EMORY
ROLLINS SCHOOL OF PUBLIC HEALTH

Department of Biostatistics and Bioinformatics

BIOS 522: Survival Analysis Methods

Reading 4:

The hazard and cumulative hazard functions

This week, we will define the hazard and cumulative hazard functions and their relationship to the survival function. We will study three key parametric distributions used for time-to-event data. We will review maximum likelihood estimation for survival data. Finally, we will estimate the cumulative hazard function using the Nelson-Aalen estimator.

Part 1. The hazard function and cumulative hazard function

Random variable review

Previously we learned about the survival function $S(t)$, where $S(t) = \Pr(T > t)$ characterizes the distribution of survival random variable T. Recall that the survival function is related to the CDF $F(t)$ by $S(t) = 1 - F(t)$.

We can also define the **probability density function** (pdf) $f(t)$ for our random variable T, where $\frac{d}{dt}F(t) = f(t)$. The pdf characterizes when failures are most likely to occur in time. Like the CDF, the pdf uniquely defines the distribution of the random variable T.

Defining the hazard function

Another way to uniquely define the distribution of a survival random variable is by its **hazard function** or **hazard rate** $h(t)$. The hazard function specifies the instantaneous rate of failure at time t given that you have survived up until time t. The hazard function is like a derivative. Conditional on surviving to time t, it is the limit of the probability of an event in the next small interval $t + \Delta$ for some infinitely small Δ, divided by the interval length Δ.

We can show that $h(t) = f(t)/S(t^-)$, where $S(t^-) = \lim_{s \uparrow t} S(s)$ is survival immediately before time t:

$$\begin{aligned} h(t) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr(t \le T < t + \Delta | T \ge t) \\ &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \frac{\Pr([t \le T < t + \Delta\,] \cap [T \ge t])}{\Pr\,(T \ge t)} \\ &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \frac{\Pr(t \le T < t + \Delta\,)}{\Pr\,(T \ge t)} \\ &= \frac{f(t)}{S(t^-)} \end{aligned}$$

The last line follows because $f(t)$ is the first derivative of $F(t)$. For continuous variables, we drop the $S(t^-)$ notation and simplify to $h(t) = f(t)/S(t)$.

Other ways to re-express the hazard function for continuous variables are:

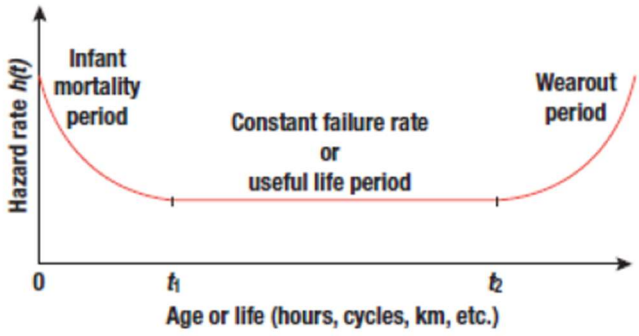
$$h(t) = -\frac{d}{dt} \ln[1 - F(t)] = -\frac{d}{dt} \ln S(t)$$

Note that this implies that we can represent the pdf as $f(t) = h(t)S(t)$,

Understanding the hazard function

The hazard function acts like a speedometer for risk. When $h(t)$ is high, you are increasingly likely to fail soon if you haven't failed already.

In studies of product failure, the hazard rate is sometimes referred to as a "bathtub curve." There is an early period where failures are common (high hazard rate), maybe because these products are faulty ("infant mortality period"). Products that last past this beginning period then have a long "useful life period" with a low hazard rate. Eventually, the hazard rate increases as the product begins to wear out from overuse.



Relationship to incidence rates

Recall that an incidence rate is the number of events per total person-time. Person-time could be measured in years, months, days, etc. Although

sometimes the terms incidence rate and hazard rate are used interchangeably, the hazard rate refers to the theoretical limit approached by an incidence rate as the unit of time approaches zero. While incidence rates are typically assumed to be constant throughout a time period, hazard rates can vary continuously as a function of time, like in the "bathtub curve" example above.

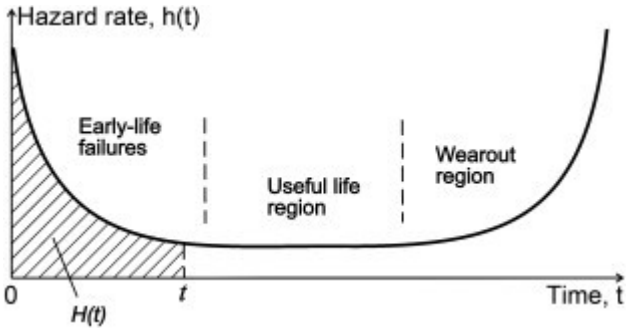
The cumulative hazard function

Another way to uniquely define the distribution of a survival random variable is by the **cumulative hazard function**. The cumulative hazard function for a survival random variable T is:

$$H(t) = \int_0^t h(u)du$$

The cumulative hazard $H(t)$ is the area under the hazard function between the time origin and time t. It sums up the accumulated hazard through time t.

In the figure below, the area of the shaded region $H(t)$ is the cumulative hazard function through time t.



Several properties follow from the definition of the cumulative hazard:

- $H(0) = 0$. At time 0, you have accumulated no hazard.
- $S(t) = e^{-H(t)}$ (*this gives us a convenient way to calculate* $S(t)$)
- As a result of the above, $H(t) = -\log S(t)$ (*this also comes in handy*)

Part 2. Parametric distributions for time-to-event data

Though many of the methods we use in survival analysis are non-parametric or semi-parametric, we can also assume that our random variable T follows some parametric distribution. We cover three distributions in this class. Each is continuous, defined by one or two parameters, and non-negative ($T \geq 0$):

- $T \sim Exponential(\lambda)$ or $Exp(\lambda)$
- $T \sim Weibull(\lambda, \gamma)$
- $T \sim LogLogistic(\lambda, \gamma)$

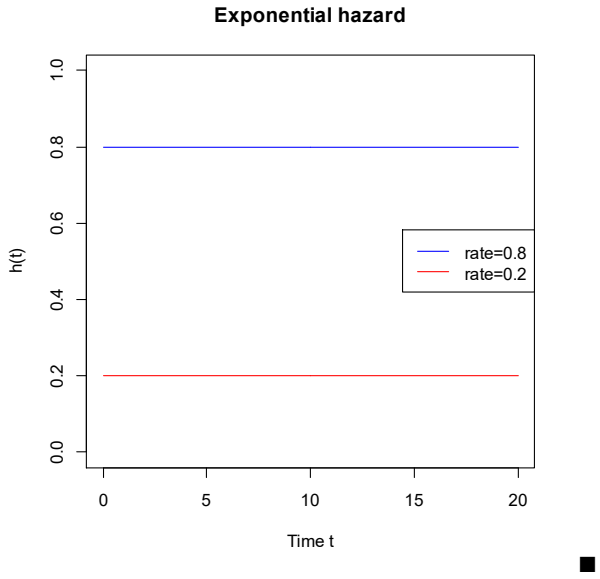
The exponential distribution

The **exponential distribution** is the simplest of the parametric survival time distributions. It has a constant hazard function:

$$h(t) = \lambda$$

λ is called the **rate parameter** and has units of time^{-1}. Note, sometimes instead of the rate parameter, the exponential distribution is summarized by the **scale parameter** $\sigma = 1/\lambda$ which is the inverse of the rate parameter and describes the mean survival time.

Example: The hazard function is plotted for the exponential distribution assuming two different values of the parameter λ ($\lambda = 0.8$ and $\lambda = 0.2$). It is constant at all times t.



■

The cumulative hazard function for the exponential distribution is:

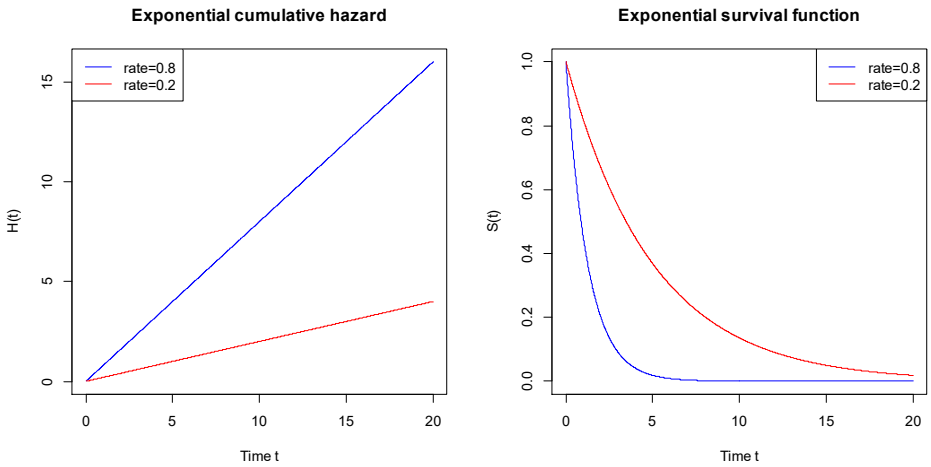
$$H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t$$

Note that this is a function for a straight line with origin 0.

The survival function can be obtained from the cumulative hazard function:

$$S(t) = e^{-H(t)} = e^{-\lambda t}$$

Example: the cumulative hazard function (left) and survival function (right) are plotted for the exponential distribution assuming two different values of the parameter λ ($\lambda = 0.8$ and $\lambda = 0.2$).



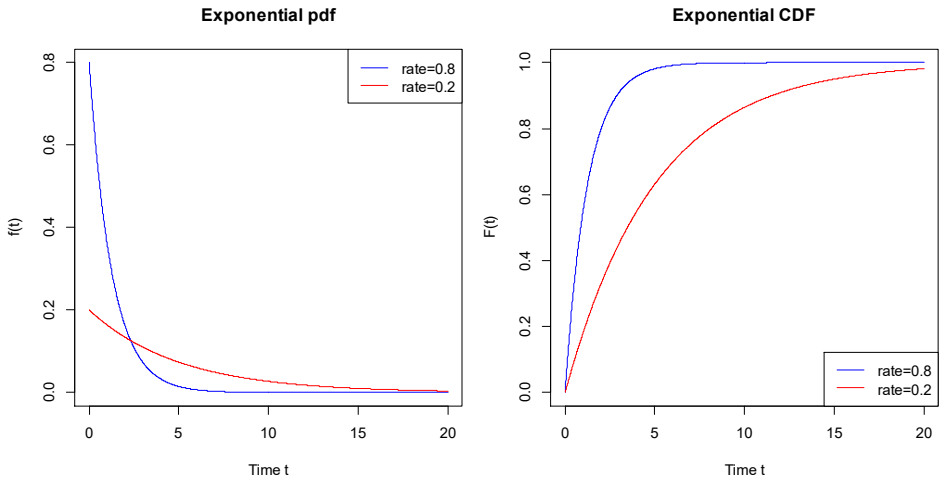
Note that although the cumulative hazard functions are straight lines, the survival functions are curved. A higher rate parameter λ results in worse survival (shorter survival times, faster decline in survival).

■

The following are also true for $T \sim Exp(\lambda)$:

pdf	$f(t) = \lambda e^{-\lambda t}$
CDF	$F(t) = 1 - e^{-\lambda t}$
Mean	$E[T] = 1/\lambda$
Variance	$Var[T] = 1/\lambda^2$
Median	$t_{0.50} = \log(2)/\lambda$

Example: the pdf (left) and CDF (right) are plotted for the exponential distribution assuming two different values of the parameter λ ($\lambda = 0.8$ and $\lambda = 0.2$).



Recall that the pdf is the product of the hazard and survival functions, $f(t) = h(t)S(t)$, and it displays which failure times are most likely to occur.

When $\lambda = 0.8$, most failures occur immediately ($t < 5$). We observe relatively few failures after $t = 5$ because when we examine the survival function, we see that most everyone has already failed; survival at that time is very low. In contrast, examine the pdf when $\lambda = 0.2$. Since everyone does not immediately fail, events are distributed across a longer time interval.

Note that, in both examples, a constant hazard does not translate into a constant pdf. That is because the hazard defines the instantaneous rate of failure among people who are still alive at time t. At later times when fewer people are alive, we observe fewer events during each successive time interval.

■

The Weibull distribution

The **Weibull distribution** is another common parametric survival time distribution. It has two parameters: λ is the rate parameter and γ is the **shape parameter**. Sometimes, the two parameters of the Weibull distribution will be summarized by σ and γ, where $\sigma = 1/\lambda$ is called the **scale parameter**.

The hazard function for the Weibull distribution is

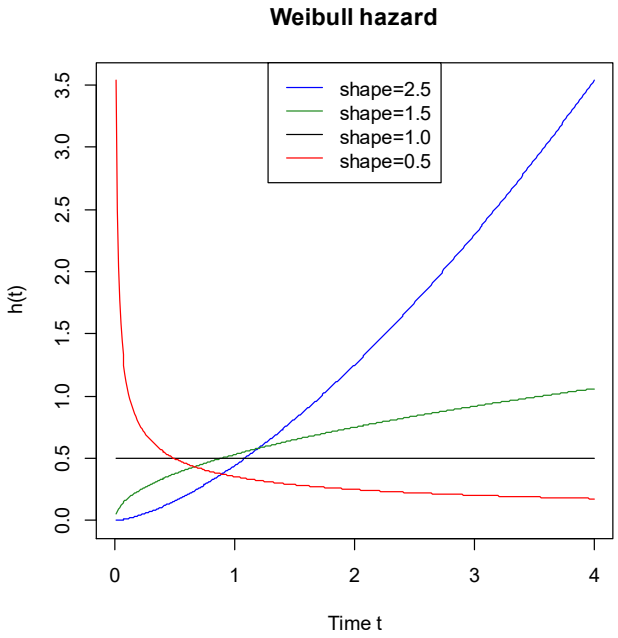
$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$$

We can see from the equation $h(t)$ that the hazard varies as a function of time t. Depending on the value of the shape parameter γ, it is either increasing, decreasing, or constant over time:

$$h(t) \text{ is } \begin{cases} \text{decreasing,} & \text{if } \gamma < 1 \\ \text{constant,} & \text{if } \gamma = 1 \\ \text{concave increasing,} & \text{if } 1 < \gamma < 2 \\ \text{linearly increasing,} & \text{if } \gamma = 2 \\ \text{convex increasing,} & \text{if } \gamma > 2 \end{cases}$$

When $\gamma = 1$, $h(t) = \lambda$, and the Weibull distribution simplifies to an exponential distribution. The exponential distribution is thus a "special case" of the Weibull. Equivalently, the Weibull is a "generalization" of the exponential distribution, with another parameter added to relax the constant hazard assumption.

Example: the hazard functions are plotted for the Weibull distribution assuming four different values of the shape parameter γ ($\gamma = 2.5, 1.5, 1.0, 0.5$). All curves have the same rate parameter $\lambda = 0.5$.



■

The appropriate shape parameter will depend upon the event being studied. For example, initially high but rapidly decreasing hazard might be appropriate for post-surgical outcomes, in which the rate of complications is highest immediately after the procedure.

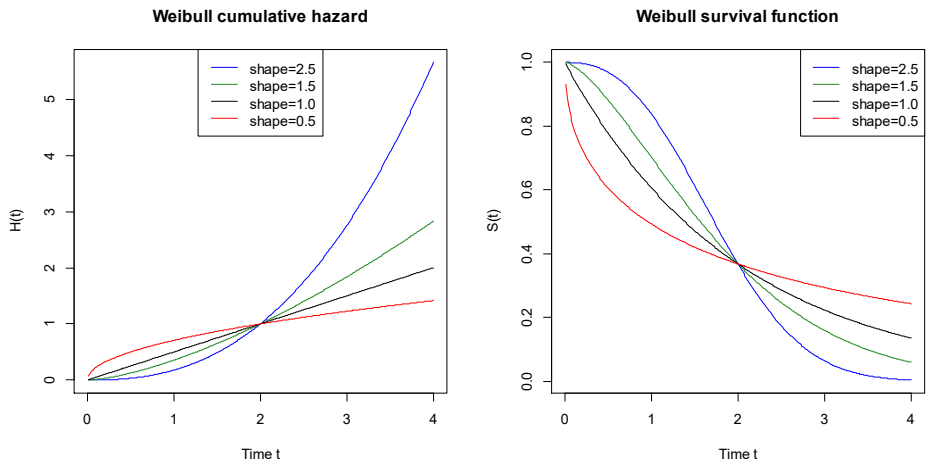
The Weibull cumulative hazard function is

$$H(t) = \int_{u=0}^{u=t} \lambda\gamma(\lambda u)^{\gamma-1}\, du = \lambda^{\gamma} \int_{u=0}^{u=t} \gamma u^{\gamma-1}\, du = (\lambda t)^{\gamma}$$

The Weibull distribution has survival function

$$S(t) = \exp\left(-H(t)\right) = \exp\left(-(\lambda t)^{\gamma}\right)$$

Example: the cumulative hazard functions (left) and survival functions (right) are plotted for the Weibull distribution assuming four different values of the shape parameter γ ($\gamma = 2.5,\ 1.5,\ 1.0,\ 0.5$). All curves have the same rate parameter $\lambda = 0.5$.



■

The log-logistic distribution

The **log-logistic distribution** has two parameters: λ, the rate parameter, and γ, the shape parameter. Sometimes, the distribution is summarized by the scale parameter $\sigma = 1/\lambda$ and the shape parameter γ. The distribution is called log-logistic because it is the distribution of a random variable whose logarithm has a logistic distribution.

$$T \sim \text{Log-logistic}(\lambda, \gamma)$$

$$\log(T) \sim \text{Logistic}(\lambda, \gamma)$$

(Though not discussed further, the log-normal distribution is similarly defined, where $\log(T)$ follows a normal distribution. This ensures non-negative T.)

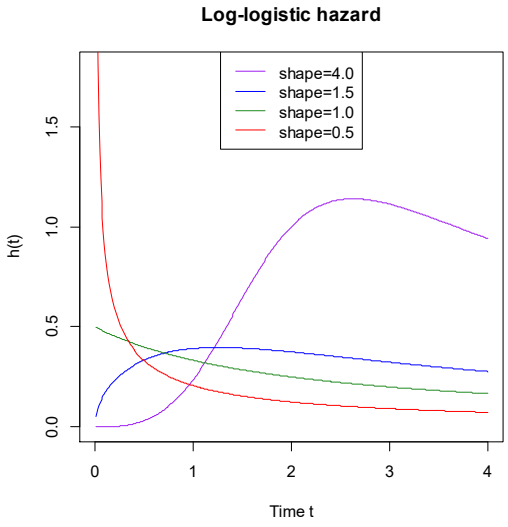
The log-logistic hazard function is:

$$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1 + (\lambda t)^{\gamma}}$$

There are three general shapes that the hazard function can take depending on the shape parameter γ:

$$h(t) \text{ is } \begin{cases} \text{decreasing from } \infty, & \text{if } \gamma < 1 \\ \text{decreasing from } \lambda, & \text{if } \gamma = 1 \\ \text{increasing then decreasing}, & \text{if } \gamma > 1 \end{cases}$$

Example: the hazard functions are plotted for the log-logistic distribution assuming four different values of the parameter γ ($\gamma = 4.0,\ 1.5,\ 1.0,\ 0.5$). All curves have the same rate parameter $\lambda = 0.5$.



■

The log-logistic cumulative hazard function is:

$$H(t) = \log\left(1 + (\lambda t)^{\gamma}\right)$$

The log-logistic survival function is:

$$\begin{aligned} S(t) &= \exp\big(-H(t)\big) \\ &= \exp(-\log(1 + (\lambda t)^{\gamma})) \\ &= \frac{1}{1 + (\lambda t)^{\gamma}} \end{aligned}$$

Part 3. Likelihood and parameter estimation

Overview of maximum likelihood estimation

Given a sample of right censored data (T_i^*, δ_i) for $i = 1, \ldots, n$, we are interested in making inference about the underlying population from which they were drawn. If we assume the data follow a parametric distribution, we want to estimate the values of the parameters as these define the mean survival, median survival, hazard function, density function, survival function, and so on. **Maximum likelihood estimation** is a general method for using data to estimate parameters such as those used in the exponential, Weibull, and log-logistic distributions.

Likelihood function for right-censored data

The likelihood function is the probability of observing the data at a particular value of the parameter(s). For right-censored data, there are two types of data points:

- If T_i^* is a failure time ($\delta_i = 1$), the probability of observing this failure time can be expressed by the pdf at time T_i^*.
- If T_i^* is a censoring time ($\delta_i = 0$), then we do not directly observe when this person failed, but we do observe that this person survived at least up until T_i^*. Thus, the probability of observing this censoring time can be expressed by the survival function at time T_i^*.

Consider a distribution with a single parameter λ. The **likelihood function** $L(\lambda)$ evaluated at λ can be conveniently expressed as follows:

$$L(\lambda) = \prod_{i=1}^{n} [h(T_i^*|\lambda)]^{\delta_i} S(T_i^*|\lambda)$$

It may take a minute to see why this works – but it's very handy. If $\delta_i = 1$, we leverage the fact that the hazard function times the survival function equals the pdf. If $\delta_i = 0$, the hazard function drops out, leaving only the survival function.

The above is a general likelihood function for right-censored data. It accommodates both failure times and censoring times. To use this likelihood function to fit a desired parametric distribution, we plug in the hazard function and survival function specific to that distribution.

We can evaluate the likelihood function at many different values of λ. The value of λ that maximizes the likelihood function is known as the **maximum likelihood estimate (MLE)** $\hat{\lambda}$. It is our best estimate of λ given the data.

Likelihood function for the exponential distribution

Let's use the above expression to calculate the likelihood function for the exponential distribution. The hazard function for the exponential distribution is $h(t) = \lambda$. The survival function for an exponential distribution is $S(t) = e^{-\lambda}$. Thus, our likelihood function is as follows:

$$L(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda T_i^*}$$

For a failure time $\delta_i = 1$, their likelihood contribution is the exponential pdf evaluated at the failure time T_i^*, i.e., $\lambda e^{-\lambda T_i^*}$. For a censoring time $\delta_i = 0$, their likelihood contribution is the exponential survival probability evaluated at the censoring time T_i^*, i.e., $e^{-\lambda T_i^*}$.

For the exponential distribution, we are able to calculate a closed form for the maximum likelihood estimator $\hat{\lambda}$ (proof not shown here). This form is:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i^*}$$

We can recognize this MLE as the incidence rate. It is the total number of events divided by the person-time. Thus, when we assume a constant hazard rate, as

we do for an exponential, our best estimate of the hazard rate is exactly the incidence rate.

Part 4. The Nelson-Aalen estimator

Another non-parametric estimator of the survival function

Returning to our discussion of the cumulative hazard function, the properties of $H(t)$ give us a nice result. Given an estimate of the cumulative hazard function $H(t)$, we can estimate the survival function using the relationship $S(t) = e^{-H(t)}$.

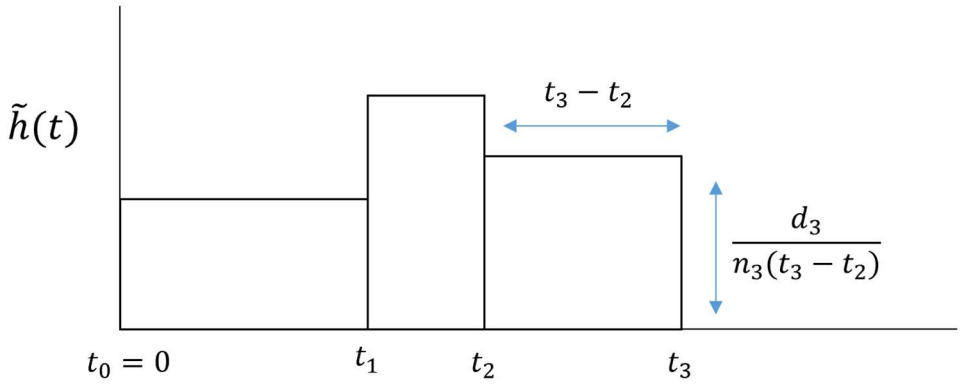
To create another nonparametric estimator of $S(t)$, we can start with a nonparametric estimator of $H(t)$. One such example is the **Nelson-Aalen estimator**.

To derive the Nelson-Aalen estimator, start by recalling that the cumulative hazard function is the area under the hazard function up until time t. One way we can approximate the hazard function is by dividing time into intervals and assuming that the hazard is constant within that interval.

Consider a data set with ordered, unique failure times $t_1, \dots, t_J$. We can define the same intervals as the Kaplan-Meier estimator, $(0, t_1], (t_1, t_2], (t_2, t_3]$ and so on. Within each interval, what is a good estimate of the hazard? Since we are assuming the hazard is constant in that interval, the incidence rate is a reasonable estimate. The numerator is the number of events d_j at the end of the interval, and the denominator is the person-time at risk, calculated by the number at risk n_j times the length of that interval $(t_j - t_{j-1})$. The estimated hazard $\tilde{h}(t_j)$ is as follows:

$$\tilde{h}(t_j) = \frac{d_j}{n_j(t_j - t_{j-1})}$$

Thus, we can imagine our hazard function as a set of non-overlapping intervals, each with a horizontal line placed at the interval's incidence rate. To estimate the cumulative hazard function, we need to calculate the area under that hazard up to time t. We can visualize the area under the curve as being constructed of many small rectangles placed between unique failure times $t_1, t_2, t_3, \dots$. The width of each interval is the time between the failure times. The height of each interval is the incidence rate during that interval.



The contribution of each interval to the cumulative hazard is the area of the rectangle. This is the height (incidence rate) by the length $(t_j - t_{j-1})$. This simplifies nicely:

$$\frac{d_j}{n_j(t_j - t_{j-1})} \times (t_j - t_{j-1}) = \frac{d_j}{n_j}$$

The Nelson-Aalen estimator of the cumulative hazard function $H(t)$ is thus the sum of the area of each rectangle:

$$\tilde{H}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}$$

*Example: In **Lecture 2**, we calculated survival for a cohort of hemophiliacs with AIDS diagnosis using the Kaplan-Meier method. Recall that their right-censored survival times (in months) were: 2, 3+, 6, 6, 8, 10+, 15, 15, 16, 27, 30, 32.*

Using the same data, use the Nelson-Aalen estimator to obtain a point estimate of the cumulative hazard 7 months after primary AIDS diagnosis.

We are asked to estimate $\tilde{H}(t = 7)$. We use the same time intervals as used for the Kaplan-Meier estimator.

Unique failure/ censoring time	**Number at risk n_j during (t_{j-1}, t_j)**	**Number of deaths d_j at t_j**	**Number censored c_j at t_j**	**Cumulative hazard contribution $\frac{d_j}{n_j}$**	**Nelson-Aalen estimate**
$t_0 = 0$					$t = [0,2)$ $\tilde{H}(t) = 0$
$t_1 = 2$	$n_1 = 12$	$d_1 = 1$	$c_1 = 0$	$\frac{d_1}{n_1} = \frac{1}{12}$	$t = [2,3)$ $\tilde{H}(t) = 0.083$
$t_2 = 3$	$n_2 = 11$	$d_2 = 0$	$c_2 = 1$	$\frac{d_2}{n_2} = \frac{0}{11}$	$t = [3,6)$ $\tilde{H}(t) = 0.083$
$t_3 = 6$	$n_3 = 10$	$d_3 = 2$	$c_3 = 0$	$\frac{d_3}{n_3} = \frac{2}{10}$	$t = [6,8)$ $\tilde{H}(t) = 0.283$

$$\tilde{H}(t=7) = \sum_{j:t_j \leq 7} \frac{d_j}{n_j} = \frac{1}{12} + \frac{0}{11} + \frac{2}{10} = 0.283$$

■

To obtain an estimate of $S(t)$ based on the Nelson-Aalen cumulative hazard estimate $\tilde{H}(t)$, we can calculate the following:

$$\tilde{S}(t) = \exp\left(-\tilde{H}(t)\right)$$

This is referred to as the **Nelson-Aalen estimator of the survival function,** the **Breslow estimator,** or the **Fleming-Harrington estimator**. This returns a similar, though slightly higher, estimate of the survival function as compared to the Kaplan-Meier estimator. Neither is fundamentally better than the other. They are simply different approaches to estimating the same quantity.

The Kaplan-Meier estimator and Nelson-Aalen estimator of the survival function are asymptomatically equivalent. Both estimators are consistent. Under suitable regularity conditions, both estimators weakly converge to a Gaussian process.

Example: Use the Breslow estimator to obtain a point estimate of the survival function 7 months after primary AIDS diagnosis.

The Breslow estimator is based on the Nelson-Aalen cumulative hazard.

$$\tilde{S}(t) = \exp\left(-\tilde{H}(t)\right)$$

$$\tilde{S}(t=7) = \exp\left(-\tilde{H}(t=7)\right) = \exp(-0.283) = 0.753$$

An estimated 75.3% of hemophiliacs are alive 7 months after primary AIDS diagnosis. This is similar, although not identical, to the Kaplan-Meier estimator of survival (i.e. 73.3%).

■

The variance of the Nelson-Aalen estimator follows from a Poisson argument, as proposed by Tsiatis (1981):

$$\widehat{Var}\left(\tilde{H}(t)\right) = \sum_{j:t_j \leq t} \frac{d_j}{n_j^2}$$

A Wald 95% confidence interval for $H(t)$ is:

$$\tilde{H}(t) \pm 1.96 \sqrt{\widehat{Var}\left(\tilde{H}(t)\right)}$$

This can produce confidence intervals with negative lower bounds, outside the allowed range of $H(t)$. The normal approximation is improved if one works instead with the log of the cumulative hazard. This transformation also ensures that that the resulting bounds are positive.

A log-transformed 95% confidence interval for $H(t)$ is:

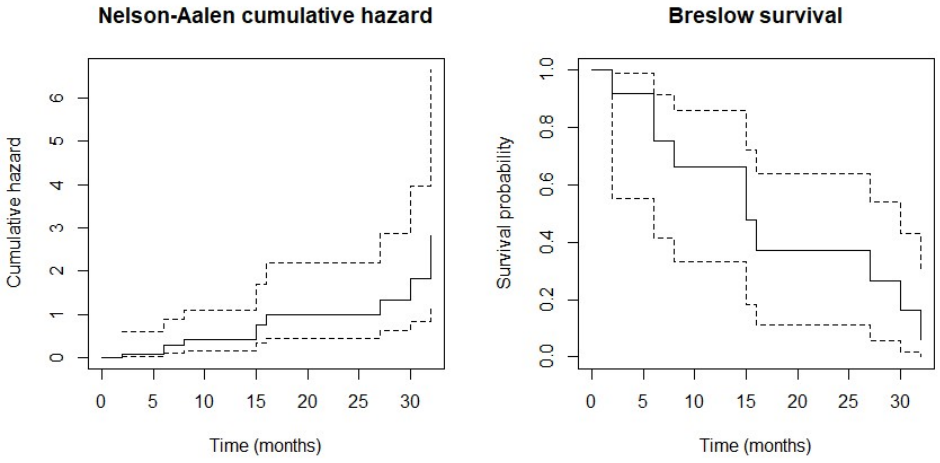
$$\widetilde{H}(t)\exp\left(\frac{\pm 1.96\sqrt{\widehat{Var}\left(\widetilde{H}(t)\right)}}{\widetilde{H}(t)}\right)$$

Given a confidence interval for the cumulative hazard calculated by either the Wald method or by the log transformation, we can transform the upper and lower bounds to obtain a confidence interval for the Breslow estimator of the survival function. For a confidence interval for $H(t)$ with bounds $\left(\widetilde{H}_L(t), \widetilde{H}_U(t)\right)$, the transformed confidence interval for survival is:

$$\left(e^{-\widetilde{H}_U(t)}, e^{-\widetilde{H}_L(t)}\right)$$

Note the flipped order of the bounds. The upper bound for cumulative hazard corresponds to the lower bound for survival.

Example: The Nelson-Aalen cumulative hazard curve and Breslow survival curve are plotted below for the AIDS dataset.



■

Part 5. Looking ahead

The hazard function is a key function for modeling time-to-event data. Next week, we will learn about the **Cox proportional hazards model**. In this model, the dependent variable is the hazard function, with the effects of covariates modeled as factors that multiply the hazard function