

## Lecture 5

### Survival analysis

Let  $T$  be a non-negative random variable that represents the time to event (for example,  $T$  might be unemployment time, or time from a treatment to death of a patient, or time from production to failure of a device, etc.). Denote  $F(t) = P(T \leq t)$  the c.d.f. of  $T$  and assume that  $F$  has a p.d.f.  $f$ . A central concept of the survival analysis is the *hazard function*

$$\begin{aligned} h(t) &= \lim_{td \rightarrow 0} \frac{P(t \leq T \leq t + td | T \geq t)}{td} \\ &= \lim_{td \rightarrow 0} \frac{1}{td} \frac{P(t \leq T \leq t + td)}{P(T \geq t)} \\ &= \frac{f(t)}{1 - F(t)}, \end{aligned}$$

That is,  $h(t)$  is the instantaneous rate of failure at a time  $T = t$  under the condition of survival to the time  $t$  (loosely speaking,  $h(t)$  is the probability density of failure at time  $t$ , given survival to then). If  $T$  is a discrete random variable,  $h(t) = P(T = t | T > t) = P(T = t) / \{1 - F(t)\}$ . Function

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

is called *survivor function* of  $T$ . Next, define the *cumulative hazard function* by

$$H(t) = \int_0^t h(s) ds = \int_0^t \frac{f(s)}{1 - F(s)} ds = \int_0^t \frac{F'(s)}{1 - F(s)} ds = -\log\{S(t)\}.$$

Thus the survivor function may be written as  $S(t) = \exp\{-H(t)\}$  and  $f(t) = h(t)S(t) = h(t) \exp\{-H(t)\}$ .

### Examples

1. Exponential distribution:  $h(t) = \lambda$  and  $S(t) = \exp(-\lambda t)$
2. Weibull distribution:  $h(t) = \alpha \lambda^\alpha t^{\alpha-1}$  and  $S(t) = \exp\{-(\lambda t)^\alpha\}$
3. Log-logistic distribution:  $h(t) = \alpha \lambda^\alpha t^{\alpha-1} \{1 + (\lambda t)^\alpha\}^{-1}$  and  $S(t) = \{1 + (\lambda t)^\alpha\}^{-1}$

Ideally, we would have independent realisations of  $T$ :  $t_1, \dots, t_n$ . However, in practice the failure time can not always be observed due to various reasons. This phenomenon is called *censoring*. The simplest form of censoring occurs when  $T$  is observed until

some pre-determined time  $c$ . If  $T < c$ , we observe the value  $t_i$  of  $T$ , if  $T > c$ , we only know that  $T$  survived beyond  $c$ . This is called *Type I censoring*. *Type II censoring* arises when  $n$  independent variables are observed until there have been  $r$  failures, so only  $0 < T_{(1)} < \dots < T_{(r)}$  are observed. This type of censoring has an open-ended random trial time and is therefore impractical and is rarely used. Under *random censoring* the  $j$ th subject in the study has a random censoring time  $C_j$  drawn from some distribution  $G$ , independent of  $T_j$ . These are all examples of *right-censoring*. Left-censoring (the time of origin is not known) is less common.

Hence, under censoring one rather deals with  $Y_j = \min\{T_j, C_j\}$ , while it is known if  $Y_j = T_j$ . That is, a pair  $(y_j, \delta_j)$  is observed, where  $\delta_j = 1$  if  $y_j$  is the survival time and  $\delta_j = 0$ , if  $y_j$  is the censoring time. Note that the assumption of independence of  $T$  and  $C$  is crucial.

Now assume that  $T$  has a continuous distribution  $F$  and there are  $n$  data points available  $(y_1, \delta_1), \dots, (y_n, \delta_n)$ , where  $y_i = \min\{t_i, c_i\}$ . Assume that  $F(x) = F(x; \theta)$  is a some parametric distribution and that censoring variables  $C_i$  (independent on  $T_i$ ) have c.d.f.  $G$  and p.d.f.  $g$ , which are independent on  $\theta$ . Hence, the likelihood contribution from  $y_i$  is

$$\begin{cases} f(y_i; \theta)\{1 - G(y_i)\}, & \text{if } \delta_i = 1 \\ S(y_i; \theta)g(y_i), & \text{if } \delta_i = 0. \end{cases}$$

Since  $G$  and  $g$  are independent on  $\theta$ , the likelihood becomes

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(y_i; \theta)^{\delta_i} S(y_i; \theta)^{1-\delta_i},$$

while log-likelihood is

$$\ell(\theta) = \sum_{i=1}^n [\delta_i \log\{f(y_i; \theta)\} + (1 - \delta_i) \log\{S(y_i; \theta)\}].$$

This includes Type I censoring, for which  $G$  puts all its probability at  $c$ . Note that we can represent the log-likelihood as

$$\ell(\theta) = \sum_{i=1}^n [\delta_i \log\{h(y_i; \theta)\} - H(y_i; \theta)].$$

For exponential distribution  $h(t; \lambda) = \lambda$  and  $H(t; \lambda) = \lambda t$ , so that the log-likelihood becomes

$$\ell(\lambda) = \sum_{i=1}^n \{\delta_i \log(\lambda) - \lambda y_i\} = \log(\lambda) \sum_{i=1}^n \delta_i - \lambda \sum_{i=1}^n y_i,$$

implying

$$\hat{\lambda}_{ML} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n y_i}.$$

In particular, if all observations are censored (=no failures), the estimator is zero. To calculate the (asymptotic) variance of a maximum likelihood estimator, we would need to calculate the Fisher Information  $E\{J(\lambda)\}$ , where  $J(\lambda) = -\partial^2 \ell(\lambda) / \partial \lambda^2$  is the observed information. However, this is not possible without some assumption on  $G$ . In practice one approximates  $\widehat{\text{var}}(\hat{\lambda}) = J(\hat{\lambda})^{-1} = \hat{\lambda}^2 / \sum_{i=1}^n \delta_i$ . In particular, this can be used to build an approximate confidence interval for  $\lambda$  (using asymptotic normality of maximum likelihood estimators) as

$$\left[ \hat{\lambda}(1 - z_{\alpha/2}/\sqrt{r}), \hat{\lambda}(1 + z_{\alpha/2}/\sqrt{r}) \right], \quad r = \sum_{i=1}^n \delta_i.$$

The assumption of a constant hazard function is often unrealistic (for example, the instantaneous failure rate (hazard) of a technical device usually grows with the time from being put into service). A commonly used parametric distribution for modelling lifetimes with monotone hazard is the Weibull distribution. Values for  $\lambda$  and  $\alpha$  can be estimated by the maximum likelihood similarly to the exponential distribution, however, this has to be done numerically.

### Kaplan-Meier and Fleming-Harrington estimator

Often it is unclear which parametric model would be appropriate for the data (if any). A standard tool for initial data inspection, for suggesting plausible models and for checking their fit is a nonparametric estimator of the survivor function. If there were no censored observations, then we could estimate  $\hat{S}(t) = n^{-1} \sum_{i=1}^n \mathbb{I}(T_i > t)$ . For censored observations we have the likelihood

$$\mathcal{L}(S) = \prod_{i=1}^n f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}.$$

Let  $0 \leq \tau_1 < \tau_2 < \dots$  be the ordered uncensored failure times. Let  $r_i$  denote the number of units that are still in risk at  $\tau_i$  (=not failed yet or censored) and  $d_i$  the number of units that fail at  $\tau_i$ . It can be shown that the function  $\mathcal{L}$  is maximized by the piecewise constant function  $\hat{S}_{\text{KM}}$  defined by

$$\hat{S}_{\text{KM}}(t) = \prod_{\{j: \tau_j < t\}} \left( 1 - \frac{d_j}{r_j} \right).$$

This is the *Kaplan-Meier* estimator for the survivor function  $S$ .

A further estimator for  $S$  is the *Fleming-Harrington* estimator  $\hat{S}_{\text{FH}}$ . It is a plug in estimator defined by  $\hat{S}_{\text{FH}} = \exp\{-\hat{H}(t)\}$ , where  $\hat{H}$  is the *Nelson-Aalen* estimator for  $H$ .

It is defined by

$$\hat{H}(t) = \sum_{\{j:\tau_j < t\}} \frac{d_j}{r_j}.$$

Observe that

$$\hat{S}_{\text{FH}} = \exp\{-\hat{H}(t)\} = \prod_{\{j:\tau_j < t\}} \exp\left(-\frac{d_j}{r_j}\right).$$

Since  $1 - x \approx \exp(-x)$  for small  $x$ , the estimators  $\hat{S}_{\text{FH}}$  and  $\hat{S}_{\text{KM}}$  are quite similar, if many items are still at risk.

We next aim for computing **confidence bands** for the true survivor function  $S$ . To this end, assume that  $\hat{S}$  is an estimator for  $S$  (e.g. the Kaplan-Meier or the Fleming-Harrington estimator) and let  $\widehat{\text{var}}(\log \hat{S})$  be some estimate for the variance of  $\log \hat{S}$ . Then, by the delta-method, it follows that  $\text{var}\{\hat{S}(t)\} \approx \hat{S}^2 \widehat{\text{var}}(\log \hat{S})$  and hence an approximate confidence band can be constructed by

$$\left[ \hat{S}(t) - z_{\alpha/2} \hat{S}(t) \sqrt{\widehat{\text{var}}\{\log \hat{S}(t)\}}, \hat{S}(t) + z_{\alpha/2} \hat{S}(t) \sqrt{\widehat{\text{var}}\{\log \hat{S}(t)\}} \right].$$

The main problem with this is the fact that the upper and lower bounds may be larger than 1 and smaller than 0, respectively. As a way out, one considers confidence bands for the statistic  $\log\{-\log(\hat{S})\}$  (that has the range  $\mathbb{R}$ ) and obtains again by the delta-method that

$$\text{var}[\log\{-\log \hat{S}(t)\}] \approx \frac{1}{\log^2 \hat{S}(t)} \widehat{\text{var}}(\log \hat{S}).$$

Setting  $B^\pm = \log(-\log \hat{S}(t)) \pm z_{\alpha/2} \log^{-1} \hat{S}(t) \sqrt{\widehat{\text{var}}(\log \hat{S})}$  an approximate confidence band (contained in  $[0, 1]$ ) is given by

$$[\exp(-\exp(B^-)), \exp(-\exp(B^+))].$$

As an example we obtain for the Kaplan-Meier estimator by Greenwood's formula

$$\widehat{\text{var}}\{\log \hat{S}_{\text{KM}}(t)\} \approx \sum_{\{j:\tau_j < t\}} \frac{d_j}{r_j(r_j - d_j)}.$$

Often we wish to decide whether or not two (or more) samples stem from the same survivor function or not. The **log-rank test** is such a simple test procedure. We will now assume that the failure times  $\tau_1 < \tau_2 < \dots < \tau_k$  are realizations of *two* random variables  $T_1$  and  $T_2$  corresponding to two groups of items (patients). For each observed failure time  $\tau_j$  we consider the contingency table

Groups	failure at time $\tau_j$	items at risk at time $\tau_j$
1	$d_{1j}$	$r_{1j}$
2	$d_{2j}$	$r_{2j}$
1 + 2	$d_j$	$r_j$

Under the null-hypothesis that  $T_1 = T_2$  the expected number of failures at time  $\tau_j$  in group 1 and 2 are hypergeometrically distributed with parameters  $r_j, r_{1j}, d_j$  and  $r_j, r_{2j}, d_j$ , respectively. Thus, mean and variance of the number of failures in group 1 and 2 can be computed as

$$e_{1j} = \frac{d_j}{r_j} r_{1j} \quad \text{and} \quad e_{2j} = \frac{d_j}{r_j} r_{2j},$$

and

$$v_{1j} = v_{2j} = \frac{d_j r_{1j} r_{2j} (r_j - d_j)}{r_j^2 (r_j - 1)}.$$

Under the null-hypothesis, the statistic

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

is  $\chi^2$ -distributed with 1 degree of freedom.

An important question in practice is: Is the assumption of a Weibull distributed survivor time justified? An indication to the answer of this question can be obtained as follows: Observe that under the assumption that  $T$  is Weibull distributed one has

$$\log\{-\log S(t)\} = \alpha \log(t) + \log(\lambda), \quad t > 0.$$

Now let  $\hat{S}(t)$  be a nonparametric estimate for  $S$  (e.g. the Kaplan-Meier estimator  $\hat{S}_{\text{KM}}$ ). Then the plot  $\log\{-\log \hat{S}(t)\}$  against  $\log(t)$  should approximately be a straight line with slope  $\alpha$  and intercept  $-\log(\lambda)$ .