Lecture 4
PARAMETRIC SURVIVAL MODELS

Some Parametric Survival Distributions (defined on $t \geq 0$):

- The **Exponential** distribution (1 parameter)

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t} \ (\lambda > 0) \\ S(t) &= \int_t^\infty f(u)du \\ &= e^{-\lambda t} \\ \lambda(t) &= \frac{f(t)}{S(t)} \\ &= \lambda \quad \textbf{constant hazard!} \\ \Lambda(t) &= \int_0^t \lambda(u)\, du \\ &= \int_0^t \lambda\, du \\ &= \lambda t \end{aligned}$$

Check: Does $S(t) = e^{-\Lambda(t)}$?

median: solve $0.5 = S(\tau) = e^{-\lambda\tau}$:

$$\Rightarrow \tau = \frac{\log 2}{\lambda}$$

mean:

$$\int_0^\infty u\lambda e^{-\lambda u} du = \frac{1}{\lambda}$$

- The **Weibull** distribution (2 parameters)
  Generalizes exponential:

$$\begin{aligned} S(t) &= e^{-\lambda t^\kappa} \quad (\lambda, \kappa > 0) \\ f(t) &= \frac{-d}{dt} S(t) = \kappa\lambda t^{\kappa-1} e^{-\lambda t^\kappa} \\ \lambda(t) &= \kappa\lambda t^{\kappa-1} \\ \Lambda(t) &= \int_0^t \lambda(u) du = \lambda t^\kappa \end{aligned}$$

λ - the *scale* parameter
κ - the *shape* parameter

The Weibull distribution is convenient because of simple forms. It includes several hazard shapes:

$\kappa = 1 \rightarrow$ constant hazard (i.e. Exponential)

$0 < \kappa < 1 \rightarrow$ decreasing hazard

$\kappa > 1 \rightarrow$ increasing hazard

- **Rayleigh** distribution
  Another 2-parameter generalization of exponential:

$$\lambda(t) = \lambda_0 + \lambda_1 t$$

- **log-normal**, **log-logistic**:
  Distributions for T obtained by specifying for $\log T$ convenient family of distributions, e.g.

  $\log T \sim$ normal (non-monotone hazard)

  $\log T \sim$ logistic (density $\frac{e^t}{(1+e^t)^2}$)

- **piecewise exponential**:
  Let $0 = t_0 < t_1 < ... < t_K < t_{K+1} = \infty$,

$$\lambda(t) = \lambda_k, \;\; \lambda_k > 0$$

  if $t_{k-1} \leq t < t_k$.

- **compound exponential**

  $T \sim \exp(\lambda)$, $\lambda \sim g$

$$f(t) = \int_0^\infty \lambda e^{-\lambda t} g(\lambda) d\lambda$$

- **Generalized gamma, generalized F, inverse Gaussian**...

Likelihood for Right-censored Data

- **Observed data:** (X_i, δ_i), $i = 1, ..., n$

- **Right censoring:** recall that $X_i = \min(T_i, C_i)$

- **Parametric distribution:** eg. T_i follows an exponential distribution with a parameter λ, and we can write:

  $$T_i \sim \text{Exp}(\lambda)$$

  (What is the interpretation of λ?)

For right-censored data, each observation has one of two possible contributions to the likelihood:

(a) if it is an **event** at X_i ($\delta_i = 1$) $\Rightarrow$ contribution is

$$L_i = f(X_i) = \underbrace{S(X_i)}_{\text{survive to } X_i} \cdot \underbrace{\lambda(X_i)}_{\text{fail at } X_i} = e^{-\lambda X_i} \cdot \lambda$$

Note that this is the usual likelihood term for uncensored data that we have known;

(b) if it is **censored** at X_i ($\delta_i = 0$) $\Rightarrow$ contribution is

$$L_i \;=\; \underbrace{S(X_i)}_{\text{survive to } X_i} \;=\; e^{-\lambda X_i}$$

This is because the information we have from the data is that $T_i > X_i$, and this is $P(T_i > X_i)$.

The **likelihood** is the product over all of the observations:

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} L_i \\
&= \prod_{i=1}^{n} \underbrace{f(X_i)^{\delta_i}}_{\text{events}} \; \underbrace{S(X_i)^{(1-\delta_i)}}_{\text{censorings}} \\
&= \prod_{i=1}^{n} \lambda(X_i)^{\delta_i} \cdot S(X_i)
\end{aligned}$$

[Recall $f(t) = \lambda(t)S(t)$.]

Maximum Likelihood for Exponential

Plug in the likelihood terms for Exp distribution:

$$\begin{aligned}\mathcal{L}(\lambda) &= \prod_i \left(\lambda e^{-\lambda X_i}\right)^{\delta_i} \quad \left(e^{-\lambda X_i}\right)^{(1-\delta_i)} \\ &= \prod_i \lambda^{\delta_i} \left(e^{-\lambda X_i}\right)\end{aligned}$$

How do we use the likelihood?

- first take the log
- then take the partial derivative with respect to λ
- then set to zero and solve for $\hat{\lambda}$
- this gives us the **maximum likelihood estimator**

The log-likelihood is:

$$\begin{aligned}\log \mathcal{L}(\lambda) &= \log \left[\prod_i \lambda^{\delta_i} e^{-\lambda X_i}\right] \\ &= \sum_i \left[\delta_i \log(\lambda) - \lambda X_i\right] \\ &= \log(\lambda) \sum_i \delta_i - \lambda \sum_i X_i\end{aligned}$$

We set $\frac{\partial \log \mathcal{L}}{\partial \lambda} = \sum \delta_i / \lambda - \sum X_i = 0$ and solve:

$$\Rightarrow \quad \hat{\lambda} = \frac{d}{t}$$

where $d = \sum \delta_i$ is the total number of deaths (or events), and $t = \sum X_i$ is the total 'person-time' contributed by all individuals.

What happens if there is no censoring?

Using the 2nd derivative of the log-likelihood (how)

$$\widehat{Var}(\hat{\lambda}) = \left(\frac{d}{\hat{\lambda}^2} \right)^{-1} = \frac{d}{t^2}$$

Read about the **observed Fisher information** (Efron and Hinkley 1978).

Exponential model: Mean and Median

Mean Survival Time

For the exponential distribution, $E(T) = 1/\lambda$. The estimate is

$$\overline{T} = 1/\hat{\lambda} \;=\; \frac{t}{d}$$

Median Survival Time

This is the value M at which $S(t) = e^{-\lambda t} = 0.5$, so $M =$ median $= \frac{\log 2}{\lambda}$. The estimate is

$$\hat{M} = \frac{\log 2}{\hat{\lambda}} \;=\; \log 2 \cdot \frac{t}{d}$$

Hand Calculations using events and follow-up:

Nursing home example:

By adding up "LOS" for males to get t_1 and for females to get t_0, we obtain: **For the females:**

- $n_0 = 1173$
- $d_0 = 902$
- $t_0 = 310754$

What is the estimate of λ_0, its variance, mean and median survival?

For the males:

- $n_1 = 418$
- $d_1 = 367$
- $t_1 = 75457$

What is the estimate of λ_1, its variance, mean and median survival?

How to choose one parametric distribution versus another?

- qualitative shape of hazard function
- explicit simple forms for $f(t), S(t)$, and $\lambda(t)$, and simple interpretation
- technical convenience for estimation and inference, availability of software
- how well a model fits the data

One can usually distinguish between a one-parameter model (like the exponential) and two-parameter (like Weibull or log-Normal) in terms of the adequacy of fit to a dataset by, for example, testing for the additional parameter.

There are also graphical methods using the Kaplan-Meier estimate of survival.

Without a lot of data, it may be hard to distinguish between the fits of various 2-parameter models (i.e., Weibull vs log-normal)

Checking parametric assumptions

As mentioned before, these estimates can be used to check parametric assumptions, such as exponential and Weibull.

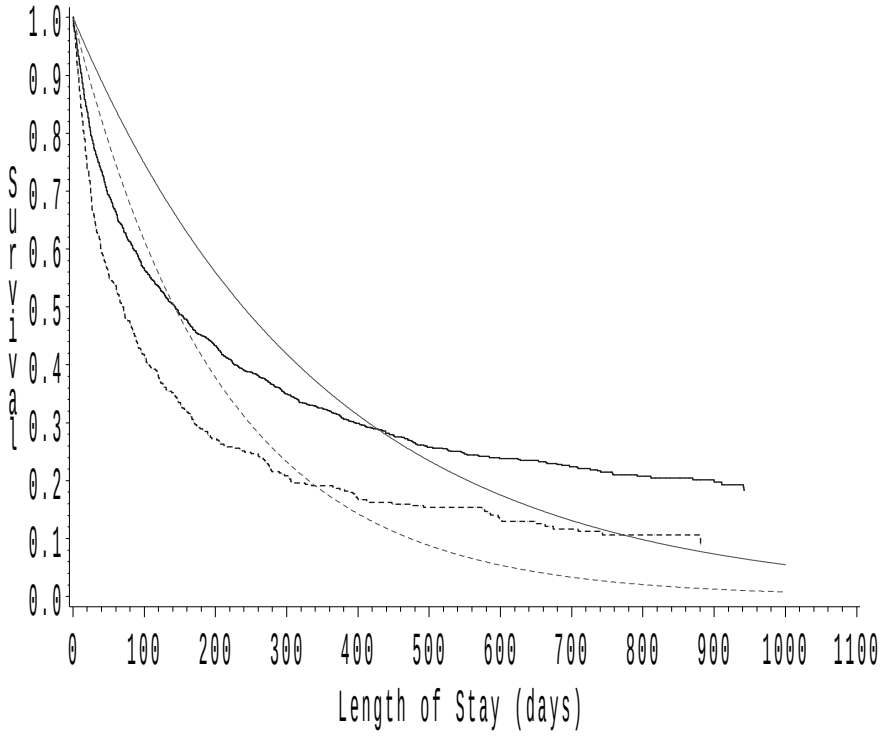
The idea is (almost always) to compare the nonparametric estimate to what is obtained under the parametric assumption.

Example: nursing home data

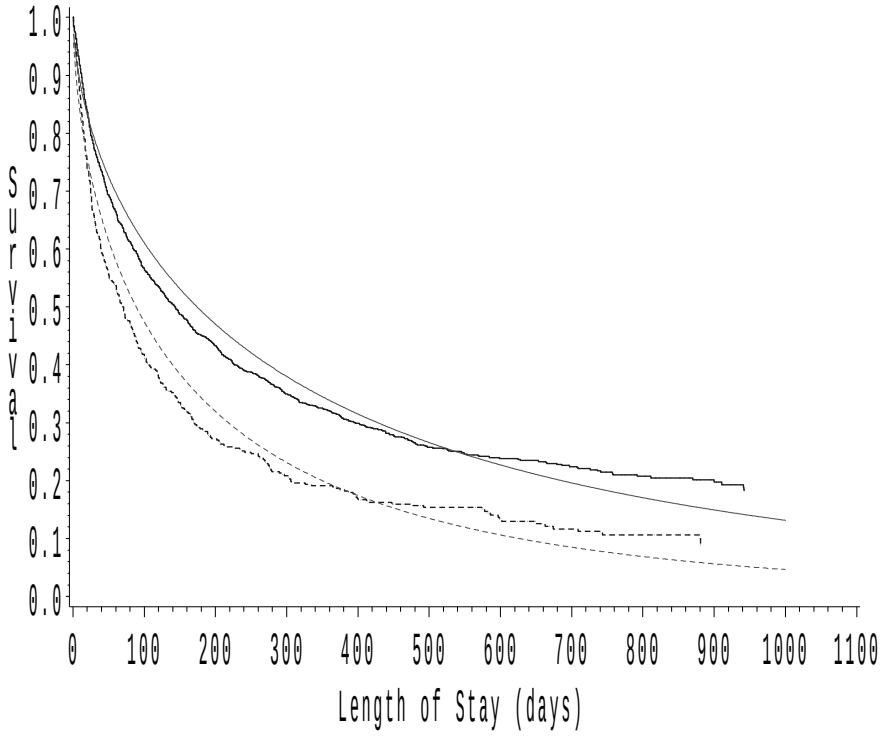
We can see how well the Exponential model fits by comparing the survival estimates for males and females under the exponential model, i.e., $P(T \geq t) = e^{(-\hat{\lambda}_z t)}$, to the Kaplan-Meier survival estimates:

We can see how well the Weibull model fits by comparing the survival estimates, $P(T \geq t) = e^{(-\hat{\lambda}_z t^{\hat{\kappa}})}$, to the Kaplan-Meier survival estimates.

Predicted Survival for Exponential model vs Kaplan – Meier



Predicted Survival for Weibull model vs Kaplan – Meier



Note that direct comparison of survival curves are sometimes less informative. Better ways are to:

- $-\log(\hat{S}(t))$ vs t
- $\log[-\log(\hat{S}(t))]$ vs $\log(t)$

Why are these useful?

If T is exponential, then $S(t) = \exp(-\lambda t))$

$$\begin{aligned} \text{so} \quad \log(S(t)) &= -\lambda t \\ \text{and} \quad \Lambda(t) &= \lambda\, t \end{aligned}$$

a straight line in t with slope λ and intercept=0

If T is Weibull, then $S(t) = \exp(-\lambda t^{\kappa})$

$$\begin{aligned} \text{so} \quad \log(S(t)) &= -\lambda t^{\kappa} \\ \text{then} \quad \Lambda(t) &= \lambda t^{\kappa} \\ \text{and} \quad \log(-\log(S(t))) &= \log(\lambda) + \kappa * \log(t) \end{aligned}$$

a straight line in $\log(t)$ with slope κ and intercept $\log(\lambda)$.

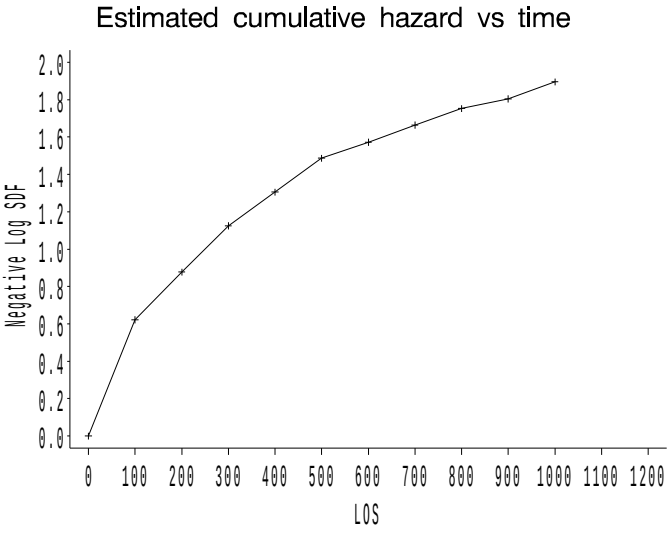


Figure 1: Nursing home data: $\hat{\Lambda}(t)$ vs t for checking exponential distribution

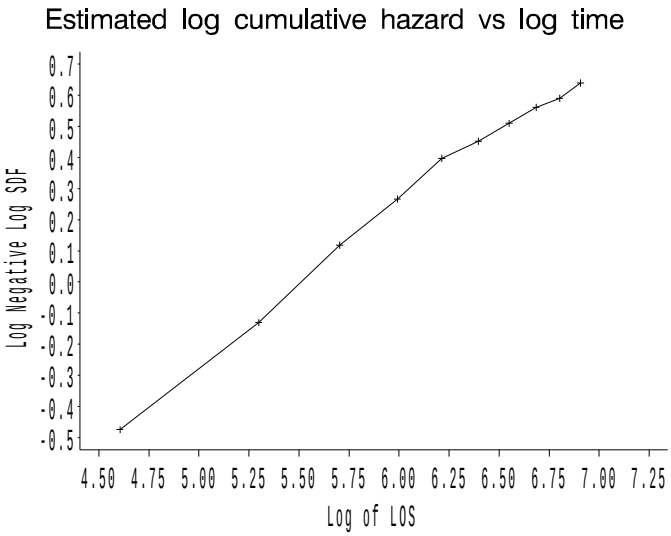


Figure 2: Nursing home data: $\log[-\log(\hat{S}(t))]$ vs $\log(t)$ for checking Weibull distribution

Likelihood for Right-censored and Left Truncated Data

- **Observed data:** (Q_i, X_i, δ_i), $i = 1, ..., n$

Without assuming a distribution for Q, the observed data likelihood is conditional upon the fact that $Q_i < T_i$ $(i = 1, ..., n)$.

For any $t > Q_i$,

$$S(t|Q_i) \equiv P(T > t|T > Q_i) = \frac{P(T > t)}{P(T > Q_i)} = \frac{S(t)}{S(Q_i)},$$

so

$$f(t|Q_i) = -\frac{d}{dt}S(t|Q_i) = \frac{-S'(t)}{S(Q_i)} = \frac{f(t)}{S(Q_i)}.$$

The likelihood is then

$$\begin{aligned}
\mathcal{L} &= \prod_i f(t|Q_i)^{\delta_i} S(t|Q_i)^{1-\delta_i} \\
&= \prod_i \left\{\frac{f(X_i)}{S(Q_i)}\right\}^{\delta_i} \left\{\frac{S(X_i)}{S(Q_i)}\right\}^{1-\delta_i} \\
&= \prod_{i=1}^{n} \frac{\lambda(X_i)^{\delta_i} \cdot S(X_i)}{S(Q_i)}
\end{aligned}$$

Exercise:

a) Derive the maximum likelihood estimate (MLE) and estimate its variance for the Weibull distribution with right-censored data.

b) Simulate a data set of $n = 100$ from the Weibull distribution with $\lambda = 1$, $\kappa = 2$, and uniform $(0, \tau)$ censoring. Play with values of τ so that about 20% of the observations are censored. Print out the data set and give a brief summary of it. Save your dataset for future use.

c) Find an R function/package that can calculate the MLE of the Weibull distribution [Hint: try mle() in package 'stats4', or you may write your own function], apply it to the data you generated in part b); be sure to provide the standard error of your estimate.