QMWS - Survival Analysis
Non-parametric survival methods

Instructor: Kevin McGregor

York University
Department of Mathematics and Statistics

Non-parametric statistics

There are several common techniques in survival analysis that fall under the category of **non-parametric statistics**. What does this mean?

Parametric statistics: This means that we assume some parametric form (i.e. a probability distribution) for our data:

- E.g. $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$

Non-parametric statistics: We do **not** assume any particular probability distribution for our data.

Non-parametric survival analysis

There are two main non-parametric methods for survival analysis that we will cover:

Kaplan-Meier estimator: This gives an estimate of the **survival** function $S(t)$.

- By far the most common method used in survival analysis.
- The Kaplan-Meier estimate is the default method in R for plotting the survival curve.

Nelson-Aalen estimator: This gives an estimate of the **cumulative hazard** function $\Lambda(t)$.

- This can then be transformed to $S(t)$.

Both of these estimators handle right-censored data.

Notation

There is a particular notation used in both the Kaplan-Meier estimator and the Nelson-Aalen estimator.

In both cases, we will index our observations by the **observed event times**. That is, assume that, among n individuals in the study, we observe k events (and have $n - k$ right-censored events). The **observed** event times are:

$$t_1 < t_2 < \cdots < t_k$$

Note that we assume **discrete time**, so that **multiple** events can happen at each of these times. The number of events occurring at each time (respectively) is:

$$d_1, d_2, \ldots, d_k$$

Notation

Finally, at each time t_j, for $j = 1, \ldots, k$, we have the number of individuals **at risk**:

$$n_1, n_2, \ldots, n_k$$

Each n_j is the number of individuals still in the study at time t_j. The two ways an individual is no longer "at risk" is if (a) they have an event, or (b) they are right censored.

- n_1 is the sample size
- $n_j = n_{j-1} - d_{j-1} - \#$ censored in $[t_{j-1}, t_j)$

Example

Example: Time to relapse (weeks) for 21 children with acute leukemia who are on a drug called 6-MP (Freireich et al. 1963): 10, 7, 32+, 23, 22, 6, 16, 34+, 32+, 25+, 11+, 20+, 19+, 6, 17+, 35+, 6, 13, 9+, 6+, 10+

t_j	d_j	n_j
6	3	21
7	1	
10	1	
13	1	
16	1	
22	1	
23	1	

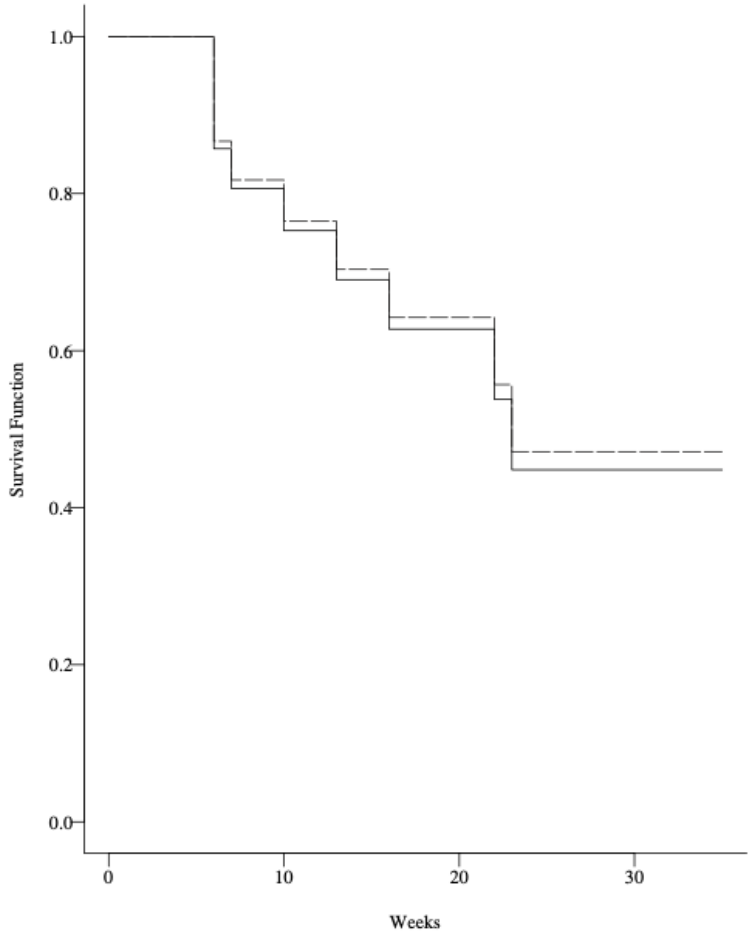
Kaplan-Meier estimator

The Kaplan-Meier estimator (also known as the "Product-Limit" estimator) is the most common way of estimating a survival curve $S(t)$.

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

Kaplan-Meier estimator

The Kaplan-Meier curve is a step function, with drops at each t_j. Here's the KM curve for the 6-MP example (solid line):



Klein, John P., and Melvin L. Moeschberger. Survival analysis: techniques for censored and truncated data. Vol. 1230. New York: Springer, 2003.

Variance of Kaplan-Meier estimator

An important formula called **Greenwood's formula** approximates the variance of the Kaplan-Meier curve:

$$\widehat{var}\left[\widehat{S}(t)\right] = \widehat{S}(t)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

This formula is derived using the **delta method**. Thus, it is only an **approximation** of the variance.

Confidence interval: Kaplan-Meier estimator

Using Greenwood's formula, we can get a pointwise confidence interval for the Kaplan-Meier estimator. That is, we have a $(1-\alpha) \times 100\%$ confidence interval at each value of t:

$$\widehat{S}(t) \pm z_{\alpha/2} \sqrt{\widehat{var}\left[\widehat{S}(t)\right]}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Cumulative hazard

Recall the hazard function:

$$\lambda(t) = P(t < T_i < t + \delta | T_i > t)$$

for arbitrarily small δ.

Cumulative hazard

An important related function is called the **cumulative hazard** function:

$$\Lambda(t) = \int_0^t \lambda(s)\, ds$$

There is a relationship between the survival function and the cumulative hazard:

$$S(t) = \exp\{-\Lambda(t)\}$$

Nelson-Aalen estimator

Another popular non-parametric estimator in survival analysis is the **Nelson-Aalen** estimator.

The Nelson-Aalen estimator estimates the **cumulative hazard** function $\Lambda(t)$.

We use the same notation as we did for the Kaplan-Meier estimate:

- Discrete event times $t_1 < t_2 < \cdots < t_k$, multiple events at each time are possible.
- n_j is the number at risk at time t_j
- d_j is the number of events at time t_j

Nelson-Aalen estimator

The **Nelson-Aalen estimator** for cumulative hazard is given by:

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

Thus, the survival function can be estimated as:

$$\begin{aligned} \widehat{S}(t) &= \exp\left\{-\widehat{\Lambda}(t)\right\} \\ &= \exp\left\{-\sum_{j:t_j \leq t} \frac{d_j}{n_j}\right\} \end{aligned}$$

Variance

The variance of the Nelson-Aalen estimator can be estimated as:

$$\widehat{var}\left[\widehat{\Lambda}(t)\right] = \sum_{j:t_j \leq t} \frac{(n_j - d_j)d_j}{(n_j - 1)n_j^2}$$

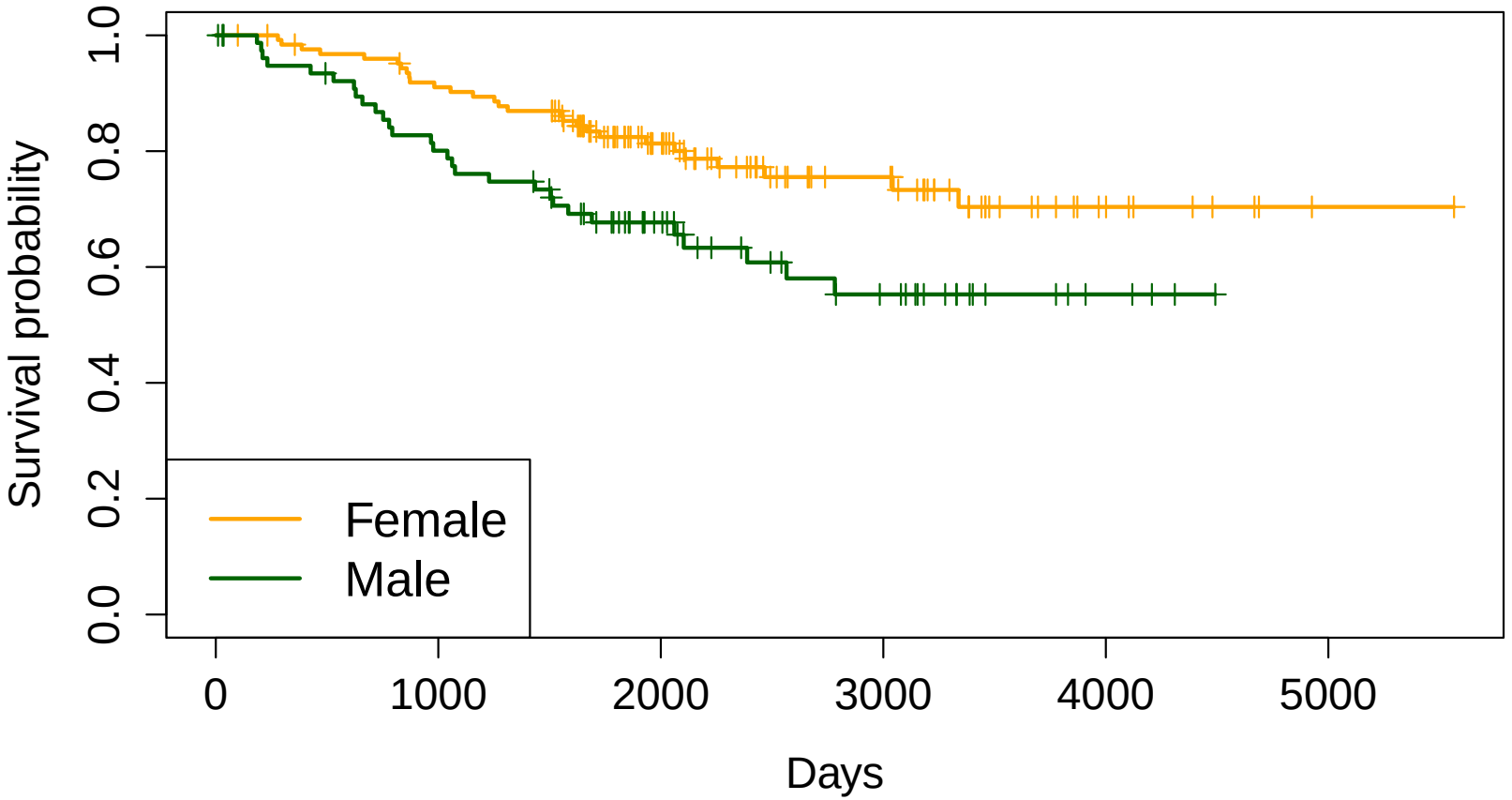
A $100 \times (1 - \alpha)\%$ confidence interval for $\Lambda(t)$ is given by:

$$\widehat{\Lambda}(t) \pm z_{\alpha/2}\sqrt{\widehat{var}\left[\widehat{\Lambda}(t)\right]}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ tail of the standard normal distribution.

Comparing curves

Comparing survival curves for males/females in the Melanoma dataset.



Creating multiple curves

When creating multiple Kaplan-Meier curves, we split the usual n_j (number at risk) and d_j (number of events) vectors into multiple groups.

Let's consider the simplest case of 2 groups.

- Among the 2 groups we have **observed** event times:

$$t_1 < t_2 < \cdots < t_k$$

- d_j is the total number of events at time t_j (across **both** groups)
- n_j is the total number at risk at time t_j (across **both** groups)

Creating multiple curves

- we have d_{1j} and d_{2j}, which are the numbers of events in groups 1 and 2 at time t_j
  - $d_{1j} + d_{2j} = d_j$
- We have n_{1j} and n_{2j}, which are the number of at risk individuals in groups 1 and 2 at time t_j
  - $n_{1j} + n_{2j} = n_j$

	Events			At risk		
Time	Grp 1	Grp 2	Total	Grp 1	Grp 2	Total
t_1	d_{11}	d_{21}	d_1	n_{11}	n_{21}	n_1
t_2	d_{12}	d_{22}	d_2	n_{12}	n_{22}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_k	d_{1k}	d_{2k}	d_k	n_{1k}	n_{2k}	n_k

Creating multiple curves

We can simply create two separate Kaplan-Meier curves using the within-group event and at risk counts:

Survival curve for group 1:

$$\widehat{S}_1(t) = \prod_{t_j : t_j \leq t} \left(1 - \frac{d_{1j}}{n_{1j}}\right)$$

Survival curve for group 2:

$$\widehat{S}_2(t) = \prod_{t_j : t_j \leq t} \left(1 - \frac{d_{2j}}{n_{2j}}\right)$$

Log-rank test

How do we compare the curves? We can test for differences between the two survival curves using the **log-rank test**, also known as the **Mantel-Cox** test.

We have the following hypotheses in the log-rank test:

$$H_0: \quad S_1(t) = S_2(t) \text{ for all } t > 0$$

$$H_1: \quad S_1(t) \neq S_2(t) \text{ for some } t > 0$$

Log-rank test: distribution

What are the **expected** values of the d_{ij}, for $i = 1, \ldots, p$ under the null hypothesis?

$$e_{ij} = \mathbb{E}[d_{ij}] = \frac{n_{ij}}{n_j} d_j$$

Let V_j be the **variance-covariance matrix** for the d_{ij} $i = 1, \ldots, p$ under the null hypothesis at time t_j. Diagonal terms:

$$(V_j)_{ii} = \text{var}[d_{ij}] = d_j \frac{n_{ij}}{n_j} \frac{n_j - n_{ij}}{n_j} \frac{n_j - d_j}{n_j - 1}$$

Off-diagonal terms for elements i and r (e.g. covariances):

$$(V_j)_{ir} = -d_j \frac{n_{ij}}{n_j} \frac{n_{rj}}{n_j} \frac{n_j - d_j}{n_j - 1}$$

Log-rank test: test statistics

Our test statistic in the log-rank test is then:

$$X^2 = \frac{\left(\sum_{j=1}^{k} d_{1j} - \sum_{j=1}^{k} e_{1j}\right)^2}{\sum_{j=1}^{k} v_{1j}}$$

It can be shown that this is approximately distributed as chi-squared with one degree of freedom under the null hypothesis:

$$X^2 \sim \chi_1^2 \qquad \text{(approximately under } H_0\text{)}$$

Thus, we reject H_0 at level α if $X^2 > \chi_1^2(\alpha)$, where $\chi_1^2(\alpha)$ is the upper-α quantile of the χ_1^2 distribution.

Advantages/disadvantages

Log-rank test advantages:

- Simple calculation.
- Easy to interpret.
- Easily extends to p survival curve comparisons.

Log-rank test disadvantages:

- Will often not detect difference in two survival curves if they cross.
- Assumes censoring is not related to survival time.
- Assumes probability censoring is not significantly different between groups.