WAKE FOREST
UNIVERSITY
Department of Statistical Sciences

Part IIa. Basic Quantities

Sarah Lotspeich
STA779: *Applied Survival Analysis (Spring 2023)*

Assumptions

Before we get started, there are a few key assumptions we will make throughout this class.

- **Non-negative failure times**: If X denotes time to the event of interest (or failure time), then $T \geq 0$.
  - For now, X is a random variable from a *homogeneous population* (this will be relaxed later).
- **Types of failure times**: The failure-time variable X can be discrete (i.e., it takes on a finite or countable number of value) or continuous (i.e., it can take on any value in an interval like $(0, \infty)$).
  - Discrete Xs arise most often due to rounding (e.g., from continuous time to integer days) or because the variable was originally collected in integer units (e.g., lifetimes).
- **Non-negative censoring times**: If C denotes the censoring variable of X, then $C \geq 0$, too. (Depending on the type of censoring, there will be further constraints on C later.)

Also, a "failure" is not necessarily a bad thing. For example, the event of interest might be end of treatment.

Characterizing the event time distribution

There are four functions which characterize the distribution of event time X, each used to illustrate different aspects. Conveniently, if we know any one, we can know them all.

1. Probability density (or mass) function:
2. Survival function: The probability of an individual surviving to time x.
3. Hazard (rate) function: The chance that an individual who survived to x experiences an event in the next instant in time.
4. Mean residual life: The expected time to the event, given that the event has not yet occurred by time x.

Another useful quantity can be derived from these: the cumulative hazard function, which is the total risk of the event accumulated up to time x.

The probability mass and density functions

- Recall that upper-case letters like X are **random variables**, whereas lower-case letters like x denote non-random, **possible values** of their upper-case counterparts.
- The **probability mass function (PMF)** and **probability density function (PDF)** describe the relative likelihood of observing all of the possible values x.
- For discrete X, which takes on possible values $x_1 < \cdots < x_p$, the PMF is defined

$$p(x) = \Pr(X = x), \tag{2.1}$$

  for $x \in \{x_1, \ldots, x_p\}$ and $p(x) = 0$ otherwise.
- For continuous X, which could possibly take on any value $x \in [0, \infty)$, the PDF is defined

$$f(x) = \lim_{\Delta \to 0} \frac{\Pr(x \leq X < x + \Delta)}{\Delta}, \tag{2.2}$$

  and $f(x) = 0$ otherwise.

Reminders about the PMF and PDF

- Importantly, $f(x)$ is a density, *not* a probability.
- However, $f(x)\,dx$ can be thought of as the **"approximate" probability** that the event will occur at time x (unconditionally on whether the individual is "alive" just prior to x).
- If $f(x)$ is the PDF for a continuous failure-time variable X, then
  1. Nonnegative: $f(x) \geq 0$ for all $0 \leq x < \infty$ and
  2. Integrates to 1: $\int_0^\infty f(x)\,dx = 1$.
- If $p(x)$ is the PMF for a discrete failure-time variable X, then it possesses the same two properties (except it sums to 1 rather than integrates).

The survival function

As its name might suggest, the star of this show is often going to be the survival function, defined as

$$S(x) = \Pr(X > x), \tag{2.3}$$

and interpreted as the **probability of an individual surviving beyond time x**, or, generally, experiencing the event after time x.

- Rate of decline in $S(x)$ depends on risk of experiencing the event across time, but this is better explained by other quantities we will discuss.
- Still, survival functions are popular in describing survival data and can be useful in comparing mortality patterns.

Example survival curve in the `gbsg` data

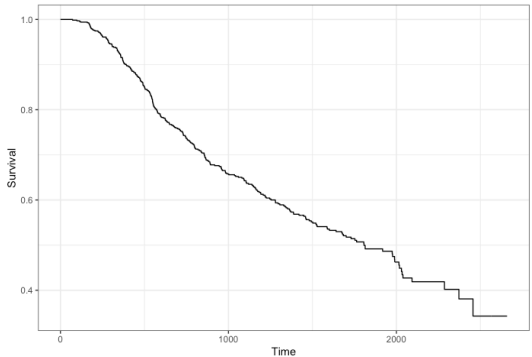

Figure: Survival curve of $n = 686$ patients with node positive breast cancer from a 1984–1989 trial conducted by the German Breast Cancer Study Group (GBSG). The event was defined as death or disease recurrence. These data are available in the `survival` package in R as `gbsg`.

Find the following from the graph:

- By what time had half of the patients either died or experienced disease recurrence?
- What is the survival probability 1000 days from study start?
- What is the probability of patients either dying or experiencing disease recurrence within 1000 days of the study start?

Properties of the survival function

We will discuss many types of survival curves for X in this class, but they must all have the same basic properties:

1. $S(x)$ is a monotone, nonincreasing function,
2. $S(x) \to 1$ as $x \to 0$, and
3. $S(x) \to 0$ as $x \to \infty$.

You might notice that these properties follow pretty closely as the opposites of the properties of the CDF.

Connecting $S(x)$ to probability quantities

To start, suppose that X is continuous. In this case, the survival function has convenient connections to the **cumulative distribution function (CDF)** $F(x)$

$$S(x) = \Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F(x) \tag{2.4}$$

or the **probability density function (PDF)** $f(x)$

$$S(x) = \int_x^\infty f(t)\,dt. \tag{2.5}$$

And thanks to (2.5), you can even get the back to the PDF from the survival function as

$$f(x) = -\frac{d}{dx}S(x). \tag{2.6}$$

Your turn: Survival function for the `lung` data

`lung`: *Let X be survival time for patients with advanced lung cancer. Using the `lung` data from $n = 228$ patients in the North Central Cancer Treatment Group (available in the `survival` package), I found that the distribution of X is well approximated by an exponential distribution with rate $\beta = \exp(-6)$. Note: $f(x) = \beta \exp(-\beta x)$.*

Derive the survival function for X.

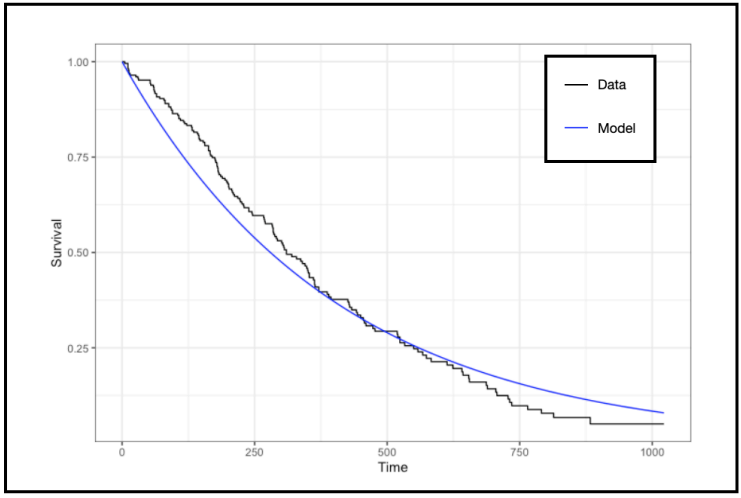

Figure: Comparison of the observed (data) and estimated (exponential model) survival curves in the `lung` data.

Survival functions for discrete X

- Now consider a discrete failure time X, which can take on a values x_j ($j = 1, 2, \dots$).
- Recall that the **probability mass function (PMF)** describes the relative likelihood that X takes on any of these potential values x_j and is denoted $p(x_j) = \Pr(X = x_j)$.
- For ease of notation, suppose that the possible values are ordered, i.e., $x_1 < x_2 < \cdots$.
- In place of (2.5), we can **calculate the survival function for discrete X** from its PMF as

$$S(x) = \sum_{x_j > x} p(x_j). \tag{2.7}$$

- When X is discrete, $S(x)$ is a nonincreasing **step function**.
- Think about it: If an event occurs *exactly* at time x, does it contribute to $F(x)$ or $S(x)$?

Rewriting $S(x)$ in terms of conditional survival

We could also write the survival function at x_j ($j = 1, \ldots, p$) as the **product of the conditional survival probabilities** for all intervals up to it:

$$
\begin{aligned}
S(x_j) &= \Pr(X > x_j) = \Pr(X \geq x_{j+1}) = \Pr(X \geq x_{j+1}, \ldots, X \geq x_2, X \geq x_1) \\
&= \Pr(X \geq x_{j+1} | \ldots, X \geq x_2, X \geq x_1) \cdots \Pr(X \geq x_2 | X \geq x_1) \Pr(X \geq x_1) \\
&= \frac{\Pr(X \geq x_{j+1}, \ldots, X \geq x_2, X \geq x_1)}{\Pr(X \geq x_j, \ldots, X \geq x_2, X \geq x_1)} \cdots \frac{\Pr(X \geq x_2, X \geq x_1)}{\Pr(X \geq x_1)} \Pr(X \geq x_1) \\
&= \frac{\Pr(X > x_j)}{\Pr(X > x_{j-1})} \cdots \frac{\Pr(X > x_1)}{\Pr(X > x_0)} \Pr(X > x_0) \\
&= \frac{S(x_j)}{S(x_{j-1})} \cdots \frac{S(x_1)}{S(x_0)} S(x_0) = \prod_{k=1}^{j} \frac{S(x_k)}{S(x_{k-1})},
\end{aligned}
\tag{2.8}
$$

where $S(x_0) \equiv 1$.

The hazard function

The hazard function estimates the chance that an individual who already survived to time x will experience an event in the next *instant* in time, and is defined as

$$h(x) = \lim_{\Delta \to 0} \frac{\Pr(x \leq X < x + \Delta | X \leq x)}{\Delta}. \tag{2.9}$$

- There is only **one condition (nonnegative):** $h(x) \geq 0$ for all $0 \leq x < \infty$.
- Can view $h(x)\,dx$ as the **"approximate" probability** that the event will occur at time x (conditionally on the individual being "alive" just prior to x).[1]
- Particularly useful in determining appropriate distributions based on the context or in describing how the chance of experiencing the event changes over time. (Generally **more informative** about the underlying mechanism than the survival, too.)

[1] Think about it: How does this differ from the approximate probability on slide 4?

Different shapes of hazards

Based on the context of your data, you might believe that the hazard rate for the occurrence of the event has a particular shape.

1. Increasing (ex: machine parts wearing out over time)
2. Decreasing (ex: patient experience after an organ transplant)
3. Constant (ex: being struck by a meteor)
4. Bathtub-shaped (ex: higher mortality rates among infants and elderly)
5. Hump-shaped (ex: patient outcomes after surgery)

There are other shapes, but these are the most common. **Specific distributions** capture these different shapes for us.

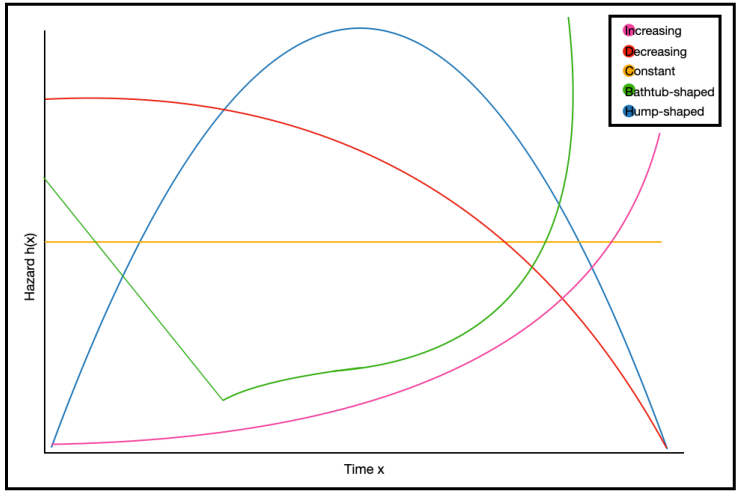

Figure: Illustration of common shapes of hazard functions.

Your turn: Identifying shapes

For each of the failure-time variables, which of the **5 common shapes of hazards** do you think would be most appropriate?

1. X_1: time from vaccination to peak antibody response
2. X_2: time from arrival at bus stop to getting on bus
3. X_3: time from eating at a restaurant to getting food poisoning
4. X_4: time from start of semester to thinking about dropping a class (hopefully not this one!)
5. X_5: time from birth to arthritis diagnosis

Connecting $h(x)$ to probability quantities

Again, suppose for now that X is continuous. In this case,

$$h(x) = \frac{f(x)}{S(x)}. \tag{2.10}$$

Proof:

Connecting $h(x)$ to probability quantities

Proof (Continued):

Connecting $h(x)$ to other survival quantities

- Having proven (2.10), we can proceed to relate $h(x)$ to just $S(x)$.
- For a left-continuous survival function $S(x)$, we can show that

$$f(x) = -\frac{d}{dx} S(x). \tag{2.11}$$

- Begin by considering the end result (cheating, I know):

$$-\frac{d}{dx}\{\ln S(x)\} = -\frac{\frac{d}{dx}S(x)}{S(x)} = -\frac{-f(x)}{S(x)} = \frac{f(x)}{S(x)} = h(x). \checkmark \tag{2.12}$$

Hazard functions for discrete X

- Now consider a **discrete failure time** X, which can take on a distinct values $x_1 < \cdots < x_p$ (no ties).
- In place of (2.10), we can **calculate the hazard function for discrete X** as

$$h(x_j) = \Pr(X = x_j | X \geq x_j) = \frac{\Pr(X = x_j, X \geq x_j)}{\Pr(X \geq x_j)} = \frac{p(x_j)}{S(x_{j-1})} = \frac{p(x_j)}{\sum_{x_k > x_{j-1}} p(x_k)}, \quad (2.13)$$

  for $j = 1, \ldots, p$ and $h(x) = 0$ otherwise. Note: $S(x_0) \equiv 1$.

- Can also get the hazard function in terms of **survival alone**:

$$h(x_j) = 1 - \frac{S(x_j)}{S(x_{j-1})} \quad (2.14)$$

  for $j = 1, \ldots, p$. (How??)

Defining discrete survival in terms of hazard

Recall that (2.8) defines $S(x)$ in terms of the conditional survival probabilities in the following way:

$$S(x_j) = \prod_{k=1}^{j} \frac{S(x_k)}{S(x_{k-1})}.$$

Now, manipulating (2.14), we have $S(x_k)/S(x_{k-1}) = 1 - h(x_k)$, which we can plug in to get

$$S(x_j) = \prod_{k=1}^{j} \{1 - h(x_k)\} \tag{2.15}$$

for $j = 1, \ldots, p$.

Properties of hazard functions

For continuous failure-times X, we have the following named properties based on the behavior of their hazard function $h(x)$.

1. **Increasing failure-rate (IFR)**: if $h(x)$ is nondecreasing for $x \geq 0$
2. **Increasing failure-rate on the average (IFRA)**: if $H(x)/x$ is nondecreasing for $x > 0$
3. **Decreasing failure-rate (DFR)**: if $h(x)$ is nonincreasing for $x \geq 0$

The cumulative hazard function

A (very) related quantity to the hazard function is the cumulative hazard function, defined as

$$H(x) = \int_0^x h(t)\,dt, \tag{2.16}$$

i.e., the total chance (or hazard) of the event that the individual accumulated up to time x.

- If X is discrete, we instead define

$$H(x_j) = \sum_{k=1}^{j} h(x_k), \tag{2.17}$$

for $j = 1, \ldots, p$.

Connecting $H(x)$ to other survival quantities

For a continuous failure-time X, we can relate the cumulative hazard function to the survival function as follows using the connections defined thus far:

$$\begin{aligned} H(x) &= \int_0^x h(t)\,dt = \int_0^x \frac{f(t)}{S(t)}\,dt \\ &= \int_0^x \frac{-\frac{d}{dt}S(t)}{S(t)}\,dt = -\ln\{S(t)\}\Big|_{t=0}^{x} \\ &= \ln\{S(x)\} + \ln\{S(0)\} = \ln\{S(x)\} + \ln(1) = \ln\{S(x)\} \\ \rightarrow S(x) &= \exp\{-H(x)\}, \end{aligned} \tag{2.18}$$

where $\exp(t) = e^t$.

- Conversely, the hazard function can also be calculated as the **rate of change of the cumulative hazard**, i.e., $h(x) = \frac{d}{dx}H(x)$, which agrees with (2.12).

Alternative formula for discrete $H(x)$

Initially (i.e., in (2.17)), we defined the cumulative hazard function for a discrete X as

$$H(x_j) = \sum_{k=1}^{j} h(x_k), \tag{2.17}$$

for $j = 1, \ldots, p$. This formula, proposed by **Kaplan and Meier**, is *one* way to define $H(x)$, but then $S(x_j) \neq \exp\{-H(x_j)\}$.[2]

Alternatively, **Cox** defined the following so that (2.18) could be used for discrete X as well:

$$H(x_j) = -\sum_{k=1}^{j} \ln\{1 - h(x_k)\}. \tag{2.19}$$

In fact, (2.17) **is an approximation of** (2.19) when $h(x_k)$ is small, i.e.,

$$-\frac{\ln\{1 - h(x_k)\}}{h(x_k)} \to 1 \text{ when } h(x_k) \to 0.$$

[2] This is the one our book uses.

Your turn: Hazard function for the `lung` data

`lung` (revisited): *Recall that X is survival time for patients with advanced lung cancer, where $X \sim expo(\beta = \exp(6))$.*

Derive the hazard and cumulative hazard functions for X.

Does X possess any of the properties from the previous slide?

Measures of central tendency in survival

There are a few **measures of central tendency** which can be used to summarize the distribution of failure time X.

1. Mean residual life[3]:
2. Mean survival:
3. Median survival:
4. pth quantile:

[3] The mean residual life is a conditional function, so only people living to at least time x are considered in its calculation. For the other measures, everyone contributes.

Measures of central tendency in survival

There are a few **measures of central tendency** which can be used to summarize the distribution of failure time X.

1. Mean residual life[3]: at time x, the mean time to event *given* that the individual was event-free up until just before x.
   - Ex: According to a study in the *Journal of Women & Aging*, women with heart disease who have already lived to be 50 years old are expected longer than men who do ($\text{mrl}(50) = 7.9$ vs. 6.7 more years).
2. Mean survival:
3. Median survival:
4. pth quantile:

[3] The mean residual life is a conditional function, so only people living to at least time x are considered in its calculation. For the other measures, everyone contributes.

Measures of central tendency in survival

There are a few **measures of central tendency** which can be used to summarize the distribution of failure time X.

1. Mean residual life[3]:
2. Mean survival: at time 0, the mean time to event.
   - Ex: According to the Centers for Disease Control and Prevention (CDC), the life expectancy from birth (i.e., mean survival) in the United States was calculated to be $\mathrm{E}(X) = 76.1$ years in 2021.
3. Median survival:
4. pth quantile:

[3]The mean residual life is a conditional function, so only people living to at least time x are considered in its calculation. For the other measures, everyone contributes.

Measures of central tendency in survival

There are a few **measures of central tendency** which can be used to summarize the distribution of failure time X.

1. Mean residual life[3]:
2. Mean survival:
3. Median survival: the time $\phi_{0.5}$ at which exactly 50% of individuals survive.
   - Ex: According to a study in *Movement Disorders Clinical Practice*, the median survival for a person with Huntington's disease was $\phi_{0.5}^{(d)} = 24$ years from diagnosis and $\phi_{0.5}^{(s)} = 35$ years from symptom onset
4. pth quantile:

[3] The mean residual life is a conditional function, so only people living to at least time x are considered in its calculation. For the other measures, everyone contributes.

Measures of central tendency in survival

There are a few **measures of central tendency** which can be used to summarize the distribution of failure time X.

1. Mean residual life[3]:
2. Mean survival:
3. Median survival:
4. pth quantile: the time ϕ_p ($p \in [0, 1]$) at which exactly $100p\%$ of individuals survive.
   - Ex: According to a study in *Circulation*, the $p = 0.55$th quantile for survival (among people < 50 years old) following coronary artery surgery was $\phi_{0.55} = 20$ years.

[3] The mean residual life is a conditional function, so only people living to at least time x are considered in its calculation. For the other measures, everyone contributes.

The mean residual life function

The mean residual life function summarizes the distribution of X by the expected time to the event, given that it has not yet occurred by time x:

$$\mathrm{mrl}(x) = \mathrm{E}(X - x | X > x). \tag{2.15}$$

- In mortality studies, $\mathrm{mrl}(x)$ can literally be interpreted as a person's **expected remaining lifetime** having already lived to age x.
- Answers questions like "if the person survived to time x, how much longer are they expected to live?"

Definition of $\mathbf{mrl}(x)$

For a continuous[4] failure-time X, the mean residual life at x is equal to

$$\operatorname{mrl}(x) = \frac{\int_x^\infty (t-x) f(t)\,dt}{S(x)} = \frac{\int_x^\infty S(t)\,dt}{S(x)}, \tag{2.20}$$

i.e., the **area under X's survival curve to the right of x**.

On Homework 1, you will work out where these convenient formulas comes from. To do so, you'll need to use some of the between-quantity relationships we've established and integration by parts (sad, I know).

[4] Formulas for discrete failure-time X are scarce. We will focus on continuous for now.

Mean survival

By the **definition of expectation**, the mean survival (or lifetime) is defined as

$$\mathrm{E}(X) = \int_0^\infty x f(x)\,dx \text{ if } X \text{ is continuous or} \tag{2.21}$$

$$\mathrm{E}(X) = \sum_{j=1}^{p} x_j p(x_j) \text{ if } X \text{ is discrete.} \tag{2.22}$$

- Also, since $\mathrm{mrl}(0) = \mathrm{E}(X)$ it follows that for a continuous X

$$\mathrm{E}(X) = \int_0^\infty S(x)\,dx, \tag{2.23}$$

  i.e., that the mean life/survival is the **total area under the survival curve**.

pth quantiles

Another method to capture "snapshots" of the distribution of the failure-time X is with quantiles. Let ϕ_p denote the pth quantile (aka the $100p$th percentile), where $p \in [0, 1]$.

- Recall from Probability that ϕ_p can be found by **inverting the CDF** in the following way:

$$\Pr(X \leq \phi_p) = F(\phi_p) \stackrel{\text{def}}{=} p \rightarrow \phi_p = F^{-1}(p). \tag{2.24}$$

- We're going to do the same thing now, except with the survival function!

$$\Pr(X \leq \phi_p) = 1 - S(\phi_p) \stackrel{\text{def}}{=} p \rightarrow \phi_p = S^{-1}(1 - p). \tag{2.25}$$

- When X is discrete, we can't always solve for ϕ_p exactly, so instead we define

$$\phi_p = \inf\{x : S(x) \leq 1 - p\}. \tag{2.26}$$

- The median survival (or lifetime) is simply the special case of the $p = 0.5$th quantile.

Your turn: Summarizing center in the `lung` data

`lung` (revisited): *Recall that X is survival time for patients with advanced lung cancer, where $X \sim expo(\beta = \exp(6))$.*

Derive the mean residual life function for X.

Your turn: Summarizing center in the `lung` data

Find the mean and median lifetimes.

What do you notice about the mean lifetime and mean residual life?