Modelos probabilísticos y estadísticos en fiabilidad

Miriam Jarauta Baigorri

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: Javier López Lorente
28 de junio de 2019

Abstract

Reliability theory, understood as the discipline which studies the probability that a system works satisfactorily, is a discipline with many applications, especially in the fields of Engineering and Medicine. Its mathematical development is based fundamentally on probability and statistical techniques, the study of which is the aim of this work.

We begin with a short summary of the history of reliability (Introduction). In Chapter 2 we consider the static performance of a system and its components, that is, we consider that they can be either in operation or in failure, without analysing their evolution over time. First, we define the structure function and the reliability function, which indicate the state and the probability of functioning of the system from the state and the reliability of its components. Then we analyze their properties and some techniques for their computation. We will also see some examples of the most commonly used systems: series, parallel, k-out-of-n.

In Chapter 3 we consider that the components of the system develop over time, so that they start operating at $t = 0$ and fail at a random time T. Throughout the chapter we study several functions of the lifetime distribution of an item.

- Survival function, $S(t)$, which represents the probability that an item survives beyond time t.
- Hazard function, $h(t)$, which determines the failure rate of an item in the interval $[t, t+\Delta t)$ with small $\Delta t \geq 0$, given that it has not failed before.
- Cumulative hazard function, $H(t)$, which indicates the accumulated "risk" until time t.
- Mean residual life function, $L(t)$, which calculates the expected remaining life of an item that has already survived until time t.

Based on the previous functions we will define some of the most used distribution classes in reliability. These classes are: increasing failure rate (IFR), decreasing failure rate (DFR), bathtub-shaped failure rate (BF), increasing failure rate on the average (IFRA), decreasing failure rate on the average (DFRA), increasing mean residual life (IMRL) and decreasing mean residual life (DMRL). We then explain three of the most used distributions in reliability: exponential, Weibull and geometric. In the last part of this chapter we make an introduction to the competing risks model, which studies, for an item, the probability to survive an instant t when it is subject to different risks that can cause its failure.

Finally, in Chapter 4 we focus on the most important statistical techniques used in reliability. In this chapter we consider censored data sets, in other words, sets in which not all lifetimes are known and, for some of them, we only have an upper and/or lower limit. We will analyze the most two common types of censoring: type I and type II. In the first one, all the items are observed, at most, until a specified time. In the second one, items are observed until the failure of a predetermined number of them. In the first part of the chapter we show how to make parametric inference with censored data. Taking the exponential distribution as an example, we will see how to perform point and interval estimation for its parameter. Later we consider nonparametric estimation of the survival function. For this we will define the Kaplan-Meier estimator and see how it can be used to find confidence intervals.

Índice general

Capítulo 1

Introducción

La fiabilidad de un sistema se define como la probabilidad que tiene de cumplir adecuadamente su propósito durante un período de tiempo determinado y unas condiciones ambientales específicas [10]. En sus orígenes fue utilizada para evaluar la mortalidad derivada de las epidemias. También fue usada por las compañías de seguros para determinar los riesgos de sus pólizas de seguro de vida.

Antes de la Segunda Guerra Mundial, el concepto de fiabilidad era en gran medida intuitivo, subjetivo y cualitativo. El uso de métodos actuariales tanto para estimar la supervivencia de pacientes sometidos a distintos tratamientos como para estudiar la fiabilidad de sistemas comenzó a principios del siglo XX. Un enfoque más matemático y formal de la fiabilidad surgió de las demandas de la tecnología moderna y, en particular, de las necesidades de la Segunda Guerra Mundial, debido a la aparición de sistemas militares complejos [2].

En 1939, Waloddi Weibull propuso una distribución para describir la resistencia a la rotura de materiales, que más tarde llevaría su nombre [14]. En este artículo, que tiene más de 10000 citas en Google Scholar [6], considera distribuciones de la forma $F(x) = 1 - e^{-(x-x_u)^m/x_0}$, argumentando que esta es la expresión más simple que permite expresar la supervivencia de un sistema sujeto a fallos de múltiples causas. Aunque reconoce la ausencia de una base teórica sólida para esta distribución, afirma que *"la experiencia ha demostrado que, en muchos casos, se ajusta a los datos mejor que otras distribuciones conocidas"*.

En 1951, Epstein y Sobel empezaron a trabajar con la distribución exponencial como modelo probabilístico para estimar el tiempo de vida de dispositivos [5]. Una razón fundamental de la popularidad de la distribución exponencial, además de su simplicidad, es que corresponde a los tiempos entre ocurrencias en procesos de Poisson. Aunque desde un principio se vio que la pérdida de memoria de la distribución exponencial no es razonable en muchos tipos de sistemas, se sigue estudiando dentro de la teoría de la fiabilidad, entre otros motivos, porque puede servir como una primera aproximación a los datos y como base para modelos más complejos. [12]

Un hito en el desarrollo de la fiabilidad fue el establecimiento en 1951 del Rome Air Development Center en Rome (estado de Nueva York, EEUU), uno de cuyos objetivos era realizar estudios de fiabilidad. Durante más de dos décadas, este centro fue el pionero de la investigación en fiabilidad. A partir de los años 80, con la proliferación de aparatos electrónicos, programas de software,... el estudio de la fiabilidad adquirió más importancia y se establecieron un gran número de centros de investigación dedicados a esta disciplina. [13]

Además de la ingeniería, el otro gran campo en el que se aplica la fiabilidad es en las Ciencias Actuariales y la Medicina. Bajo esta perspectiva, el objetivo principal es la estimación de la función de supervivencia, es decir, la probabilidad de que el tiempo de vida de un individuo sea superior a una

cantidad dada. A partir de ella se pueden construir tablas de mortalidad, calcular la esperanza de vida o el tiempo medio hasta la recidiva de una enfermedad. La primera tabla de mortalidad fue construida en 1663 por John Graunt (ver, por ejemplo, [7]), pero el desarrollo matemático más importante comenzó después de la publicación por Edward L. Kaplan y Paul Meier del estimador que lleva su nombre en [8] (artículo con más de 55000 citas en Google Scholar [6]). Este estimador permite estimar la función de supervivencia de forma no paramétrica en muestras completas y en muestras censuradas, en las que no se conocen todos los tiempos de vida de los artículos de la muestra. Otro hito importante fue la introducción del modelo de Cox de riesgos proporcionales [3] en 1972, que permite la incorporación de covariables para estimar la función de supervivencia. A partir de entonces ha habido un gran número de investigadores dedicados a la mejora de estos modelos.
Desde finales de los años ochenta y principios de los noventa, el análisis de supervivencia se estableció como el método estadístico estándar en la investigación biomédica. En las Facultades de Medicina, el análisis de supervivencia forma parte esencial de los planes de estudios de bioestadística, y es ampliamente utilizado en el análisis de datos en importantes revistas médicas como la JAMA y la revista New England Journal of Medicine. [11]

Nuestro objetivo en el presente trabajo será estudiar las diferentes técnicas de probabilidad y estadística que se usan en fiabilidad. En el capítulo 2 estudiaremos los sistemas independientemente del tiempo. Veremos cómo es su composición, su función estructural y de fiabilidad, los tipos que hay y las técnicas utilizadas para calcularlas. En el capítulo 3 estudiaremos la variable aleatoria que representa el tiempo de vida, así como las diferentes características de las distribuciones que son importantes en fiabilidad. Además haremos una breve introducción de los modelos de riesgos competitivos. Por último, dedicaremos el capítulo 4 a hacer una introducción a los métodos estadísticos usados en fiabilidad, analizando cómo a través de los datos (que pueden estar censurados) se puede hacer estimación de los parámetros o de las funciones de supervivencia.

Debido a las limitaciones de tiempo y espacio, en el trabajo no hemos incluido algunos temas importantes dentro de la fiabilidad. Entre otros, cabe destacar los modelos de vida acelerada, en los que el tiempo hasta el fallo de las componentes depende de covariables, o los sistemas reparables, en los que cuando una componente falla puede ser reparada y para cuyo análisis se utiliza la teoría de procesos estocásticos. Durante la preparación del trabajo se han consultado un alto número de artículos y libros sobre fiabilidad. Los libros que se han usado de forma más continuada han sido las referencias [2], [9], [10].

Capítulo 2

Sistemas coherentes

A lo largo de este capítulo, estudiaremos las funciones estructurales de los diferentes tipos de sistemas más comunes, definiremos la función de fiabilidad y comentaremos los distintos métodos utilizados para calcularla. También analizaremos la importancia que tiene cada una de las componentes en el sistema.

2.1. Funciones estructurales

Un sistema es un conjunto de n componentes. Estas componentes pueden estar en uno de los dos estados: funcionamiento o fallo. En este capítulo consideramos el comportamiento estático de las componentes y el sistema; es decir, no analizamos su evolución en el tiempo. Para describir la forma en la que las componentes están relacionadas dentro del sistema utilizaremos la función estructural que define el estado del sistema como una función del estado de las n componentes.

Definición. El estado de un sistema de n componentes viene dado por $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ con $x_i \in \{0,1\}$ para $i = 1, \ldots, n$. El valor x_i representa el estado de la componente i:

$$x_i = \begin{cases} 0 & \text{si la componente i no funciona} \\ 1 & \text{si la componente i funciona} \end{cases}$$

Puesto que todo sistema de n componentes viene dado por su función estructural, pasaremos a definir esta función.

La función estrutural es $\phi : \{0,1\}^n \to \{0,1\}$ y se interpreta como

$$\phi(\mathbf{x}) = \begin{cases} 0 & \text{si el sistema está en estado de fallo} \\ 1 & \text{si el sistema está en estado de funcionamiento} \end{cases}$$

Una vez definida la función estructural asociada a un sistema ya tenemos los conocimientos suficientes para podemos definir y analizar las propiedades de los sistemas coherentes.

Definición. Dado un sistema con función estructura ϕ, se dice que la componente i es irrelevante si $\phi(1_i, \mathbf{x}) = \phi(0_i, \mathbf{x})\ \forall \mathbf{x} \in \{0,1\}^n$, siendo $(1_i, \mathbf{x})$ y $(0_i, \mathbf{x})$ los vectores estado del sistema donde la única diferencia entre ellos se encuentra en el estado de la componente i, con $x_i = 1$ y $x_i = 0$ respectivamente.

En otras palabras, una componente se llama irrelevante si su estado no influye en el estado del sistema. En la siguiente definición usaremos el orden coordenada a coordenada en $\{0,1\}^n$. Es decir, la función ϕ es no decreciente si $\phi(\mathbf{x}) \leq \phi(\mathbf{y})\ \forall \mathbf{x}, \mathbf{y}$ tal que $x_i \leq y_i,\ i = 1, \ldots, n$.

Definición. Un sistema es coherente si ϕ es no decreciente y no hay componentes irrelevantes.

Algunas propiedades de un sistema coherente son:

1. $\phi(\mathbf{0}) = 0$ y $\phi(\mathbf{1}) = 1$
2. $\prod_{i=1}^{n} x_i \leq \phi(\mathbf{x}) \leq 1 - \prod_{i=1}^{n}(1-x_i) \quad \forall\, \mathbf{x} \in \{0,1\}^n$

Entre los sistemas coherentes, se encuentran los sistemas en serie y los sistemas en paralelo, los cuales serán definidos en la siguiente subsección.

2.1.1. Tipos de sistemas

En esta subsección, describiremos los distintos tipos de sistemas más comunes asociando a cada uno de ellos su función estructural.

- **Sistema en serie**: El sistema funciona cuando todas sus componentes funcionan.
  Por lo tanto,

$$\phi(\mathbf{x}) = \begin{cases} 0 & \text{si existe algún i tal que } x_i = 0 \\ 1 & \text{si } x_i = 1 \quad \forall i = 1,...,n \end{cases} = \text{mín}\{x_1, x_2, ..., x_n\} = \prod_{i=1}^{n} x_i$$

  Un ejemplo de sistema en serie podría ser nuestro sistema cardiovascular, el cual está formado por el corazón, las arterias y las venas. El fallo en el funcionamiento de alguna de sus partes conlleva el fallo del sistema.

- **Sistema en paralelo**: El sistema funciona cuando al menos una de sus componentes funciona. Por tanto,

$$\phi(\mathbf{x}) = \begin{cases} 0 & \text{si } \ x_i = 0 \quad \forall i = 1,...,n \\ 1 & \text{si existe algun i tal que } x_i = 1 \end{cases} = \text{máx}\{x_1, x_2, ..., x_n\} = 1 - \prod_{i=1}^{n}(1-x_i)$$

  Un ejemplo de sistema en paralelo podrían ser nuestros riñones, cuando uno de los dos falla la persona puede seguir viviendo normalmente con un único riñon.

- **Sistema "k-out-of-n"**: El sistema funciona si al menos k de sus n componentes funcionan. Por tanto,

$$\phi(\mathbf{x}) = \begin{cases} 0 & \text{si } \sum_{i=1}^{n} x_i < k \\ 1 & \text{si} \sum_{i=1}^{n} x_i \geq k \end{cases}$$

  Considerando un caso sencillo, por ejemplo, cuando $n = 3$ y $k = 2$, la función estructural viene expresada como:

$$\phi(\mathbf{x}) = 1 - (1 - x_1x_2)(1 - x_1x_3)(1 - x_2x_3)$$

  Un ejemplo de este sistema sería un puente colgante, el cual solo necesita k de los n cables disponibles para sujetarlo.

Como se puede apreciar, los sistemas en serie y en paralelo son casos especiales del sistema "k-out-of-n" y ambos sistemas se pueden combinar para dar un sistema más complejo.

2.2. Funciones de fiabilidad

En esta sección nos centraremos en definir la función de fiabilidad y analizar las diferentes técnicas utilizadas para calcularla. La fiabilidad de un sistema es la probabilidad de que el sistema funcione. Para hallarla, supondremos que cada componente es aleatoria y funciona o no con una cierta probabilidad, de forma independiente al resto de componentes.

Definición. La variable aleatoria que indica el estado de la componente i, X_i, es

$$X_i = \begin{cases} 0 & \text{si la componente i no funciona} \\ 1 & \text{si la componente i funciona} \end{cases}$$

para $i = 1, ..., n$.

Entonces, $\mathbf{X} = (X_1, X_2, ..., X_n)$, denominado el vector estado aleatorio del sistema, denotará el estado de las n componentes.

Una vez definidos los valores que puede tomar el estado de una componente nos centraremos en el concepto de fiabilidad.

Definición. El vector de fiabilidad de un sistema de n componentes viene dado por $\mathbf{p} = (p_1, p_2, ..., p_n)$ con $p_i = P[X_i = 1]$ para $i = 1, 2, ..., n$. La función de fiabilidad es una función $r : [0,1]^n \to [0,1]$ definida por $r(\mathbf{p}) = P[\phi(\mathbf{X}) = 1]$.

En otras palabras, la fiabilidad p_i es la probabilidad de que la componente i funcione mientras que la función $r(\mathbf{p})$ es la probabilidad de que el sistema funcione (fiabilidad del sistema).

2.2.1. Técnicas para calcular la fiabilidad de un sistema

Una vez dada la función de fiabilidad, introduciremos varias técnicas que se pueden utilizar para calcularla.

Definición de $r(\mathbf{p})$

Esta técnica es útil cuando se conoce la función estructural del sistema y esta función no es muy compleja, por ejemplo, para los sistemas en serie o en paralelo.

Ejemplo 1. Calculamos la fiabilidad de un sistema en paralelo de n componentes usando la definición de $r(\mathbf{p})$ y la suposición de independencia.

$$r(\mathbf{p}) = P[\phi(\mathbf{X}) = 1] = P[1 - \prod_{i=1}^{n}(1 - X_i) = 1] = 1 - \prod_{i=1}^{n}(1 - p_i)$$

Valor esperado de $\phi(\mathbf{X})$

Esta técnica se basa en el hecho de que $r(\mathbf{p}) = P[\phi(\mathbf{X}) = 1] = E[\phi(\mathbf{X})]$ ya que $\phi(\mathbf{X})$ es una v.a. Bernoulli.

Ejemplo 2. Calculamos la fiabilidad del sistema de la Figura 2.1 usando la esperanza y la suposición de independencia. Como la función estructural de este sistema es

$$\phi(\mathbf{X}) = X_1[1 - (1 - (1 - (1 - X_2)(1 - X_3)))(1 - X_4)]X_5$$

entonces

$$\begin{aligned} r(\mathbf{p}) &= E[\phi(\mathbf{X})] = E[X_1(1 - (1 - (1 - (1 - X_2)(1 - X_3)))(1 - X_4))X_5] \\ &= E[X_1]\Big(E[X_4] + E[X_3] - E[X_3]E[X_4] + E[X_2] - E[X_2]E[X_4] - E[X_2]E[X_3] + E[X_2]E[X_3]E[X_4]\Big)E[X_5] \\ &= p_1p_2p_5 + p_1p_3p_5 + p_1p_4p_5 - p_1p_2p_3p_5 - p_1p_3p_4p_5 - p_1p_2p_4p_5 + p_1p_2p_3p_4p_5 \end{aligned}$$

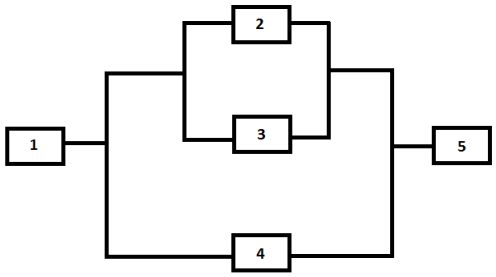


Figura 2.1: Sistema de cinco componentes

Técnica del vector de trayectoria

En este apartado, definiremos los vectores de trayectoria y explicaremos cómo se utilizan estos vectores en el cálculo de la fiabilidad.

Definición. Dado $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$. Se dice que $\mathbf{x} < \mathbf{y}$ si $x_i \leq y_i \ \forall i$, y $x_i < y_i$ para algún i.

Definición. Sea ϕ la función estructural de un sistema coherente de n componentes. Un vector estado $\mathbf{x}$ es un vector de trayectoria si $\phi(\mathbf{x}) = 1$.

En otras palabras, todo vector estado que conlleva el funcionamiento del sistema es un vector de trayectoria. Una vez definido el vector de trayectoria pasamos a explicar en que consiste la técnica del vector de trayectoria. Esta técnica únicamente suma las probabilidades correspondientes a los vectores de trayectoria del sistema. Así, la fiabilidad del sistema es

$$r(\mathbf{p}) = P[\mathbf{X} \text{ es un vector de trayectoria}]$$

Ejemplo 3. Consideramos el sistema de la Figura 2.1. Este sistema tiene siete vectores de trayectoria:

$$(1,1,0,0,1),(1,1,1,1,1),(1,1,1,0,1),(1,1,0,1,1)(1,0,1,0,1),(1,0,1,1,1) \text{ y } (1,0,0,1,1).$$

Entonces, la fiabilidad de dicho sistema puede calcularse usando la técnica del vector de trayectoria:

$$\begin{aligned} r(\mathbf{p}) &= p_1p_2(1-p_3)(1-p_4)p_5 + p_1p_2p_3p_4p_5 + p_1p_2p_3(1-p_4)p_5 + p_1p_2(1-p_3)p_4p_5 \\ &\quad + p_1(1-p_2)p_3(1-p_4)p_5 + p_1(1-p_2)p_3p_4p_5 + p_1(1-p_2)(1-p_3)p_4p_5 \\ &= p_1p_2p_5 + p_1p_3p_5 + p_1p_4p_5 - p_1p_2p_3p_5 - p_1p_3p_4p_5 - p_1p_2p_4p_5 + p_1p_2p_3p_4p_5 \end{aligned}$$

Técnica del vector de corte

En este apartado, definiremos los vectores de corte y explicaremos cómo se utilizan estos vectores en el cálculo de la fiabilidad.

Definición. Sea ϕ la función estructural de un sistema coherente de n componentes. Un vector estado $\mathbf{x}$ es un vector de corte si $\phi(\mathbf{x}) = 0$.

En otras palabras, todo vector estado que provoca el fallo del sistema es un vector de corte. Una vez definido el vector de corte pasamos a explicar en que consiste la técnica del vector de corte, la cual es análoga a la técnica anterior. Para esta técnica la fiabilidad del sistema es

$$r(\mathbf{p}) = 1 - P[\mathbf{X} \text{ es un vector de corte}]$$

Ejemplo 4. Consideramos el sistema de la Figura 2.2. Este sistema tiene tres vectores de corte:

$$(0,0,0,0),(0,1,0,0) \text{ y } (1,0,0,0)$$

Entonces, aplicando la técnica del vector de corte obtenemos que:

$$\begin{aligned} r(\mathbf{p}) &= 1-[(1-p_1)(1-p_2)(1-p_3)(1-p_4)+(1-p_1)p_2(1-p_3)(1-p_4)+p_1(1-p_2)(1-p_3)(1-p_4)] \\ &= p_4+p_3-p_3p_4+p_1p_2-p_1p_2p_4-p_1p_2p_3+p_1p_2p_3p_4 \end{aligned} \tag{2.1}$$

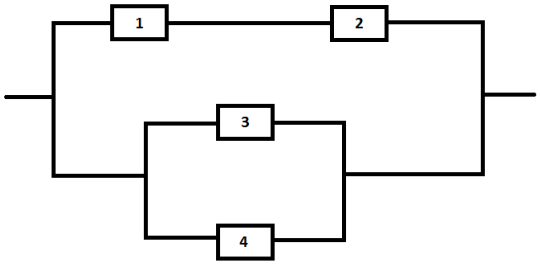


Figura 2.2: Sistema de cuatro componentes

Cabe destacar que la suma de las probabilidades de los vectores de corte nos da la probabilidad que tiene el sistema de fallar y la suma de las probabilidades de los vectores de trayectoria nos da la probabilidad que tiene el sistema de no fallar. Por lo tanto, la suma de ambas probabilidades debe sumar 1. Estos dos últimos métodos son poco eficaces cuando el número de componentes en el sistema aumenta.

Descomposición

Para usar el método de descomposición hay que identificar una componente clave en el sistema y condicionar el estado del sistema al estado de esa componente clave. Usando ese argumento de condición, la fiabilidad del sistema es

$$\begin{aligned} r(\mathbf{p}) &= \text{P[el sistema funciona | la componente clave funciona]} \cdot \text{P[la componente clave funciona]} \\ &\quad + \text{ P[sistema funciona | la componente clave no funciona]} \cdot \text{P[componente clave no funciona]} \\ &= P[\text{sistema A funciona}]\ P[\text{la componente clave funciona}] \\ &\quad + P[\text{sistema B funciona}]\ P[\text{la componente clave no funciona}] \end{aligned}$$

donde el sistema A, es el sistema en estudio con la componente clave sustituida por una componente perfecta y el sistema B es el sistema en estudio con la componente clave reemplazada por una componente que no funciona.

Esta expresión también puede reescribir como

$$r(\mathbf{p}) = r(1_i,\mathbf{p})p_i + r(0_i,\mathbf{p})(1-p_i),$$

donde i es la componente clave y $r(\alpha,\mathbf{p})$ denota la función de fiabilidad cuando la componente i tiene fiabilidad α y las otras componentes tienen fiabilidades $p_1,p_2,...,p_{i-1},p_{i+1},...,p_n$. Cualquier componente puede ser usada como la componente clave, aunque es más fácil elegir una que ocupe una posición destacada en el sistema.

Ejemplo 5. Considerando el sistema de ocho componentes cuyo diagrama de bloques se muestra en la Figura 2.3 y usando como componente clave la componente 2, los sistemas A y B que se obtiene son

los que aparecen en la Figura 2.4. Usando la fórmula de descomposición, la fiabilidad del sistema es

$$\begin{aligned} r(\mathbf{p}) &= P[\text{sistema A funciona}] \cdot P[\text{la componente 2 funciona}] \\ &\quad + P[\text{sistema B funciona}] \cdot P[\text{la componente 2 no funciona}] \\ &= \big(p_1[1-(1-(1-(1-p_4)(1-p_5))p_6)(1-p_3p_7)]p_8\big) \cdot p_2 + (p_1p_3p_7p_8) \cdot (1-p_2) \\ &= p_1p_3p_7p_8 + p_1p_2p_4p_6p_8 + p_1p_2p_5p_6p_8 - p_1p_2p_4p_5p_6p_8 \\ &\quad - p_1p_2p_3p_4p_6p_7p_8 - p_1p_2p_3p_5p_6p_7p_8 + p_1p_2p_3p_4p_5p_6p_7p_8 \end{aligned} \tag{2.2}$$

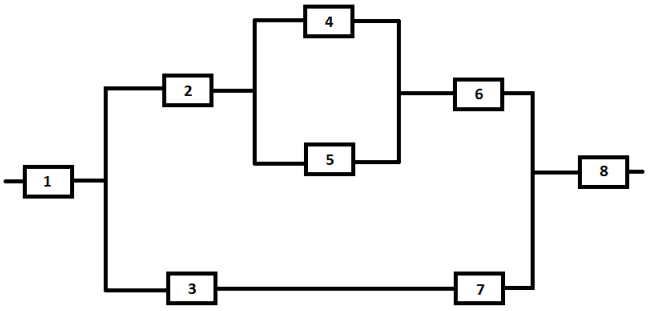
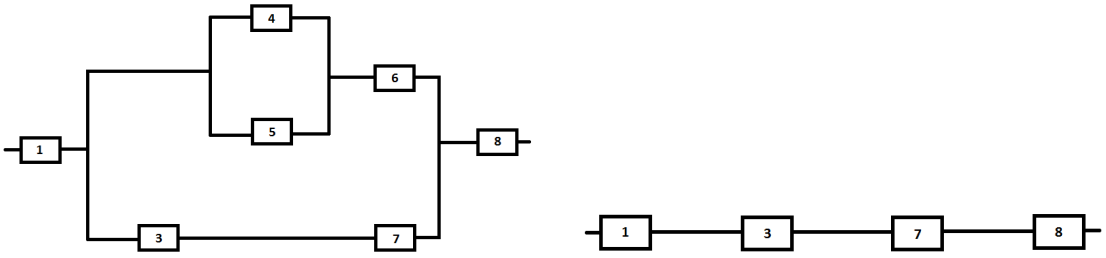


Figura 2.3: Sistema de ocho componentes



(a) Sistema A (b) Sistema B

Figura 2.4: Sistemas A y B condicionados al estado de la componente 2

2.3. Importancia de las componentes del sistema

Como hemos estudiado en las secciones anteriores, tanto la función estructural de un sistema como su fiabilidad dependen del estado y fiabilidad de sus componentes. Como es de esperar, los sistemas formados por más de una componente poseen componentes con mayor importancia que otras. En esta sección nos centraremos en medidas que calculen la importancia de cada componente en el rendimiento global del sistema.

2.3.1. Importancia estructural de las componentes

Para un sistema en serie, todas las componentes son igual de importantes, ya que el fallo de alguna de ellas provoca el fallo del sistema. Para sistemas con estructuras más complejas, [1] propone una medida para evaluar la importancia estructural de una componente.

Definición. La importancia estructural de la componente i en un sistema coherente de n componentes es

$$I_\phi(i) = \frac{1}{2^{n-1}} \sum_{\{\mathbf{x}|x_i=1\}} [\phi(1_i,\mathbf{x}) - \phi(0_i,\mathbf{x})]$$

para i=1,...,n.

El sumatorio recorre todos los vectores estado $\mathbf{x}$ tales que $x_i = 1$ y cuenta el número de esos vectores para los que $\phi(1_i, \mathbf{x}) - \phi(0_i, \mathbf{x}) = 1$, es decir, cuenta únicamente los vectores estado tales que el fallo en la componente i resulta en el fallo del sistema. Así, para cualquier sistema coherente, $0 < I_\phi(i) \leq 1$ para i=1,...,n.

Ejemplo 6. Calcular la importancia estructural de todas las componentes del sistema representado en la Figura 2.2.

Puesto que el sistema está formado por 4 componentes, hay 16 posibles vectores estado pero el sumatorio usado en calcular la importancia estructural de la componente i únicamente recorrerá 8 de esos vectores, aquellos en los que $x_i = 1$. Para hallar la importancia estructural de la componente 1, $I_\phi(1)$, el sumatorio recorrerá los vectores estado $(1,0,0,0), (1,0,0,1), (1,0,1,0), (1,1,0,0), (1,0,1,1), (1,1,0,1)$, $(1,1,1,0)$ y $(1,1,1,1)$. Por lo tanto,

$$I_\phi(1) = \frac{1}{8}[(0-0)+(1-1)+(1-1)+(1-0)+(1-1)+(1-1)+(1-1)+(1-1)] = \frac{1}{8}.$$

Análogamente, obtenemos: $I_\phi(2) = 1/8$, $I_\phi(3) = 3/8$, $I_\phi(4) = 3/8$.

2.3.2. Importancia de la fiabilidad de las componentes

Mientras la importancia estructural indicaba la importancia de una componente en el sistema debido a la posición que ocupaba en él, la importancia de la fiabilidad combina la posición y la fiabilidad para indicar la importancia de cada componente en la fiabilidad del sistema.

Definición. La importancia de la fiabilidad de la componente i en un sistema coherente de n componentes es

$$I_r(i) = \frac{\partial r(\mathbf{p})}{\partial p_i}$$

para i=1,...,n.

La expresión anterior es función de $\mathbf{p} = (p_1, \ldots, p_n)$. Para un valor $\mathbf{p}$ determinado, la componente con la $I_r(i)$ más grande es la componente que provocará el mayor incremento en la fiabilidad del sistema cuando aumente su fiabilidad. La importancia de la fiabilidad de la componente i satisface $0 < I_r(i) < 1$ para i=1,...,n.

Ejemplo 7. Calcular la importancia de la fiabilidad de todas las componentes del sistema representado en la Figura 2.2 cuando $\mathbf{p}$ =(0.5, 0.2, 0.7, 0.1).

Tomando como función de fiabilidad la función (2.1) obtenemos que, en el caso de la componente 1:

$$I_r(1) = \frac{\partial r(\mathbf{p})}{\partial p_1} = p_2 - p_2p_4 - p_2p_3 + p_2p_3p_4 = 0.054$$

Análogamente, $I_r(2)$ =0.135, $I_r(3)$ =0.81 e $I_r(4)$ =0.27.

Capítulo 3

Distribuciones de tiempos de vida

Hasta ahora no habíamos tenido en cuenta el tiempo en el estudio de los sistemas. El estado de un sistema era determinado a partir del estado de cada una de sus componentes y la posición que ocupaban estas en el sistema. En este capítulo las componentes evolucionarán en el tiempo, de forma que comenzarán en funcionamiento en el tiempo 0 y fallarán en un tiempo aleatorio T. A lo largo del capítulo supondremos que el tiempo T es no negativo con función de distribución $F(t) = P[T \leq t]$, $\forall t \in \mathbb{R}$. En el caso de que T sea absolutamente continua denotaremos f a su función de densidad y en el caso discreto, con valores en $t_1, t_2, \ldots$ con $0 \leq t_1 < t_2 < \cdots$, $p(t_j) = P[T = t_j]$, $j = 1, 2, \ldots$ será su función de masa de probabilidad.

3.1. Distribuciones de tiempos de fallo

En esta sección introduciremos cuatro funciones relacionadas con la distribución del tiempo de vida de un artículo. Definiremos estas funciones tanto para distribuciones continuas como para distribuciones discretas.

3.1.1. Función de supervivencia $S(t)$

La función de supervivencia, también conocida como función de fiabilidad, es la probabilidad que tiene un artículo de sobrevivir al instante t:

$$S(t) = P[T > t] = 1 - P[T \leq t] = 1 - F(t) \quad t \geq 0.$$

Toda función de supervivencia satisface las siguientes condiciones:

$$S(0) = 1 \qquad \lim_{t \to +\infty} S(t) = 0 \qquad S(t) \text{ es no creciente y continua por la derecha.}$$

Cabe destacar que cuando disponemos de una población muy grande cuyos artículos tienen distribuciones de tiempos de vida idénticas, $S(t)$ puede ser interpretada como la fracción de artículos de la población que se espera que sobrevivan al instante t.

La función de supervivencia condicional, $S_{T|T>a}(t)$, es la función de supervivencia de un artículo que ha sobrevivido al tiempo a; es decir,

$$S_{T|T>a}(t) = P[T > t \,|\, T > a] = \frac{P[T > t,\, T > a]}{P[T > a]} = \frac{P[T > t]}{P[T > a]} = \frac{S(t)}{S(a)} \quad t \geq a.$$

En el caso en el que la v.a. T sea continua, la función de supervivencia viene definida por la siguiente expresión:

$$S(t) = \int_t^{+\infty} f(x)\, dx$$

En el caso en el que la v.a. T sea discreta, la función de supervivencia, función escalonada no creciente, viene dada por

$$S(t) = P[T > t] = \sum_{j|t_j>t} p(t_j) \quad t \geq 0.$$

3.1.2. Función de riesgo $h(t)$

La función de riesgo o tasa de fallo es la más utilizada para analizar el tiempo de supervivencia de un artículo. En el caso discreto se define como la probabilidad de que el fallo ocurra en un instante dado que no ha ocurrido antes. Esto es

$$h(t_j) = P[T = t_j | T > t_{j-1}] = \frac{P[T = t_j]}{P[T > t_{j-1}]} = \frac{p(t_j)}{S(t_{j-1})} \quad j = 1,2,\ldots$$

En el caso continuo se define como

$$h(t) = \frac{f(t)}{S(t)} \quad t \geq 0.$$

Puede interpretarse como la densidad de probabilidad de fallo en el tiempo t dado que no ha fallado antes. En efecto, considerando la probabilidad de fallo entre los tiempos t y $t+\Delta t$:

$$P[t < T \leq t+\Delta t] = \int_t^{t+\Delta t} f(\tau)\, d\tau = S(t) - S(t+\Delta t)$$

y condicionándola al hecho de que el artículo estaba en funcionamiento en el momento t tenemos que

$$P[t < T \leq t+\Delta t | T > t] = \frac{P[t < T \leq t+\Delta t]}{P[T > t]} = \frac{S(t) - S(t+\Delta t)}{S(t)}.$$

Como la longitud del intervalo $[t, t+\Delta t]$ es Δt, al dividir por Δt la expresión anterior, obtenemos que la tasa media de fallo en ese intervalo es:

$$\frac{S(t) - S(t+\Delta t)}{S(t)\Delta t}$$

Por último, hacemos $\Delta t \to 0$ para obtener la tasa de fallo instantáneo y dar la función de riesgo como

$$h(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t+\Delta t)}{S(t)\Delta t} = -\frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)} \quad t \geq 0.$$

Una vez definida la función de riesgo, cabe destacar que toda función de riesgo satisface las siguientes condiciones:

$$\int_0^{+\infty} h(t)dt = \infty \qquad h(t) \geq 0 \quad \forall t \geq 0.$$

Clases de distribuciones

La forma de la función de riesgo indica cómo los artículos envejecen. Dicha función determina la cantidad de riesgo al que un artículo está sometido en el momento t. Según la función de riesgo, $h(t)$, las tres clases de distribuciones más utilizadas son: IFR (tasa de fallo creciente), DFR (tasa de fallo decreciente) y BF (tasa de fallo en forma de bañera).

Definición. Una distribución de tiempo de vida es IFR (DFR) si $h(t)$ es no decreciente (no creciente) en t.

Definición. Una distribución de tiempo de vida es BT si $h(t)$ es no creciente hasta un punto t y a partir de ese punto es no decreciente.

A la clase IFR pertenecen los artículos que tienen más probabilidad de fallar con el paso del tiempo, es decir, aquellos artículos que se desgastan o degradan (elementos mecánicos). En cambio, a la clase DFR pertenecen los artículos que tienen mayor probabilidad de fallar nada más ponerlos en funcionamiento (programa de ordenador). Y por último, a la clase BT pertenecen aquellos artículos que tienen una tasa de fallo alta al principio, baja en el medio y nuevamente alta al final de la vida útil (aparatos electrónicos complejos).

3.1.3. Función de riesgo acumulada $H(t)$

Una vez definida la función anterior podemos pasar a definir la función de riesgo acumulada, la cual representa el "riesgo" acumulado hasta el instante t.

En el caso discreto, es la suma de las probabilidades de fallo en cada uno de los momentos t_j siendo $t_j \leq t$ y viene dada por:

$$H(t) = \sum_{j|t_j \leq t} h(t_j)$$

Por otro lado, en el caso continuo, viene dada por:

$$H(t) = \int_0^t h(\tau) d\tau.$$

Así, en este caso se tiene que $H(t) = -\log S(t)$.

Cabe destacar que toda función de riesgo acumulada satisface las siguientes condiciones:

$$H(0) = 0 \qquad \lim_{t \to +\infty} H(t) = \infty \qquad H(t) \text{ es no decreciente.}$$

Clasificación de distribuciones

Una vez dada la función de riesgo acumulada podemos definir $\gamma(t) = H(t)/t$, la cual indica la tasa media de fallo en el intervalo $[0,t)$. Según $\gamma(t)$ las clases de distribuciones más utilizadas son: IFRA (tasa media de fallo creciente) y DFRA (tasa media de fallo decreciente).

Definición. Una distribución de tiempo de vida es IFRA (DFRA) si $\gamma(t)$ es no decreciente (no creciente) en t para $t \geq 0$.

La clase IFRA indica que la tasa media de fallo es mayor con el paso del tiempo y la DFRA, en cambio, que es más grande nada más ponerlos en funcionamiento.

3.1.4. Función de vida residual media $L(t)$

La función de vida residual media, $L(t)$, es la vida media que le queda a un artículo dado que ha sobrevivido al tiempo t y viene dada por:

$$L(t) = E[T - t | T > t] \quad t \geq 0.$$

Para el caso continuo, calcularemos $L(t)$ sabiendo que $E[T] = \int_0^{+\infty} t f_T(t) dt$.
En primer lugar, calculamos $E[T|T > t]$. Para ello necesitamos la función de densidad condicionada, la cual viene dada por:

$$f_{T|T>t}(\tau) = \frac{f(\tau)}{S(t)} \quad \tau \geq t.$$

Así obtenemos que

$$L(t) + t = E[T|T > t] = \int_t^{\infty} \tau f_{T|T>t}(\tau) d\tau = \int_t^{\infty} \tau \frac{f(\tau)}{S(t)} d\tau$$

$$\Longrightarrow L(t) = \frac{1}{S(t)} \int_t^{\infty} \tau f(\tau) d\tau - t \quad t \geq 0.$$

En el caso discreto, la función de vida residual media viene dada por:

$$L(t) = \frac{1}{S(t)} \left[\sum_{j|t_j > t} t_j p(t_j) \right] - t \quad t \geq 0.$$

Por último, cabe destacar que toda función de vida residual media satisface las siguientes propiedades:

$$L(t) \geq 0 \qquad L'(t) \geq -1 \qquad \int_0^{\infty} \frac{dt}{L(t)} = \infty.$$

Clases de distribuciones

Según $L(t)$ las clases de distribuciones más comunes son: IMRL (vida residual media creciente) y DMRL (vida residual media decreciente).

Definición. Una distribución de tiempo de vida es IMRL (DMRL) si $L(t)$ es no decreciente (no creciente) $\forall t$.

El primer caso se daría cuando la esperanza de vida que le queda a un artículo es cada vez mayor y, en cambio, el segundo caso se daría cuando la esperanza de vida que le queda a un artículo es menor con el paso de los años.

En la Figura 3.1 se muestran las relaciones existentes entre las distintas clases antes definidas.

Figura 3.1: Relaciones entre las diferentes clases de distribuciones

3.2. Distribuciones de tiempos de vida de un sistema

Una vez vistas las funciones que definen la distribución del tiempo de vida de un artículo nos centraremos en utilizar estas funciones para describir la distribución de un sistema formando por n componentes. Para ello, supondremos que las componentes tienen tiempos de vida aleatorios e independientes.

Utilizando la función de fiabilidad definida en el Capítulo 2 y las diferentes funciones dadas en la sección anterior podemos definir la distribución del tiempo de vida del sistema mediante $S(t)$, $f(t)$, $h(t)$, $H(t)$ o $L(t)$ a partir de las distribuciones de tiempos de vida de las componentes definidas por $S_i(t)$, $f_i(t)$, $h_i(t)$, $H_i(t)$ o $L_i(t)$ para $i = 1,\ldots,n$.

Para encontrar la fiabilidad de un sistema en cualquier momento t, las funciones de supervivencia de las componentes pueden ser usadas en la función de fiabilidad:

$$S(t) = r(S_1(t), S_2(t), \ldots, S_n(t)).$$

A partir de la función de supervivencia podemos calcular el resto de funciones que definen la distribución del tiempo de vida del sistema.

Ejemplo 8. Disponemos del sistema de la Figura 2.3 formado por ochos componentes con funciones de supervivencia

$$S_1(t) = e^{-3t},\ S_2(t) = e^{-4t},\ S_3(t) = e^{-6t},\ S_4(t) = e^{-t},$$

$$S_5(t) = e^{-2t},\ S_6(t) = e^{-3t},\ S_7(t) = e^{-2t} \text{ y } S_8(t) = e^{-t} \quad t \geq 0.$$

La función de fiabilidad del sistema de la Figura 2.3 viene dada en (2.2), por tanto la función de supervivencia del sistema se obtiene sin más que sustituir p_i por $S_i(t)$ para $i = 1,\ldots,8$, es decir,

$$S(t) = 2e^{-12t} + e^{-13t} - e^{-14t} - e^{-20t} - e^{-21t} + e^{-22t} \quad t \geq 0.$$

3.3. Distribuciones importantes en fiabilidad

En esta sección estudiaremos distribuciones tanto continuas como discretas con un importante papel en fiabilidad y en la modelización de los tiempos de vida.

3.3.1. Distribución exponencial

La distribución exponencial es una distribución continua cuya función de densidad esta definida por:

$$f(t) = \lambda e^{-\lambda t} \quad \forall t \geq 0$$

donde $\lambda > 0$ es la tasa de fallo.

Su media y varianza vienen dadas por:

$$\mu = \frac{1}{\lambda} \quad \text{y} \quad \sigma^2 = \frac{1}{\lambda^2}$$

Cabe destacar que dicha distribución es la única distribución continua con la propiedad de ausencia de memoria ($P[T \geq t] = P[T \geq t+s|T \geq s] \quad t \geq 0, s \geq 0$).

3.3.2. Distribución Weibull

La distribución Weibull es una distribución continua cuya función de densidad esta definida por:

$$f(t) = \kappa \lambda^{\kappa} t^{\kappa-1} e^{-(\lambda t)^{\kappa}} \quad \forall t \geq 0$$

donde $\lambda > 0$ y $\kappa > 0$.

Su media y varianza vienen dadas por:

$$\mu = \Gamma\left(1+\frac{1}{\kappa}\right)\frac{1}{\lambda} \quad \text{y} \quad \sigma^2 = \left[\Gamma\left(1+\frac{2}{\kappa}\right) - \left[\Gamma\left(1+\frac{1}{\kappa}\right)\right]^2\right]\frac{1}{\lambda^2}$$

donde Γ es la función gamma. Cabe destacar que la distribución exponencial es un caso particular de la distribución Weibull con $\kappa = 1$.

3.3.3. Distribución geométrica

La distribución geométrica es una distribución discreta cuya función de masa de probabilidad esta definida por:

$$P[T = j] = (1-p)^j p$$

para $j = 0, 1, 2, \ldots$ y $0 < p < 1$.

Esta distribución es la versión discreta de la distribución exponencial y, como ella, tiene pérdida de memoria. Representa la situación en la que, en cada instante de tiempo , la probabilidad de que la componente falle es igual a p, independiente de lo ocurrido hasta entonces.

Su media y varianza vienen dadas por:

$$\mu = \frac{(1-p)}{p} \quad \text{y} \quad \sigma^2 = \frac{(1-p)}{p^2}$$

En las tablas 3.1 y 3.2, respectivamente, se muestran las funciones definidas en la sección 3.1 y las clases a las que pertenecen estas tres distribuciones.

Tabla 3.1: Distribuciones de tiempos de vida

Distribución	$S(t)$	$h(t)$	$H(t)$	$L(t)$
Exponencial	$e^{-\lambda t}$	λ	λt	$\frac{1}{\lambda}$
Weibull	$e^{-(\lambda t)^\kappa}$	$\kappa\lambda^\kappa t^{\kappa-1}$	$(\lambda t)^\kappa$	$\frac{e^{(\lambda t)^\kappa}}{\lambda\kappa}\Gamma\left(\frac{1}{\kappa}\right)\left[1-I\left(\frac{1}{\kappa},(\lambda t)^\kappa\right)\right]$
Geométrica $t=0,1,2,\ldots$	$(1-p)^{j+1}$	p	$(j+1)p$	$\frac{1}{p}$

(*) $I(y,x)=\dfrac{1}{\Gamma(y)}\displaystyle\int_0^x u^{y-1}e^{-u}du$

Tabla 3.2: Propiedades de las distribuciones

Distribución	IFR	DFR	BT	IFRA	DFRA	IMRL	DMRL
Exponencial	SÍ, $\forall\lambda$	SÍ, $\forall\lambda$	NO	SÍ, $\forall\lambda$	SÍ, $\forall\lambda$	SÍ, $\forall\lambda$	SÍ, $\forall\lambda$
Weibull	SÍ, para $\kappa\geq 1$	SÍ, para $\kappa\leq 1$	NO	SÍ, para $\kappa\geq 1$	SÍ, para $\kappa\leq 1$	SÍ, para $\kappa\leq 1$	SÍ, para $\kappa\geq 1$
Geométrica	SÍ, $\forall p$	SÍ, $\forall p$	NO	SÍ, $\forall p$	SÍ, $\forall p$	SÍ, $\forall p$	SÍ, $\forall p$

3.4. Riesgos competitivos

En este apartado vamos a analizar el comportamiento de artículos que están sometidos a distintos riesgos. Cada uno de los riesgos genera un tiempo de fallo y el tiempo de fallo del artículo será el primer instante de fallo de alguna de sus componentes. Un ejemplo en el que aplicaríamos dicho modelo sería en el estudio del tiempo de vida de un ser humano, en el que la muerte puede producirse por varias causas, un accidente, un cáncer, una enfermedad de corazón, etc.

Podemos ver los riesgos competitivos como un sistema en serie de k componentes. Los tiempos de fallo de las componentes $T_1,T_2,\ldots,T_k$ los supondremos independientes y el tiempo de vida del sistema será $T=\text{mín}\{T_1,\ldots,T_k\}$, por lo que $S_T(t)=\prod_{j=1}^k S_{T_j}(t)$ y, en consecuencia, $h_T(t)=\sum_{j=1}^k h_{T_j}(t)$. Por simplicidad nos centraremos en el caso continuo.

Definición. La probabilidad neta de fallo en $[a,b)$ a causa del riesgo j se define como $q_j(a,b)=P[T_j\in[a,b)\mid T_j\geq a]$. Es decir, es la probabilidad que tiene un artículo de fallar en $[a,b)$ debido al riesgo j si el riesgo j es el único al que esta sometido y suponiendo que ha sobrevivido hasta el tiempo a.

Proposición 3.1.

$$q_j(a,b)=P[a\leq T_j<b\mid T_j\geq a]=1-e^{-\int_a^b h_{T_j}(t)dt}\quad\text{para } i=1,\ldots,k.$$

Demostración.

$$q_j(a,b) = P[a \le T_j < b \mid T_j \ge a] = 1 - P[T_j \ge b \mid T_j \ge a]$$
$$= 1 - \frac{S_{T_j}(b)}{S_{T_j}(a)} = 1 - \frac{e^{-H_{T_j}(b)}}{e^{-H_{T_j}(a)}} = 1 - e^{-\int_a^b h_{T_j}(t)dt}.$$

□

Definición. La probabilidad bruta de fallo en $[a,b)$ debido al riesgo j se define como $Q_j(a,b) = P[T_j \in [a,b),\ T_j < T_i\ \forall i \neq j \mid T \ge a]$. Es decir, es la probabilidad que tiene un artículo de fallar en $[a,b)$ a causa del riesgo j en presencia de los otros riesgos a los que esta sometido y suponiendo que ha sobrevivido hasta el tiempo a.

Proposición 3.2.

$$Q_j(a,b) = P[a \le T_j < b,\ T_j < T_i\ \forall i \neq j \mid T \ge a] = \int_a^b h_{T_j}(t) e^{-\int_a^t h_T(x)dx} dt \quad \text{para } j = 1,\dots,k$$

Demostración. Sea $T' = \min\{T_1,\dots,T_{j-1},T_{j+1},\dots,T_k\}$. Se tiene

$$Q_j(a,b) = P[a \le T_j < b,\ T_j < T_i\ \forall i \neq j \mid T \ge a] = \int_a^b P[T_i > t,\ \forall i \neq j \mid T \ge a,\ T_j = t]\ \frac{f_{T_j(t)}}{S_{T_j}(a)}\ dt =$$
$$= \int_a^b P[T' > t \mid T' \ge a,\ T_j = t]\ \frac{f_{T_j}(t)}{S_{T_j}(a)}\ dt = \int_a^b \frac{f_{T_j}(t)}{S_{T_j}(a)}\ \frac{\prod_{i\neq j} S_{T_i}(t)}{\prod_{i\neq j} S_{T_i}(a)}\ dt =$$
$$= \int_a^b h_{T_j}(t) \frac{S_T(t)}{S_T(a)}\ dt = \int_a^b h_{T_j}(t) e^{-\int_a^t h_T(x)\,dx} dt$$

□

La probabilidad de fallo debido al riesgo j viene dada por $\pi_j = P[T_j = T] = Q_j(0,\infty)$ para $j = 1,\dots,k$ y puesto que el fallo ocurre por alguna de las causas, $\sum_{j=1}^k \pi_j = 1$.

Si el fallo se ha debido a la componente j, se puede definir el tiempo de vida bruto de la componente j, Y_j. Su función de supervivencia viene dada por

$$S_{Y_j}(y_j) = P[T \ge y_j \mid T_j = T] = \frac{P[T \ge y_j,\ T_j = T]}{\pi_j}.$$

Las vidas netas pueden ser interpretadas como los tiempos de vida posibles y las vidas brutas como los tiempos de vida observados, por tanto, la distribución de $T_1,\dots,T_k$ determina la distribución de $Y_1,\dots,Y_k$. Cuando los riesgos son independientes, la distribución de las vidas netas puede ser determinada a partir de la distribución de los tiempos de vida brutos usando

$$h_{T_j}(t) = \frac{\pi_j f_{Y_j}(t)}{\sum_{i=1}^k \pi_i S_{Y_i}(t)} \quad t \ge 0$$

para $j = 1,2,\dots,k$.

Ejemplo 9. Consideramos que el artículo en estudio está sometido a $k = 2$ riesgos, es decir, su fallo puede ser provocado por dos causas diferentes. A las v.a. T_1 y T_2 las denotaremos como las vidas netas para las causas 1 y 2 respectivamente. Por ejemplo, consideramos que el artículo bajo estudio es un

móvil y la causa 1 sería la caída del móvil y la causa 2 podría ser cualquier otra cosa (que se moje, que no funcione la batería...). En este caso, T_1 será el tiempo de vida del móvil si el único riesgo al que esta sometido es el de caerse y T_2 el tiempo de vida del móvil si este está pegado a la mesa y no puede caerse, por lo que su fallo será debido a la causa 2. El tiempo de vida observado de ese artículo será $T = \text{mín}\{T_1, T_2\}$. Además, suponemos que T_1 y T_2 son independientes y siguen una distribución Weibull de parámetros $(\lambda = 1, \kappa = 2)$ y $(\lambda = 2, \kappa = 2)$ respectivamente.

Así,

$$S_{T_1}(t) = e^{-t^2} \qquad f_{T_1}(t) = 2te^{-t^2} \qquad h_{T_1}(t) = 2t$$
$$S_{T_2}(t) = e^{-4t^2} \qquad f_{T_2}(t) = 8te^{-4t^2} \qquad h_{T_1}(t) = 8t$$

para $t \geq 0$. Utilizando la Proposición 3.1 obtenemos que las probabilidades netas de fallo en el intervalo $[a,b)$ son

$$q_1(a,b) = 1 - e^{-\int_a^b 2t\, dt} = 1 - e^{-(b^2-a^2)}; \qquad q_2(a,b) = 1 - e^{-4(b^2-a^2)}$$

para $a < b$. Y por la Proposición 3.2 que las probabilidad brutas de fallo debido al primer o segundo riesgo en el intervalo $[a,b)$ son:

$$Q_1(a,b) = \int_a^b 2te^{-\int_a^t 10x\, dx} dt = -\frac{1}{5}\left(e^{-5(b^2-a^2)} - 1\right); \qquad Q_2(a,b) = -\frac{4}{5}\left(e^{-5(b^2-a^2)} - 1\right)$$

para $a < b$. La probabilidad de fallo debido al riesgo 1, π_1, y la probabilidad de fallo debido al riesgo 2, π_2, se obtienen tomando $a = 0$ y $b = \infty$ en $Q_1(a,b)$ y en $Q_2(a,b)$ respectivamente. Así, $\pi_1 = 1/5$ y $\pi_2 = 4/5$.

Una vez calculadas π_1 y π_2 pasaremos a calcular la función de supervivencia del primer tiempo de vida bruto, $S_{Y_1}(y_1)$, que se corresponde con un móvil que falla porque se ha caído y sobre el que esta presente el riesgo 2. La función de supervivencia de Y_1 es

$$S_{Y_1}(y_1) = P[T \geq y_1 \mid T_1 = T] = 5\int_{y_1}^{\infty}\int_{x_1}^{\infty} 16x_1x_2e^{-x_1^2}e^{-4x_2^2}\, dx_2\, dx_1 = e^{-5y_1^2} \quad y_1 \geq 0.$$

Análogamente, $S_{Y_2}(y_2) = e^{-5y_2^2} \quad y_2 \geq 0$.

Nota 1. Un caso contrario a los modelos de riesgos competitivos serían los modelos multiplicativos, donde el tiempo de fallo, T, viene dado por $T = \text{máx}\{T_1, T_2, \ldots, T_k\}$, es decir, como un sistema en paralelo.

Capítulo 4

Métodos estadísticos utilizados en el análisis de tiempos de vida

En este capítulo haremos un breve resumen de algunos de los procedimientos estadísticos más usados en fiabilidad. El punto de partida será un conjunto de n tiempos de fallo independientes e idénticamente distribuidos. Al contrario que en muchas aplicaciones estadísticas puede ocurrir que no todas las variables sean observadas (muestras censuradas) lo que requiere métodos específicos para su estudio. Denotando $(T_1,\ldots,T_n)$ a los tiempos de fallo y $(t_1,\ldots,t_n)$ a sus realizaciones nos podemos encontrar en dos situaciones. En la primera de ellas, todos los valores $(t_1,\ldots,t_n)$ son observados; en este caso se habla de muestra completa y las técnicas a utilizar son las básicas de un primer curso de inferencia estadística, por lo que no las desarrollaremos aquí. La otra situación aparece cuando no todos los tiempos son observados; y para aquellos valores no observados únicamente tenemos cotas inferiores y/o superiores de sus valores. En esta situación se habla de muestras censuradas. Hay varios tipos de censura y en este capítulo trabajaremos con los dos más importantes, que pasamos a describir.

- **Censura tipo I:** Cada una de las componentes es observada, como máximo, hasta un tiempo fijo. Así, la componente i se observa hasta que falla, T_i, (si es antes del tiempo fijado c_i) o hasta c_i para $i = 1,\ldots,n$.
  Cabe destacar que en este tipo de censura podemos encontrarnos con dos situaciones: que todos los tiempos de censura sean iguales, es decir, que $c_1 = c_2 = \ldots = c_n$, o que cada artículo tenga un tiempo de censura diferente. En el primer caso, todos los artículos de la muestra han sido puestos en observación al mismo tiempo y el experimento termina en el instante c_1; lo que suele ser común en problemas de ingeniería. Por el contrario, en modelos de supervivencia (medicina) la otra situación es más habitual; por ejemplo, para un caso médico en el que queremos conocer el tiempo de recidiva de un cáncer después de la intervención, los pacientes que padecen dicho cáncer no han sido operados al mismo tiempo, por lo que se ponen en observación en instantes distintos.

- **Censura tipo II:** Todos los artículos han sido puestos en observación al mismo tiempo y la prueba finaliza después de que un número predeterminado de fallos, r, ha ocurrido. Por tanto, los datos observados son los r tiempos de vida más pequeños en una muestra de tamaño n.

4.1. Estimación paramétrica

Los métodos utilizados en la estimación paramétrica están basados en distribuciones conocidas salvo por un número finito de parámetros, es decir, requieren conocer la distribución de tiempo de vida a la que se ajustan los artículos en estudio para después poder estimar los parámetros desconocidos que determinan dicha distribución. En esta sección, nos centraremos en la estimación de los parámetros de una muestra censurada ya que para la estimación de los parámetros de una muestra completa se utiliza la teoría de muestras aleatorias simples.

4.1.1. Función de verosimilitud

Consideraremos una muestra de tamaño n con tiempos de fallo $t_1, t_2, \ldots, t_n$ independientes e idénticamente distribuidos y cuya distribución de tiempo de vida tiene como función de densidad continua $f(t)$. Suponemos que la función de distribución tiene un vector $\theta = (\theta_1, \ldots, \theta_p)^T$ de parámetros desconocidos asociados a él, donde p es el número de parámetros.

En una muestra censurada de tipo I, en la que no todos los datos han sido observados y los datos de fallo son independientes, denotaremos por $c_1, \ldots, c_n$ a los tiempos de censura correspondientes y después separaremos las n observaciones en dos conjuntos disjuntos $U = \{i \mid t_i \leq c_i\}$ y $C = \{i \mid t_i > c_i\}$. El primero contendrá los índices de los artículos que han fallado durante la observación y el otro los de las unidades cuyos tiempos de vida son superiores al tiempo de censura predeterminado. Por tanto, la función de verosimilitud viene definida por:

$$L(\mathbf{x}, \theta) = \prod_{i \in U} f(t_i, \theta) \prod_{i \in C} S(c_i, \theta)$$

donde $\mathbf{x} = (x_1, \ldots, x_n)$ e $x_i = \text{mín}\{t_i, c_i\}$ para $i = 1, \ldots, n$

En una muestra censurada de tipo II, los datos observados son $(t_{(1)}, \ldots, t_{(r)})$, es decir, los r primeros estadísticos ordenados de $(t_1, \ldots, t_n)$. Por tanto, la verosimilitud de la muestra es

$$L(t_{(1)}, t_{(2)}, \ldots, t_{(r)}, \theta) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(t_{(i)}, \theta)\, S(t_{(r)}, \theta)^{n-r} \tag{4.1}$$

Como podemos observar, esta función no coincide con la función de verosimilitud para una muestra censurada de tipo I y es debido a que los estadísticos de orden no son ni independientes ni idénticamente distribuidos. Dicha función de verosimilitud se obtiene a partir de la teoría de estadísticos ordenados y coincide con la función de densidad conjunta marginal de $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$ la cual se obtiene integrando respecto a $T_{(n)}, T_{(n-1)} \ldots, T_{(r+1)}$ la función de densidad conjunta

$$f_{T_{(1)}, T_{(2)}, \ldots, T_{(n)}}(t_{(1)}, t_{(2)}, \ldots, t_{(n)}) = n! \prod_{i=1}^{n} f(t_{(i)}).$$

4.1.2. Distribución exponencial en muestras censuradas

La distribución exponencial es popular debido a su gran cantidad de aplicaciones y su facilidad a la hora de realizar inferencia. En esta subsección trataremos de encontrar el estimador máximo verosímil del parámetro λ y su intervalo de confianza, IC, en una muestra censurada. Suponemos que tenemos una muestra aleatoria de n artículos cuyos tiempos de vida $T_1, T_2, \ldots, T_n$ son independientes e idénticamente distribuidos, y siguen una distribución exponencial de parámetro λ.

Censura Tipo II

Particularizando (4.1) a este caso, tenemos

$$\log L(t_{(1)}, \cdots, t_{(r)}, \lambda) = \log\left(\frac{n!}{(n-r)!}\right) + \sum_{i=1}^{r} \log(\lambda e^{-\lambda t_{(i)}}) + (n-r)\log(e^{-\lambda t_{(r)}})$$

$$= \log\left(\frac{n!}{(n-r)!}\right) + r\log\lambda - \lambda \sum_{i=1}^{n} x_i$$

donde $x_i = t_{(i)}$, $i = 1, \ldots, r$ y $x_i = t_{(r)}$, $i > r$.

Así, obtenemos que el estimador máximo verosímil es $\hat{\lambda} = \dfrac{r}{\sum_{i=1}^{n} x_i}$.

Notemos que $\sum_{i=1}^{n} x_i = \sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}$ es la suma de los tiempos que han estado funcionando las n componentes durante el experimento. A continuación, estudiaremos la distribución de esta cantidad. Definimos

$$W_1 = n\,T_{(1)}$$

$$W_i = (n-i+1)(T_{(i)} - T_{(i-1)}) \quad i = 2,\ldots,r \tag{4.2}$$

y obtenemos el siguiente resultado.

Proposición 4.1. *Sean $T_{(1)} < \cdots < T_{(r)}$ los r primeros fallos observados en una muestra aleatoria de tamaño n con distribución exponencial de parámetro λ. Entonces, las cantidades $W_1,\ldots,W_r$ dadas por* (4.2) *son independientes e idénticamente distribuidas, con distribución exponencial de parámetro λ.*

Demostración. Bajo un modelo exponencial, la función de densidad conjunta de $T_{(1)},\ldots,T_{(r)}$ es

$$f(t_{(1)},\ldots,t_{(r)}) = \frac{n!}{(n-r)!}\left(\prod_{i=1}^{r}\lambda e^{-\lambda t_{(i)}}\right)(e^{-\lambda t_{(r)}})^{n-r}, \quad \text{con } 0 < t_{(1)} < \cdots < t_{(r)}.$$

Tomando el cambio de variable (4.2) y usando que $T_{(i)} = \frac{W_i}{n-i+1} + \frac{W_{i-1}}{n-i+2} + \ldots + \frac{W_2}{n-1} + \frac{W_1}{n}$ obtenemos que

$$\sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)} = \sum_{i=1}^{r} W_i.$$

Además, el Jacobiano viene dado por

$$det\left(\frac{\partial(w_1,\ldots,w_r)}{\partial(t_{(1)},\ldots,t_{(r)})}\right) = \frac{n!}{(n-r)!}.$$

Aplicando la fórmula de cambio de variable se obtiene que la función de densidad de $(W_1,W_2,\ldots,W_r)$ viene dada por

$$f_{(W_1,\ldots,W_r)}(w_1,\ldots,w_r) = \lambda^r e^{-\lambda\sum_{i=1}^{r} w_i} \quad w_i > 0,$$

lo que demuestra el resultado. □

Corolario 4.2. *Bajo las condiciones de la Proposición 4.1,*

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}$$

tiene una distribución tal que $2\lambda\sum_{i=1}^{n} X_i \sim \chi^2_{2r}$.

Demostración. Basta notar que $\sum_{i=1}^{n} X_i$ tiene la misma distribución que la suma de r v.a. $exp(\lambda)$ independientes, esto es, $\Gamma(\lambda, r)$ por lo que $2\lambda\sum_{i=1}^{n} X_i \sim \chi^2_{2r}$. □

Así, utilizando que $2\lambda\sum_{i=1}^{n} X_i = \frac{2r\lambda}{\hat{\lambda}} \sim \chi^2_{2r}$ obtenemos que un intervalo de confianza de la tasa de fallo λ al $100(1-\alpha)\,\%$ es

$$\frac{\hat{\lambda}\chi^2_{2r,\alpha/2}}{2r} < \lambda < \frac{\hat{\lambda}\chi^2_{2r,1-\alpha/2}}{2r}$$

Censura Tipo I

Analizaremos el caso en el que todos los c_i son iguales. El análisis de las muestras censuradas de tipo I es similar al análisis de las muestras censuradas de tipo II puesto que, como vimos anteriormente, la función de verosimilitud correspondiente al tipo I es proporcional a la del tipo II, por lo que la expresión del estimador máximo verosímil coincide.
Una de las discrepancias que se encuentran entre el análisis de una censura y otra está en el tiempo total bajo observación, ya que en este caso hemos supuesto que todos los tiempos de censura c_i son iguales, es decir, $c_1 = c_2 = \ldots = c_n = c$ y el número de fallos, r, hasta el momento c es una variable aleatoria. En este tipo, la suma de los tiempos que han estado funcionando las unidades es

$$\sum_{i=1}^{n} x_i = \sum_{i\in U} t_i + \sum_{i\in C} c_i = \sum_{i=1}^{r} t_{(i)} + (n-r)c.$$

Otra de las discrepancias aparece a la hora de calcular el intervalo de confianza de λ ya que en este caso $2\lambda\sum_{i=1}^{n} X_i$ ya no sigue una distribución χ^2. Sin embargo, como en este tipo la censura tiene lugar entre el instante $t_{(r)}$ y $t_{(r+1)}$ y vimos en el apartado anterior que $2\lambda\sum_{i=1}^{n} X_i \sim \chi^2_{2r}$ si $c = t_{(r)}$ y $2\lambda\sum_{i=1}^{n} X_i \sim \chi^2_{2r+2}$ si $c = t_{(r+1)}$ se considera que aproximadamente $2\lambda\sum_{i=1}^{n} X_i \sim \chi^2_{2r+1}$.

Por tanto, un intervalo de confianza aproximado de la tasa de fallo, λ, al $100(1-\alpha)\,\%$ es

$$\frac{\hat{\lambda}\chi^2_{2r+1,\alpha/2}}{2r} < \lambda < \frac{\hat{\lambda}\chi^2_{2r+1,1-\alpha/2}}{2r}$$

Comparación de dos poblaciones exponenciales con censura de tipo II

Para comparar dos poblaciones exponenciales con censura de tipo II calcularemos el IC de λ_1/λ_2. Para la población i, $i = 1,2$, tenemos n_i unidades que observamos hasta el tiempo del r_i-ésimo fallo. Así nuestro datos serán

$$t_{1(1)}, t_{1(2)}, \ldots, t_{1(r_1)}$$

$$t_{2(1)}, t_{2(2)}, \ldots, t_{2(r_2)}$$

Definimos como antes $x_{ij} = t_{i(j)}$ para $j < r_i$, $x_{ij} = t_{i(r_i)}$ para $j \geq r_i$, $j = 1,\ldots,n_i$, $i = 1,2$. Como ambos conjuntos presentan censura de tipo II, $2\lambda_1\sum_{i=1}^{n_1} X_{1i} \sim \chi^2_{2r_1}$ y $2\lambda_2\sum_{i=1}^{n_2} X_{2i} \sim \chi^2_{2r_2}$, entonces

$$\frac{2\lambda_1\sum_{i=1}^{n_1} X_{1i}/2r_1}{2\lambda_2\sum_{i=1}^{n_2} X_{2i}/2r_2} = \frac{r_2\lambda_1\sum_{i=1}^{n_1} X_{1i}}{r_1\lambda_2\sum_{i=1}^{n_2} X_{2i}} = \frac{\lambda_1\hat{\lambda_2}}{\lambda_2\hat{\lambda_1}} \sim F_{2r_1,2r_2}$$

Por lo que un intervalo de confianza de λ_1/λ_2 al $100(1-\alpha)\,\%$ es

$$\frac{\hat{\lambda_1}}{\hat{\lambda_2}}F_{2r_1,2r_2,\alpha/2} < \frac{\lambda_1}{\lambda_2} < \frac{\hat{\lambda_1}}{\hat{\lambda_2}}F_{2r_1,2r_2,1-\alpha/2} \tag{4.3}$$

Con un razonamiento similar se construye un test de hipótesis para $H_0 : \lambda_1 = \lambda_2$ frente a $H_1 : \lambda_1 \neq \lambda_2$ en el que se rechaza H_0 si 1 no se encuentra entre los extremos del intervalo (4.3).

4.2. Estimación no paramétrica

No siempre es posible conocer la forma de la distribución del tiempo de vida asociada a un conjunto de datos por lo que necesitamos de la estimación no paramétrica. En esta sección, estimaremos la función de supervivencia y calcularemos un intervalo de confianza aproximado de $S(t)$ de una muestra completa y una muestra censurada utilizando métodos no paramétricos.

4.2.1. Muestras completas

Cuando disponemos de un conjunto de datos en los que todos los tiempos de fallo son conocidos podemos estimar la función de supervivencia, $S(t)$, como el número de artículos que no han fallado en el momento t, $n(t)$, dividido por el número total de artículos en la muestra, n. Por tanto, un estimador no paramétrico de la función de supervivencia viene dado por

$$\hat{S}(t) = \frac{n(t)}{n} \quad t \geq 0,$$

el cual se conoce como función de supervivencia empírica. Es una función escalonada no creciente que disminuye un escalón de tamaño d/n en cada tiempo de vida observado t, siendo d el número de artículos que fallan en ese momento t.

Sobrevivir al momento t puede considerarse como una variable Bernoulli para cada uno de los n artículos bajo estudio. Así, el número de artículos que sobreviven al momento t, $n(t)$, sigue una distribución $Bin(n, S(t))$, donde el éxito se define como la probabilidad de sobrevivir al momento t.

Por tanto, su media y varianza vienen dadas por:

$$E[\hat{S}(t)] = S(t) \quad \text{y} \quad V[\hat{S}(t)] = \frac{S(t)(1-S(t))}{n}$$

Cuando n es grande la distribución Binomial se aproxima a una distribución Normal. En consecuencia, el intervalo de confianza aproximado de $S(t)$ al $100(1-\alpha)\,\%$ es

$$\hat{S}(t) - z_{1-\alpha/2}\sqrt{\frac{\hat{S}(t)(1-\hat{S}(t))}{n}} < S(t) < \hat{S}(t) + z_{1-\alpha/2}\sqrt{\frac{\hat{S}(t)(1-\hat{S}(t))}{n}}.$$

4.2.2. Muestras censuradas de Tipo I

En este apartado calcularemos el estimador de la función de supervivencia de una muestra con censura de tipo I. Cada individuo de la muestra tendrá un tiempo de vida independiente al resto de individuos y los tiempos de censura no serán necesariamente iguales.

Aunque hasta el momento hemos utilizado la verosimilitud para estimar los parámetros desconocidos de una distribución, en este apartado la utilizaremos para estimar la función $S(t)$. La idea principal es escribir la verosimilitud de la muestra en términos de la función de supervivencia y, a partir de allí, encontrar la función $\hat{S}(t)$ que la maximice.

Nos centraremos en el caso discreto. Recordemos que la función de riesgo en este caso viene definida por

$$h(t_j) = \frac{p(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})} \quad j = 1, \ldots$$

donde $t_0 = -1$, por lo que $S(t_0) = 1$.

Notar que $S(t_j)/S(t_{j-1})$ es la probabilidad de sobrevivir a t_j, sabiendo que ha sobrevivido a t_{j-1}. Como es obvio, para que un artículo sobreviva a un cierto tiempo t debe haber sobrevivido a los diferentes tiempos de fallo $t_1 < t_2 < \cdots$ que tienen lugar antes del instante t. Así, la función de supervivencia vendrá definida como un producto de probabilidades condicionadas

$$S(t) = \prod_{j:t_j \leq t} (1 - h(t_j)) \tag{4.4}$$

Una vez dada la definición anterior pasaremos a construir la función de verosimilitud. Consideraremos que hay k instantes distintos, $t_1 < t_2 < \ldots < t_k$, en los que se producen fallos y denotaremos por d_j al número de fallos observados, b_j al número de censuras observadas, es decir, artículos que han sobrevivido a t_j pero no sabemos si han sobrevivido a t_{j+1} y $n_j = n(t_j)$ al número de artículos en riesgo (en los que están incluidos los artículos censurados entre t_j y t_{j+1}) en cada momento t_j para $j = 1, \ldots, k$. Supondremos que los valores $t_1, \ldots, t_k$ son los que pueden tomar la variable.

Una vez definida la notación, pasamos a describir las dos maneras en las que los individuos pueden contribuir en la construcción de la verosimilitud:

- Si el individuo falla en el momento t_j, contribuye con la función de masa de probabilidad, $p(t_j)$.
- Si el individuo es censurado entre t_j y t_{j+1}, contribuye con la función de supervivencia, $S(t_j)$.

Entonces, dado que que las observaciones son independientes, la función de verosimilitud viene dada por:

$$L(S) = \prod_{j=1}^{k} p(t_j)^{d_j} S(t_j)^{b_j}. \tag{4.5}$$

Una vez construida la función de verosimilitud pasamos a buscar la función $\hat{S}(t)$ que la maximice.
Se tiene de (4.4) que

$$S(t_j) = \prod_{i=1}^{j}(1 - h(t_i)) \quad \text{y} \quad p(t_j) = h(t_j)S(t_{j-1}) = h(t_j)\prod_{i=1}^{j-1}(1 - h(t_i)).$$

Sustituyendo en la función de verosimilitud (4.5) obtenemos:

$$L(h(t_1), h(t_2), \ldots, h(t_k)) = \prod_{j=1}^{k} h(t_j)^{d_j}(1 - h(t_j))^{b_j} \left(\prod_{l=1}^{j-1}(1 - h(t_l))^{d_j + b_j} \right).$$

Finalmente, desarrollando, agrupando para los mismos valores de j y usando que $n_j = \sum_{i=j}^{k}(d_i + b_i)$ la verosimilitud se puede reescribir como

$$L(S) = L(h(t_1), h(t_2), \ldots, h(t_k)) = \prod_{j=1}^{k} h(t_j)^{d_j}(1 - h(t_j))^{n_j - d_j}$$

Es decir, la verosimilitud queda escrita en función de k parámetros $(h(t_1), \ldots, h(t_k))$ y se expresa como producto de variables separadas. Tomando logaritmos y derivando con respecto a cada parámetro, obtenemos que $\hat{h}(t_j) = d_j/n_j$ para $j = 1, \ldots, k$. Así, el estimador máximo verosímil de $S(t)$ es

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{4.6}$$

(donde el producto vacío se define como 1), el cual recibe el nombre de estimador Kaplan-Meier.

Notar que si $d_k = n_k$, entonces $\hat{S}(t_k) = 0$ y, por tanto $\hat{S}(t) = 0 \ \ \forall t > t_k$, por ser $\hat{S}$ no creciente. Es decir, en el último instante observado todos los artículos han fallado por lo que estimaremos como 0 la probabilidad de sobrevivir a ese instante. Sin embargo, si $d_k < n_k$, $\hat{S}(t_k) > 0$ y, en este caso, el estimador de Kaplan-Meier no se usa para valores mayores que t_k. Esto ocurre porque sabemos que ha habido artículos que han superado el tiempo t_k pero no tenemos información de cuándo han fallado.

Para el caso continuo se procede de manera análoga a la anterior y se llega a la misma expresión (4.6).

Ahora, una vez calculado el estimador máximo verosímil de la función de supervivencia, $\hat{S}(t)$, pasaremos a calcular un intervalo de confianza aproximado de $S(t)$. Para ello, tenemos que calcular una estimación de la varianza de $\hat{S}(t)$, $\widehat{Var}[\hat{S}(t)]$.

En primer lugar, cabe destacar que el número de muertes en el instante t_j, $d_j = n_j\hat{h}(t_j)$, tiene distribución $Bin(n_j, h(t_j))$. Por tanto,

$$E(\hat{h}(t_j)) = h(t_j), \quad Var(\hat{h}(t_j)) = \frac{h(t_j)(1-h(t_j))}{n_j}$$

y, se puede probar que, $\hat{h}(t_i)$ y $\hat{h}(t_j)$ para $i \neq j$ son independientes asintóticamente (para muestras grandes).

Usaremos el método delta para aproximar la varianza de $\hat{S}(t)$ y su distribución asintótica. Para ello usamos el siguiente resultado. (Ver, por ejemplo, pp 230–231 de [4])

Teorema 4.3. *Sea (X_n) una sucesión de variables aleatorias tal que*

$$\frac{\sqrt{n}(X_n - \mu)}{\sigma} \underset{d}{\to} N(0,1)$$

y $g : \mathbb{R} \to \mathbb{R}$ derivable, tal que $g'(\mu) \neq 0$, entonces

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \underset{d}{\to} N(0,1).$$

Una vez dados los resultados anteriores podemos calcular $\widehat{Var}[\hat{S}(t)]$.
Tomando logaritmos en (4.4) obtenemos que

$$\log(\hat{S}(t)) = \sum_{j:\, t_j \leq t} \log(1 - \hat{h}(t_j)).$$

Suponiendo que los $1 - \hat{h}(t_j)$ son independientes,

$$Var(\log(\hat{S}(t))) = Var\left(\sum_{j:\, t_j \leq t} \log(1 - \hat{h}(t_j))\right) = \sum_{j:\, t_j \leq t} Var[\log(1 - \hat{h}(t_j))]. \tag{4.7}$$

Aplicando el método delta para calcular la varianza de $\log(1 - \hat{h}(t_j))$, con $g(x) = \log(1-x)$ obtenemos que

$$Var\left[\log(1 - \hat{h}(t_j))\right] \approx \left(\frac{-1}{1 - h(t_j)}\right)^2 Var\left[1 - h(t_j)\right] = \frac{h(t_j)}{n_j(1 - h(t_j))} \approx \frac{\hat{h}(t_j)}{n_j(1 - \hat{h}(t_j))}$$

Entonces, por (4.7) se tiene

$$Var(\log \hat{S}(t)) \approx \sum_{j:\, t_j \leq t} \frac{\hat{h}(t_j)}{n_j(1 - \hat{h}(t_j))} = \sum_{j:\, t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

Como queremos obtener la varianza de $\hat{S}(t)$ aplicamos nuevamente el método delta, donde en este caso $g(x) = \log(x)$ y obtenemos que

$$Var\left(\log \hat{S}(t)\right) \approx \left(\frac{1}{\hat{S}(t)}\right)^2 Var(\hat{S}(t)) \Longrightarrow Var(\hat{S}(t)) \approx [\hat{S}(t)]^2 \sum_{j:\, t_j \leq t} \frac{d_j}{n_j(n_j - d_j)},$$

la cual es conocida como la fórmula de Greenwood.

Una vez obtenida la fórmula de Greenwood y el estimador de Kaplan-Meier podemos calcular un intervalo de confianza aproximado para $S(t)$ al $100(1-\alpha)\,\%$ el cual viene definido por:

$$\hat{S}(t) - z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{S}(t))} < S(t) < \hat{S}(t) + z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{S}(t))}$$

Bibliografía

[1] Z. W. BIRNBAUM, *On the importance of different components in a multi-component system*, (No. TR-54), Washington Univ Seattle Lab of Statistical Research, 1968.

[2] W. R. BLISCHKE Y D. N. P. MURTHY, *Reliability: Modeling, Prediction, and Optimization*, 1.ª ed., John Wiley & Sons, 2000.

[3] D. R. COX, *Regression models and life-tables*, Journal of the Royal Statistical Society: Series B, **34** (2) (1972), 187–202.

[4] M.J. CROWDER, A.C. KIMBER, R.L. SMITH Y T.J. SWEETING, *Statistical Analysis of Reliability Data*, 1.ª ed., Chapman & Hall , 1991.

[5] B. EPSTEIN Y M. SOBEL, *Life testing*, Journal of the American Statistical Association, **48** (263) (1953), 486–502.

[6] GOOGLE SCHOLAR, `http://scholar.google.es/` (Consultado el 24-06-2019).

[7] D. V. GLASS, *Graunt's life table*, Journal of the Institute of Actuaries, **76** (1) (1950), 60–64.

[8] E. L. KAPLAN Y P. MEIER, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association, **53** (282) (1958), 457–481.

[9] J. F. LAWLESS, *Statistical Models and Methods for Lifetime Data*, 2.ª ed., John Wiley & Sons, 2003.

[10] L. M. LEEMIS, *Reliability: Probabilistic Models and Statistical Methods*, 1.ª ed., Prentice Hall, 1995.

[11] Z. MA Y A. W. KRINGS, *Survival Analysis Approach to Reliability, Survivability and Prognostics and Health Management (PHM)*, IEEE Aerospace Conference, (2008), 1–20, IEEE.

[12] A. W. MARSHALL Y I. OLKIN, *Life Distributions*, Springer Series in Statistics, Springer, New York, NY, 2007.

[13] J. MCLINN, *A short history of reliability*, The Journal of the Reliability Information Analysis Center, (2011), 8–15.

[14] W. WEIBULL, *A statistical distribution function of wide applicability*, Journal of Applied Mechanics, **18** (3) (1951) , 293–297.