Màster universitari en Estadística i Investigació Operativa

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH — UNIVERSITAT DE BARCELONA

Lifetime Data Analysis, Course 2020/21

Exercises Topics 1 and 2: Solutions.

Exercise 1 ($0.6 + 0.6 + 0.8 = 2$ points)

It is said that the survival time T (in days) after a bone marrow transplantation follows a lognormal distribution with parameters $\rho = 0.042$ and $\tau = 2.084$.

(a) Which are the expected value and the median of T? What do you observe?

(b) Compute the probabilities that a person survives 100, 200, and 300 days after a bone marrow transplantation?

(c) Draw the hazard function of T and interpret its shape.

Solution:

(a) Expected value and median of T:

$$E(T) = \exp(-\ln(0.042) + \frac{1}{2} \cdot 2.084^2) = 208.85 \text{ (days)},$$

$$S(t_{0.5}) = 0.5 \Longleftrightarrow t_{0.5} = \exp\left(\Phi^{-1}(0.5) \cdot 2.084 - \ln(0.042)\right) = 23.81 \text{ (days)}.$$

The expected value is far larger than the median, which implies a right-skewed distribution.

(b) Survival probabilities for 100, 200, and 300 days, respectively:

$$S(100) = 1 - \Phi\Big(\frac{\ln(0.042 \cdot 100)}{2.084}\Big) = 1 - \Phi(0.6886) = 0.25,$$

$$S(200) = 1 - \Phi(1.0212) = 0.15,$$

$$S(300) = 1 - \Phi(1.2158) = 0.11.$$

(c) The hazard function of T, $\lambda(t) = f(t)/S(t)$, is shown in Figure 1. The instantaneous death risk is relatively high just after bone marrow transplantation and decreases thereafter. This explains the large difference between the mean and the median. The R code used to draw the figure is shown on page 6.

Exercise 2 ($0.8 + 0.8 + 0.4 = 2$ points)

(a) Given a random variable T with hazard function $\lambda(t)$, show that the following equality holds:

$$\mathrm{P}(T > t + s | T > t) = \exp\Big(-\int_t^{t+s} \lambda(u)\, du\Big).$$

(b) If $\lambda(t)$ is an increasing function and given $s > 0$, is the probability $\mathrm{P}(T > t + s | T > t)$ increasing or decreasing in t?

(c) What does the result imply for the mean residual lifetime of an 80 years old person compared to that of a 70 years old person?

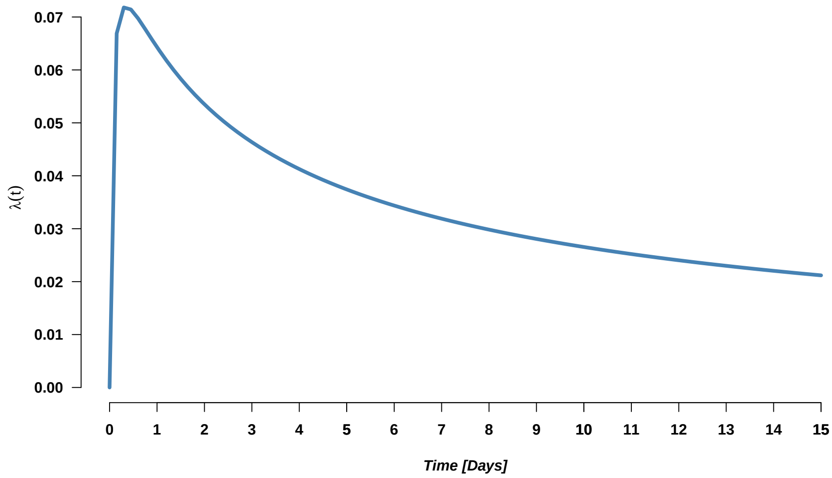

Figure 1: Hazard function of the lognormal distribution with parameters $\rho = 0.042$ and $\tau = 2.084$.

Solution:

(a)

$$\begin{aligned} P(T > t+s | T > t) &= \frac{P(T > t+s)}{P(T > t)} = \frac{S(t+s)}{S(t)} = \frac{\exp\big(-\Lambda(t+s)\big)}{\exp\big(-\Lambda(t)\big)} \\ &= \frac{\exp\big(-\int_0^{t+s}\lambda(u)\,du\big)}{\exp\big(-\int_0^{t}\lambda(u)\,du\big)} = \exp\Big(-\int_t^{t+s}\lambda(u)\,du\Big). \end{aligned}$$

(b) If λ is an increasing function, $P(T > t+s | T > t)$ is a **decreasing** function in t for each $s > 0$. One way to show this is the following:
The derivative of $f(t) = \exp\big(g(t)\big) = \exp\big(-\int_t^{t+s}\lambda(u)\,du\big)$ is

$$\begin{aligned} f'(t) &= \exp\big(g(t)\big)\cdot g'(t) = \exp\big(g(t)\big)\cdot(-1)\cdot\frac{\partial}{\partial t}\int_t^{t+s}\lambda(u)\,du \\ &= \underbrace{-\exp\big(g(t)\big)}_{<0}\cdot\underbrace{\big(\lambda(t+s)-\lambda(t)\big)}_{>0\ (s>0)} < 0. \end{aligned}$$

Hence, f is decreasing in t.

(c) The previous result implies that the mean residual lifetime of an 80 years old person is shorter than the one of a 70 years old person.

Exercise 3 ($0.6 + 0.6 + 0.8 = 2$ points)

Let T be the elapsed time between two Olympic records of a certain event in one of the Olympic sports. Since the Olympic Games are held only every four years, T is a discrete variable with the following values of its survival function:

t (Years)	$S(t)$	t (Years)	$S(t)$
$0 \le t < 4$	1.00	$20 \le t < 24$	0.26
$4 \le t < 8$	0.65	$24 \le t < 28$	0.15
$8 \le t < 12$	0.47	$28 \le t < 32$	0.10
$12 \le t < 16$	0.34	$32 \le t < 36$	0.06
$16 \le t < 20$	0.30	$36 \le t$	0.00

(a) Which is the probability density function of T?

(b) Which is the hazard function of T?

(c) Calculate the mean and the median residual lifetime of T at 8, 16, and 24 years.

Solution:

The values of the probability density function and the hazard function of T are calculated as follows:

(a)

$$p(t_j) = S(t_{j-1}) - S(t_j). \tag{1}$$

(b)

$$\lambda(t_j) = \frac{p(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}. \tag{2}$$

Applying formulas (1) and (2), the values in Table 1 are obtained.

In addition, for all $t \neq t_j,\ j = 0, 1, \ldots, 9$, the values of the probability density and the hazard functions are

$$p(t) = \lambda(t) = 0.$$

See Figure 2 for a graphical representation of both functions.

Table 1: Values of the survival, probability density, and hazard functions (Exercise 3).

t (Years)	$S(t)$	j	t_j	$p(t_j)$	$\lambda(t_j)$
$0 \le t < 4$	1.00	0	0	0.0	0.0
$4 \le t < 8$	0.65	1	4	0.35	0.35
$8 \le t < 12$	0.47	2	8	0.18	0.28
$12 \le t < 16$	0.34	3	12	0.13	0.28
$16 \le t < 20$	0.30	4	16	0.04	0.12
$20 \le t < 24$	0.26	5	20	0.04	0.13
$24 \le t < 28$	0.15	6	24	0.11	0.42
$28 \le t < 32$	0.1	7	28	0.05	0.33
$32 \le t < 36$	0.06	8	32	0.04	0.4
$36 \le t$	0.00	9	36	0.06	1.0

(c) The mean residual lifetime is obtained with the following formula:

$$\text{mrl}(t) = \frac{1}{S(t)}\Big((t_{i+1} - t)S(t_i) + \sum_{j \ge i+1} (t_{j+1} - t_j)S(t_j)\Big) \quad \text{for } t_i \le t < t_{i+1}.$$

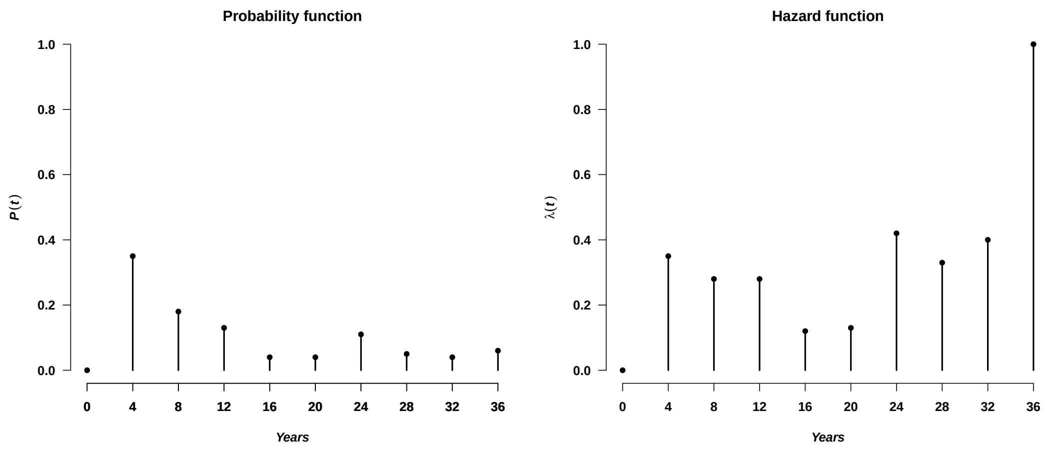

Figure 2: Probability and hazard functions corresponding to Exercise 3.

Hence,

$$
\begin{aligned}
\mathrm{mrl}(8) &= \frac{1}{S(8)} \cdot 4 \cdot (0.47 + \cdots + 0.06) = 14.3,\\
\mathrm{mrl}(16) &= 11.6,\\
\mathrm{mrl}(24) &= 8.3.
\end{aligned}
$$

The median residual lifetimes are:

$$
\begin{aligned}
&\mathrm{medrl}(8) = \ldots\\
&\frac{S(20)}{S(8)} = \frac{S(12+8)}{S(8)} = 0.55\\
&\frac{S(24)}{S(8)} = \frac{S(16+8)}{S(8)} = 0.32\\
\Longrightarrow\ &\mathrm{medrl}(8) = 16.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
\mathrm{medrl}(16) &= 8,\\
\mathrm{medrl}(24) &= 8.
\end{aligned}
$$

Exercise 4 ($5 \times 0.4 = 2$ points)

A study is carried out on the effects of a new drug for HIV-infected patients. The objective consists of studying the time (T) from the start of the therapy —patients are supposed to take the new drug daily— until the viral load falls below a certain threshold U. Four weeks after the first day of the therapy, patients start to visit a health care centre, where the viral load is measured, and from then once every week.

In each of the following situations, identify the type of censoring and indicate the time interval that contains T taking into account that T is measured in days. Which is the contribution to the likelihood function in each case?

(a) Eight weeks after the start of the therapy, the first patient visits the health care centre for the last time. By then, the viral load is still above U.

(b) In the case of another patient, the viral load is above U at the first visit, but below U at the second visit.

(c) One patient dies in week 7. It is known that before his death, the viral load was always above U.

(d) Patient 4 has bad luck: at her first visit, the viral load cannot be determined. After that, she decides not to return to the health care centre.

(e) The viral load of the last patient is below U already at the first visit.

Solution:

Following, T denotes the time (in days) from the start of the therapy until the viral load falls below the threshold U and L denotes the contribution to the likelihood function. All intervals are considered half-open, but they could be also considered closed intervals.

(a) **Right-censored** observation:

$$T \in (56, \infty) \quad \Longrightarrow \quad L = S(56).$$

(b) **Interval-censored** observation:

$$T \in (28, 35] \quad \Longrightarrow \quad L = S(28) - S(35).$$

(c) **Right-censored** observation:

$$T \in (42, \infty) \quad \Longrightarrow \quad L = S(42).$$

Right censoring would be informative, if the cause of death was related to T. In that case, we would deal with competing risks.

(d) Since no information on T is available, we deal with **missing data**. A right-censored observation at time 0 does not contribute any information since $S(0) = 1$.

(e) **Left-censored** observation:

$$T \in (0, 28] \quad \Longrightarrow \quad L = 1 - S(28).$$

Exercise 5 (0.8 + 1.2 = 2 points)

In a study on the onset of breast cancer in postmenopausal women, eight 50-years-old women were followed throughout a period of 10 years. The presence of breast cancer was checked yearly and the following (censored) times until cancer onset were observed:

$$(55, 56],\ (58, 59],\ (52, 53],\ (59, 60],\ (54, 55],\ \geq 60,\ \geq 60,\ \geq 60.$$

Defining T as the time from age 50 until cancer onset:

(a) Which is the expression of the likelihood function assuming that T follows a Weibull distribution with parameters ρ and k?

(b) Which is the maximum likelihood estimator of ρ if we assume $k = 1$ and substitute the censoring intervals by the corresponding midpoints?

Solution:

The likelihood function has the following expression:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{5} \big(S(L_i) - S(R_i)\big) \cdot \prod_{i=6}^{8} S(C_i),$$

where C_i denotes the right-censored times until breast cancer onset and $(L_i, R_i]$ the censoring interval in the case of an interval-censored observation. Hence:

$$L(\boldsymbol{\theta}) = \big(S(5) - S(6)\big)\big(S(8) - S(9)\big)\big(S(2) - S(3)\big)\big(S(9) - S(10)\big)\big(S(4) - S(5)\big)S(10)^3.$$

(a) If T follows a Weibull distribution with parameters ρ and k, its survival function is $S(t) = e^{-(\rho t)^k}$ and the expression of the likelihood function is:

$$L(\rho, k) = e^{-3(\rho\cdot 10)^k}\big(e^{-(\rho\cdot 5)^k} - e^{-(\rho\cdot 6)^k}\big)\cdot\cdots\cdot\big(e^{-(\rho\cdot 4)^k} - e^{-(\rho\cdot 5)^k}\big).$$

(b) Let now $m_i = \frac{1}{2}(l_i + r_i)$, $i = 1, \ldots, 5$ be the interval midpoints and $k = 1$. Hence, the expressions of the likelihood and its natural logarithm are the following:

$$L(\rho) = \prod_{i=1}^{5} f(m_i)\prod_{i=6}^{8} S(c_i) = \prod_{i=1}^{5} \rho e^{-\rho\cdot m_i}\prod_{i=6}^{8} e^{-\rho\cdot c_i} = \rho^5\prod_{i=1}^{5} e^{-\rho\cdot m_i}\prod_{i=6}^{8} e^{-\rho\cdot c_i},$$

$$l(\rho) = 5\ln(\rho) - \sum_{i=1}^{5}\rho\cdot m_i - \sum_{i=6}^{8}\rho\cdot c_i = 5\ln(\rho) - \rho\Big(\sum_{i=1}^{5} m_i + \sum_{i=6}^{8} c_i\Big).$$

The first and second derivatives of $l(\rho)$ are

$$l'(\rho) = \frac{\partial}{\partial\rho}l(\rho) = \frac{5}{\rho} - \Big(\sum_{i=1}^{5} m_i + \sum_{i=6}^{8} c_i\Big),$$

$$l''(\rho) = -\frac{5}{\rho^2} < 0,$$

and, hence, the maximum likelihood estimator of ρ is

$$\hat{\rho}_{\mathrm{ML}} = \frac{5}{\sum_{i=1}^{5} m_i + \sum_{i=6}^{8} c_i} = \frac{5}{30.5 + 30} = 0.083.$$

Appendix: R Code of Figure 1

Exercise 1

```
## Figure 1
## --------
windows(width = 10, height = 7)
par(las = 1, font = 2, font.axis = 2, font.lab = 4, mar = c(5, 5, 2, 2),
    bty = "n")
curve(dlnorm(x, -log(0.042), 2.084) / plnorm(x, -log(0.042), 2.084,
                                             lower.tail = FALSE),
      from = 0, to = 15, lwd = 4, col = "steelblue", xlab = "Time [Days]",
      ylab = expression(lambda(t)))
axis(1, at  =  1:15)

# Alternative using the function hlnorm() of the eha package
# ----------------------------------------------------------
library(eha)
par(las = 1, font = 2, font.axis = 2, font.lab = 4, mar = c(5, 5, 2, 2),
    bty = "n")
curve(hlnorm(x, -log(0.042), 2.084), from = 0, to = 15, lwd = 4,
      col = "steelblue", xlab = "Time [Days]",
      ylab = expression(lambda(t)))
axis(1, at  =  1:15)
```