# 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 $F(t) = P(T \leq t)$.
- The survival function of $T$ is then defined as

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

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

```
> library(tidyverse)
```

# Survivor, hazard and cumulative hazard functions

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

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

- $\Pr(t \leq T < t + \delta | T \geq t)$ is a conditional probability.
- The conditional probability is then expressed as a probability per nit time by dividing by the time interval, $\delta$, 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 $\mathrm{E}\{\mathrm{I}(\cdot)\} = \Pr(\cdot)$.
  - 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:

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

- $h(t)$ 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 $\lambda$ 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

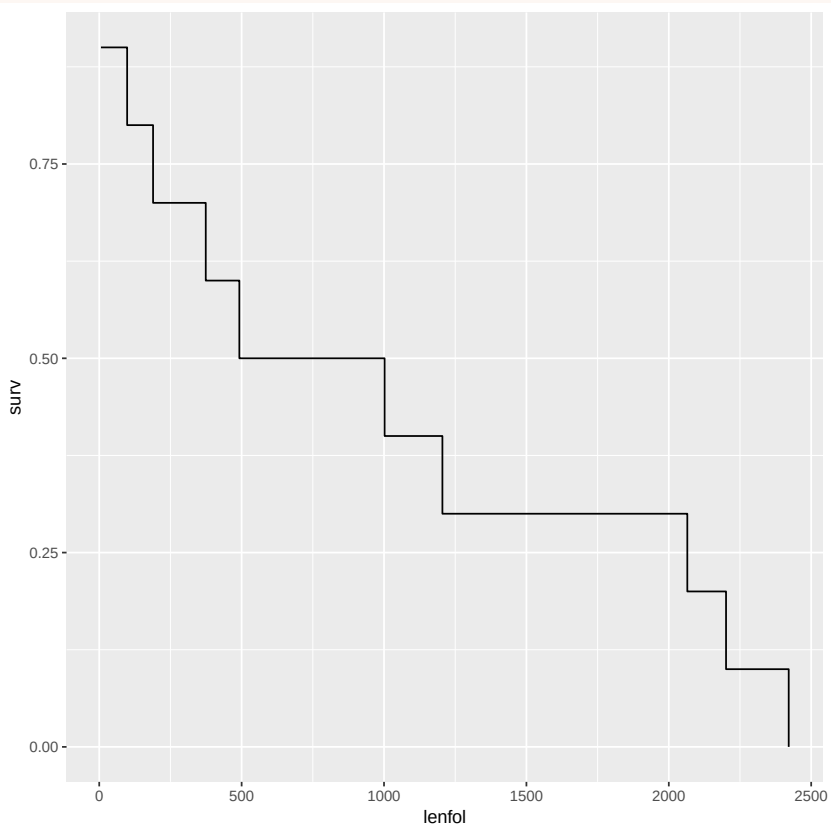
- The empirical estimates can be easily computed with `ecdf`.

```
> 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.

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

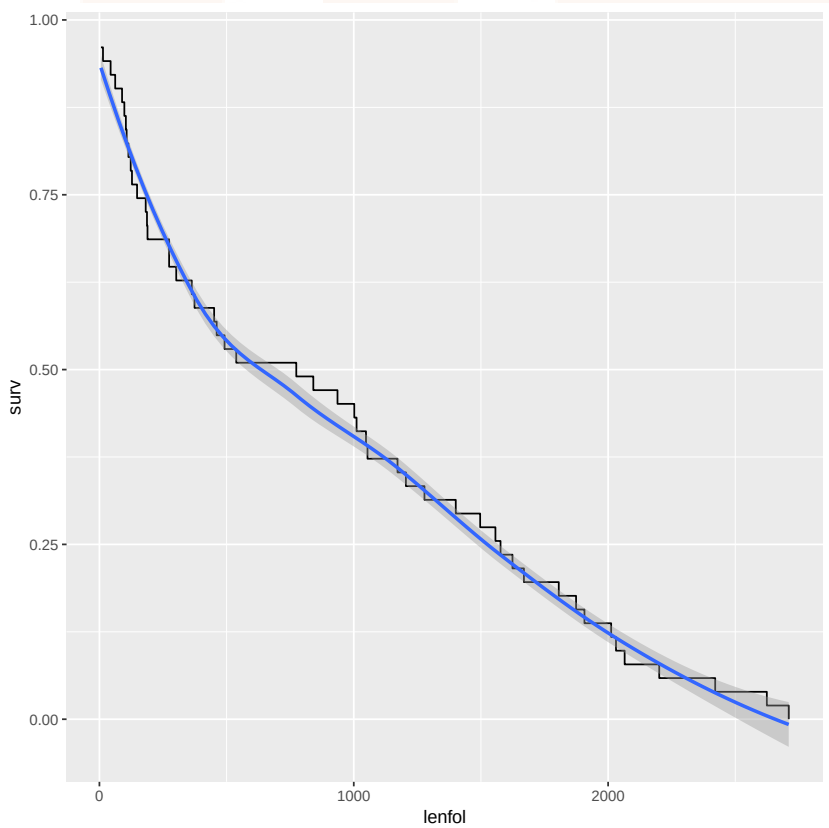



- The $\hat{S}_e(t)$ is 1 at $t = 0$ and 0 at the final death time.
- The $\hat{S}_e(t)$ is assumed to be constant between adjacent death times.

# Empirical 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

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