STAT 6390: Analysis of Survival Data

Textbook coverage: Chapter 2

Steven Chiou

Department of Mathematical Sciences,
University of Texas at Dallas

Survival, hazard and cumulative hazard functions

- Define T as the random variable of the actual (uncensored, untruncated) survival time of an individual.
- 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 $S(t)$?

Survival, hazard and cumulative hazard functions

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

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

- $\mathrm{P}(t \leq T < t + \mathrm{d}t | T \geq t)$ is a conditional probability.
- The conditional probability is then expressed as a probability per unit time by dividing by the time interval, $\mathrm{d}t$, 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.

Survival, 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 $\mathrm{d}t = 1$.
  - $\lim_{\mathrm{d}t \to 0}$ 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.

Survival, hazard and cumulative hazard functions

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

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

- $h(t)$ 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.

Survival, 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 $(0, t)$.

Survival, hazard and cumulative hazard functions

- It is possible for $H(t) > 1$, $h(t) > 1$, or $f(t) > 1$.
- $F(t)$, $S(t)$ are bounded in $[0, 1]$.
- $F(t)$ and $H(t)$ are 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 survival 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 none of survival times are censored.
- In this case, the survival 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 survival function*.
- Similarly, $\hat{F}_e(t) = 1 - \hat{S}_e(t)$ is called the *empirical cumulative distribution function*.

Empirical survival 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 survival function

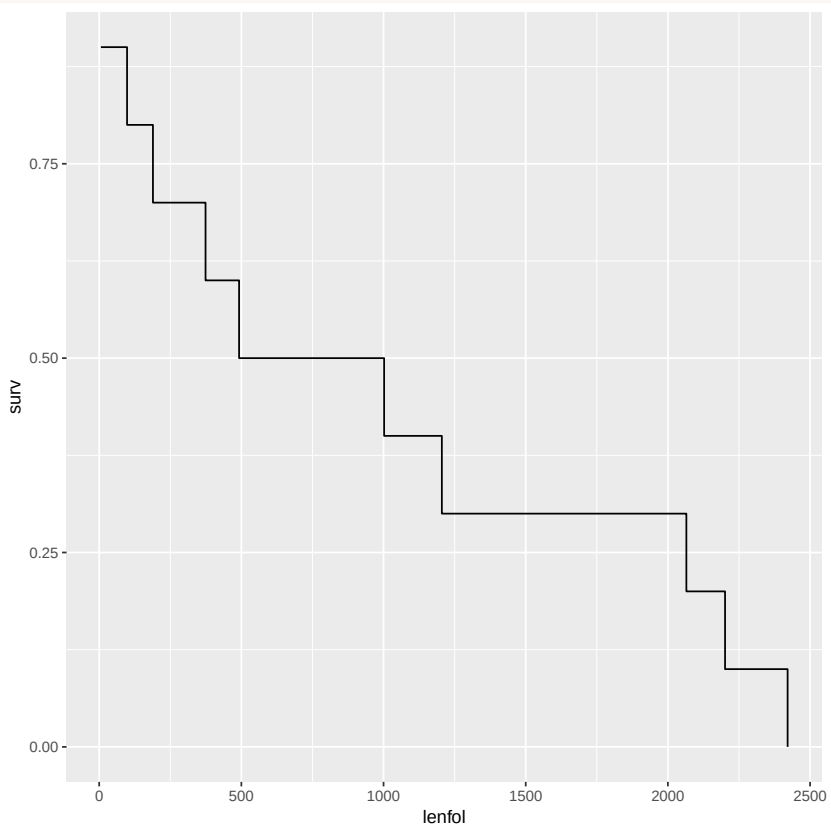
- The empirical estimates can be easily computed with `ecdf`.

```
> 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 survival function

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

```
> whas10 %>% ggplot(aes(lenfol, surv)) + geom_step(size = 1.2)
```

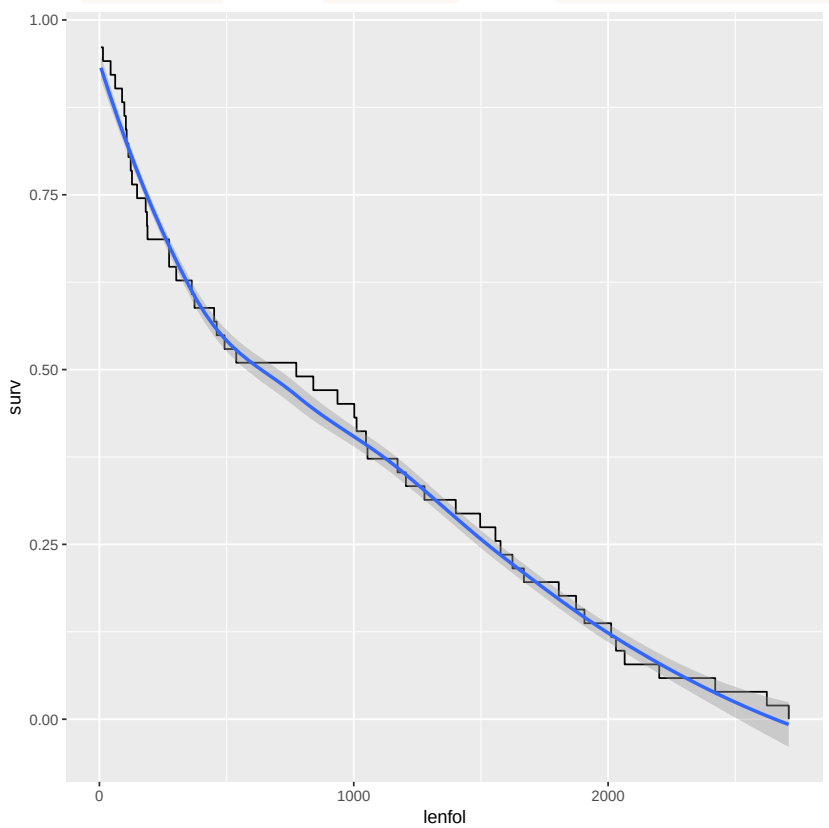


- The $\hat{S}_e(t)$ is 1 at $t = 0$ and 0 at the final death time.
- The $\hat{S}_e(t)$ is assumed to be constant between adjacent death times.

Empirical survival 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

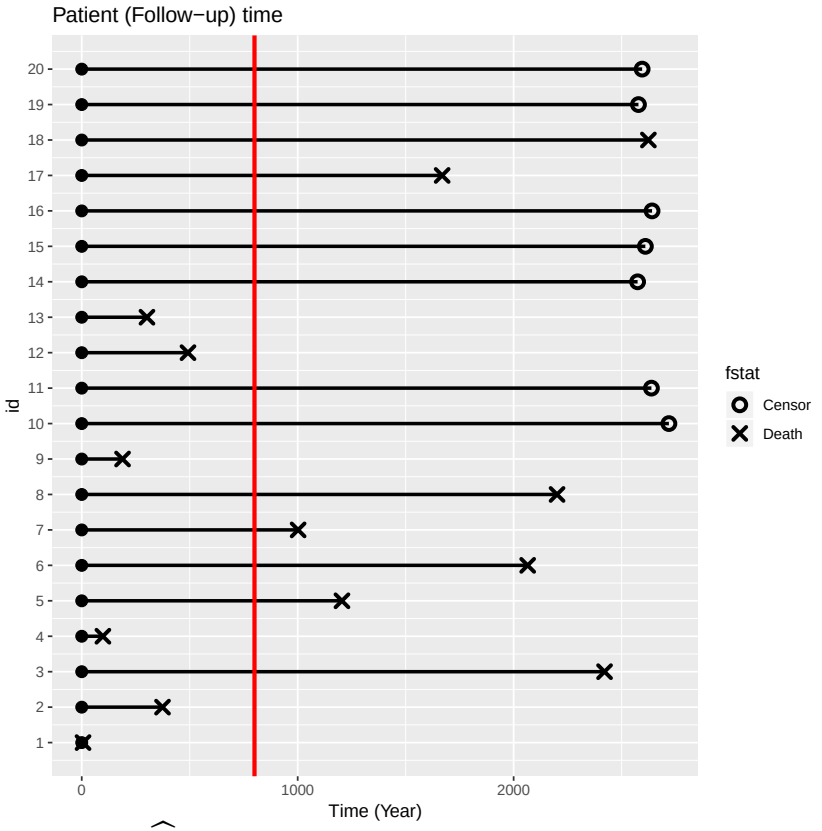
- With censoring, the same idea can be applied with proper adjustment.
- Kaplan-Meier estimator is the default estimator used by many packages.
- The basic idea is to decompose $\mathrm{P}(T > t)$ by conditioning on prior times.
- Suppose a sample size of n, $\mathrm{P}(T > t)$ can be decomposed as

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

  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

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

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

Kaplan-Meier estimator

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

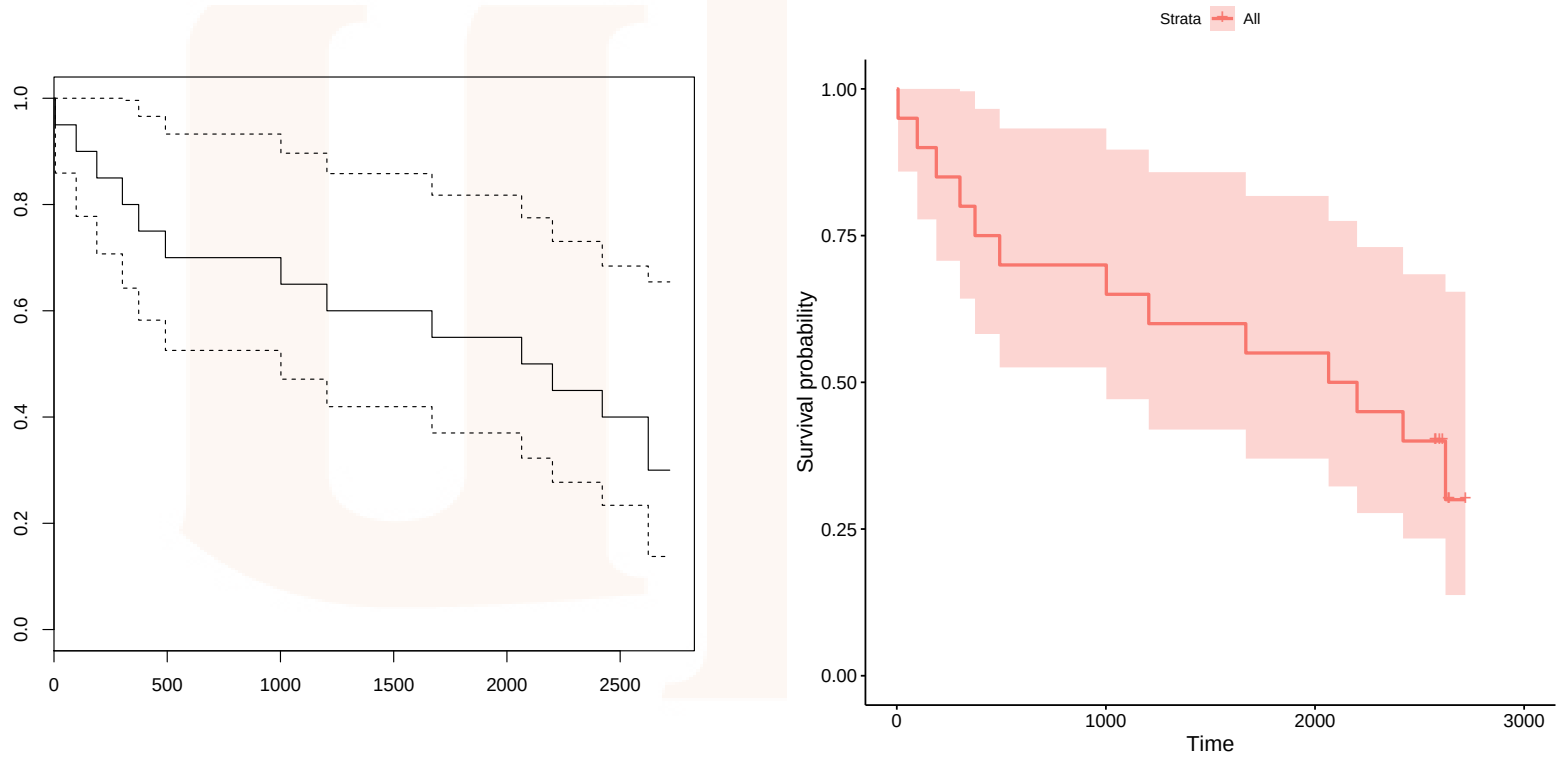
```
> library(survival)
> km <- survfit(Surv(lenfol, fstat) ~ 1, data = whas100, subset = id <= 20)
> summary(km)
Call: survfit(formula = Surv(lenfol, fstat) ~ 1, data = whas100, subset = id <=
    20)

 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`.

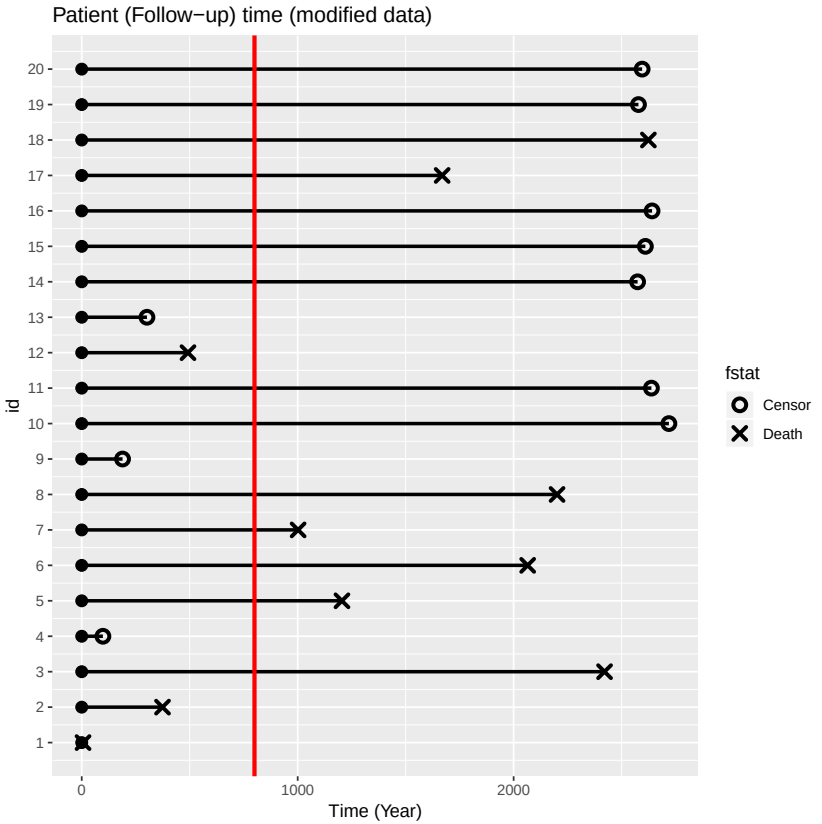
```
> library(survminer)
> plot(km)
> ggsurvplot(km)
```



- 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

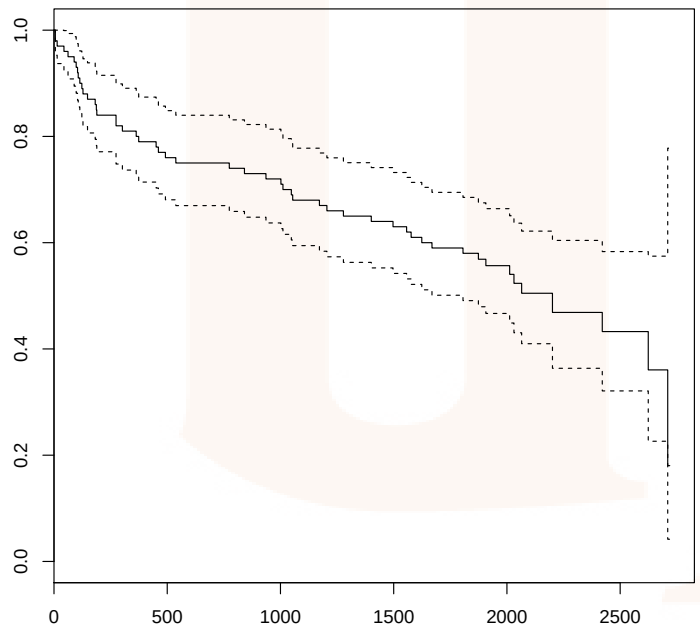
- In this modified data, $t = 98, 189, 302$ are considered as censored.

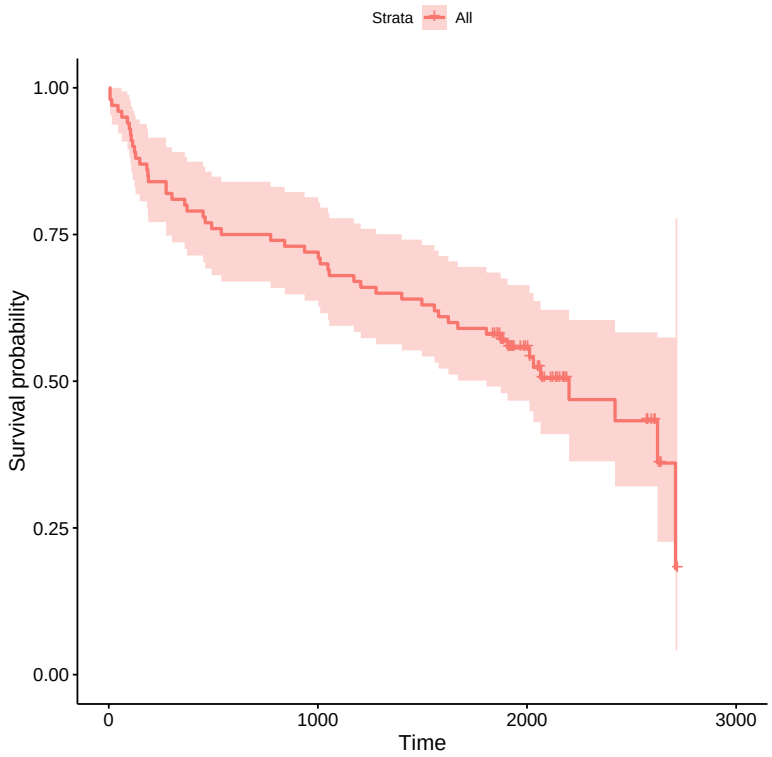
$$\begin{aligned}\widehat{S}_{KM}(800) &= \mathrm{P}(T > 800) = \\ &= \mathrm{P}(T > 0) \times \mathrm{P}(T > 6 | T > 0) \times \mathrm{P}(T > 374 | T > 6) \times \mathrm{P}(T > 492 | T > 374) \\ &= 1 \times \frac{19}{20} \times \frac{15}{16} \times \frac{14}{15} \approx 83.1\%\end{aligned}$$

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 $(t_i, c_i), i = 1, 2, \ldots, n$.
- Suppose 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 $\widehat{S}_{KM}(t)$:

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

- Problems with this estimator?

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.

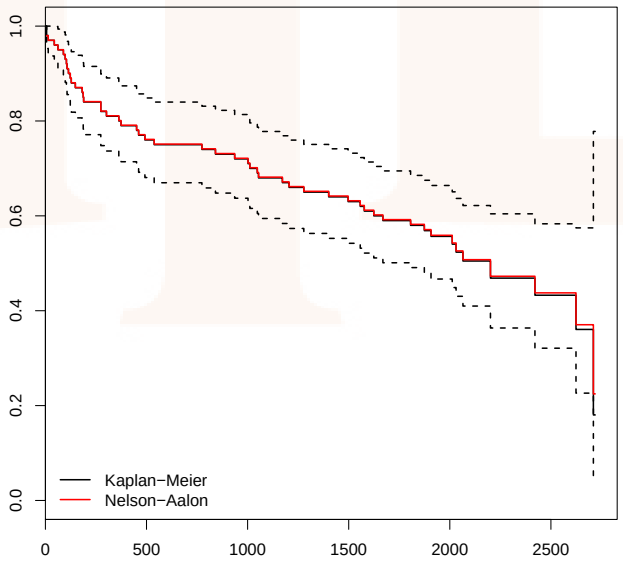
- More details will be given in Chapter 3.

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).
- Better to adopt methods for data with interval censoring.

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 $\widehat{S}_{KM}(t)$ needs to lie 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

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

- 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 fixed but d_i is random.
- $n_i - d_i$ 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

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

- By the Delta method (again),

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

- Altogether, this gives

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

- This result is known as the *Greenwood's formula*, or the log transformation.
- 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

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

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

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

- 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.

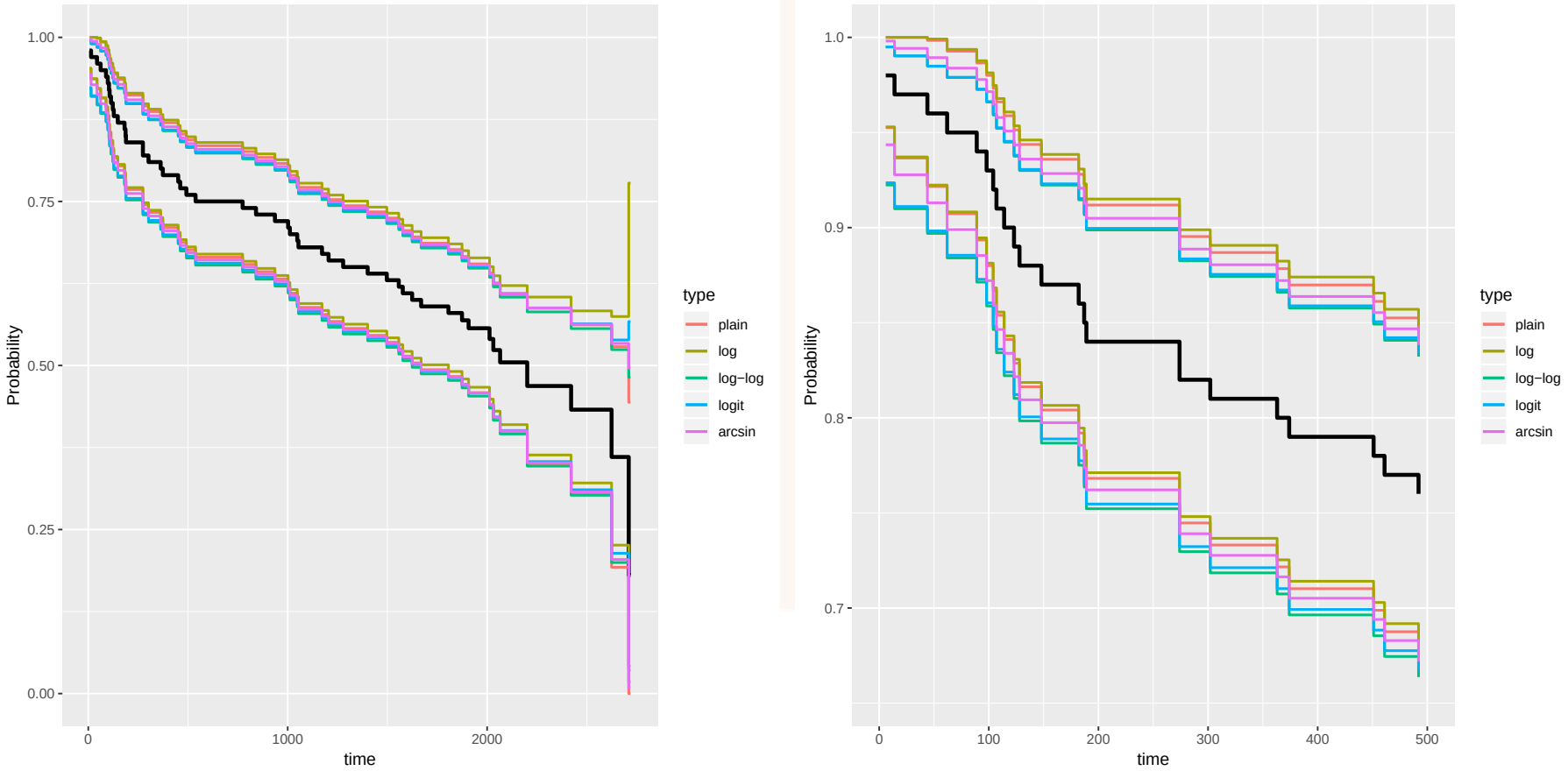
$$\begin{aligned}
\texttt{plain} \quad & g(x) = x \\
\texttt{log} \quad & g(x) = \log(x) \\
\texttt{log-log} \quad & g(x) = \log\{-\log(x)\} \\
\texttt{logit} \quad & g(x) = \log\left(\frac{x}{1-x}\right) \\
\texttt{arcsin} \quad & g(x) = \arcsin\sqrt{(x)}
\end{aligned}$$

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

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

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 by 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):

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

- Like the variance, the larger the SIQR, the more dispersed is the survival time distribution.

*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
```

- Median follow-up time does not always 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?

Counting processes

- As is seen from before, the focus in survival analysis is on observing the occurrence of events over time.
- Such occurrences constitute point (counting) process; counting number of events as they come along.
- There is a very neat collection of theories for counting processes.
- More in-depth details can be found in Fleming and Harrington (2011); Kalbfleisch and Prentice (2011).

Counting processes

- Appendix 2 gives a short introduction of counting processes, in the case of right censoring.
- We start by focusing on a *single type* of event without censoring.
- For a given time t, let $N(t)$ be the number of events over the time period $(0, t]$, then $N(t)$ is a counting process.
- The counting process $N(t)$ is *continuous from the right*, with jump size 1.

Poisson process

- A well-known example of a counting process is the homogeneous Poisson process.
- Jumps occur randomly and independently of each other (independent increment property).
- A homogeneous Poisson process is described by its *intensity* λ.
- The λ is the probability of occurrence of an event in a small interval divided by the length of the interval.
- We will apply this idea to modeling counting process.

Counting processes

- Suppose the intensity function that describes $N(t)$ is $\lambda(t)$.
- Under our assumption that a subject can experience at most one event, consider a small time interval $[t, t + \mathrm{d}t)$, $\lambda(t)$ is

$$\lambda(t)\mathrm{d}t = \mathrm{P}(\mathrm{d}N(t) = 1|\text{past}),$$

  where $\mathrm{d}N(t)$ denotes the # of events in $[t, t + dt)$, or $N(t + \mathrm{d}t) - N(t)$.
- Formally speaking, the "past" represents the *filtration* of the process up to but not including time t.

Counting processes

- Under our assumption, $\mathrm{d}N(t)$ is binary, and

$$\lambda(t)\mathrm{d}t = \mathrm{E}\{\mathrm{I}(\mathrm{d}N(t) = 1)|\text{past}\} = \mathrm{E}\{\mathrm{d}N(t)|\text{past}\}.$$

- Then it follows

$$\mathrm{E}(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t|past) = 0. \tag{5}$$

- If we define a new process

$$M(t) = N(t) - \int_0^t \lambda(s)\mathrm{d}s, \tag{6}$$

  then we have $\mathrm{E}(\mathrm{d}M(t)|past) = 0$.
- It turns out $M(t)$ is a zero-mean martingale.

Counting processes

- Definition (6) can be written as

$$\mathrm{d}N(t) = \lambda(t)\mathrm{d}t + \mathrm{d}M(t), \tag{7}$$

  reflecting the relationship *observation = signal + noise*.
- Think of $M(t)$ as the sum of random errors.
- Let

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s,$$

  (5) and (6) implies $\Lambda(t)$ is the cumulative expected # of events in $(0, t]$, or $\Lambda(t) = \mathrm{E}\{N(t)\}$.
- $\Lambda(t)$ versus $H(t)$?

Counting processes

- Now we will look at the scenario when there is only *one event type* and a subject can experience *at most one event* (no ties in event times).
- In this case, the counting process $N(t) = \mathrm{I}(T \leq t)$, where T is the survival time with hazard $h(t)$.
- The $N(t)$ defined above has a jump size 1 at T.
- We then have

$$\mathrm{P}(\mathrm{d}N(t) = 1|\text{past}) = \begin{cases} h(t)\mathrm{d}t & \text{if } T \geq t \\ 0 & \text{if } T < t \end{cases} = h(t)\mathrm{I}(T \geq t)\mathrm{d}t.$$

- The relationship above implies the intensity process $\lambda(t) = h(t) \cdot \mathrm{I}(T \geq t)$.

Counting processes

- Building onto our assumption, now assume we have *n independent subjects*, each with T_i, $i = 1, \ldots, n$, and hazard $h_i(t)$.
- Suppose $h_i(t) = h(t)$ for all i.
- From the last example, we have

$$N_i(t) = \mathrm{I}(T_i \leq t) \text{ and } \lambda_i(t) = h_i(t)\mathrm{I}(T_i \geq t).$$

- Define the *aggregated* process by adding together the individual processes:

$$N(t) = \sum_{i=1}^{n} N_i(t).$$

- This process counts the # of individuals who experienced the event by t.
- Is $N(t)$ here a proper counting process?

Counting processes

- Assuming continuous survival times, and the aggregated processs jumps one unit at a time,

$$\mathrm{E}(\mathrm{d}N(t)|\text{past}) = \sum_{i=1}^{n} \mathrm{E}(\mathrm{d}N_i(t)|\text{past})$$

  implies the aggregated intensity satisfies $\lambda(t)\mathrm{d}t = \sum_{i=1}^{n} \lambda_i(t)\mathrm{d}t$ and

$$\lambda(t) = h(t)\,Y(t),$$

  where $Y(t) = \sum_{i=1}^{n} \mathrm{I}(T_i \geq t)$ is the number of individuals at risk right before t, e.g., $Y(t) = n - N(t^-)$.
- Plug $\lambda(t)$ into Equation (7) gives the Nelson-Aalen estimator for $H(t)$.

Derivation of $\widehat{H}_{NA}(t)$

- Plug $\lambda(t)$ into Equation (7), we have

$$\mathrm{d}N(t) = h(t)Y(t)\mathrm{d}t + \mathrm{d}M(t)$$

- Intergrating both side after diving by $Y(t)$ gives

$$\int_0^t \frac{1}{Y(t)}\mathrm{d}N(t) - \int_0^t h(t)\mathrm{d}t = \int_0^t \frac{1}{Y(t)}\mathrm{d}M(t)$$

- The right-hand side is a stochastic integral with respect to a zero-mean martingale.
- The first term in the left-hand side gives $\widehat{H}_{NA}(t) = \sum_{t_{(i)} \le t} \frac{1}{n_i}$.
- Then $\mathrm{E}\left\{\widehat{H}_{NA}(t) - \widehat{H}(t)\right\} = 0$ and $\widehat{H}_{NA}(t)$ is an unbiased estimator of $\widehat{H}(t)$.
- The variance estimator can be derived using the *variation processes* or with transformation through the Delta methods.

Counting processes

- Consider the same assumptions but allowing ties in event times.
- Two common approaches to deal with tied event times depend on how tied event times arise.
  1. Two event times coincide due to rounding.

$$\widehat{H}_{NA}(t) = \sum_{t_{(i)} \le t} \sum_{k=1}^{d_i} \frac{1}{n_i - k}.$$

  2. Event times are genuinely discrete, so that tied event times are real and not due to rounding.

$$\widehat{H}_{NA}(t) = \sum_{t_{(i)} \le t} \frac{d_i}{n_i}.$$

- More discussion on large number properties with tied event times can be found in Anderson et al. (1993).

Counting processes

- For nonparametric or semi-parametric methods, we do not need to have a distributional assumption on the censoring time, C.
- Rather, we will assume independent censoring:

$$\mathrm{P}(t \leq \tilde{T}_i < t + \mathrm{d}t, \Delta_i = 1 | \tilde{T}_i \geq t, \text{past}) = \mathrm{P}(t \leq T_i < t + \mathrm{d}t | T_i \geq t),$$

  where $\tilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = \mathrm{I}(T_i < C_i)$, for $i = 1, \ldots, n$.
- In the presense of right censoring, the counting process is specified as

$$N_i(t) = \mathrm{I}\{\tilde{T}_i \leq t, \Delta_i = 1\}, i = 1, \ldots, n,$$

  and $\lambda_i(t)\mathrm{d}t = \mathrm{P}(\mathrm{d}N_i(t) = 1 | \text{past})$.
- Applying the independent censoring assumption, the intensity process for $N_i(t)$ takes the form $\lambda_i(t) = h_i(t)Y_i(t)$, where $Y_i(t) = \mathrm{I}(\tilde{T}_i \geq t)$ is the at risk indicator for individual i.

Counting processes

- As in the uncensored example, the *aggregated* process is

$$N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{i=1}^{n} \mathrm{I}(\tilde{T}_i \leq t, \Delta_i = 1),$$

  and the *aggregated* intensity process is

$$\lambda(t) = \sum_{i=1}^{n} \lambda_i(t) = \sum_{i=1}^{n} h_i(t) Y_i(t).$$

- In the case where $h_i(t) = h(t)$ for all i, the intensity process takes the form $\lambda(t) = h(t)Y(t)$, where $Y(t) = \sum_{i=1}^{n} Y_i(t)$.
- The Nelson-Aalen estimator can be derived similarly.

Counting processes

- Counting processes are non-decreasing.
- Martingales associated with $M(t)$ are sub-martingales.
- Doob-Meyer decomposition guarantees the relationship

$$N(t) = h(t)Y(t) + M(t).$$

- Martingale central limit theorem.

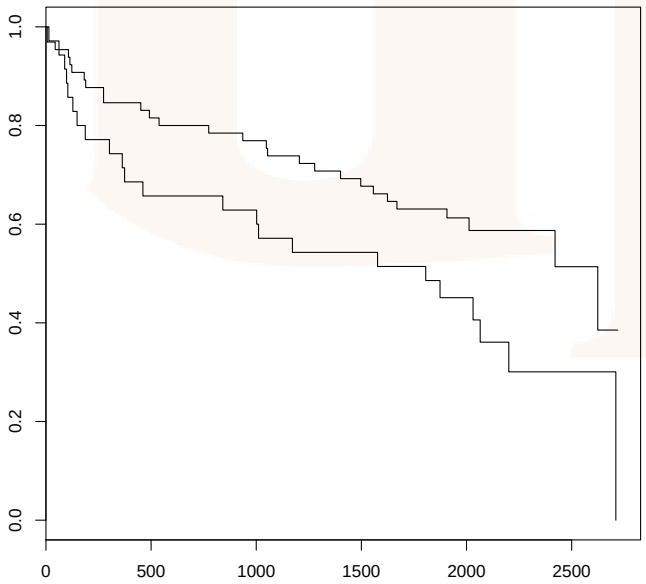
Comparison of survival functions

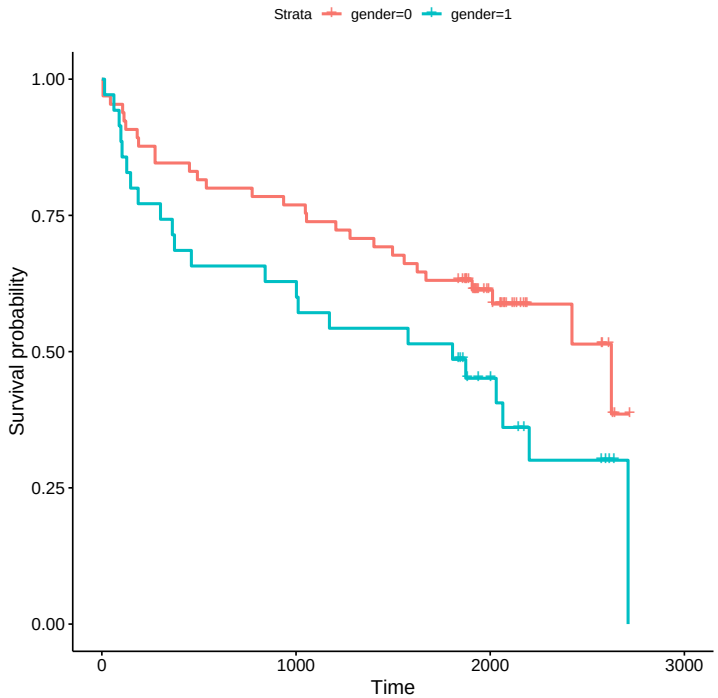
- The simplest way of comparing the survival functions is to plot them on the same axe.

```
> km <- survfit(Surv(lenfol, fstat) ~ gender, data = whas100)
> km
Call: survfit(formula = Surv(lenfol, fstat) ~ gender, data = whas100)

           n events median 0.95LCL 0.95UCL
gender=0 65     28   2624    2012      NA
gender=1 35     23   1806     841      NA
```

- `ggsurvplot` gives more informative output.





Comparison of survival functions

- In the `whas100` study, the survival function for female patients (`gender = 0`) is always greater than that for male patients.
- Two explanations for the observed difference:
  - A real difference between the survival times of the two groups.
  - The difference has been observed is the result of chance variation.
- Procedures such as the *hypothesis test* is needed to distinguish the two possible explanations.

Hypothesis testing

- As the hypothesis tests commonly encountered, the testing procedure here consists of the following major steps:
  1. State the *null hypothesis* and the *alternative hypothesis*.
  2. Formulate a *test statistic* that measures the extent to which the observed data depart from the null hypothesis.
  3. Calculate the *p*-value; the probability of obtaining a value as extreme or more extreme than the observed value, when the null hypothesis is true.

Hypothesis testing

- p-value can be interpreted as a measure of the strength of evidence against the null hypothesis.
- A p-value of 0.000 should be interpreted as $p < 0.001$.
- A one-sided hypothesis test is only appropriate when there is no interest in departures from the null hypothesis in the opposite direction.
- We will look at non-parametric procedures including the log-rank test and the Wilcoxon test.

The log-rank test

- If we are willing to assume survival data follow a particular parametric distribution, we can construct a test based on likelihood theory.
- We will derive a nonparametric test using a rank-based procedure.
- Suppose the population survival curves are $S_1(t)$ and $S_0(t)$.
- Ideally, we would like to test $H_o : S_1(t) = S_0(t)$ versus $H_a : S_1(t) > S_0(t)$, $S_1(t) < S_0(t)$, or $S_1(t) \neq S_0(t)$.
- Unfortunately, this is difficult because H_a can take a wide range of forms.
- Instead, it is more feasible to consider the *Lehman alternative*:

$$H_a : S_1(t) = [S_0(t)]^{\theta}, \text{ for } \theta > 0. \tag{8}$$

The log-rank test

- The hypothesis in (8) can be view in terms of the hazard functions:

$$H_o : \lambda_1(t) = \lambda_0(t) \text{ versus } H_a : \lambda_1(t) = \theta\lambda_0(t).$$

- Either way, we could rewrite the hypothesis test as

$$H_o : \theta = 1 \text{ versus } H_a : \theta > 1, \text{ or } \theta < 1, \text{ or } \theta \neq 1.$$

- One can interpret the results solely based on θ too.
- For example, under $H_a : \theta > 1$, then
  - $S_1(t)$ will be uniformly lower than $S_0(t)$.
  - $h_1(t)$ will be uniformly higher than $h_0(t)$.

The log-rank test

- Following the notations defined previously, we assume
  - the two independent random samples are of size N_0 and N_1.
  - D is the total number of events in the combined sample (size $N_0 + N_1$).
  - $\{t_{(1)}, \ldots, t_{(D)}\}$ is the ordered event times.
  - Let d_{0i} and d_{1i} are the number of events in group 0 and 1, respectively.
  - Let n_{0i} and n_{1i} are the number of at risk in group 0 and 1, respectively.

The log-rank test

- For the ith event time, we construct the following two-by-two table:

	Group 1	Group 0	Total
Failure	d_{1i}	d_{0i}	d_i
Non-failure	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
At risk	n_{1i}	n_{0i}	n_i

- Suppose the numbers of failures in two groups are independent, holding the margins fixed, then the distribution of d_{1i} (or d_{0i}) follows a *hypergeometric distribution*.
- This gives

$$\mathrm{E}(d_{1i}) = \frac{n_{1i} \cdot d_i}{n_i}, \text{ and } \mathrm{Var}(d_{1i}) = \frac{n_{1i} \cdot n_{0i} \cdot d_i \cdot (n_i - d_i)}{n_i^2 \cdot (n_i - 1)}.$$

The log-rank test

- Using the idea from the χ^2 tests, the test statistic has the form

$$Q = \frac{\left[\sum_{i=1}^{D} \omega_i \{d_{1i} - \mathrm{E}(d_{1i})\}\right]^2}{\sum_{i=1}^{D} \omega_i^2 \mathrm{Var}(d_{1i})}, \tag{9}$$

  where ω_i is a possibly data-dependent weight.
- When there is no tied failure times, $\mathrm{Var}(d_{1i}) = \frac{n_{1i} \cdot n_{0i}}{n_i^2}$.
- Under the null, Q follows the χ^2 distribution with one degree of freedom.
- The log-rank test sets $\omega_i = 1$ (Peto and Peto, 1972).

The log-rank test

- Different selections of ω_i were studied.
- Most common expression has the form of

$$\omega_i = \left\{ \hat{S}(t_{(i)}) \right\}^{\rho}, \text{ for } \rho \in [0, 1],$$

  where $\hat{S}(t)$ is the survival estimator from the combined sample, ignoring group.
- This is called the Fleming-Harrington $G(\rho)$ test (e.g. Cox and Oakes, 1984; Harrington and Fleming, 1982).
  - $\rho = 0$ gives the log-rank test
  - $\rho = 1$ gives the Wilcoxon test.
  - $\rho = 0.5$ gives the Tarone and Ware (1977) test.

The log-rank test

- The $G(\rho)$ test is implemented in `survdiff` of the **`survival`** package

```
> args(survdiff)
function (formula, data, subset, na.action, rho = 0)
NULL
> survdiff(Surv(lenfol, fstat) ~ gender, data = whas100)
Call:
survdiff(formula = Surv(lenfol, fstat) ~ gender, data = whas100)

          N Observed Expected (O-E)^2/E (O-E)^2/V
gender=0 65       28     34.6      1.27      3.97
gender=1 35       23     16.4      2.68      3.97

 Chisq= 4  on 1 degrees of freedom, p= 0.05
```

- The log-rank test shows `gender` is marginally significant at $\alpha = 0.05$.

The log-rank test

- Varying ρ shows similar results here.
- In general, when the tests give different results, then more than one result should be reported.
- It is important to also consult the plot to ascertain the directional effect of stratification or treatment.

The log-rank test

- As in the Pearson's χ^2 test, a continuity correction could be considered when sample sizes are too small:

$$Q^* = \frac{\left[\sum_{i=1}^{D} \omega_i \{|d_{1i} - \mathrm{E}(d_{1i})| - 0.5\}\right]^2}{\sum_{i=1}^{D} \omega_i^2 \mathrm{Var}(d_{1i})}.$$

The log-rank test

- A problem can occur if the estimated survival (hazard) functions cross one another.
- This is similar to the proportional hazard assumption in the Cox models.
- In `whas100` data, the two survival curves (by gener) crossed one another at early t, despite being so different in large t.
- Fleming et al. (1987) proposed supremum versions of the log-rank tests, but have not been implemented in any software packages.

The log-rank test

- The tests discussed here can be extended to comparison more than two survival functions simultaneously.
- K-by-K tables instead of two-by-two tables at each event time.
- When there are many groups to compare, it might be more convenient to consider regression models.

Reference

Anderson, P., Borgan, O., Gill, R., and Keiding, N. (1993). Statistical methods based on counting processes.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*, volume 21. CRC Press.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Fleming, T. R., Harrington, D. P., and O'sullivan, M. (1987). Supremum versions of the log-rank and generalized wilcoxon statistics. *Journal of the American Statistical Association* **82,** 312–320.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69,** 553–566.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)* pages 185–207.

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.

Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64,** 156–160.