3 0.100
 9     9 8/18/1995  2/23/1996      5    189     1    76      0  27.1 0.7
10    12 5/26/1995  9/29/1996     11    492     1    83      0  24.7 0.5
```

Empirical survivor function

- The empirical survivor function is a non-increasing step function.

```
> whas20 %>% ggplot(aes(lenfol, surv)) + geom_step(size = 1.2)
```

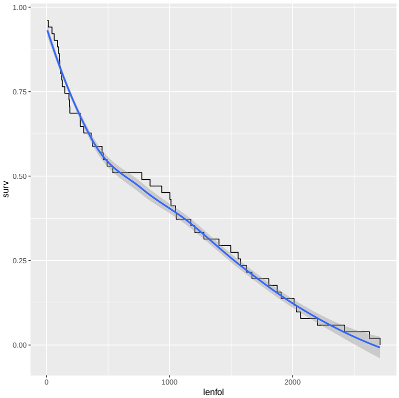


- The $\hat{S}_e(t)$ is 1 at $t = 0$ and 0 at the final death time.
- The $\hat{S}_e(t)$ is assumed to be constant between adjacent death times.

Empirical survivor function

- Putting everything together, we could plot the empirical survival curve for all the uncensored subjects in `whas100`:

```
> whas100 %>% filter(fstat > 0) %>% mutate(surv = 1 - ecdf(lenfol)(lenfol)) %>%
+     ggplot(aes(lenfol, surv)) + geom_step() + geom_smooth()
```



- The pipeline between `ggplot` is "+" instead of "`%>%`".

Kaplan-Meier estiamtor

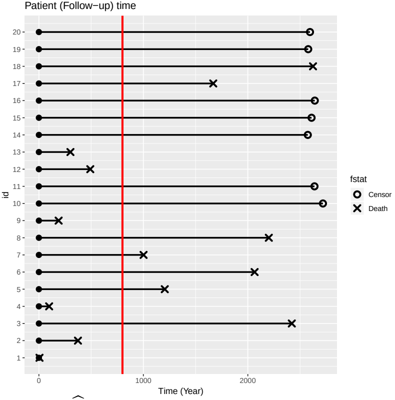
- With censoring, the same idea can be applied with proper adjustment.
- Kaplan-Meier estimator is the default estimator used by many packages.
- The basic idea is to decompose $\mathrm{P}(T > t)$ by conditioning on prior times.
- Suppose a sample size of n, $\mathrm{P}(T > t)$ can be decomposed as

$$\widehat{S}_{KM}(t) \doteq \mathrm{P}(T > t) = \mathrm{P}(T > t_{(0)}) \cdot \mathrm{P}(T > t_{(1)}|T > t_{(0)}) \cdot \mathrm{P}(T > t_{(2)}|T > t_{(1)}) \cdot \ldots \cdot \mathrm{P}(T > t|T > t_{(i)}),$$

  for a series of time intervals $0 \doteq t_{(0)} < t_{(1)} < \ldots < t_{(i)} < t$ for some $i \leq n$.
- In general, the series $\{t_{(1)}, \ldots, t_{(m)}\}$ denotes the m ordered death times.

Kaplan-Meier estiamtor

Suppose we want $\mathrm{P}(T > 800)$ among the first 20 patients in `whas1000`.



- There are 6 events before $t = 800$.
- The events occured at

$t_{(0)}$	$t_{(1)}$	$t_{(2)}$	$t_{(3)}$	$t_{(4)}$	$t_{(5)}$	$t_{(6)}$
0	6	98	189	302	374	492

$$
\begin{aligned}
\widehat{S}_{KM}(800) &= \mathrm{P}(T > 800) = \\
&= \mathrm{P}(T > 0) \times \mathrm{P}(T > 6 | T > 0) \times \mathrm{P}(T > 98 | T > 6) \times \ldots \times \mathrm{P}(T > 492 | T > 374) \\
&= 1 \times \frac{19}{20} \times \frac{18}{19} \times \frac{17}{18} \times \frac{16}{17} \times \frac{15}{16} \times \frac{14}{15} = \frac{14}{20} = 70\%
\end{aligned}
$$

Why does $\widehat{S}_{KM}(800) = \hat{S}_e(800)$ here?

Kaplan-Meier estimator

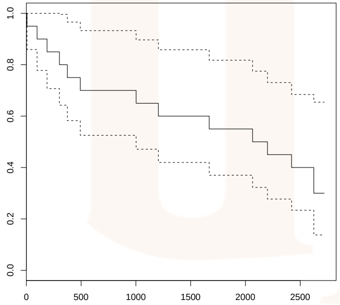
- The Kaplan-Meier estimator can be obtained with the `survfit` function.

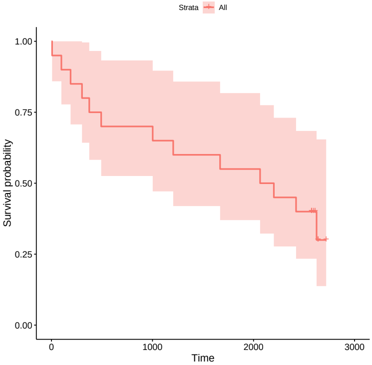
```
> library(survival)
> km <- survfit(Surv(lenfol, fstat) ~ 1, data = whas100, subset = id <= 20)
> summary(km)
Call: survfit(formula = Surv(lenfol, fstat) ~ 1, data = whas100, subset = id <=
    20)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     20       1     0.95  0.0487        0.859        1.000
   98     19       1     0.90  0.0671        0.778        1.000
  189     18       1     0.85  0.0798        0.707        1.000
  302     17       1     0.80  0.0894        0.643        0.996
  374     16       1     0.75  0.0968        0.582        0.966
  492     15       1     0.70  0.1025        0.525        0.933
 1002     14       1     0.65  0.1067        0.471        0.897
 1205     13       1     0.60  0.1095        0.420        0.858
 1669     12       1     0.55  0.1112        0.370        0.818
 2065     11       1     0.50  0.1118        0.323        0.775
 2201     10       1     0.45  0.1112        0.277        0.731
 2421      9       1     0.40  0.1095        0.234        0.684
 2624      4       1     0.30  0.1194        0.138        0.654
```

Kaplan-Meier estimator

- The Kaplan-Meier curve can be plotted with `plot` or `ggsurvplot`.

```
> library(survminer)
> plot(km)
> ggsurvplot(km)
```

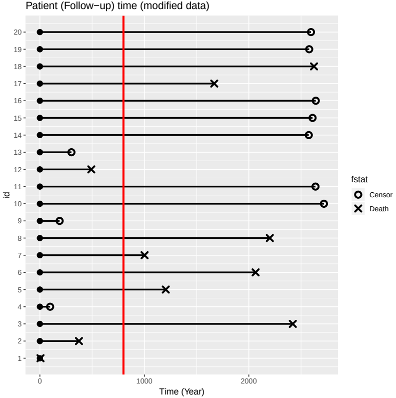




- Since **`survminer`** depends on the newest version of **`survMisc`**, you might need to update the latter to be able to use `ggsurvplot`.

Kaplan-Meier estimator

Suppose we want $\mathrm{P}(T > 800)$ based on the following modified data:



- There are 3 events before $t = 800$.
- The events occured at

$t_{(0)}$	$t_{(1)}$	$t_{(2)}$	$t_{(3)}$
0	6	374	492

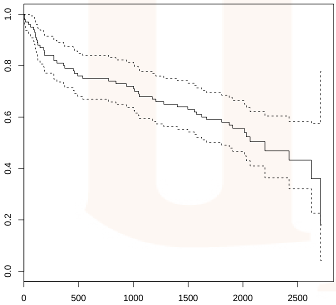
- In this modified data, $t = 98, 189, 302$ are considered as censored.

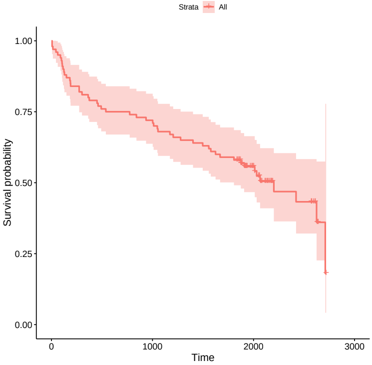
$$
\begin{aligned}
\widehat{S}_{KM}(800) &= \mathrm{P}(T > 800) = \\
&= \mathrm{P}(T > 0) \times \mathrm{P}(T > 6 | T > 0) \times \mathrm{P}(T > 374 | T > 6) \times \mathrm{P}(T > 492 | T > 374) \\
&= 1 \times \frac{19}{20} \times \frac{15}{16} \times \frac{14}{15} \approx 83.1\%
\end{aligned}
$$

Kaplan-Meier estimator

- The Kaplan-Meier estimator for the whole data is

```
> library(survival)
> km <- survfit(Surv(lenfol, fstat) ~ 1, data = whas100)
> plot(km)
> ggsurvplot(km)
```





- If the last observed time corresponds to a censored observation, then the estimate of the survival function does not go to zero.

Kaplan-Meier estimator

- Suppose we have a sample of n independent observations $(t_i, c_i), i = 1, 2, \ldots, n$.
- There are m deaths and $m \leq n$.
- The series $\{t_{(1)}, \ldots, t_{(m)}\}$ are the m ordered death times.
- The Kaplan-Meier estimator has the form

$$\widehat{S}_{KM}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i} = \prod_{t_{(i)} \leq t} 1 - \frac{d_i}{n_i},$$

  where n_i is the number of individual who are alive at $t_{(i)}$ (at risk), and d_i is the number of individual who died at $t_{(i)}$.
- A potential problem with the Kaplan-Meier estimator is when n_i is small and $n_i = d_i$ occurs at early time.

Nelson-Aalon estimator

- An alternative estimate of $\widehat{S}_{KM}(t)$ is the *Nelson-Aalon estimator*:

$$\widehat{S}_{NA}(t) = \prod_{t_{(i)} \le t} \exp\left(-\frac{d_i}{n_i}\right).$$

- The main idea is to see d_i/n_i as the event rate, i.e., $h(t_{(i)}) = d_i/n_i$.
- Recall the relationship $h(t) = f(t)/S(t)$ and think of d_i/n and n_i/n are raw estimates of $f(t)$ and $S(t)$.
- By the similar argument, we have

$$\widehat{H}_{NA}(t) \doteq H(t) = \sum_{t_{(i)} \le t} d_i/n_i, \text{ and } S(t) = e^{-\widehat{H}_{NA}(t)} = \widehat{S}_{NA}(t).$$

- $\widehat{S}_{NA}(t)$ and $\widehat{S}_{KM}(t)$ are derived differently, but both based on d_i and n_i.
- In general $\widehat{S}_{NA}(t) \ge \widehat{S}_{KM}(t)$ but $\widehat{S}_{NA}(t) \approx \widehat{S}_{KM}(t)$.

Nelson-Aalon estimator

- $\widehat{S}_{NA}(t)$ has slightly nicer properties and is more stable.
- If the interest is in estimating the cumulative hazard function, $H(t)$, we can use either the $\widehat{H}_{NA}(t)$, or $\widehat{H}_{KM}(t) = -\log \widehat{S}_{KM}(t)$.
- The $\widehat{H}_{KM}(t)$ follows directly form from $\widehat{S}_{KM}(t)$:

$$\widehat{H}_{KM}(t) = -\sum_{t_{(i)} \leq t} \log\left(\frac{n_i - d_i}{n_i}\right).$$

Nelson-Aalon estimator

- $\widehat{S}_{NA}(t)$ can be obtained with `coxph` of the **survival** package.

```
> args(coxph)
function (formula, data, weights, subset, na.action, init, control,
    ties = c("efron", "breslow", "exact"), singular.ok = TRUE,
    robust = FALSE, model = FALSE, x = FALSE, y = TRUE, tt, method = ties,
    ...)
NULL
```

- `coxph` refers to "Cox proportional hazard model" that has the form

$$h(t) = h_0(t)e^{X^\top \beta}, \tag{4}$$

where X is the covariate matrix, β is the regression coefficient, and $h_0(t)$ is called the *baseline hazard* function.

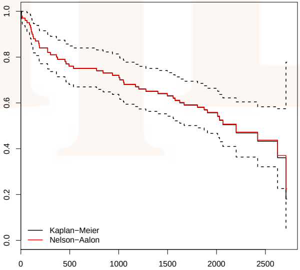
- More details will be given in Chapter 3.

Nelson-Aalon estimator

- For now, we will assume $\beta = 0$ in (4), which implies $h(t) = h_0(t)$.
- We will use $h_0(t)$ to obtain $\widehat{S}_{NA}(t)$.

```
> cox <- coxph(Surv(lenfol, fstat) ~ 1, data = whas100)
> H0 <- basehaz(cox)
> str(H0)
'data.frame': 95 obs. of  2 variables:
 $ hazard: num  0.0201 0.0303 0.0406 0.051 0.0616 ...
 $ time  : num  6 14 44 62 89 98 104 107 114 123 ...
```

```
> plot(km)
> lines(H0$time, exp(-H0$hazard), 's', col = 2)
```



Life-table estimates

- When dataset is large, the $\widehat{S}_{KM}(t)$ and $\widehat{S}_{NA}(t)$ can be obtained with intervals of time, rather than exact time points.
  - The series $\{t_{(1)}, \ldots, t_{(m)}\}$ represents intervals.
  - d_i represents the number of individual who died in $t_{(i)}$.
  - n_i represents the number of individual who are alive in $t_{(i)}$.
- Potential problem with censoring?
- Adjustments under uniform assumption (p25).

Inference on $\widehat{S}_{KM}(t)$

- The 95% confidence interval (CI) does not follow the usual form of

$$\text{PE} \pm 1.96 \times \text{SE}.$$

- This is mainly because $\widehat{S}_{KM}(t)$ lies between 0 and 1.
- Two common methods to obtain the 95% CI for $\widehat{S}_{KM}(t)$ are the log and log-log transformations.
- The idea is to derive the standard errors on the transformed scale first, then back-transform these back.

The Delta Method

- We need the Delta method to estimate the standard errors.
- The Delta method states that

$$\mathrm{Var}\{g(X)\} \approx \mathrm{Var}(X) \cdot \{g'(x_0)\}^2,$$

where $g'(x_0)$ is the 1st derivative of $g(\cdot)$ evaluates at constant x_0.

The Delta Method

- A special case of the Delta method is when $g(\cdot) = \log(\cdot)$.
- Setting $g(\cdot) = \log(\cdot)$, we have

$$\mathrm{Var}\{f(X)\} \approx \frac{\mathrm{Var}(X)}{x_0^2}.$$

Inference on $\widehat{S}_{KM}(t)$

- We will first look at the log transformation.
- Recall

$$\widehat{S}_{KM}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i}.$$

- The variance of log-transformed $\widehat{S}_{KM}(t)$ gives

$$\mathrm{Var}\left\{\log \widehat{S}_{KM}(t)\right\} = \mathrm{Var}\left\{\sum_{t_{(i)} \leq t} \log\left(\frac{n_i - d_i}{n_i}\right)\right\} = \sum_{t_{(i)} \leq t} \mathrm{Var}\left\{\log\left(\frac{n_i - d_i}{n_i}\right)\right\}.$$

- We assume independence between observations in the risk sets.
- For convenience, let's write $p_i = (n_i - d_i)/n_i$, and $\hat{p}_i$ when n_i and d_i are known.

Inference on $\widehat{S}_{KM}(t)$

- The key is to estimate $\mathrm{Var}\{\log(p_i)\}$ with the Delta method.
- For each $t_{(i)}$, n_i is a fixed constant d_i is random.
- $n_i - d_i$ is the risk set size and can be assumed to follow the binomial distribution with parameters n_i and $1 - d_i/n_i$. Then

$$\mathrm{Var}(p_i) = \frac{\mathrm{Var}(n_i - d_i)}{n_i^2} = \frac{\frac{d_i}{n_i} \cdot \left(1 - \frac{d_i}{n_i}\right)}{n_i}.$$

- With the Delta method, we have

$$\mathrm{Var}\{\log(p_i)\} \approx \frac{\mathrm{Var}(p_i)}{\hat{p}_i} = \frac{d_i}{n_i \cdot (n_i - d_i)}.$$

Inference on $\widehat{S}_{KM}(t)$

- From the above result, we have

$$\mathrm{Var}\left\{\log \widehat{S}_{KM}(t)\right\} \approx \sum_{t_{(i)} \leq t} \frac{d_i}{n_i \cdot (n_i - d_i)}.$$

- By the Delta method,

$$\mathrm{Var}\left\{\log \widehat{S}_{KM}(t)\right\} \approx \mathrm{Var}\{\widehat{S}_{KM}(t)\} \cdot \frac{1}{\widehat{S}^2_{KM}(t)}.$$

- Altogether, this gives

$$\mathrm{Var}\{\widehat{S}_{KM}(t)\} \approx \widehat{S}^2_{KM}(t) \cdot \sum_{t_{(i)} \leq t} \frac{d_i}{n_i \cdot (n_i - d_i)}.$$

- This result is known as the *Greenwood's formula*.
- This estimator can be obtained from a counting process approach.

Inference on $\widehat{S}_{KM}(t)$

- The Greenwood formula is the default method for `survfit`.
- With the Greenwood formula, the $100(1-\alpha)\%$ condifdence interval of $\widehat{S}_{KM}(t)$ can be obtained using the usual form of PE $\pm Z_{\alpha/2} \times$ SE.
- The bounds can still be outside of $[0, 1]$.
- An alternative approach is to consider the log-log transformation.

Inference on $\widehat{S}_{KM}(t)$

- By the Delta method, we have

$$\mathrm{Var}\left[\log\left\{-\log\widehat{S}_{KM}(t)\right\}\right] \approx \frac{1}{\{-\log\widehat{S}_{KM}(t)\}^2}\cdot\sum_{t_{(i)}\leq t}\frac{d_i}{n_i\cdot(n_i-d_i)}.$$

- This imples that the $100(1-\alpha)\%$ condifdence interval can be constructed by inversing

$$\log\{-\log\widehat{S}_{KM}(t)\}\pm Z_{\alpha/2}\times \mathrm{SE}\left[\log\{-\log\widehat{S}_{KM}(t)\}\right]$$

- Since $-\log$ of a survival function gives the cumulative hazard function, e.g., $-\log S(t) = H(t)$, the log-log approach called the “log hazard” approach.

Inference on $\widehat{S}_{KM}(t)$

- Types of CI can be specified with `conf.type` in `survfit`.

```
> ?survfit.coxph
```

- Some options are available for `conf.type` depending on $g(\cdot)$ used in the Delta method.

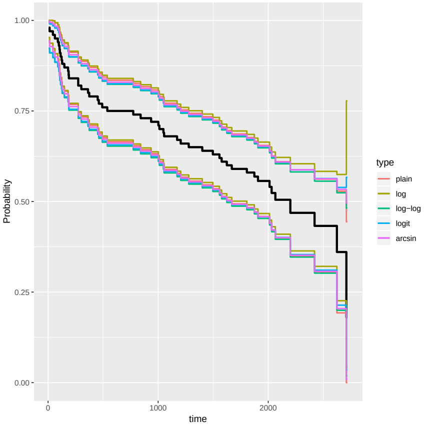
$$\begin{aligned} \texttt{plain} \quad g(x) &= x \\ \texttt{log} \quad g(x) &= \log(x) \\ \texttt{log-log} \quad g(x) &= \log\{-\log(x)\} \\ \texttt{logit} \quad g(x) &= \log\left(\frac{x}{1-x}\right) \\ \texttt{arcsin} \quad g(x) &= \arcsin\sqrt{(}x) \end{aligned}$$

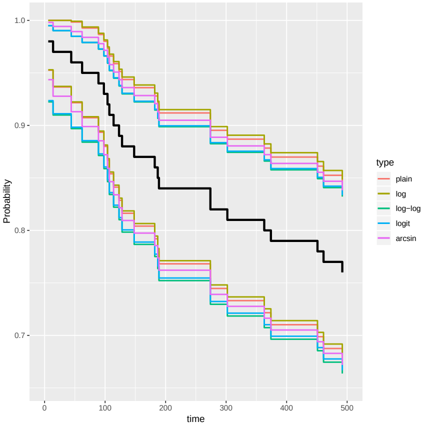
- In addition, Peto et al. (1977) proposed to estimate $\mathrm{Var}\{\widehat{S}_{KM}(t)\}$ from

$$\mathrm{Var}\{\widehat{S}_{KM}(t)\} = \frac{\widehat{S}_{KM}(t) \cdot (1 - \widehat{S}_{KM}(t))}{n_i}$$

Inference on $\widehat{S}_{KM}(t)$

- The following depict $\widehat{S}_{KM}(t)$ with the five `conf.type`'s for `whas100`.
- The CI's are quite close to each other.





*Inference on the median and percentiles

- Given a $\widehat{S}(t)$ ($\widehat{S}_{KM}(t)$ or $\widehat{S}_{NA}(t)$), the estimate of the pth percentile is

$$\hat{t}_p = \min\{t : \hat{S}(t) \leq (p/100)\}.$$

- The Delta method can be used again to obtain $\mathrm{Var}(\hat{t}_p)$.
- Setting $g(\cdot) = S(\cdot)$, we have the relationship

$$\mathrm{Var}\{\widehat{S}(\hat{t}_p)\} \approx \mathrm{Var}(\hat{t}_p) \cdot \{f(\hat{t}_p)\}^2,$$

where $f(t) = \mathrm{d}\widehat{S}(t)/\mathrm{d}t$.

- The only unknown in the above equation is $f(t)$, which can be approximated with linear interpolation:

$$\hat{f}(\hat{t}_p) \approx \frac{\widehat{S}(\hat{u}_p) - \widehat{S}(\hat{l}_p)}{\hat{l}_p - \hat{u}_p},$$

where $\hat{u}_p < \hat{t}_p < \hat{l}_p$, $\hat{u}_p = \max\{t : \widehat{S}(t) \geq p/100 + \epsilon\}$ and $\hat{l}_p = \min\{t : \widehat{S}(t) \leq p/100 - \epsilon\}$, for some small constant ϵ.

*Inference on the median and percentiles

- Replacing the unknown quantities with their empirical estimates, the $100(1-\alpha)\%$ CI can be obtained through $\hat{t}_p \pm Z_{\alpha/2} \times \mathrm{SE}(\hat{t}_p)$.
- Clinicians are specifically interested in median survival (follow-up) time, because survival time data often tend to skewed to the right.
- Other quantity of interest is the semi-interquartiles range (SIQR):

$$\text{SIQR} = \frac{t_{75} - t_{25}}{2}.$$

*Inference on the median and percentiles

- Median survival time is printed by `survfit`.

```
> survfit(Surv(lenfol, fstat) ~ 1, data = whas100)
Call: survfit(formula = Surv(lenfol, fstat) ~ 1, data = whas100)

      n  events  median 0.95LCL 0.95UCL
    100      51    2201    1806      NA
```

- However, median follow-up time does not usually exists.

```
> survfit(Surv(lenfol, fstat) ~ 1, data = whas100[81:100,])
Call: survfit(formula = Surv(lenfol, fstat) ~ 1, data = whas100[81:100,
    ])

      n  events  median 0.95LCL 0.95UCL
     20       9      NA    1577      NA
```

- A more practical approach?

Reference

Peto, R., Pike, M., Armitage, P., Breslow, N. E., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., and Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British journal of cancer* **35,** 1.