Chapter 2

Descriptive Methods for Survival Data

Describing Survival Data

- The first step of any statistical analysis should be describing and visualizing the data
  - Measures of location and spread
  - Graphs
- In survival analysis, start by estimating the cumulative distribution function (CDF)
  - Let T be the survival time
  - The CDF is $F(t) = P(T \leq t)$
  - It's more common to work with the survival function, S, where $S(t) = P(T > t)$

Notation

- Have a sample of *n* independent observations

  $(t_1, c_1), (t_2, c_2), \ldots (t_n, c_n)$, where

  - *t* is the survival time
  - *c* is an indicator variable for whether the observation was censored
- Observations are ordered so that $t_1 < t_2 < \ldots t_n$
- n_i is the number of subjects at risk of dying at time t_i
- d_i is the number of deaths that occur at time t_i

Kaplan-Meier Estimation

- The Kaplan-Meier estimator is the most common method of estimating the survival function
  - KM uses conditional probability

- Recall: $P(A \cap B) = P(B|A)P(A)$
  - Let A be the event $T \geq t_i$ (survival to *at least* time t_i)
  - Let B be the event $T > t_i$ (survival *past* time t_i)

- $P(A \cap B) = P(B) = P(T > t_i) = S(t_i)$
- Since there are no events between t_{i-1} and t_i,

$$P(A) = P(T \geq t_i) = P(T > t_{i-1}) = S(t_{i-1})$$

Kaplan-Meier Estimation

- The K-M estimate of the survival function is

$$\hat{S}(t_i) = P(B) = P(B|A)P(A)$$

$$= P(T > t_i | T \geq t_i) S(t_{i-1})$$

$$= \prod_{k=1}^{i} P(T > t_k | T \geq t_k)$$

- How do we estimate $P(T > t_i \mid T \geq t_i)$?

Kaplan-Meier Estimation

- At time t_i, there are n_i subjects at risk. d_i of these subjects died and n_i - d_i survived, so we can estimate this conditional probability as

$$\hat{p}_i = \frac{n_i - d_i}{n_i}$$

- So the K-M estimate of the survival function is

$$\hat{S}(t_i) = \prod_{k=1}^{i} P(T > t_k | T \geq t_k)$$

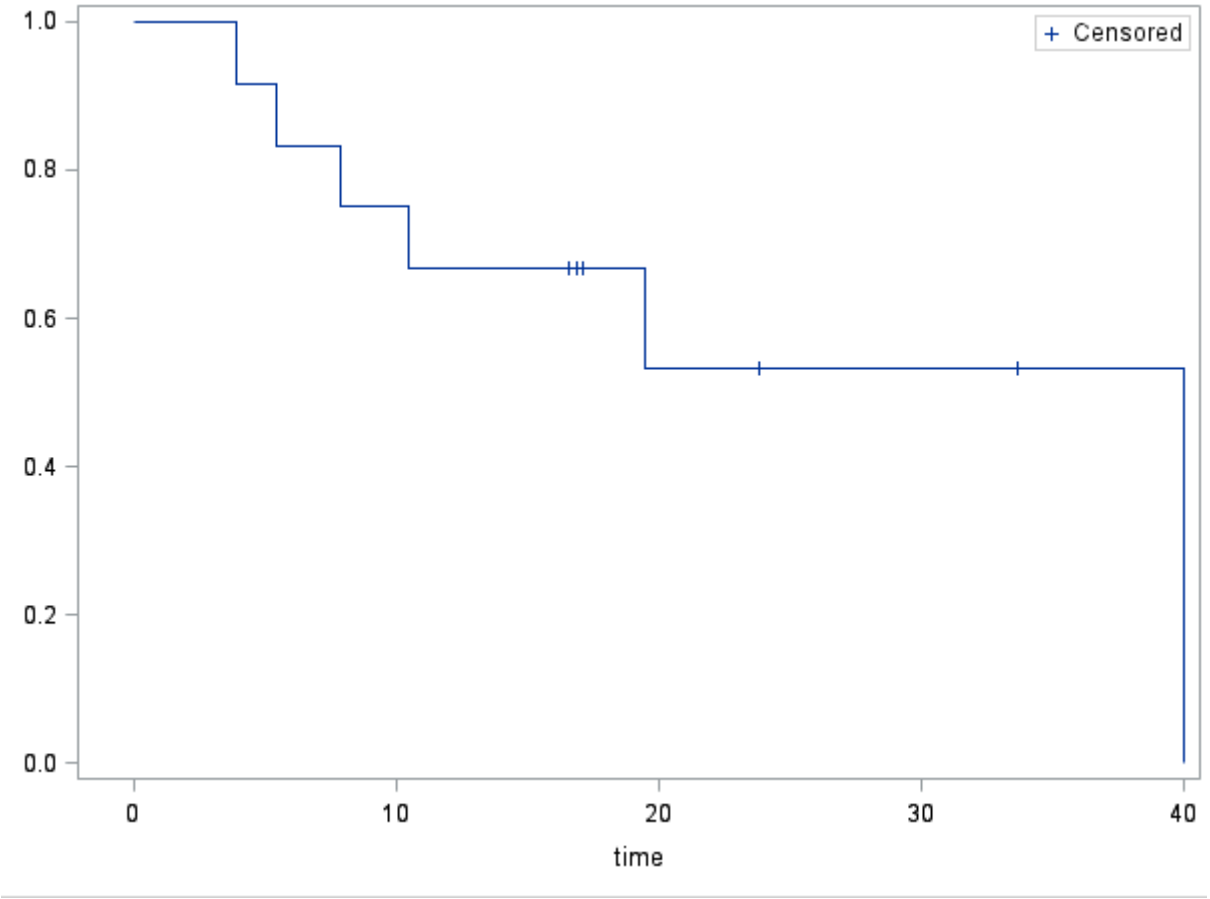
$$= \prod_{k=1}^{i} \frac{n_k - d_k}{n_k}$$

Kaplan-Meier Estimation

- The K-M estimates the survival function with a step function, with the steps at the non-censored survival times.
  - For $0 \leq t < t_1$, $\hat{S}(t) = 1$
  - For $t > t_n$, $\hat{S}(t)$ is undefined

Kaplan-Meier Estimation

- A typical K-M plot



Kaplan-Meier Estimation

- Once we've estimated the survival function, S, we can use it to estimate the percentiles of the survival time distribution
- The median survival time is a common measure of location
  - The time at which the probability of surviving past that time is exactly 50%
  - Estimate the median survival time as

$$\hat{t}_{50} = \min\{t : \hat{S}(t) \leq 0.50\}$$

- In general, the p^{th} percentile is estimated as either

$$\hat{t}_p = \min\left\{t : \hat{S}(t) \leq \frac{p}{100}\right\} \quad \textbf{(Book)}$$

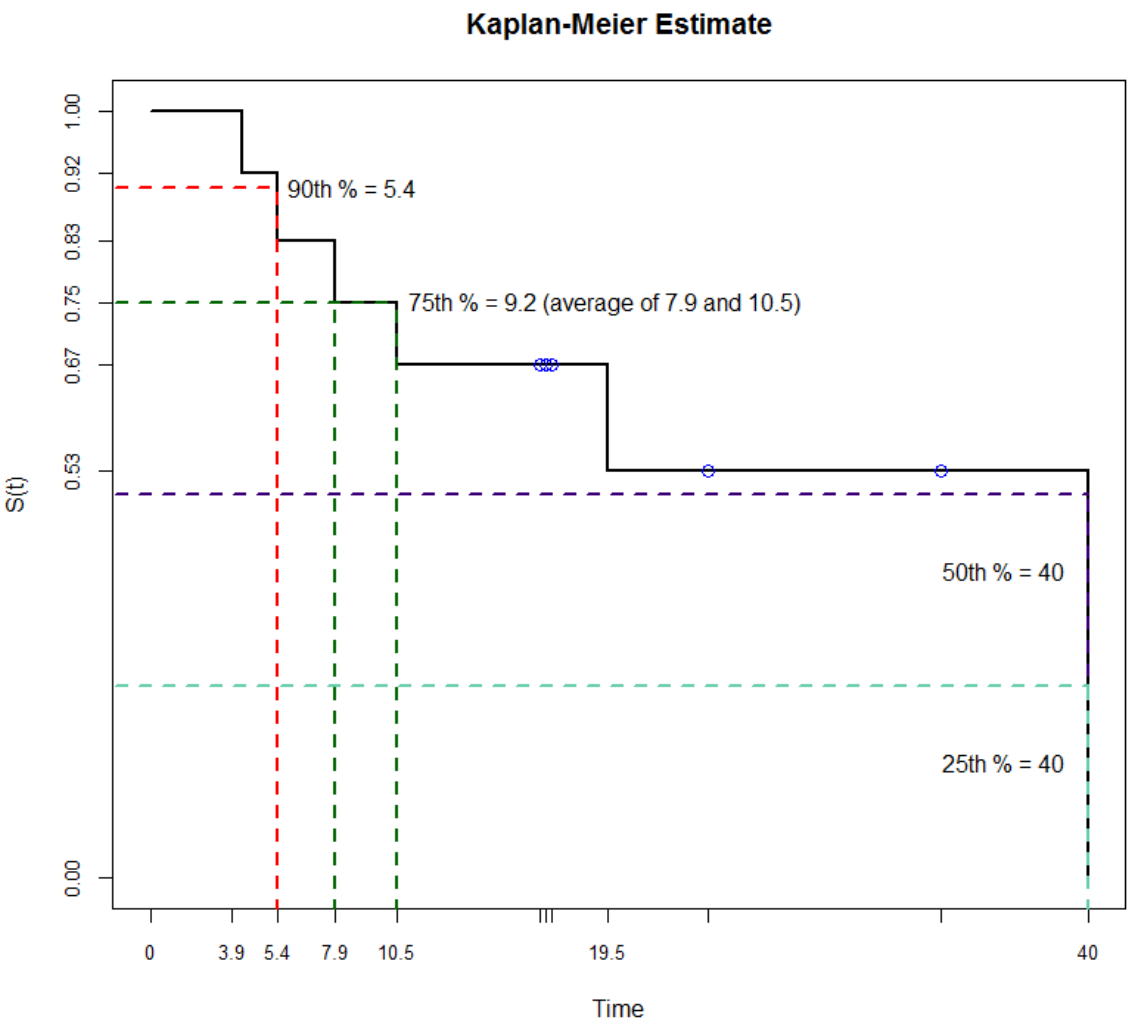
or

$$\hat{t}_p = \min\left\{t : \hat{S}(t) \leq \frac{100 - p}{100}\right\} \quad \textbf{(SAS)}$$

Kaplan-Meier Estimation

- Another way to think of estimating $\hat{t}_p$ is to draw a horizontal line at $y = \frac{p}{100}$ and find the first place that line crosses the KM curve.

- In this example, t = 7.9 is the first time the line y = 0.75 crosses the curve.
  - Note that the KM estimate is exactly 0.75 for all times within the interval [7.9, 10.5).
  - SAS reports 9.2 (the midpoint of this interval) as the estimate for this percentile.
  - Either 7.9 or 9.2 is an acceptable estimate.



Kaplan-Meier Estimation

- For a non-negative, continuous random variable

$$\mu = \int_0^\infty S(u)\, du$$

- Since $\hat{S}(t)$ is undefined after the largest survival time, use $\mu(t^*)$ to estimate the mean survival time, where

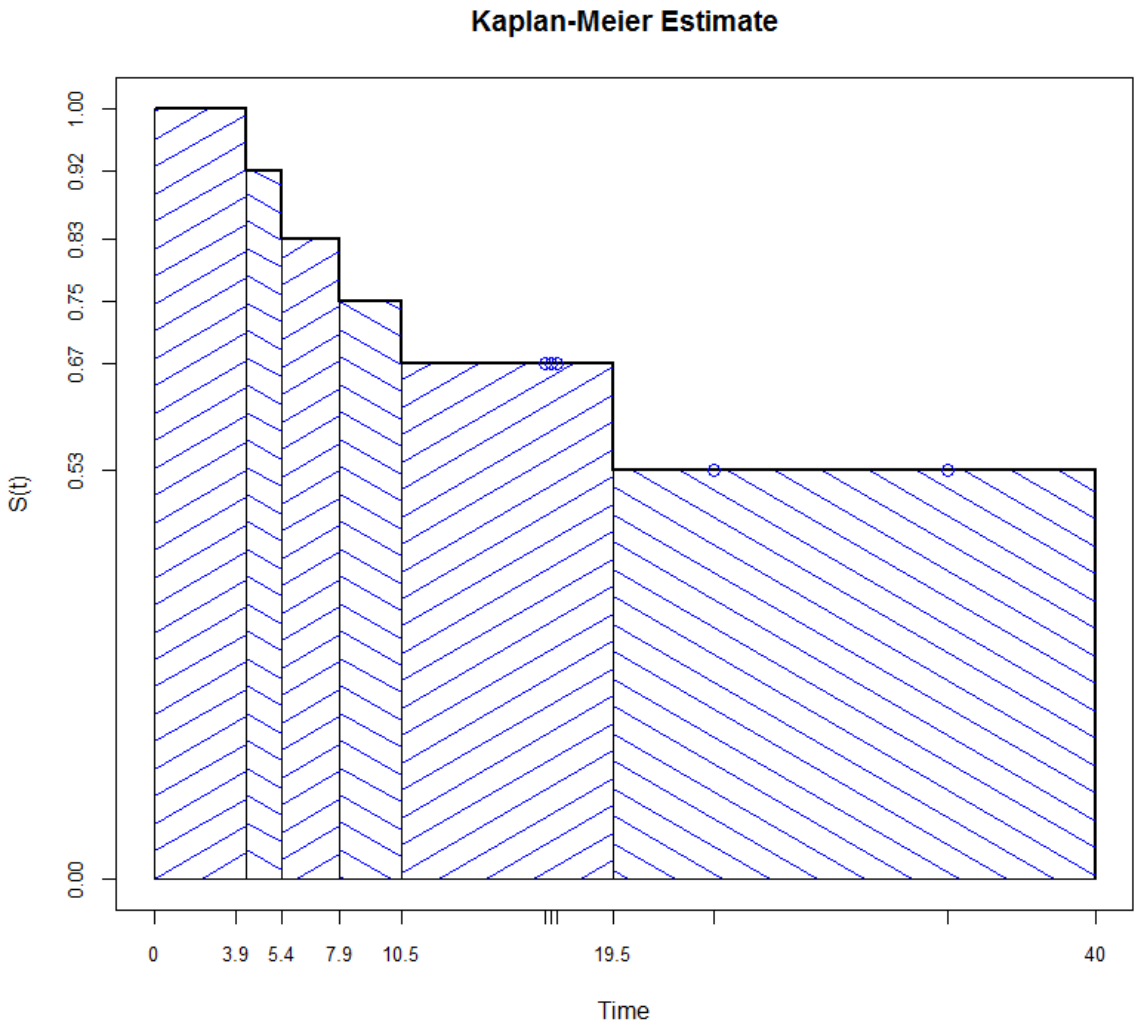
$$\mu(t^*) = \int_0^{t^*} S(u)\, du\,,$$

and t* can either be the largest survival time or the largest *non-censored* survival time.

  - SAS uses the largest non-censored survival time

Kaplan-Meier Estimation

- To estimate the mean survival time using the KM estimate, calculate the area under the curve between 0 and t*.
  - Usually the median, and not the mean, survival time is reported



Standard Error

- There is uncertainty in any estimate, need to calculate the standard error to quantify this uncertainty.
- For a given time t, what is $Var\left(\hat{S}(t)\right)$?
  - Variance of a sum is easier to calculate than the variance of a product, so take the natural log

$$log\left(\hat{S}(t)\right) = \sum_{t_i \le t} log\left(\frac{n_i - d_i}{n_i}\right) = \sum_{t_i \le t} \log(\hat{p}_i)$$

$$Var\left(log\left(\hat{S}(t)\right)\right) = \sum_{t_i \le t} Var(\log(\hat{p}_i))$$

- Using the Delta method several times, we can derive an estimate of the variance of $\hat{S}(t)$

Standard Error

- Greenwood's estimator of the variance of $\hat{S}(t)$:

$$\hat{V}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{t_i<t} \frac{d_i}{n_i(n_i - d_i)}$$

- Using this estimate of the variance to form confidence intervals could lead to endpoints less than zero or greater than one.

Confidence Intervals

- Kalbfleisch and Prentice suggested estimating the variance of a function of $\hat{S}(t)$, and then transforming to calculate a confidence interval
  - Use log(-log($\hat{S}(t)$)), the log-log survival function

$$\hat{V}\left(log\left(-log\left(\hat{S}(t)\right)\right)\right) = \frac{1}{\left(log\left(\hat{S}(t)\right)\right)^2}\sum\frac{d_i}{n_i(n_i - d_i)}$$

- A 100(1-α)% confidence interval for the log-log survival function is

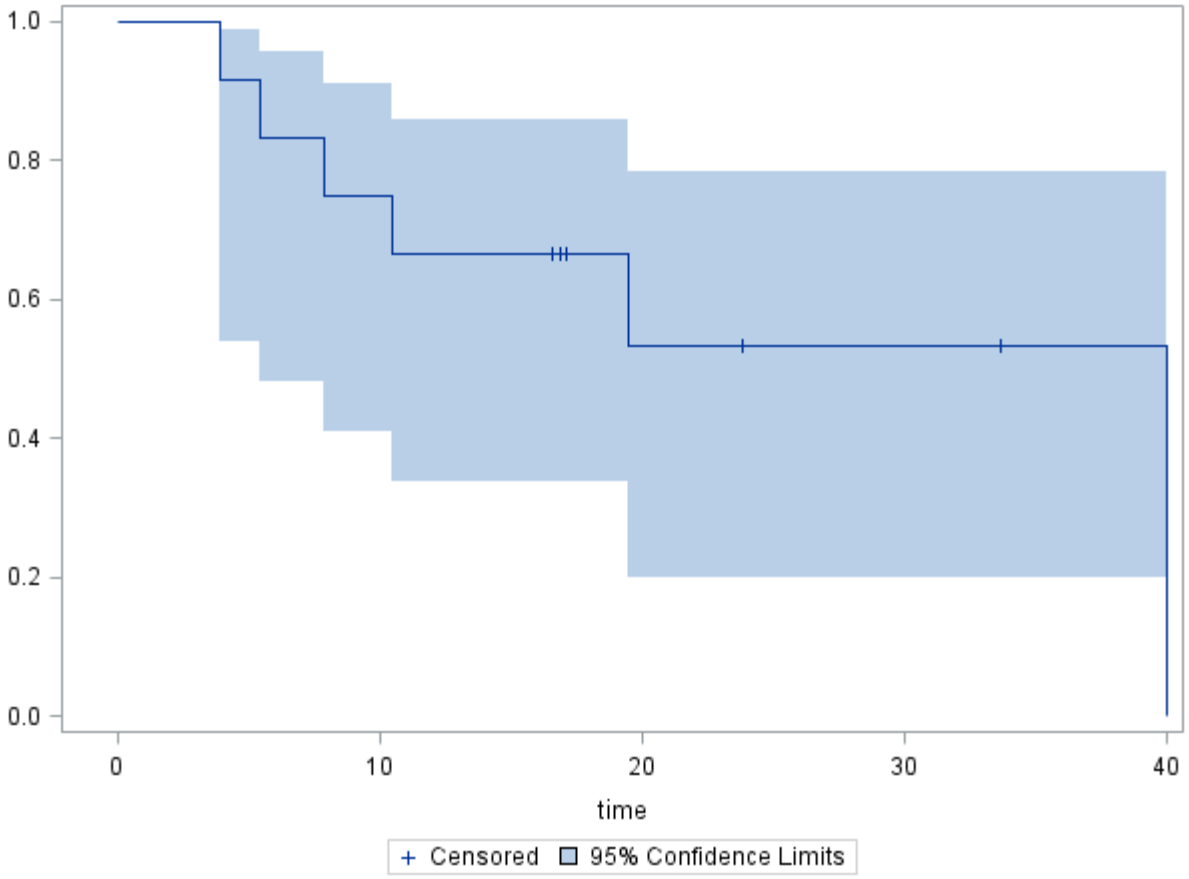
$$\log(-\log(\hat{S}(t))) \pm z_{1-\alpha/2}\sqrt{\hat{V}(\log(-\log(\hat{S}(t)))}$$

- The CI for $\hat{S}(t)$ is then found using the transformation

$$f(x) = e^{-e^x}$$

Confidence Intervals

- Can construct point-wise confidence intervals at each point on the survival curve



Confidence Intervals

- To calculate a confidence interval for the p^{th} percentile, estimate variance of $\hat{t}_p$:

$$\hat{V}(\hat{t}_p) \approx \frac{\hat{V}\left(\hat{S}(\hat{t}_p)\right)}{\left(\hat{f}(\hat{t}_p)\right)^2}$$

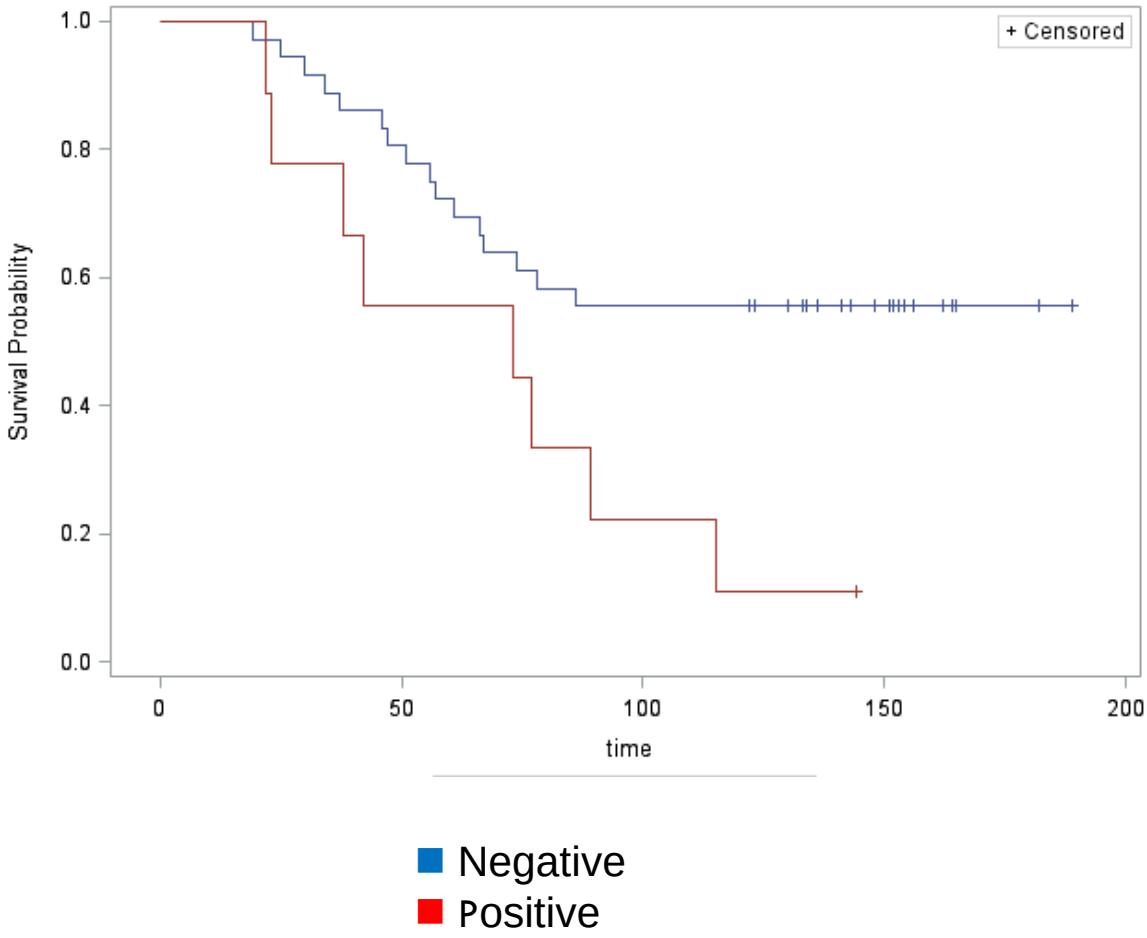
where $\hat{V}(\cdot)$ is Greenwood's estimator and $\hat{f}(\cdot)$ is an estimate of the density function of the survival time. (See p.37 of textbook for more details).

- A 100(1-α)% CI for $\hat{t}_p$ is

$$\hat{t}_p \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{t}_p)}$$

Comparing Two Survival Distributions

- How can we test for differences in the survival functions?
  - Example: Comparing survival for breast cancer patients who were either positive or negative for metastasis in lymph nodes



Comparing Two Survival Distributions

- Want to test H_0: $S_1(t) = S_2(t)$ vs H_1: $S_1(t) \neq S_2(t)$
- Motivation:
  - At a given survival time, t_i, can create a contingency table:

	Group 1	Group 2	
Died	d_{1i}	d_{2i}	d_i
Survived	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
Number at Risk	n_{1i}	n_{2i}	n_i

  - Treat the margins of each contingency table as fixed
    - d_{1i} follows a hypergeometric distribution

$$P(d_{1i} = k) = \frac{\binom{d_i}{k}\binom{n_i - d_i}{n_{1i} - k}}{\binom{n_i}{n_{1i}}}$$

Comparing Two Survival Distributions

- The test statistic is

$$Q = \frac{\left[\sum_i w_i(d_{1i} - \hat{e}_{1i})\right]^2}{\sum_i w_i^2 \hat{v}_{1i}}$$

where $\hat{e}_{1i}$ and $\hat{v}_{1i}$ are the expected value and variance of d_{1i}, respectively

$$\hat{e}_{1i} = \frac{n_{1i} d_i}{n_i} \quad , \quad \hat{v}_{1i} = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 \, (n_i - 1)}$$

and $w_i = w(t_i)$ is a weighting function

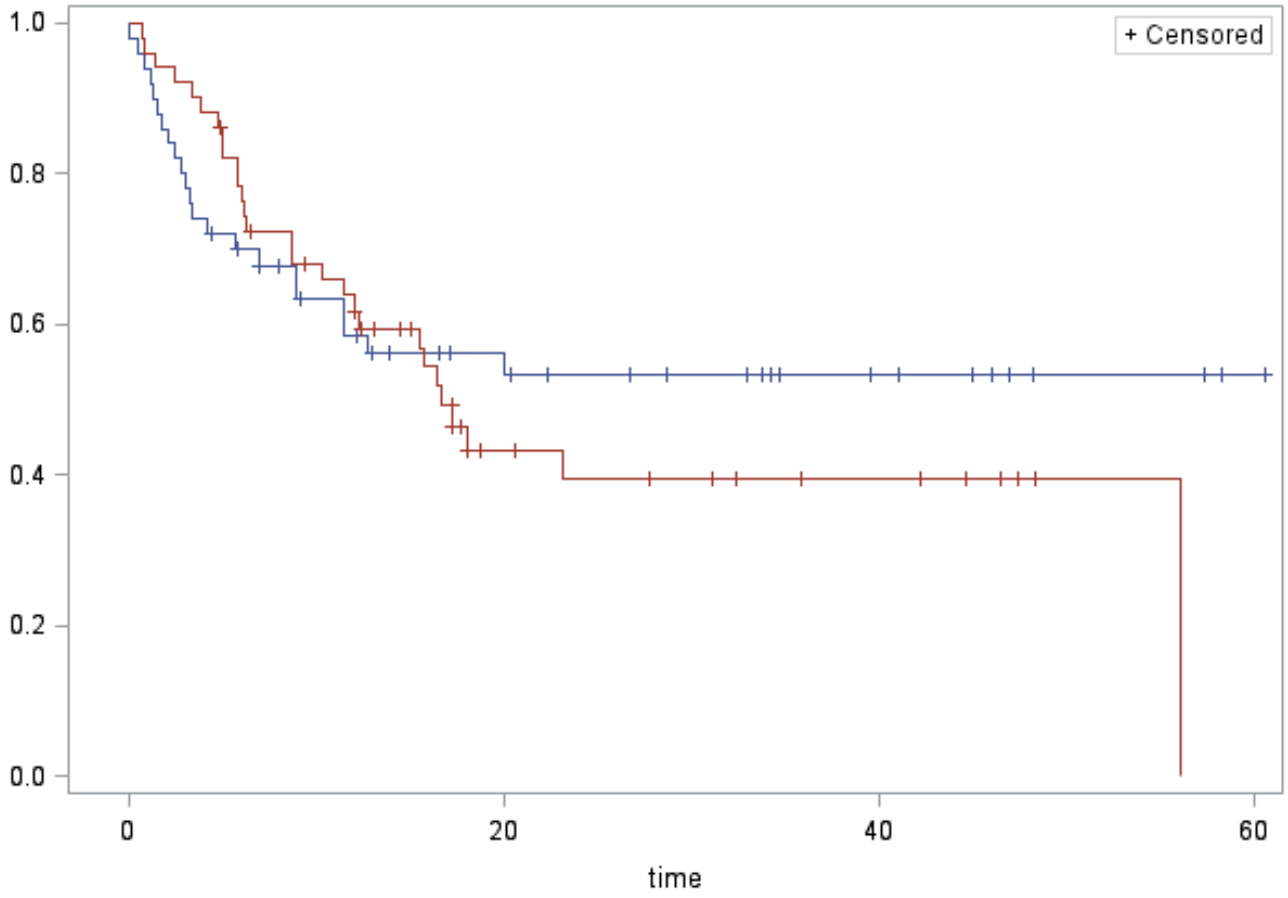
- If H_0 is true, Q follows a χ^2 distribution with 1 degree of freedom
- The (two-sided) p-value is $P(\chi_1^2 > Q)$

Comparing Two Survival Distributions

- Different choices of the weighting function lead to different tests
  - Log-rank test: $w(t_i) = 1$
  - Gehan test or Wilcoxon test: $w(t_i) = n_i$
  - Tarone and Ware: $w(t_i) = f(n_i)$ for some function *f*
- Log-rank and Wilcoxon tests are the most common choices
  - Log-rank test weights all survival times equally
  - The Wilcoxon test gives more weight to the earlier part of the survival curve where more subjects are at risk

Comparing Two Survival Distributions

- These tests should not be used when the survival curves cross



Comparing K Survival Distributions

- Can generalize these tests to comparing the survival distributions of K groups
  - H_0: $S_1(t) = S_2(t) = \ldots = S_K(t)$ vs H_1: Not H_0

	Group 1	Group 2	...	Group k	...	Group K	
Died	d_{1i}	d_{2i}	...	d_{ki}	...	d_{Ki}	d_i
Survived	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	...	$n_{ki} - d_{ki}$	...	$n_{Ki} - d_{Ki}$	$n_i - d_i$
Number at Risk	n_{1i}	n_{2i}	...	n_{ki}	...	n_{Ki}	n_i

- At each time t_i, calculate the expected number of deaths in Group k under H_0:

$$\hat{e}_{ki} = \frac{d_i n_{ki}}{n_i}, k = 1, \ldots, K-1$$

Comparing K Survival Distributions

- Use vector notation to extend the test to K groups

$$d_i^T = \left(d_{1i}, d_{2i}, \ldots d_{K-1,i}\right)$$
$$\hat{e}_i^T = \left(\hat{e}_{1i}, \hat{e}_{2i}, \ldots \hat{e}_{K-1,i}\right)$$

  - Let $\hat{V}_i$ be the *(K-1)* x *(K-1)* covariance matrix of d_i^T:
    - The diagonal elements of $\hat{V}_i$ are

$$\hat{v}_{kk,i} = \frac{n_{ki}(n_i - n_{ki})d_i(n_i - d_i)}{n_i^2(n_i - 1)}, k = 1, \ldots K-1$$

    - The off-diagonal elements of $\hat{V}_i$ are

$$\hat{v}_{jk,i} = \frac{n_{ji}n_{ki}d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \text{for } j \neq k$$

Comparing K Survival Distributions

- The test statistic is

$$Q = \left[\sum_i W_i(d_i - \hat{e}_i)\right]^T \left[\sum_i W_i \hat{V}_i W_i\right]^{-1} \left[\sum_i W_i(d_i - \hat{e}_i)\right]$$

where W_i is a diagonal matrix of weights.

  - Under H_0, Q follows a χ^2 distribution with *K-1* degrees of freedom
  - The two-sided p-value is $P(\chi^2_{K-1} > Q)$
  - Note, that for the simple case (K=2), this test statistic is the same as before.

Functions of Survival Time

- So far we have focused on the survival function, S:

$$S(t) = P(T > t) = 1 - F(t)$$

where F(t) is the cdf of the survival time, T.

- Let the pdf of T be f(t)

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$$

  - f(t) is also called the unconditional failure rate

Functions of Survival Time

- Subjects who have died before time *t* are no longer at risk during (t, t+Δt), so it makes sense to think instead of a conditional failure rate
  - Conditional failure rate: instantaneous risk of an event occurring at time *t*, given subject has survived to time *t*.
  - Also called the hazard function:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t}$$

    - h(t) > 0
    - Note: h(t) is not a probability

Functions of Survival Time

- Can also define the cumulative hazard function, H

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

  - H(t) is the total amount of risk accumulated up to time t

Functions of Survival Time

- Note S(t), f(t), and h(t) are equivalent ways of describing any continuous probability distribution.
- If we know one of them, we can get the others:

$$f(t) = -\frac{d}{dt}S(t)$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log\big(S(t)\big)$$

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

Estimating the Cumulative Hazard Function

- Instead of using the Kaplan-Meier estimator $\hat{S}(t)$ to estimate the survival function, can instead estimate H(t)
- The Nelson-Aalen estimate:

$$\tilde{H}(t) = \sum_{t_i<t} \frac{d_i}{n_i}$$

$$Var\left(\tilde{H}(t)\right) = \sum_{t_i<t} \frac{d_i}{n_i^2}$$

Estimating the Cumulative Hazard Function

- The Nelson-Aalen estimator of the survival function is therefore

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

  - A 100(1-α)% CI for H(t) is

$$\tilde{H}(t) \pm z_{1-\alpha/2}\sqrt{Var\left(\tilde{H}(t)\right)}$$

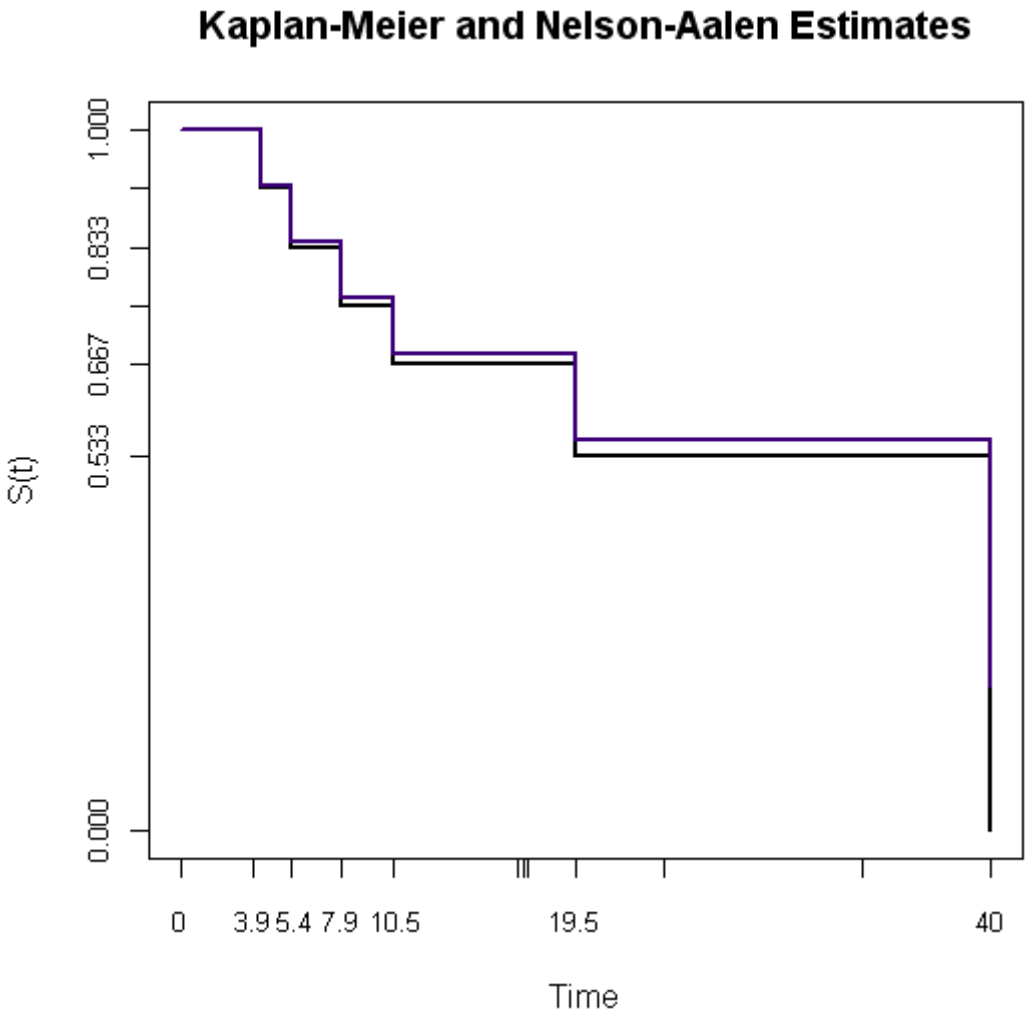
  - A confidence interval for $\tilde{S}(t)$ can be found by exponentiating the CI for $\tilde{H}(t)$

Nelson-Aalen **Kaplan-Meier**

- $\tilde{S}(t) \geq \hat{S}(t)$ for all t, but the two estimates are often very close

Estimating the Cumulative Hazard Function

- Previous example:
  - Survival times: 3.9, 5.4, 7.9, 10.5, 16.6+, 16.9+, 17.1+, 19.5, 23.8+, 33.7+, 33.7+, 40



The Hazard Function

- The hazard function is useful for describing the way that the risk of an event changes with time.
- In contrast to S(*t*), the graph of h(*t*) can start anywhere and can increase, decrease, remain constant, or change constantly over time.

The Hazard Function

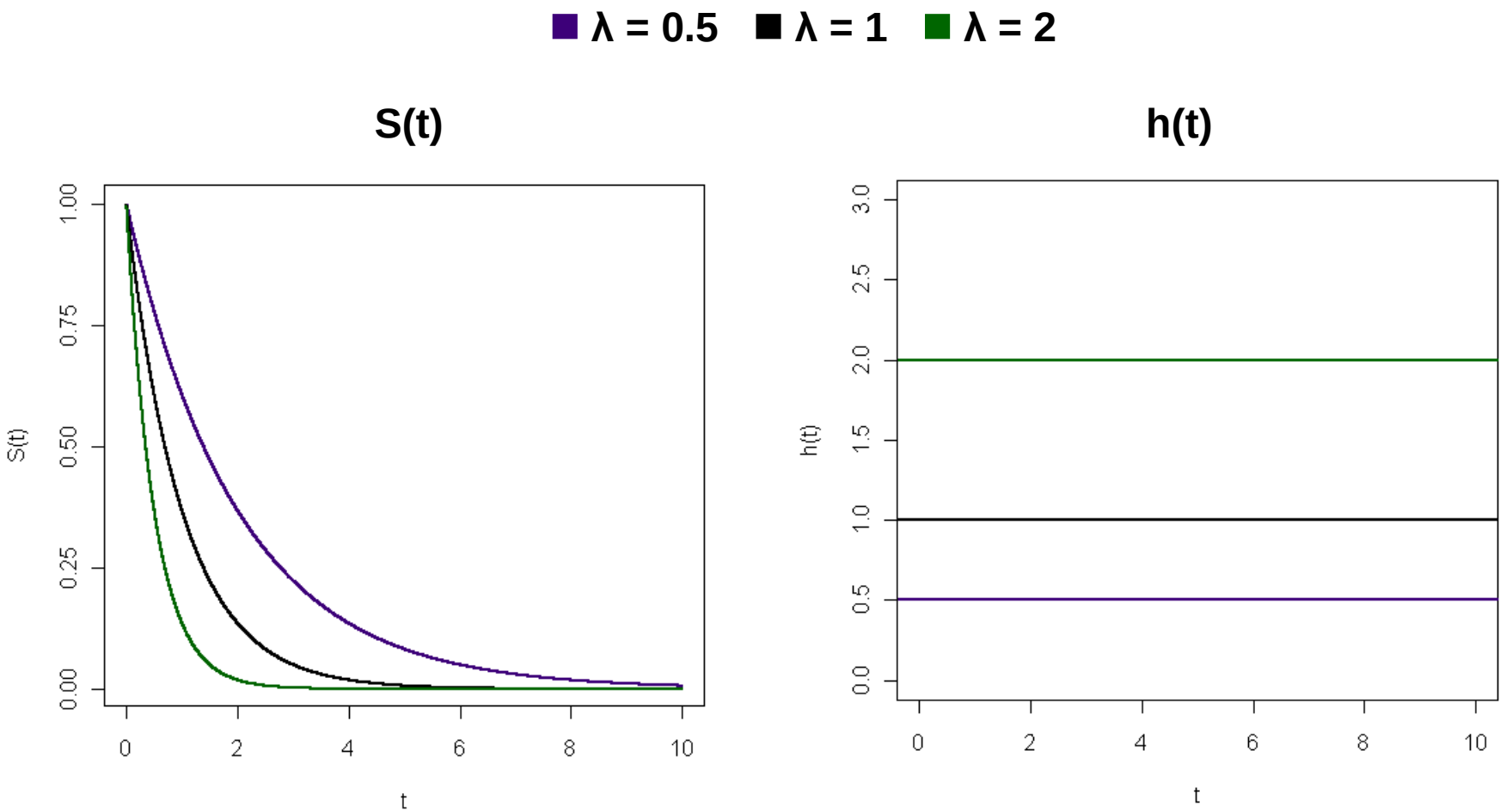
- The simplest hazard function is one that says that the hazard is constant over time

$$h(t) = \lambda$$

which implies

- $S(t) = e^{-\lambda t}$
- $f(t) = \lambda e^{-\lambda t}$, i.e. T follows an exponential distribution with mean $1/\lambda$

The Hazard Function



The Hazard Function

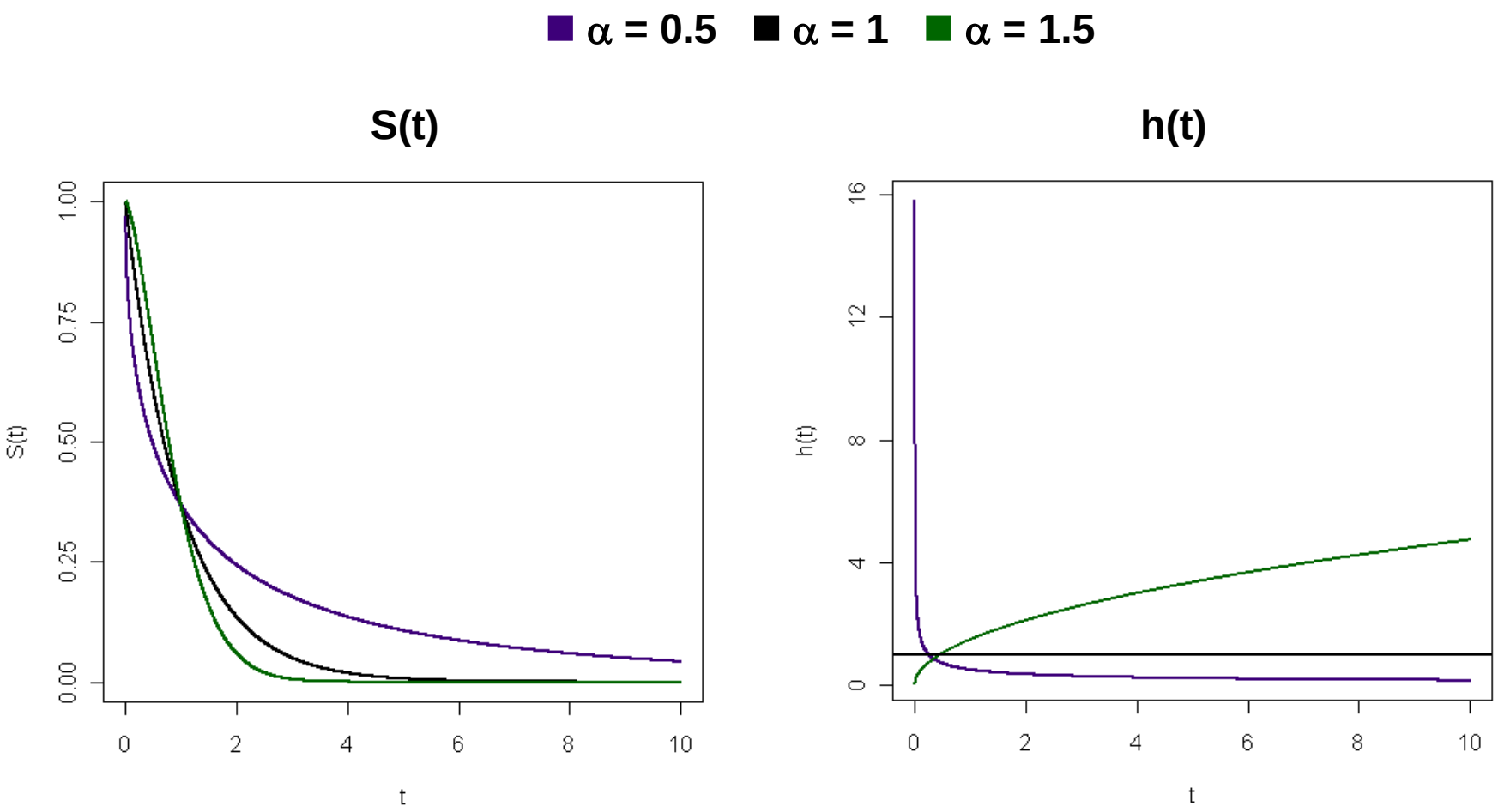
- The next step up in complexity is to let the hazard be some function of t.
- For example, one can use:

$$h(t) = \lambda\alpha t^{\alpha-1}$$
$$\lambda > 0, \alpha > 0$$

which implies

- f(t) = $\lambda\alpha t^{\alpha-1}e^{-\lambda t^{\alpha}}$, i.e. T follows a Weibull distribution
- S(t) = $e^{-\lambda t^{\alpha}}$
  - If α = 1, the hazard is constant.
  - If α > 1, the hazard increases with time.
  - If α < 1, the hazard decreases with time.

The Hazard Function



The Hazard Function

- Note that the survival functions have the same basic shape, but the hazard functions are very different.
- The hazard function usually gives more information regarding the underlying mechanism of failure than the survival function.
- For this reason, it is usually preferred to use the hazard function to summarize survival data.