Temas de matemáticas

Análisis de supervivencia: conceptos y modelos básicos

Luis Rincón

Análisis de supervivencia: conceptos y modelos básicos

Luis Rincón
Departamento de Matemáticas
Facultad de Ciencias UNAM
Circuito Exterior de CU
04510 México CDMX

Proyecto PAPIME PE102321 "Estadística y Simulación".

La presente obra se financió con recursos del proyecto PAPIME PE102321
"Estadística y Simulación" de la DGAPA, UNAM

Análisis de Supervivencia: conceptos y modelos básicos
1a edición digital, 2024

D.R. 2024. Universidad Nacional Autónoma de México
Facultad de Ciencias. México 04510, Ciudad de México
editoriales@ciencias.unam.mx
tienda.fciencias.unam.mx

ISBN: 978-607-30-9984-4

Diseño de portada: Laura Uribe

Este libro ha sido dictaminado por pares académicos y sometido a aprobación del Comité Editorial de la Facultad de Ciencias de la UNAM.

Impreso y hecho en México.

Prólogo

Todo lo que inicia acaba. El tiempo de vida de un ser humano es el ejemplo más cotidiano que tenemos de un proceso que comienza y, eventualmente, termina; la duración de ese tiempo es inexorablemente azaroso. En el análisis de supervivencia se buscan crear modelos matemáticos para responder preguntas acerca del tiempo de vida de un individuo, o el tiempo de vida útil de un aparato, cuando la información a partir de la cual se construye el modelo se encuentra incompleta, es decir, cuando algunos de los datos han sido censurados.

Este texto contiene una introducción a algunos temas básicos del análisis estadístico de datos de supervivencia. Se tratan temas como la censura, se definen las principales funciones que caracterizan a la distribución de un tiempo de vida, se estudian algunos modelos paramétricos y sus funciones de verosimilitud, se estudia también el estimador de Kaplan-Meier y se expone el modelo de riesgos proporcionales de Cox. El texto contiene material suficiente para ser cubierto en medio semestre, pues está dirigido a estudiantes de la carrera de Actuaría en la Facultad de Ciencias de la UNAM, quienes deben estudiar estos temas como parte de la asignatura obligatoria Estadística III del Plan de Estudios vigente a la fecha de escritura de este trabajo.

Al final del texto se incluye una amplia bibliografía, la cual refleja la existencia de una necesidad de exponer los resultados del análisis de supervivencia a públicos de diferentes áreas y también de distintos niveles de conocimiento de la probabilidad y la estadística. Se tienen exposiciones elementales como J. Box-Steffensmeier y B. S. Jones [8], A. B. Cantor [11] y D. G. Kleinbaum y M. Klein [34], de nivel intermedio como R. C. Elandt-johnson y N. L. Johnson [19], O. Korosteleva [35], S. Guo [24], J. P. Klein y M. L. Moeschberger [33], J. F. Lawless [36], D. London [39], X. Liu [42], y P. J. smith [47], y tratados más avanzados que requieren la teoría de martingalas como O. O. Aalen, Ø. Borgan y H. K. Gjessing [2], P. K. Andersen *et al* [4], T. R. Fleming y D. P. Harrington [20], y J. D. Kalbfleisch y R. L. Prentice [30]. El presente trabajo puede clasificarse en el nivel intermedio con énfasis en algunos aspectos matemáticos de los modelos. Presupone que el lector ha

tomado por lo menos un curso semestral de probabilidad y otro de estadística a nivel licenciatura. Nuestro interés principal es la exposición lógica y justificada de las fórmulas y los procedimientos.

Debe advertirse que no se hace demasiado énfasis en el uso de programas de cómputo, aunque en la última parte del trabajo fue necesario utilizar `R` para ilustrar algunos conceptos. Herramientas computacionales tales como `R`, `S`, `Python`, `SAS`, `STATA`, `SPSS`, `PSPP` son imprescindibles en situaciones prácticas. En particular, para el uso de `R` en el análisis de supervivencia se puede consultar, por ejemplo, A. Coghlan [14] ó D. F. Moore [44].

Por último, agradezco a la DGAPA UNAM por el apoyo otorgado a través del proyecto PAPIME PE102321 "Estadística y simulación", mediante el cual pudo ser posible la edición de este libro. Agradezco muy sinceramente a los profesores que fungieron como árbitros de este trabajo por tomarse el tiempo para revisar con cuidado este material y elaborar comentarios y sugerencias muy valiosas que permitieron corregir errores, subsanar varias deficiencias y lograr mejoras substanciales. Agradezco también a la Comisión de Publicaciones del Departamento de Matemáticas y al Comité de Publicaciones de la Facultad de Ciencias de la UNAM por el excelente trabajo editorial llevado a cabo.

Luis Rincón
Agosto 2024
Ciudad Universitaria · UNAM

Contenido

1. Tiempos de vida, censura y truncamiento **7**
- 1.1. Tiempos de vida . . . 7
- 1.2. Censura . . . 9
- 1.3. Truncamiento . . . 21
- 1.4. Distribuciones truncadas . . . 24
- 1.5. Ejercicios . . . 27

2. Funciones básicas **35**
- 2.1. Función de distribución . . . 36
- 2.2. Función de supervivencia . . . 38
- 2.3. Función de riesgo . . . 46
- 2.4. Función tiempo medio de vida restante . . . 60
- 2.5. Equivalencia de las funciones básicas . . . 67
- 2.6. Ejercicios . . . 74

3. Modelos paramétricos y funciones de verosimilitud **99**
- 3.1. Distribuciones paramétricas . . . 99
- 3.2. Verosimilitud para datos censurados . . . 110
- 3.3. Verosimilitud para censura por la derecha . . . 113
- 3.4. Verosimilitud para datos truncados . . . 125
- 3.5. Ejercicios . . . 126

4. Modelos no paramétricos **139**
- 4.1. Tablas de mortalidad . . . 139
- 4.2. Método actuarial . . . 147
- 4.3. Interpolación en tablas de mortalidad . . . 155

4.4. Función de supervivencia empírica 161
4.5. Estimador de Kaplan-Meier 167
4.6. Estimador de Nelson-Aalen 190
4.7. Prueba log-rank 194
4.8. Ejercicios 210

5. Modelos con covariables **223**
5.1. Covariables 223
5.2. Modelos para incluir covariables 226
5.3. Modelo de Cox 228
5.4. Estimación de coeficientes 237
5.5. Estimación de la función de riesgo base 243
5.6. Algunas pruebas de hipótesis 262
5.7. Ejercicios 266

Apéndice A **273**
A.1. Transformación de un tiempo de vida 273
A.2. Método delta 278

Bibliografía **285**

Índice alfabético **291**

Capítulo 1

Tiempos de vida, censura y truncamiento

Una primera definición establece que un tiempo de vida es el tiempo aleatorio que transcurre entre el momento del nacimiento de un individuo y el momento de su fallecimiento. Este concepto se puede extender con facilidad a muchos y muy variados contextos en donde hay un inicio y un término. El análisis de supervivencia es el estudio y modelación de los tiempos de vida a través de la probabilidad y la estadística. Sin embargo, a diferencia de otros estudios estadísticos, los datos que provienen de observaciones de tiempos de vida con frecuencia presentan censura o truncamiento. En este capítulo introductorio definiremos con mayor precisión estos primeros conceptos.

1.1. Tiempos de vida

Un tiempo de vida es una longitud de tiempo en el cual, por ejemplo, un ser vivo permanece con vida, o bien, un proceso o característica, permanece activo. De esta manera, existe un instante de inicio o nacimiento, y existe otro instante posterior de fin, falla o muerte del ser vivo, del proceso, o de la característica en observación. De forma más abstracta, tenemos la siguiente definición.

Definición 1.1 *Un tiempo de vida es el tiempo que transcurre entre la ocurrencia de dos eventos sucesivos.*

Los dos eventos de interés deben estar bien definidos. Se ha indicado antes que el primer evento puede ser el nacimiento de una persona y el segundo evento puede ser su fallecimiento. Este es, efectivamente, el tiempo de vida usual de un ser humano y es el ejemplo más cercano que tenemos. Sin embargo, muchos otros ejemplos de tiempos de vida, en un sentido más general, pueden proporcionarse. Algunos de ellos se muestran en la Tabla 1.1 que aparece abajo. A pesar de la generalidad que puede darse a la definición de tiempo de vida, usaremos convenientemente las siguientes tres expresiones familiares: individuo, nacimiento y muerte. Incluso, a un tiempo cualquiera $x > 0$ le podemos llamar la edad de un individuo.

Individuo	Primer evento (Nacimiento)	Segundo evento (Fallecimiento)
Persona	Primer día de trabajo	Último día de trabajo
Mujer	Nacimiento	Edad de 1er. embarazo
Persona enferma	Operación quirúrgica	Sanación
Foco	Inicio de uso	Descompostura
Maquinaria	Puesta en operación	Cese de operación
Persona enferma	Detección de cáncer	Fallecimiento por cáncer
Persona con adicción	Inicio de tratamiento	Recaimiento
Disco duro	Inicio de uso	Descompostura
Persona	Nacimiento	1era. visita al dentista
⋮	⋮	⋮

Tabla 1.1: Ejemplos de tiempos de vida.

A los tiempos de vida se les llama también tiempos de falla. Estos tiempos son positivos y los consideraremos aleatorios. Adoptando el modelo de variable aleatoria para representar un tiempo de vida, nos interesa estimar

su distribución de probabilidad. Esta distribución es desconocida y no necesariamente es una de las distribuciones paramétricas usuales. Puede ser cualquier distribución con soporte dentro del intervalo $(0, \infty)$.

Consideraremos también que el tiempo se puede medir de manera discreta o continua. En consecuencia, la variable aleatoria que representa un tiempo de vida será de alguno de estos dos tipos. Llevaremos a cabo la estimación de la distribución de un tiempo de vida a partir de una serie de observaciones independientes de esta variable aleatoria.

Definición 1.2 *Se le llama datos de supervivencia a las observaciones de los tiempos de vida de un conjunto de individuos.*

Un aspecto particularmente importante que a menudo surge cuando se recolectan datos de supervivencia es la censura. Esto significa que algunas de las observaciones sólo contienen información parcial de la variable de interés. Explicaremos este tema en la siguiente sección.

1.2. Censura

El término censura debe entenderse en el sentido de información incompleta de una o varias observaciones o mediciones de una variable. No implica la connotación negativa que surge cuando una persona, o institución, oculta información, o no permite tener acceso a ella. Se trata de un término técnico que explicaremos a continuación. Aparece no sólo en datos de supervivencia, sino también en cualquier estudio estadístico en donde se llevan a cabo observaciones.

Definición 1.3 *Un conjunto de datos de supervivencia están censurados cuando algunos de ellos no están completamente especificados.*

Esta definición es bastante general y tal vez poco informativa, pero a través

de los tipos particulares de censura que veremos más adelante, se entenderá de mejor manera la posible falta de especificación de los datos. En ocasiones, al conjunto de datos que no están completamente especificados se les llama también datos faltantes ó datos incompletos.

Es importante enfatizar que, si bien los datos de supervivencia que estudiaremos pueden presentar censura, supondremos que éstos no poseen otras deficiencias o inconsistencias que las bases de datos reales con frecuencia presentan. Tales deficiencias pueden ser la pérdida parcial, la manipulación con determinados fines, o bien, problemas con el medio electrónico en el que se almacena la información, por ejemplo.

Es necesario también advertir que existen muchos tipos de censura y en la literatura pueden aparecer referidos con otros nombres alternativos a los que aquí se utilizan. Los tipos de censura que aparecen en la Figura 1.1 son los que definiremos en este trabajo.

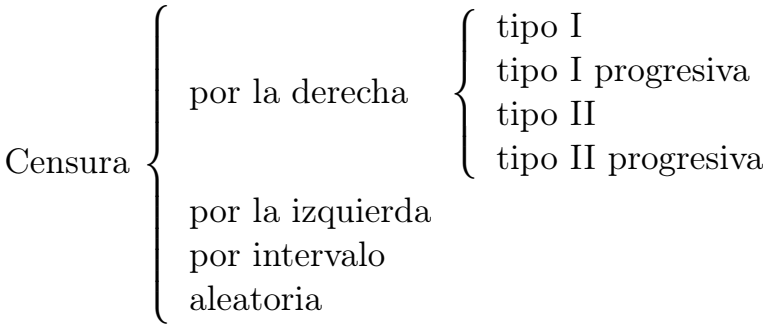

Figura 1.1: Algunos tipos de censura.

Censura por la derecha

Supongamos que para un individuo se ha presentado y registrado la ocurrencia de un primer evento (su nacimiento, por ejemplo). Este individuo se mantiene en observación hasta que ocurre un segundo evento (su fallecimiento, por ejemplo). Recordemos que el tiempo que transcurre entre el primer y el segundo evento es su tiempo de vida, el cual es modelado por

la variable aleatoria $X > 0$. Sea $C > 0$ otra variable aleatoria positiva, no necesariamente independiente de X. Se usa la letra C pues esta variable representará un tiempo de censura. Este es el momento aleatorio en el cual el individuo deja de ser observado.

Definición 1.4 *Se dice que una observación de un tiempo de vida X está censurada por la derecha si no se conoce con exactitud el valor tomado por X, únicamente se sabe que $X > C$, para alguna variable aleatoria C llamada tiempo de censura.*

De esta manera, se presenta la censura por la derecha para una observación del tiempo de vida X de un individuo cuando éste ha sido observado con vida hasta un cierto tiempo $C > 0$, y después ya no es posible continuar su observación. Sólo se cuenta con la información parcial, o incompleta, de que el momento del fallecimiento ocurre después del tiempo C, es decir, no se conoce el valor exacto, sólo que $X \in (C, \infty)$. La variable aleatoria tiempo de censura C puede ser producto de varias cuestiones azarosas y en algunos casos supondremos que es independiente del tiempo de vida en estudio. Las diferentes maneras en las que se puede especificar a la variable C da lugar a varios tipos de censura por la derecha.

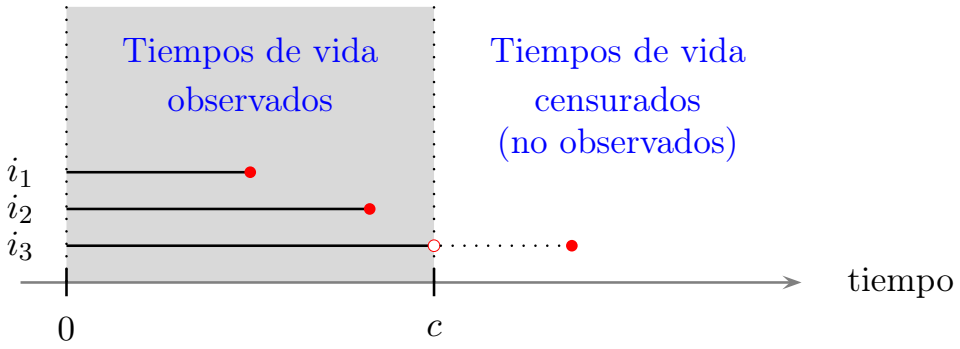

Figura 1.2: Censura por la derecha.

La censura por la derecha se puede representar gráficamente como en la Fi-

gura 1.2. Se tienen aquí 3 individuos denotados por las letras: i_1, i_2 e i_3, y como tiempo de censura una misma constante c. Los fallecimientos que ocurren antes del tiempo c son observados, mientras que los fallecimientos que ocurren después de c no son observados. Muy posiblemente los tres individuos tienen tiempos de nacimiento diferentes en una línea natural del tiempo, sin embargo en la gráfica se han colocado los tres tiempos de nacimiento en un mismo tiempo común $t = 0$.

Bajo una posible censura por la derecha al tiempo C, la observación de un tiempo de vida X se puede escribir como

$$\text{mín}\{X, C\} = \begin{cases} X & \text{si } X \in (0, C], \\ C & \text{si } X \in (C, \infty). \end{cases}$$

Observe que si x representa el valor, posiblemente desconocido, de un tiempo de vida completo, lo que se observa o registra bajo la censura por la derecha al tiempo c es el número mín $\{x, c\}$.

- **Notación 1.** Un conjunto de n tiempos de vida observados bajo el esquema de posible censura por la derecha se puede escribir como la colección de parejas de números:

  $$(x_1, \delta_1), (x_2, \delta_2), \ldots, (x_n, \delta_n),$$

  en donde

  $$\delta_i = \begin{cases} 1 & \text{si } x_i \text{ es un dato no censurado,} \\ 0 & \text{si } x_i \text{ es un dato censurado.} \end{cases}$$

  La letra δ proviene de la palabra en inglés *death*. El valor $\delta_i = 1$ indica que el fallecimiento del individuo i fue observado al tiempo x_i, se trata de un dato no censurado. En cambio, el valor $\delta_i = 0$ indica que el dato registrado x_i se debió a una censura pues en ese momento se dejó de observar al individuo i. Por ejemplo, podemos tener el siguiente conjunto de datos

  $$(x_1, 0), (x_2, 1), (x_3, 1), (x_4, 0), (x_5, 1), \tag{1.1}$$

en donde los tiempos x_1 y x_4 son censurados y los datos x_2, x_3 y x_5 son tiempos de fallecimientos auténticos.

- **Notación 2.** Una manera muy sugerente de denotar un dato x censurado por la derecha es a través del símbolo $x+$, esto sugiere que el fallecimiento ocurre después de x. Por ejemplo, la colección de datos (1.1) también se puede escribir como el vector

$$(x_1+,\, x_2,\, x_3,\, x_4+,\, x_5). \tag{1.2}$$

Bajo situaciones controladas como las que se pueden crear en experimentos en laboratorios o en la industria, es poco probable no poder darle seguimiento a los individuos en observación, a menos que deliberadamente así se haya decidido. Pero los experimentos, por ejemplo, sobre la evolución de una enfermedad en las personas que se someten a algún tratamiento médico, son susceptibles a que aparezcan datos censurados por la derecha pues un cambio de residencia, o un accidente en casa, o cualquier situación inesperada puede hacer que una persona bajo observación ya no quiera, o ya no pueda, continuar con el estudio y deja de ser observada.

Censura tipo I

Este es un tipo particular de censura por la derecha. El nombre más largo pero más preciso para esta sección debería ser "Censura por la derecha tipo I". Veamos su definición.

Definición 1.5 *La censura por la derecha tipo I ocurre cuando cada individuo i tiene un tiempo de censura fijo dado por una constante $c_i > 0$. De esta forma, su tiempo de vida X_i es observado si ocurre el evento $(X_i \leqslant c_i)$ y es censurado (no observado) cuando $(X_i > c_i)$.*

Un caso aún más simple para este tipo de censura se presenta cuando todos los tiempos de censura $c_1, \ldots, c_n$ para un conjunto de n individuos es una

misma constante c. Este es el caso que se ilustra en la Figura 1.2. Además de la simplicidad en su aplicación, la censura tipo I tiene la ventaja de ayudar a controlar el tiempo y el costo de un estudio estadístico al limitar el periodo de observación de cada individuo. La observación de todos los individuos termina, a lo sumo, transcurridas c unidades de tiempo después de su nacimiento. El costo que se tiene que pagar al limitar el tiempo de observación es la obtención de información incompleta debido a la censura.

Censura tipo I progresiva

Esta es una forma controlada de continuar la observación de tiempos de vida bajo posible censura tipo I. Sea $c_1 > 0$ una constante que denota el tiempo de censura que se aplica inicialmente a todos los individuos en observación. Es decir, una primera ventana de observación es el intervalo $(0, c_1]$. Alcanzado el tiempo c_1, algunos de los sobrevivientes son censurados en ese momento (posiblemente escogidos al azar), y a los restantes se les da seguimiento hasta un segundo tiempo de censura $c_2 > c_1$. Así, una segunda ventana de observación es el intervalo $(c_1, c_2]$. Se repite el procedimiento anterior hasta que se cumpla cierta regla preestablecida que indique el fin de las observaciones. En este esquema, c_i denota el tiempo de censura en la i-ésima etapa y no del i-ésimo individuo. Véase la Figura 1.3.

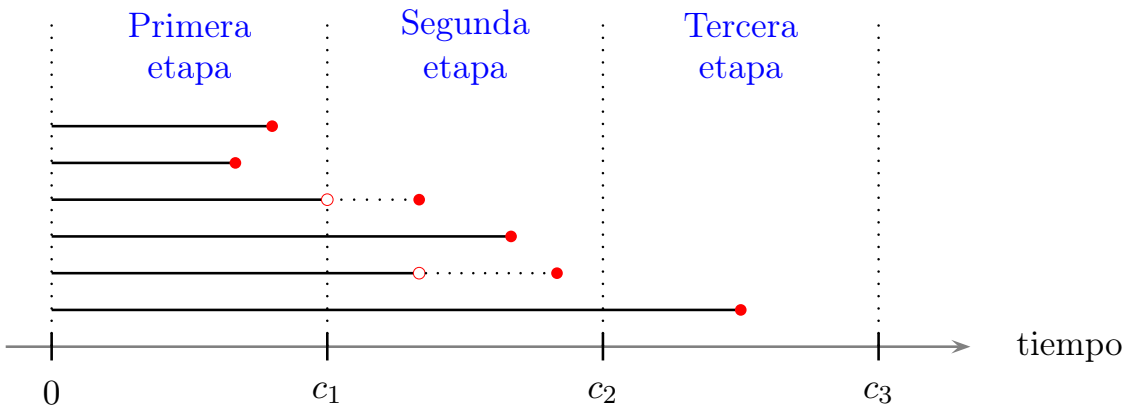

Figura 1.3: Censura tipo I progresiva.

Censura tipo II

Este es otro tipo particular de censura por la derecha y el título de la sección debería ser "Censura por la derecha tipo II". Este esquema de censura puede aplicarlo directamente el investigador que lleva a cabo el estudio estadístico cuando se cuenta con un ambiente controlado para la observación de los individuos.

Supongamos que se tienen n individuos bajo observación. Sean $X_1, \ldots, X_n$ las variables aleatorias que modelan sus tiempos de vida. Supondremos que todos los tiempos de vida inician a un mismo tiempo $t = 0$ y que los individuos pueden ser seguidos sin restricciones.

Definición 1.6 *Sea r un número entero tal que $1 \leqslant r \leqslant n$. La censura por la derecha tipo II consiste en registrar los r tiempos de vida más cortos y censurar por la derecha los $n - r$ tiempos restantes.*

Más explícitamente, si $X_{(1)}, \ldots, X_{(n)}$ son las duraciones de los tiempos de vida ordenados de menor a mayor, es decir, las estadísticas de orden de los tiempos $X_1, \ldots, X_n$, entonces se registran los valores tomados por las variables $X_{(1)}, \ldots, X_{(r)}$ y en el momento aleatorio $X_{(r)}$ se censura al resto de los individuos que aún permanezcan con vida. Es decir, los tiempos $X_{(r+1)}, \ldots, X_{(n)}$ ya no son observados, únicamente se registra que toman un valor mayor al valor tomado por la variable $X_{(r)}$, el cual es un tiempo aleatorio de censura. Observe que la censura ocurre por las características del experimento de muestreo y no por situaciones particulares de los individuos. El esquema de censura tipo II se ilustra gráficamente en la Figura 1.4.

La censura tipo II se puede aplicar en situaciones cuando continuar el seguimiento de todos los individuos hasta su fallecimiento puede tomar mucho tiempo o ser muy costoso. En este caso, es evidente que la persona que diseña el experimento estadístico impone, por conveniencia, la censura. El tiempo promedio en que concluyen las observaciones es $E(X_{(r)})$. La teoría matemática desarrollada sobre las estadísticas de orden es de gran ayuda aquí para

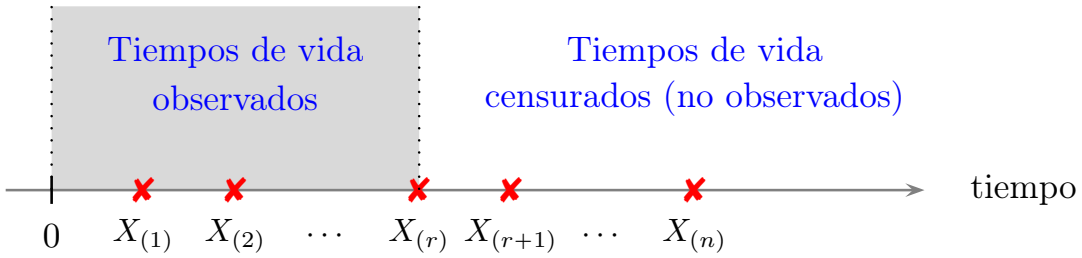

Figura 1.4: Censura por la derecha tipo II.

la estimación de la duración del periodo de observación bajo este esquema de censura.

Ejemplo 1.1 *Supongamos que tenemos una población en estudio de* $n = 10$ *individuos y deseamos llevar a cabo la observación de sus tiempos de vida bajo el esquema de censura por la derecha tipo II con* $r = 6$*. Según este esquema de censura, registraremos los primeros* 6 *fallecimientos y censuraremos los* 4 *restantes. Supongamos que los primeros* 6 *datos en meses son:* 4, 7, 8, 10, 10, 12*. Dado que al tiempo* $t = 12$ *(meses) ya se han obtenido los primeros* 6 *registros, las observaciones concluyen y los* 4 *individuos que permanecen con vida son censurados por la derecha y se les registra con el valor* 12+*. De esta manera, los* 10 *datos obtenidos del experimento son:*

$$4, 7, 8, 10, 10, 12, 12+, 12+, 12+, 12+ .$$

■

Censura tipo II progresiva

Sea n el número de tiempos de vida iniciales en el estudio. Sea r un número entero tal que $1 \leqslant r \leqslant n$. Como antes, se registran las primeras r fallas u observaciones. La censura tipo II progresiva continúa del siguiente modo: de

los restantes $n-r$ individuos, se censura un cierto número de ellos, reduciendo así la población. Se toma ahora otro valor r, menor o igual al número de individuos en la población reducida. Con estos nuevos parámetros se aplica nuevamente el procedimiento de censura tipo II. Estos experimentos de observación sucesivos continúan hasta que la muestra en estudio se extinga o hasta que se cumpla cierta regla preestablecida para el fin del experimento.

Ejemplo 1.2 *Continuando con los datos del Ejemplo 1.1, supongamos que deseamos realizar una segunda etapa de observación usando el mismo esquema de censura tipo II. De los* 4 *individuos con vida al tiempo* $t = 12$*, supongamos que* 1 *de ellos se censura y se registra con el valor* $12+$*, y que continuamos la observación de los* 3 *individuos restantes. Esta es la población reducida. Supongamos que el nuevo valor de* r *es* 1*, es decir, de los* 3 *individuos en observación, esperamos que alguno de ellos fallezca y censuramos en ese momentos a los* 2 *restantes. Supongamos que el fallecimiento ocurre al tiempo* $t = 15$*. Así, los* 10 *datos obtenidos en este esquema de censura tipo II progresiva de dos etapas son:*

$$\underbrace{4, 7, 8, 10, 10, 12, 12+}_{Etapa\ 1}, \underbrace{15, 15+, 15+}_{Etapa\ 2}.$$

■

Censura por la izquierda

Este es un tipo de censura menos frecuente que la censura por la derecha. Veamos su definición.

Definición 1.7 *Se dice que una observación de un tiempo de vida* X *está censurada por la izquierda si no se conoce con exactitud el valor tomado por* X*, únicamente se sabe que* $X < C$*, para alguna variable aleatoria* C*.*

En este caso la variable C no es un tiempo de censura. Representa simplemente un tiempo en el cual se ha hecho una consulta o valoración del

individuo con el fin de detectar la ocurrencia de su fallecimiento o falla. Clarificaremos la situación con algunos ejemplos.

Ejemplo 1.3 *Un experimento de laboratorio consiste en inyectar un cancerígeno a un grupo de ratones con el fin de estudiar el tiempo X que transcurre hasta la aparición de un tumor. La detección del tiempo exacto de aparición de un tumor es difícil llevarla a cabo, sólo es posible detectar su existencia después de una inspección quirúrgica cuando el ratón muere o es sacrificado. Pueden ocurrir distintas situaciones pero los siguientes dos ejemplos ayudan a entender la manera en la que pueden surgir datos censurados por uno y otro lado:*

- *Si un ratón muere o es sacrificado al tiempo c y se le encuentra un tumor, entonces el tiempo X de aparición del tumor está censurado por la izquierda, pues sólo se conoce que ocurre el evento $(X < c)$.*
- *Si un ratón es sacrificado al tiempo c y no se le encuentra un tumor, entonces el tiempo X de aparición del tumor está censurado por la derecha, pues sólo se sabe que $(X > c)$.*

■

Ejemplo 1.4 *Otra situación en donde puede ocurrir la censura por la izquierda es la siguiente: cuando se reporta que una luminaria (lámpara pública en una ciudad) ha fallado, su tiempo de vida X está censurado por la izquierda pues sólo se conoce que $(X < c)$, en donde c es el tiempo de reporte de la falla. No se conoce el valor exacto que tomó X, sólo se conoce que su valor se encuentra a la izquierda de c.* ■

Ejemplo 1.5 *Sean $X_1, \ldots, X_n$ los tiempos de vida de n individuos que inician todos en $t = 0$. Suponga que los individuos sólo son observados a partir del tiempo $t = c > 0$, es decir, la ventana de observación es el intervalo $[c, \infty)$. Al tiempo $t = c$, algunos tiempos de vida pueden haber concluido sin que conozcamos su valor exacto. Si x es uno de esos tiempos, sólo sabemos que $x < c$, es decir, el dato está censurado por la izquierda.* ■

Notación. Una manera muy sugerente de denotar un dato x censurado por la izquierda es a través del símbolo $x-$, esto sugiere que el tiempo de vida concluye antes de x. Por ejemplo, podemos tener la colección de los siguientes 5 datos, algunos de los cuales presentan censura por la izquierda:

$$23,\ 14-,\ 25,\ 20-,\ 17.$$

Censura por intervalo

Por diversas razones, en algunos estudios no se pueden llevar a cabo observaciones de manera continua. Los individuos bajo estudio sólo se pueden observar en tiempos específicos, de modo que la ocurrencia de un evento sólo puede conocerse que sucedió entre dos tiempos de observación sucesivos. Esta limitación produce una censura por intervalo.

Definición 1.8 *Se presenta la censura por intervalo en un tiempo de vida cuando la ocurrencia del primer evento (inicio o nacimiento), o bien la ocurrencia del segundo evento (término o fallecimiento) sólo se pueden registrar con la información de que han tomado un valor dentro de un cierto intervalo de tiempo.*

Los siguiente ejemplos ayudan a entender mejor la censura por intervalo.

Ejemplo 1.6 *Discretizar una variable continua puede considerarse como una censura por intervalo. Por ejemplo, en algunos estudios demográficos, nos interesa estimar las probabilidades de fallecimiento de las personas para cada edad en una población específica. En este caso, no es necesario conocer el tiempo exacto de fallecimiento de los individuos, sino únicamente la edad (intervalo de tiempo) en la que el suceso ocurre. Este ejemplo muestra nuevamente que la censura no necesariamente es mala, puede ser conveniente.* ■

Ejemplo 1.7 *Los pacientes no hospitalizados que se encuentran bajo tratamiento de cierta enfermedad deben valorarse en ciertos tiempos programados. Aquellos que eventualmente sanan, tienen necesariamente un tiempo de curación censurado por intervalo pues la sanación se detecta hasta el momento de una valoración. Por lo tanto, el momento exacto de la sanación ocurre en algún instante desconocido que se encuentra entre dos revisiones médicas sucesivas.* ∎

Censura aleatoria

Como se ha visto en los esquemas anteriores, en la definición de algunos tipos de censura hay tiempos aleatorios involucrados. Cuando esto es así, se dice que la censura es aleatoria.

Definición 1.9 *Se dice que ocurre una censura aleatoria para un tiempo de vida cuando el inicio o término de observación de un individuo sucede en un momento aleatorio.*

Por ejemplo, la censura por la derecha tipo II es esencialmente aleatoria, pues el momento de la censura ocurre cuando fallece el r-ésimo individuo.

§

Con esto concluimos las definiciones de los tipos de censura que aparecen en la Figura 1.1. Como hemos mencionado antes, existen muchas otras maneras en las que se puede presentar información parcial en una colección de observaciones. Puede también ocurrir la situación de alguna combinación de varios tipos de censura en un mismo conjunto de datos. Los métodos estadísticos que presentaremos en este trabajo contemplan, mayormente, la censura por la derecha, la cual es el tipo de censura más común. Un problema central en el análisis de supervivencia es estimar la distribución de un tiempo de vida a partir de observaciones que presentan algún tipo de censura.

El lector interesado en consultar otras exposiciones y ejemplos sobre el concepto de censura puede revisar cualquiera de los textos sobre análisis de supervivencia que aparecen en la bibliografía. Por ejemplo, los siguientes textos contienen por lo menos una sección separada sobre este tema: K. Bogaerts, A. Komárek y E. Lesaffre [6], G. Broström [10], D. W. Hosmer, S. Lemeshow y S. May [28], J. F. Lawless [36], J. Li y S. Ma [40], D. F. Moore [44]. Particularmente, el libro de M. Cleves *et al* [12] contiene un capítulo entero sobre el tema de censura y truncamiento. En el libro de D. R. Helsel [26] se encuentra una discusión muy completa sobre la censura, así como sobre el problema de la descripción numérica de un conjunto de datos censurados. Incluye también una discusión sobre el problema general de la estimación estadística a partir de datos censurados.

1.3. Truncamiento

En el caso de la censura, cada individuo contribuye, por lo menos, con cierta información parcial de su tiempo de vida. En el caso del truncamiento, aquellos individuos cuyos tiempos de vida no cumplen cierta condición preestablecida quedan fuera del estudio y la información de su tiempo de vida queda desechada. Veamos la definición.

Definición 1.10 *Sea X un tiempo de vida, y sean a y b dos números reales tales que $0 \leqslant a < b$. El truncamiento en datos de supervivencia ocurre cuando sólo son observados y considerados para el estudio estadístico aquellos individuos que tienen un tiempo de vida X con valor en el intervalo $(a, b]$.*

En consecuencia, las estimaciones sobre la distribución de los tiempos de vida serán para las distribuciones condicionadas a la ocurrencia del evento $(a < X < b]$. El truncamiento no debe confundirse con la censura. En ambos casos se presenta información parcial de los datos. Sin embargo, en la censura la insuficiencia de la información ocurre a nivel individual, mientras que en el truncamiento es a nivel global, pues se presenta una reducción del

conjunto completo y desconocido de observaciones al subconjunto de aquellas observaciones x que satisfagan la condición $a < x \leqslant b$. Dependiendo del problema y la variable en estudio, el intervalo de truncamiento $(a, b]$ puede incluir, o excluir, los valores extremos.

El truncamiento no es exclusivo de los datos de supervivencia. El fenómeno se presenta también en cualquier proceso de observación de variables cuantitativas de cualquier disciplina. Tenemos los siguientes dos casos particulares.

Truncamiento por la izquierda

Se toma el intervalo de truncamiento (a, b) con $a > 0$ y $b = \infty$. Sólo aquellos individuos que sobreviven un determinado tiempo $(X > a)$ son considerados en la muestra. Esto es, si se tienen n observaciones $x_1, \ldots, x_n$, entonces necesariamente todas ellas pertenecen al intervalo (a, ∞), es decir, $x_1, \ldots, x_n \in (a, \infty)$.

Ejemplo 1.8 *Se puede hacer un seguimiento hasta su fallecimiento de personas de edad avanzada que ingresan a un sistema de casas de retiro, pero para poder ingresar a estas casas es necesario tener por lo menos* 70 *años cumplidos. En consecuencia, las observaciones que se pueden efectuar están truncadas por la izquierda pues se tiene la condición* $(X \geqslant 70)$*, en donde* X *denota la edad de fallecimiento en años. Claramente el intervalo de truncamiento es* $[70, \infty)$*.* ∎

Es interesante observar que los datos del ejemplo anterior pueden presentar, además, censura por la derecha. Esto puede ocurrir, por ejemplo, cuando un individuo decide abandonar la casa de retiro. Cuando eso sucede, ya no es posible continuar el seguimiento de esa persona.

Ejemplo 1.9 *Este es un ejemplo que no pertenece al área de estudio de tiempos de vida, pero ayuda a entender el concepto de truncamiento y las razones de su origen. Suponga que se desea estimar la distribución del diámetro de ciertas partículas muy pequeñas y que para ello se utiliza un instrumento especializado. Dado que la precisión del instrumento es necesariamente finita, las partículas muy pequeñas no pueden ser registradas y sólo*

aquellas suficientemente grandes son susceptibles de ser observadas por el instrumento. Es claro que los datos estarán truncados por la izquierda pues las mediciones son mayores a cierta cantidad determinada por la precisión del instrumento de observación. ∎

Truncamiento por la derecha

Se toma el intervalo de truncamiento (a, b) con $a = 0$ y $b > 0$ finito. Sólo aquellos individuos cuyo tiempo de vida ha concluido antes de un tiempo preestablecido b son considerados en la muestra. Es decir, si se tienen n observaciones $x_1, \ldots, x_n$, entonces necesariamente todas ellas pertenecen al intervalo $(0, b)$.

Ejemplo 1.10 *El cáncer infantil se refiere a cualquier tipo de cáncer que se detecta en las personas dentro de sus primeros* 15 *años de vida. Supongamos que el tiempo de vida X se define como la edad a la que se detecta el cáncer infantil. Esta definición provoca necesariamente que la variable X esté truncada por la derecha: sólo aquellas personas con cáncer detectado dentro del rango de edad en años* $(0, 16)$ *son observados y forman parte del estudio.* ∎

Ejemplo 1.11 *Este es otro ejemplo que no pertenece al área de estudio de los tiempos de vida, pero ayuda a entender el concepto de truncamiento y lo que puede originarlo. Suponga que se desea estimar la distribución de la distancia entre la tierra y las distintas estrellas. Aquellas estrellas que se encuentran muy alejadas de nosotros no pueden ser detectadas por los instrumentos astronómicos actuales y, por lo tanto, no se puede tener registro de ellas. En consecuencia, los datos recolectados estarán necesariamente truncados por la derecha. Toda observación x pertenece a un intervalo de la forma $(0, b)$, para algún valor $b > 0$ fijo, determinado por el alcance máximo del instrumento de medición utilizado.* ∎

En la siguiente sección revisaremos algunas fórmulas sobre el truncamiento en variables aleatorias en general, es decir, no necesariamente para tiempos de vida.

1.4. Distribuciones truncadas

Se obtiene una distribución truncada cuando a una variable aleatoria X se le condiciona a tomar valores de manera restringida a un intervalo de la forma $(a, b]$, en donde $a < b$, y suponiendo que $P(a < X \leqslant b) > 0$. Este concepto se puede definir para cualquier variable aleatoria.

Para tiempos de vida, esta situación ocurre cuando las observaciones inician después de transcurridas a unidades de tiempo y concluyen hasta un tiempo máximo b. De esta manera, no hay observaciones fuera del intervalo $(a, b]$. Así, a la función de distribución de X condicionada al evento $(a < X \leqslant b)$ y denotada por $F(x \,|\, a < X \leqslant b)$ se le llama distribución truncada. Aquí tenemos la definición para cualquier variable aleatoria, discreta o continua, y que no necesariamente es un tiempo de vida.

Definición 1.11 *Sea X una variable aleatoria con función de distribución $F(x)$. Sean $a < b$ dos constantes. A la distribución de X condicionada al evento:*

a) $(a < X \leqslant b)$ *se le llama distribución doblemente truncada.*

b) $(X > a)$ *se le llama distribución truncada por la izquierda.*

c) $(X \leqslant b)$ *se le llama distribución truncada por la derecha.*

- A la distribución doblemente truncada $F(x \,|\, a < X \leqslant b)$ también se le llama bilateralmente truncada, y para que la distribución condicional esté bien definida debe ocurrir que $P(a < X \leqslant b) > 0$.

- A la distribución truncada por la izquierda $F(x \,|\, X > a)$ también se le llama truncada por abajo, y se debe cumplir que $P(X > a) > 0$.

- Finalmente, a la distribución truncada por la derecha $F(x \,|\, X \leqslant b)$ también se le llama truncada por arriba, y se debe cumplir que $P(X \leqslant b) > 0$.

Se puede comprobar que las tres distribuciones condicionales: $F(x \mid a < X \leqslant b)$, $F(x \mid X > a)$ y $F(x \mid X \leqslant b)$ son, efectivamente, funciones de distribución. A continuación veremos algunas fórmulas generales sobre las distribuciones truncadas. Estos resultados son válidos tanto en el caso continuo como en el caso discreto.

Proposición 1.1 *Sea X una variable aleatoria discreta o continua, con función de distribución $F(x)$, y con función de probabilidad o de densidad $f(x)$. Sean $a < b$ dos constantes tales que $P(a < X \leqslant b) > 0$. Entonces*

1. $$F(x \mid a < X \leqslant b) = \begin{cases} 0 & si \;\; x \leqslant a, \\ \dfrac{F(x) - F(a)}{F(b) - F(a)} & si \;\; a < x \leqslant b, \\ 1 & si \;\; x > b. \end{cases}$$

2. $$f(x \mid a < X \leqslant b) = \begin{cases} \dfrac{f(x)}{F(b) - F(a)} & si \;\; a < x \leqslant b, \\ 0 & en \; otro \; caso. \end{cases}$$

Demostración.

1. Son evidentes los casos $F(x \mid a < X \leqslant b) = 0$ para $x \leqslant a$ y $F(x \mid a < X \leqslant b) = 1$ para $x > b$. En el caso $a < x \leqslant b$, tenemos que

$$\begin{aligned} F(x \mid a < X \leqslant b) &= \frac{P(X \leqslant x, a < X \leqslant b)}{P(a < X \leqslant b)} \\ &= \frac{P(a < X \leqslant x)}{P(a < X \leqslant b)} \\ &= \frac{F(x) - F(a)}{F(b) - F(a)}. \end{aligned}$$

2. En el caso absolutamente continuo y suponiendo que $f(x)$ es la función de densidad de X, derivando la fórmula anterior se encuentra la función de densidad truncada. En el caso discreto, si x es un posible valor de

X que se encuentra en el intervalo $(a, b]$, entonces

$$\begin{aligned} f(x \,|\, a < X \leqslant b) &= P(X = x \,|\, a < X \leqslant b) \\ &= \frac{P(X = x, a < X \leqslant b)}{P(a < X \leqslant b)} \\ &= \frac{P(X = x)}{P(a < X \leqslant b)} \\ &= \frac{f(x)}{F(b) - F(a)}. \end{aligned}$$

■

A partir de las fórmulas de la proposición anterior y tomando $b \to \infty$ se pueden encontrar expresiones para $F(x \,|\, X > a)$ y $f(x \,|\, X > a)$. Si ahora b es fijo y se hace $a \to -\infty$, entonces se encuentran expresiones reducidas para $F(x \,|\, X \leqslant b)$ y $f(x \,|\, X \leqslant b)$. Estas fórmulas se muestran explícitamente en la sección de ejercicios.

Momentos de las distribuciones truncadas

El n-ésimo momento de una variable aleatoria continua X con distribución truncada al intervalo $(a, b]$ siempre existe (por ser una variable aleatoria acotada) y es

$$E(X^n \,|\, a < X \leqslant b) = \int_a^b x^n \, f(x \,|\, a < X \leqslant b)\, dx.$$

Se debe suponer aquí que $P(a < X \leqslant b) > 0$. Las expresiones para $E(X^n \,|\, X > a)$ y $E(X^n \,|\, X \leqslant b)$ son análogas, aunque encontrar estas integrales puede no ser una tarea fácil.

En particular, si X es un tiempo de vida y $x > 0$ es un valor dado, entonces cuando a la esperanza condicional $E(X \,|\, X > x)$ se le resta el valor x se obtiene la esperanza del tiempo de vida restante a la edad x. En estudios actuariales, a este tiempo promedio se le denota por el símbolo $\mathring{e}_x$ y se puede

calcular de la siguiente forma

$$\begin{aligned} \overset{\circ}{e}_x &:= E(X \,|\, X > x) - x \\ &= E(X - x \,|\, X > x) \\ &= \int_x^\infty (u - x)\, f(u \,|\, X > x)\, du \\ &= \int_0^\infty v\, f(v + x \,|\, X > x)\, dv. \end{aligned}$$

Más adelante retomaremos este tiempo promedio de vida restante, o también llamado residual, a edad x. Por otro lado, la varianza de este nuevo tiempo de vida es

$$\begin{aligned} \text{Var}(X - x \,|\, X > x) &= \text{Var}(X \,|\, X > x) \\ &= E(X^2 \,|\, X > x) - E^2(X \,|\, X > x). \end{aligned}$$

§

Hemos indicado que el truncamiento de una variable aleatoria o de su distribución se define a través de una distribución condicional. Este tema pertenece a la teoría de la probabilidad y para mayor información se puede consultar, por ejemplo, el texto de S. Guo [24].

Reiteramos que uno de los problemas centrales en el análisis de supervivencia consiste en estimar la distribución de probabilidad de un tiempo de vida a partir de datos censurados. Antes de tratar este problema, en el siguiente capítulo revisaremos algunas funciones equivalentes a través de las cuales se puede caracterizar la distribución de una variable aleatoria.

1.5. Ejercicios

Tiempos de vida

1. Proponga 3 ejemplos de tiempos de vida indicando con claridad: el individuo (no necesariamente una persona), el nacimiento (evento inicial) y el fallecimiento (evento final). Los ejemplos deben ser distintos de los que aparecen en la Tabla 1.1.

2. *Lámpara.*
Suponga que una lámpara tiene un tiempo de vida útil X con distribución $\exp(\lambda)$, es decir, su función de densidad es

$$f(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{si } \ x > 0, \\ 0 & \text{en otro caso.} \end{cases}$$

Suponga que el tiempo es medido en horas y que $\lambda = 0.00006667$. Encuentre:

   a) El tiempo promedio de vida útil de la lámpara.

   b) La probabilidad de que el tiempo de vida tome un valor entre $14,000$ hrs. y $16,000$ hrs.

   c) La probabilidad de que el tiempo de vida útil exceda el valor $15,000$ hrs.

3. *Tiempo de vida uniforme.*
Sea X un tiempo de vida con distribución $\text{unif}(0, a)$, en donde $a > 0$ es una constante.

   a) Encuentre y grafique $f(x)$.

   b) Encuentre y grafique $F(x)$.

   c) Encuentre $E(X)$ y $\text{Var}(X)$.

4. *Matrimonio.*
Sean H y M los tiempos de vida individuales de un hombre y una mujer, medido en años, a partir del momento en que se unen en matrimonio. Suponga que H y M son independientes con distribución $\exp(h)$ y $\exp(m)$, respectivamente. Encuentre la probabilidad de que:

   a) El hombre quede viudo.

   b) La mujer quede viuda.

   c) Los dos tiempos de vida concluyan a una distancia, uno del otro, de a lo sumo 1 año.

Censura por la derecha

5. Sea X un tiempo de vida y sea $C > 0$ otra variable aleatoria que representa un tiempo de censura aleatorio por la derecha. Suponga que ambas variables aleatorias tienen distribución $\exp(\lambda)$ y que son independientes. Defina

$$\delta = 1_{(X \leqslant C)} = \begin{cases} 1 & \text{si } X \leqslant C \quad \text{(No censura)}, \\ 0 & \text{si } X > C \quad \text{(Censura por la derecha)}. \end{cases}$$

Calcule:

a) $P(\delta = 1)$.

b) $P(\delta = 0)$.

Censura por la derecha tipo I

6. *Tiempo de vida exponencial.*
Sea X un tiempo de vida con distribución $\exp(\lambda)$ y sea $c > 0$ una constante que representa un tiempo de censura por la derecha. Sea $\delta = 1_{(X \leqslant c)}$ la variable que indica si se observa el valor de X o se censura. Calcule:

a) $P(\delta = 1)$.

b) $P(\delta = 0)$.

7. *Tiempo de vida uniforme.*
Sea X un tiempo de vida con distribución $\text{unif}(0, a)$, en donde $a > 0$ es una constante. Sea c otra constante tal que $0 < c < a$ y que representa un tiempo de censura por la derecha. Sea $\delta = 1_{(X \leqslant c)}$ la variable que indica si se observa el valor de X o se censura. Calcule:

a) $P(\delta = 1)$.

b) $P(\delta = 0)$.

8. *Tiempo de vida exponencial.*
Sea X un tiempo de vida con distribución $\exp(\lambda)$.

a) Sean $0 < a < b$ dos constantes. Encuentre la probabilidad de que el tiempo de vida tome un valor entre a y b.

b) Sea $c > 0$ una constante que corresponde a un tiempo de censura por la derecha. Calcule la probabilidad de que el tiempo de vida sea censurado.

9. *Tiempo de vida geométrico.*
Sea X un tiempo de vida discreto con función de probabilidad

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

en donde $0 < p < 1$ es un parámetro.

a) Sean $0 \leqslant n \leqslant m$ dos enteros. Encuentre la probabilidad de que el tiempo de vida tome un valor entre n y m, inclusive.

b) Sea $c \geqslant 0$ un entero que corresponde a un tiempo de censura por la derecha. Calcule la probabilidad de que el tiempo de vida sea censurado.

10. *Distribución del número de observaciones censuradas.*
Sean $X_1, \ldots, X_n$ tiempos de vida independientes y con idéntica función de distribución continua $F(x)$ y con soporte el intervalo $(0, \infty)$. Sea $c > 0$ una constante que representa un tiempo de censura por la derecha para todos los tiempos de vida. Encuentre:

a) La distribución del número de observaciones censuradas.

b) El número esperado de observaciones censuradas.

c) La distribución del número de observaciones no censuradas.

d) El número esperado de observaciones no censuradas.

Censura por la derecha tipo II

11. Considere el esquema de censura por la derecha tipo II junto con la notación e hipótesis usuales.

a) Escriba la expresión para la función de densidad de la r-ésima estadística de orden de una muestra aleatoria de tamaño n de una distribución continua, $1 \leqslant r \leqslant n$.

b) Escriba la expresión para calcular la esperanza de la r-ésima estadística de orden.

c) Calcule la esperanza del inciso anterior en el caso de la distribución unif$(0, 1)$. Este es el tiempo promedio en el que concluye el periodo de observación en un esquema de censura tipo II para los tiempos de vida indicados. Utilice la función beta para calcular esta esperanza.

d) ¿Qué sucede con la esperanza del inciso anterior cuando $n \to \infty$?

Censura por la izquierda

12. *Tiempo de vida exponencial.*
Sea X un tiempo de vida con distribución $\exp(\lambda)$ pero que es observado sólo cuando toma un valor en el intervalo $[c, \infty)$, para alguna constante $c > 0$. Si X no toma un valor en el intervalo indicado, entonces es censurado por la izquierda con el valor c. Calcule la probabilidad de que el tiempo sea censurado por la izquierda.

13. *Distribución del número de observaciones censuradas.*
Sean $X_1, \ldots, X_n$ tiempos de vida independientes y con idéntica función de distribución continua $F(x)$ y con soporte el intervalo $(0, \infty)$. Suponga que los tiempos sólo son observados cuando toman un valor en el intervalo $[c, \infty)$, para alguna constante $c > 0$. Los tiempos de vida que toman un valor fuera de esta ventana de observación son censurados por la izquierda por el valor c. Encuentre:

a) La distribución del número de observaciones censuradas.

b) El número esperado de observaciones censuradas.

c) La distribución del número de observaciones no censuradas.

d) El número esperado de observaciones no censuradas.

Truncamiento

14. *Truncamientos unilaterales.*
Sea X una variable aleatoria discreta o continua, con función de distribución $F(x)$, y con función de probabilidad o de densidad $f(x)$. Sean a y b dos constantes. Demuestre que:

$$a)\ F(x \,|\, X > a) = \begin{cases} 0 & \text{si } \ x \leqslant a, \\ \dfrac{F(x) - F(a)}{1 - F(a)} & \text{si } \ x > a. \end{cases}$$

$$b)\ f(x \,|\, X > a) = \begin{cases} \dfrac{f(x)}{1 - F(a)} & \text{si } \ a < x \leqslant b, \\ 0 & \text{en otro caso.} \end{cases}$$

$$c)\ F(x \,|\, X \leqslant b) = \begin{cases} \dfrac{F(x)}{F(b)} & \text{si } \ x \leqslant b, \\ 1 & \text{si } \ x > b. \end{cases}$$

$$d)\ f(x \,|\, X \leqslant b) = \begin{cases} \dfrac{f(x)}{F(b)} & \text{si } \ x \leqslant b, \\ 0 & \text{si } \ x > b. \end{cases}$$

Nota: Observe que, en estas fórmulas generales, X no necesariamente es un tiempo de vida y se debe cumplir que $P(X > a) = 1 - F(a) > 0$ ó $P(X \leqslant b) = F(b) > 0$, según sea el caso, cuando estas expresiones aparecen como el denominador de un cociente.

15. *Distribución uniforme.*
Sea X una variable aleatoria con distribución unif(A, B), en donde $A < B$. Sean a y b dos constantes tales que $A \leqslant a < b < B$.

   a) Encuentre $F(x \,|\, a < X \leqslant b)$.

   b) Encuentre $f(x \,|\, a < X \leqslant b)$.

   c) Grafique en el mismo plano $F(x)$ y $F(x \,|\, a < X \leqslant b)$.

   d) Grafique en el mismo plano $f(x)$ y $f(x \,|\, a < X \leqslant b)$.

   e) Encuentre $E(X)$ y $E(X \,|\, a < X \leqslant b)$.

16. *Distribución uniforme.*
Sea X una variable aleatoria con distribución unif$(0, a)$, en donde $a > 0$ es un parámetro. Sea c una constante tal que $0 < c < a$.

   a) Encuentre $f(x \,|\, X < c)$.

   b) Encuentre $F(x \,|\, X < c)$.

   c) Grafique en el mismo plano $f(x)$ y $f(x \,|\, X < c)$.

d) Grafique en el mismo plano $F(x)$ y $F(x \,|\, X < c)$.

e) Encuentre $E(X)$ y $E(X \,|\, X < c)$.

17. *Distribución uniforme.*
Sea X una variable aleatoria con distribución unif$(0, a)$, en donde $a > 0$ es un parámetro. Sea c una constante tal que $0 < c < a$.

a) Encuentre $f(x \,|\, X > c)$.

b) Encuentre $F(x \,|\, X > c)$.

c) Grafique en el mismo plano $f(x)$ y $f(x \,|\, X > c)$.

d) Grafique en el mismo plano $F(x)$ y $F(x \,|\, X > c)$.

e) Encuentre $E(X)$ y $E(X \,|\, X > c)$.

18. *Distribución exponencial.*
Sea X una variable aleatoria con distribución exp(λ). Sean a y b dos constantes tales que $0 \leqslant a < b \leqslant \infty$.

a) Demuestre que

$$F(x \,|\, a < X \leqslant b) = \begin{cases} 0 & \text{si } \ x \leqslant a, \\ \dfrac{1 - e^{-\lambda(x-a)}}{1 - e^{-\lambda(b-a)}} & \text{si } \ a < x \leqslant b, \\ 1 & \text{si } \ x > b. \end{cases}$$

b) Grafique en el mismo plano $F(x)$ y $F(x \,|\, a < X \leqslant b)$.

c) Demuestre que

$$f(x \,|\, a < X \leqslant b) = \begin{cases} \dfrac{\lambda\, e^{-\lambda(x-a)}}{1 - e^{-\lambda(b-a)}} & \text{si } \ a < x < b, \\ 0 & \text{en otro caso.} \end{cases}$$

d) Grafique en el mismo plano $f(x)$ y $f(x \,|\, a < X \leqslant b)$.

e) Demuestre que

$$E(X \mid \, a < X \leqslant b) = \frac{(a + 1/\lambda)e^{-\lambda a} - (b + 1/\lambda)e^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}}.$$

f) Demuestre que $\lim_{a\to 0}\lim_{b\to\infty} E(X \mid a < X \leqslant b) = 1/\lambda$.

Todas las fórmulas anteriores se reducen a expresiones más sencillas para el caso de truncamiento por abajo ($b = \infty$), y para el truncamiento por arriba ($a = 0$).

19. *Distribución uniforme discreta.*
Sea X una variable aleatoria discreta con distribución unif$\{1, \ldots, N\}$. Sean n y m dos números enteros tales que $1 \leqslant n \leqslant m \leqslant N$. Encuentre las funciones:

   a) $f(x \mid n \leqslant X \leqslant m)$.
   b) $F(x \mid n \leqslant X \leqslant m)$.

20. *Distribución geométrica.*
Sea X una variable aleatoria discreta con función de probabilidad

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

en donde $0 < p < 1$ es un parámetro. Sean $1 \leqslant n \leqslant m$ dos enteros. Encuentre las funciones:

   a) $f(x \mid n \leqslant X \leqslant m)$.
   b) $F(x \mid n \leqslant X \leqslant m)$.

21. *Distribución normal.*
Suponga que un tiempo de vida X se puede modelar mediante la distribución N(μ, σ^2) truncada al intervalo (a, ∞), con $a \geqslant 0$ constante.

   a) Encuentre $f(x \mid X > a)$.
   b) Grafique en el mismo plano $f(x)$ y $f(x \mid X > a)$
   c) Encuentre $E(X \mid X > a)$.
   d) Encuentre Var$(X \mid X > a)$.

Capítulo 2

Funciones básicas

En este capítulo estudiaremos algunas funciones asociadas a una variable aleatoria X, discreta o continua, que puede representar un tiempo de vida en el caso cuando es positiva. La distribución de esta variable aleatoria se puede caracterizar, de manera equivalente, a través de cualquiera de las siguientes funciones:

- Función de distribución

  Se define como $F(x) := P(X \leqslant x)$, para $-\infty < x < \infty$. De manera equivalente, tenemos la función de probabilidad o de densidad denotada por $f(x)$. En general, estas son las funciones más comunes para caracterizar la distribución de una variable aleatoria cualquiera, no necesariamente aquellas que representan un tiempo de vida.

- Función de supervivencia

  Se define como $S(x) := P(X > x)$, para $-\infty < x < \infty$. En el caso de tiempos de vida, esta función indica la probabilidad de que el tiempo de vida X exceda (o que el individuo sobreviva) al tiempo $x \geqslant 0$. Es claro que se cumple la identidad $S(x) = 1 - F(x)$.

- Función de riesgo

  Se define como $\lambda(x) := f(x)/S(x)$ en el caso continuo, para valores de x tales que $S(x) > 0$. En el caso de tiempos de vida, esta función está relacionada con la probabilidad condicional de que un individuo que ha alcanzado la edad x fallezca en el siguiente instante de tiempo.

- Función tiempo promedio de vida residual

  Se define como $R(x) := E(X - x \,|\, X > x)$, para $-\infty < x < \infty$, bajo la hipótesis de que esta esperanza sea finita. En el caso de tiempos de vida, esta función es el tiempo de vida promedio que le resta por vivir a un individuo de edad $x \geqslant 0$.

A continuación estudiaremos estas funciones con mayor detalle y demostraremos que son equivalentes. Según convenga, se utiliza cualquiera de ellas en el estudio de los tiempos de vida. La ventaja radica en que es suficiente estimar cualquiera de estas funciones para que la distribución del tiempo de vida quede especificada. Por otro lado, para ilustrar la forma de encontrar estas funciones, supondremos que los tiempos de vida tienen distribuciones de probabilidad conocidas, aunque, como hemos indicado antes, ese no es el caso en la mayoría de las situaciones prácticas.

Sea $X > 0$ una variable aleatoria, discreta o continua, que modela un tiempo de vida. En general, la distribución de X es desconocida y nos interesa estimarla a la luz de una serie de observaciones, posiblemente censuradas. Supondremos que se cuenta con una unidad de medida del tiempo. Esta unidad puede ser un minuto, una hora, un día, una semana, un mes, un año, etc. Si los tiempos de vida se pueden medir de manera continua, entonces el conjunto de valores de estas variables aleatorias, en general, es el intervalo $(0, \infty)$. Por otro lado, si el tiempo se mide de manera discreta, entonces los valores son $1, 2, \ldots$

2.1. Función de distribución

En la teoría general de la probabilidad y la estadística, la función de distribución es la forma más usual para caracterizar a la distribución de una variable aleatoria X. Recordemos que esta importante función se define para cualquier número real x de la siguiente forma

$$F(x) = P(X \leqslant x), \quad -\infty < x < \infty.$$

Recordemos también que esta función satisface las siguientes propiedades:

1. $\lim_{x \to \infty} F(x) = 1$.

2. $\lim_{x \to -\infty} F(x) = 0$.

3. Si $x_1 \leqslant x_2$ entonces $F(x_1) \leqslant F(x_2)$.

4. $F(x)$ es continua por la derecha, es decir, $F(x+) = F(x)$.

La expresión $F(x+)$ denota el límite por la derecha de la función F en el punto x, es decir, se define

$$F(x+) := \lim_{\epsilon \searrow 0} F(x + \epsilon).$$

La última propiedad enunciada para la función de distribución establece que la continuidad por la derecha de la función $F(x)$ se cumple siempre para cualquier punto x, pues $F(x+)$ coincide con $F(x)$. En este contexto, la expresión "$x+$" denota un límite o aproximación por la derecha y no un dato censurado por la derecha.

De manera análoga, se define el límite por la izquierda de la función F en el punto x como

$$F(x-) := \lim_{\epsilon \searrow 0} F(x - \epsilon).$$

En general, para una función de distribución cualquiera, no se cumple que $F(x-) = F(x)$ para todo valor x, es decir, no se cumple la continuidad por la izquierda en todo punto x. La expresión '$x-$" usada aquí denota un límite o aproximación por la izquierda y no un dato censurado por la izquierda.

- Caso continuo

  Cuando una variable aleatoria X es absolutamente continua, su función de distribución se puede escribir como

  $$F(x) = \int_{-\infty}^{x} f(u)\, du, \qquad -\infty < x < \infty,$$

  en donde a la función $f(x) \geqslant 0$ se le llama función de densidad. La función $f(x)$ es otra representación equivalente de la distribución de X.

- Caso discreto

  Cuando X es discreta con valores $x_1, x_2, \ldots$, se define la función de probabilidad como

$$f(x) := \begin{cases} P(X = x_i) & \text{si } \ x = x_1, x_2, \ldots \\ 0 & \text{en otro caso,} \end{cases}$$

  y la función de distribución se puede expresar como

$$F(x) = \sum_{x_i \leqslant x} f(x_i), \qquad -\infty < x < \infty.$$

Los conceptos anteriores se aplican igualmente al caso particular de nuestro interés, es decir, cuando la variable aleatoria X es positiva y representa un tiempo de vida. El lector puede encontrar una exposición de la función de distribución y sus propiedades en el texto de A. Gut [25].

2.2. Función de supervivencia

Sea X una variable aleatoria positiva que modela un tiempo de vida y sea $F(x) = P(X \leqslant x)$ su función de distribución. A la función definida como la probabilidad complementaria a $F(x)$ se le llama función de supervivencia.

Definición 2.1 *La función de supervivencia de un tiempo de vida X es*

$$S(x) := P(X > x) = 1 - F(x), \quad -\infty < x < \infty.$$

Es decir, la función de supervivencia es, simplemente, la cola de la distribución del tiempo de vida. Para cada $x > 0$, esta función proporciona la probabilidad de que la muerte o falla se presente después del valor x. Equivalentemente, es la probabilidad que el individuo sobreviva a la edad o tiempo x. De esta última interpretación surge el nombre de función de supervivencia. A la gráfica de la función $S(x)$ se le llama curva de supervivencia.

Es evidente que existe una equivalencia entre la función de distribución y la función de supervivencia. De una se puede obtener fácilmente la otra. En consecuencia, la función $S(x)$ también caracteriza de manera única a la distribución de la variable aleatoria. Las propiedades que satisface una función de supervivencia son análogas a las de una función de distribución y se enuncian a continuación. Su demostración se deja como ejercicio para el lector y para desarrollar los detalles se pueden utilizar las propiedades básicas de $F(x)$, las cuales fueron recordadas en la página 36.

Proposición 2.1 *Toda función de supervivencia $S(x)$ satisface las siguientes propiedades:*

1. $S(0) = 1$.
2. $\lim_{x\to\infty} S(x) = 0$.
3. *Si* $x_1 \leqslant x_2$ *entonces* $S(x_1) \geqslant S(x_2)$.
4. $S(x)$ *es continua por la derecha, es decir,* $S(x+) = S(x)$.

En palabras, toda función de supervivencia inicia en el valor 1 al tiempo inicial $x = 0$, y paulatinamente va decreciendo hacia el valor 0. La continuidad por la derecha se hereda de la misma propiedad para la función de distribución. Recordemos que, en este contexto, la expresión "$x+$" denota un límite o aproximación por la derecha y no un dato censurado por la derecha. Sin tener un tiempo de vida X de por medio, se puede definir una función de supervivencia a partir de las cuatro propiedades anteriores. Esto se establece a continuación.

Definición 2.2 *Una función $S(x) : \mathbb{R} \to [0, 1]$ es de supervivencia si cumple las propiedades (1), (2), (3) y (4) de la Proposición 2.1.*

El siguiente enunciado muestra la relación entre la función de supervivencia

$S(x)$ y la función de densidad $f(x)$ en el caso continuo, o de probabilidad en el caso discreto.

Proposición 2.2 *Sea X una variable aleatoria que modela un tiempo de vida y sea $f(x)$ su función de densidad o de probabilidad. Sea $S(x)$ su función de supervivencia.*

1. *Si X es continua, entonces*

$$\begin{aligned} f(x) &= -\frac{d}{dx} S(x), \qquad 0 < x < \infty, \\ S(x) &= \int_x^\infty f(u)\, du, \qquad 0 < x < \infty. \end{aligned}$$

2. *Si X es discreta, entonces*

$$\begin{aligned} f(x) &= S(x-) - S(x), \qquad 0 < x < \infty, \\ S(x) &= \sum_{u>x} f(u), \qquad 0 < x < \infty. \end{aligned}$$

Demostración. Los resultados se obtienen de manera inmediata al substituir $S(x)$ por $1 - F(x)$. ∎

Ejemplo 2.1 (Tiempo de vida exponencial)
Supongamos que X es un tiempo de vida continuo con distribución exp(λ). Es inmediato comprobar que

$$S(x) = e^{-\lambda x}, \quad 0 < x < \infty.$$

La gráfica de esta función se muestra en la Figura 2.1. Se aprecia el comportamiento decreciente, para $x \geqslant 0$, que debe presentar toda función de supervivencia. Conforme x crece, la probabilidad de sobrevivir a la edad x decrece. ∎

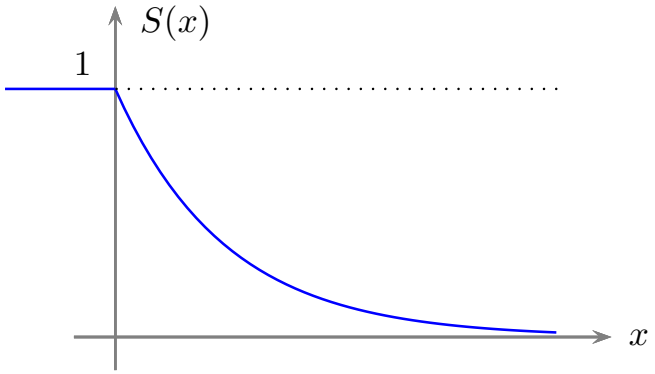

Figura 2.1: Función de supervivencia exponencial.

Ejemplo 2.2 (Tiempo de vida geométrico)
Como un ejemplo de un tiempo de vida discreto, suponga que X tiene la siguiente distribución geométrica de parámetro p, con $0 < p < 1$,

$$f(x) = \begin{cases} p(1-p)^{x-1} & si \quad x = 1, 2, \ldots \\ 0 & en\ otro\ caso. \end{cases}$$

Se puede comprobar que

$$S(x) = (1-p)^x, \quad x = 1, 2, \ldots$$

Por simplicidad en la escritura, se ha especificado la función $S(x)$ únicamente en los puntos de discontinuidad; en realidad está definida para $x \in (0, \infty)$. La gráfica de esta función se muestra en la Figura 2.2. Se observa el comportamiento decreciente en forma de escalera y se enfatiza la continuidad por la derecha de $S(x)$. Observe también que la función $S(x)$ no es continua por la izquierda en los puntos de discontinuidad. ∎

Dado que estamos considerando que los tiempos de vida son positivos, en lo sucesivo especificaremos a las funciones de supervivencia $S(x)$ únicamente para valores $x > 0$. Pueden proporcionarse más ejemplos de modelos paramétricos para tiempos de vida (variables aleatorias positivas), sin embargo, para algunos de ellos la expresión para $S(x)$ podría no ser sencilla de escribir. En la Sección 3.1 titulada "Distribuciones paramétricas", que aparece en

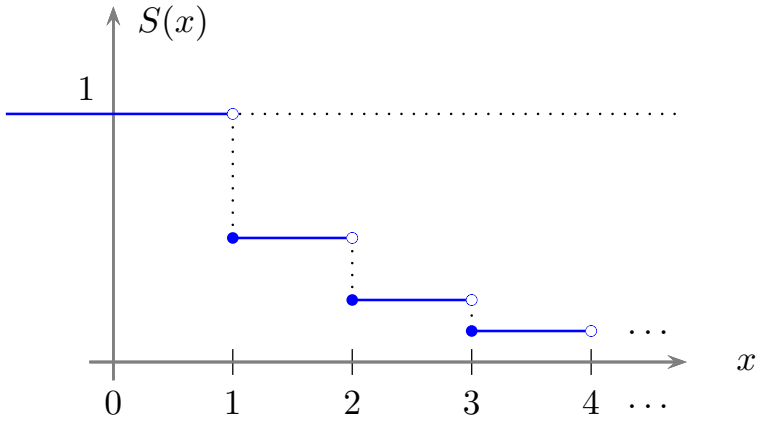

Figura 2.2: Función de supervivencia geométrica.

la página 99, se hace una revisión de las características de algunos de estos modelos paramétricos. En la Tabla 2.1 se muestra la expresión de la función de supervivencia de estos modelos.

Tiempo de vida	Función de supervivencia
unif(a, b)	$S(x) = (b - x)/(b - a), \quad a < x < b$
exp(λ)	$S(x) = \exp(-\lambda x), \quad x > 0$
Gompertz(b, c)	$S(x) = \exp(\frac{b}{c}(1 - e^{cx})), \quad x > 0$
Makeham(a, b, c)	$S(x) = \exp(\frac{b}{c}(1 - e^{cx}) - ax), \quad x > 0$
Weibull(α, λ)	$S(x) = \exp(-\lambda x^{\alpha}), \quad x > 0$
gama(γ, λ)	$S(x) = \frac{1}{\Gamma(\gamma)} \int_{\lambda x}^{\infty} u^{\gamma-1} e^{-u}\, du, \quad x > 0$
lognormal(μ, σ^2)	$S(x) = 1 - \Phi((\log x - \mu)/\sigma), \quad x > 0$
Pareto(α, γ)	$S(x) = (\gamma/x)^{\alpha}, \quad x \geqslant \gamma$

Tabla 2.1: Funciones de supervivencia de algunas distribuciones.

Función de supervivencia truncada

Veremos ahora la forma en la que cambia una función de supervivencia bajo truncamiento.

Proposición 2.3 *Sea X un tiempo de vida con función de supervivencia $S(x)$. Sean $0 \leqslant a < b$ dos constantes tales que $P(a < X \leqslant b) > 0$. La función de supervivencia truncada al intervalo $(a, b]$ es*

$$S(x \,|\, a < X \leqslant b) = \begin{cases} 1 & si \quad x \leqslant a, \\ \dfrac{S(x) - S(b)}{S(a) - S(b)} & si \quad a < x \leqslant b, \\ 0 & si \quad x > b. \end{cases}$$

Demostración. Puede comprobarse sin dificultad que, para $x \leqslant a$ y $x > b$, los valores de esta función son 1 y 0, respectivamente. Por otro lado, para $x \in (a, b]$,

$$\begin{aligned} S(x \,|\, a < X \leqslant b) &= P(X > x \,|\, a < X \leqslant b) \\ &= \frac{P(x < X \leqslant b)}{P(a < X \leqslant b)} \\ &= \frac{F(b) - F(x)}{F(b) - F(a)} \\ &= \frac{S(x) - S(b)}{S(a) - S(b)}. \end{aligned}$$

∎

Tomando $a = 0$ se puede encontrar una expresión reducida para $S(x \,|\, X \leqslant b)$. Por otro lado, para $a \geqslant 0$ fijo, haciendo $b \to \infty$ se puede encontrar también una expresión más sencilla para $S(x \,|\, X > a)$.

Puede comprobarse que las funciones truncadas $S(x \,|\, a < X \leqslant b)$, $S(x \,|\, X \leqslant b)$ y $S(x \,|\, X > a)$ son, efectivamente, funciones de supervivencia, es decir, satisfacen las cuatro propiedades que aparecen en la Definición 2.2 .

Fórmula del producto

En esta sección vamos a considerar algunas probabilidades condicionales de supervivencia. En particular, encontraremos una fórmula que resultará útil en la estimación de la función de supervivencia y que lleva el nombre de fórmula del producto. Empezaremos con tres resultados fáciles de comprobar.

Proposición 2.4 *Sea X un tiempo de vida con función de supervivencia $S(x)$. Sea $x \geqslant 0$ un tiempo o edad tal que $S(x) > 0$. Entonces, para cualquier $t > 0$,*

1. $P(x < X \leqslant x + t) = S(x) - S(x + t)$.

2. $P(x < X \leqslant x + t \,|\, X > x) = \dfrac{S(x) - S(x + t)}{S(x)}$.

3. $P(X > x + t \,|\, X > x) = \dfrac{S(x + t)}{S(x)}$.

Demostración.

1. $$\begin{aligned} P(x < X \leqslant x + t) &= P(X \leqslant x + t) - P(X \leqslant x) \\ &= F(x + t) - F(x) \\ &= [1 - S(x + t)] - [1 - S(x)] \\ &= S(x) - S(x + t). \end{aligned}$$

2. $$\begin{aligned} P(x < X \leqslant x + t \,|\, X > x) &= \frac{P(x < X \leqslant x + t)}{P(X > x)} \\ &= \frac{F(x + t) - F(x)}{1 - F(x)} \\ &= \frac{S(x) - S(x + t)}{S(x)}. \end{aligned}$$

3. $$P(X > x+t \,|\, X > x) = \frac{P(X > x+t)}{P(X > x)} = \frac{S(x+t)}{S(x)}.$$

■

La primera fórmula corresponde a la probabilidad de que un tiempo de vida concluya en el intervalo $(x, x+t]$ en términos de la función de supervivencia. La segunda fórmula es la probabilidad del mismo evento que aparece en la fórmula anterior pero condicionada al evento $(X > x)$, es decir, cuando el individuo ha sobrevivido al tiempo o edad x, suponiendo que la probabilidad de este último evento es estrictamente positiva. Puede comprobarse que esta probabilidad condicional es mayor o igual a la probabilidad de la primera fórmula, es decir,

$$P(x < X \leqslant x+t \,|\, X > x) \geqslant P(x < X \leqslant x+t).$$

La tercera fórmula corresponde a la probabilidad de que un individuo que ha alcanzado la edad x, sobreviva t unidades de tiempo adicionales. Observe que los eventos de las últimas dos probabilidades son complementarios, es decir, la suma de las probabilidades condicionales es 1.

A partir de lo anterior, tenemos la siguiente descomposición que es de utilidad para estimar la función de supervivencia.

Proposición 2.5 (Fórmula del producto) *Sea X un tiempo de vida con función de supervivencia $S(x)$. Sean $0 = x_0 < x_1 < x_2 < \cdots < x_n$ tiempos dados tales que $S(x_{n-1}) > 0$. Entonces*

$$S(x_n) = \prod_{k=0}^{n-1} \Big[1 - P(x_k < X \leqslant x_{k+1} \,|\, X > x_k)\Big]. \tag{2.1}$$

Demostración. Se inicia escribiendo a $S(x_n)$ como un producto en donde

se verifican varias cancelaciones de términos.

$$\begin{aligned} S(x_n) &= \prod_{k=0}^{n-1} \frac{S(x_{k+1})}{S(x_k)} \\ &= \prod_{k=0}^{n-1} P(X > x_{k+1} \,|\, X > x_k) \qquad (2.2) \\ &= \prod_{k=0}^{n-1} \Big[1 - P(X \leqslant x_{k+1} \,|\, X > x_k)\Big] \\ &= \prod_{k=0}^{n-1} \Big[1 - P(x_k < X \leqslant x_{k+1} \,|\, X > x_k)\Big]. \end{aligned}$$

■

Este resultado tiene una interpretación sencilla: la probabilidad de supervivencia al tiempo x_n es igual al producto de las probabilidades condicionales de que no se presente falla o muerte en ninguno de los intervalos previos, dado que se ha llegado con vida al inicio de cada intervalo. Tal vez esto pueda entenderse mejor a partir de la expresión (2.2).

En el Capítulo 4 "Modelos no paramétricos", veremos que se pueden proponer estimaciones para las probabilidades condicionales que aparecen como factores en la fórmula del producto (2.1) y, por lo tanto, se puede generar una estimación para el valor de la función de supervivencia en el tiempo x_n.

Para finalizar, observemos que todos los resultados que aparecen en esta sección se cumplen sin importar si el tiempo de vida es discreto o continuo.

2.3. Función de riesgo

Esta es otra función que se le asocia a cada tiempo de vida. Es equivalente a la función de supervivencia y resulta también conveniente para la modelación y análisis de los tiempos de vida. Separaremos nuestra exposición en el caso continuo y discreto.

Función de riesgo continua

Definición 2.3 *Sea X un tiempo de vida continuo con función de densidad $f(x)$ y función de supervivencia $S(x)$. La función de riesgo es*

$$\lambda(x) := \frac{f(x)}{S(x)}, \quad \text{para } x > 0 \text{ tal que } S(x) > 0.$$

Esta función representa la intensidad o propensión para concluir el tiempo de vida en un instante inmediatamente después de x cuando el fallecimiento o falla no ha ocurrido hasta ese momento x. Esta interpretación se obtiene del siguiente argumento: La probabilidad de que el tiempo de vida X tome un valor en el intervalo infinitesimal $(x, x + \Delta x)$, dado que al tiempo x no se ha presentado la falla o muerte, es

$$\begin{aligned} P(X \in (x, x + \Delta x) \,|\, X > x) &= \frac{P(X \in (x, x + \Delta x))}{P(X > x)} \\ &\approx \frac{f(x)\,\Delta x}{S(x)} \\ &= \lambda(x)\,\Delta x. \end{aligned}$$

En el límite cuando la longitud del intervalo infinitesimal Δx se hace cero, la probabilidad condicional considerada es la función de riesgo.

La función de riesgo $\lambda(x)$ recibe diferentes nombres en las distintas áreas de aplicación. Por ejemplo, se le llama fuerza de mortalidad (*force of mortality*) en demografía, se conoce también como tasa de falla condicional (*conditional failure rate*) en teoría de la confiabilidad, también se le llama tasa de riesgo (*hazard rate*) en teoría del riesgo, y también tasa de intensidad (*intensity rate*) en la teoría de los procesos estocásticos. La notación también puede cambiar de una disciplina a otra, por ejemplo, se le puede escribir como la función $h(x)$.

Ejemplo 2.3 (Tiempo de vida exponencial)
Sea X un tiempo de vida con distribución exp(λ). Entonces X tiene función de riesgo constante pues, para cualquier $x > 0$,

$$\lambda(x) = \frac{f(x)}{S(x)} = \frac{\lambda\, e^{-\lambda x}}{e^{-\lambda x}} = \lambda.$$

Es decir, a cualquier edad alcanzada x, la propensión o intensidad a morirse es constante igual a λ. En otras palabras, para tiempos de vida exponenciales, la vejez no incrementa la mortalidad. Esta es otra manera de expresar la propiedad de pérdida de memoria de la distribución exponencial. Puede comprobarse que esta distribución es la única distribución continua con soporte $(0, \infty)$ y con función de riesgo constante. Véase el Ejercicio 56. ∎

Ejemplo 2.4 (Tiempo de vida uniforme)
Sea X un tiempo de vida con distribución unif(a, b), con $0 < a < b$ constantes. Entonces $f(x) = 1/(b-a)$ para $a < x < b$, y puede comprobarse que $S(x) = (b-x)/(b-a)$ en el mismo intervalo. Por lo tanto,

$$\lambda(x) = \frac{f(x)}{S(x)} = \frac{1}{b-x}, \quad \text{para} \quad a < x < b.$$

La gráfica de esta función se muestra en la Figura 2.3. Observe que una persona de edad cercana al valor b tiene una propensión a morirse casi infinita pues nadie puede llegar a tener edad b. ∎

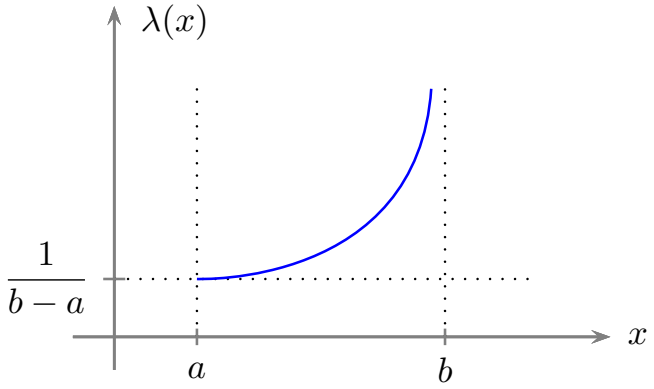

Figura 2.3: Función de riesgo de un tiempo de vida unif(a, b).

Como puede observarse en los ejemplos anteriores, la gráfica de la función de riesgo $\lambda(x)$ permite tener una idea cualitativa de las edades o tiempos de mayor y menor mortalidad.

En el caso continuo, la función de riesgo y la función de supervivencia están relacionadas como indican las expresiones siguientes, en donde la función $\log x$ denota logaritmo natural.

Proposición 2.6 *Sea X un tiempo de vida continuo con función de supervivencia $S(x)$ y función de riesgo $\lambda(x)$. Para valores x tal que $S(x) > 0$ se cumple:*

1. $\lambda(x) = -\dfrac{d}{dx}\log S(x) = -\dfrac{S'(x)}{S(x)}$.
2. $S(x) = \exp\left(-\displaystyle\int_0^x \lambda(u)\,du\right)$.

Estas igualdades son fáciles de comprobar y se deja hacer los detalles como ejercicio. Establecen la equivalencia entre $S(x)$ y $\lambda(x)$. Es decir, a partir de una de estas funciones, se determina, de manera única, la otra función a través de las fórmulas indicadas.

Ejemplo 2.5 *Como un caso general de función de riesgo continua tenemos a la función de riesgo Rayleigh, definida como el polinomio*

$$\lambda(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \quad \text{para} \quad x > 0, \tag{2.3}$$

en donde los coeficientes $a_0, a_1, \ldots, a_n$ son tales que $\lambda(x) > 0$ para todo $x > 0$. Usando la segunda fórmula de la Proposición 2.6, puede comprobarse con facilidad que la función de supervivencia asociada es

$$S(x) = \exp\{-a_0 x - a_1 x^2/2 - \cdots - a_n x^{n+1}/(n+1)\}, \quad \text{para} \quad x > 0.$$

Como casos particulares de (2.3), tenemos las siguientes funciones de riesgo: para $a > 0$ y $b \geqslant 0$,

a) $\lambda(x) = a, \quad x > 0.$

b) $\lambda(x) = ax + b, \quad x > 0.$

c) $\lambda(x) = ax^2 + b, \quad x > 0.$

d) $\lambda(x) = a(1+x)^2 + b, \quad x > 0.$

e) $\lambda(x) = \dfrac{1}{1+x}, \quad x > 0.$

Para cada una de estas funciones puede encontrarse la función de supervivencia asociada $S(x)$. ∎

Ejemplo 2.6 *Como un segundo caso general tenemos que si* $\alpha > 0$ *y* $\beta > 0$ *son dos constantes, entonces la función de riesgo exponencial se define como*

$$\lambda(x) = \alpha\, e^{\beta x}, \quad \textit{para} \quad x > 0.$$

Usando nuevamente la segunda fórmula de la Proposición 2.6, puede comprobarse que la función de supervivencia asociada es

$$S(x) = \exp\{\, \frac{\alpha}{\beta}(1 - e^{\beta x})\,\}, \quad \textit{para} \quad x > 0.$$

∎

En la Tabla 2.2 se muestran las expresiones de las funciones de riesgo de los modelos paramétricos que se revisan en la Sección 3.1 "Distribuciones paramétricas".

Caracterización de una función de riesgo

Consideremos el caso de un tiempo de vida con función de supervivencia tal que $S(x) > 0$, para todo $x > 0$. Para que una función continua $\lambda(x)$ definida sobre el intervalo $(0, \infty)$ sea la función de riesgo asociada, debe satisfacer la igualdad $\lambda(x) = -S'(x)/S(x)$, para $x > 0$ tal que $S(x) > 0$. Es decir,

$$\lambda(x)\, S(x) + S'(x) = 0.$$

Tiempo de vida	Función de riesgo
unif(a, b)	$\lambda(x) = 1/(b-x), \quad a < x < b$
exp(λ)	$\lambda(x) = \lambda, \quad x > 0$
Gompertz(b, c)	$\lambda(x) = b\, e^{cx}, \quad x > 0$
Makeham(a, b, c)	$\lambda(x) = a + b\, e^{cx}, \quad x > 0$
Weibull(α, λ)	$\lambda(x) = \alpha\, \lambda\, x^{\alpha-1}, \quad x > 0$
gama(γ, λ)	Fórmula general $\lambda(x) = f(x)/S(x), \quad x > 0$
lognormal(μ, σ^2)	Fórmula general $\lambda(x) = f(x)/S(x), \quad x > 0$
Pareto(α, γ)	$\lambda(x) = \alpha/x, \quad x \geqslant \gamma$

Tabla 2.2: Algunas distribuciones y sus funciones de riesgo.

Multiplicando esta ecuación por el factor integrante $\exp\{\int_0^x \lambda(u)\, du\}$, se encuentra que la solución para $S(x)$ es

$$S(x) = c\, \exp\{-\int_0^x \lambda(u)\, du\},$$

en donde $c > 0$ es una constante. Las propiedades que satisface $S(x)$ se pueden trasladar ahora a $\lambda(x)$. La condición $S(0) = 1$ implica que $c = 1$. La condición de que $S(x)$ es decreciente implica que $\lambda(x) \geqslant 0$. Se cumple que $S(0) = 1$ y la condición $S(\infty) = 0$ implica que la integral anterior es infinita cuando $x \to \infty$. Esto lleva a la siguiente definición.

Definición 2.4 *Una función integrable $\lambda(x)$ definida sobre $(0, \infty)$ y con soporte $\{x : \lambda(x) > 0\}$ no acotado es una función de riesgo si cumple las condiciones:*

1. $\lambda(x) \geqslant 0$.
2. $\int_0^\infty \lambda(x)\, dx = \infty$.

Se debe señalar que la definición anterior sólo aplica a funciones $\lambda(x)$ definidas sobre $(0, \infty)$. Pueden existir, sin embargo, funciones de riesgo definidas sobre intervalos de tiempo acotados, por ejemplo, la que se presenta en el Ejemplo 2.4 sobre la distribución unif(a, b).

Función de riesgo acumulado (caso continuo)

Para tiempos de vida continuos, a la integral que aparece en la segunda fórmula de la Proposición 2.6 se le llama función de riesgo acumulado. La definiremos a continuación.

Definición 2.5 *Sea X un tiempo de vida continuo con función de riesgo $\lambda(x)$ definida en un intervalo $(0, a)$ para algún valor a tal que $0 < a \leqslant \infty$. La función de riesgo acumulado es*

$$\Lambda(x) := \int_0^x \lambda(u)\, du, \quad \text{para} \quad 0 \leqslant x < a. \tag{2.4}$$

Es claro que las funciones $\lambda(x)$ y $\Lambda(x)$ son equivalentes en el sentido de que a partir de una de ellas se determina de manera única la otra. Para verificar esto se puede usar (2.4), o su expresión equivalente $\Lambda'(x) = \lambda(x)$.

Ejemplo 2.7 (Tiempo de vida exponencial)
Para un tiempo de vida X con distribución exp(λ), hemos comprobado que la función de riesgo es $\lambda(x) = \lambda$, para $x > 0$. Por lo tanto, la función de riesgo acumulado es $\Lambda(x) = \lambda x$, para $x > 0$. ∎

Ejemplo 2.8 (Tiempo de vida uniforme)
Para un tiempo de vida X con distribución unif(a, b), hemos encontrado que la función de riesgo es $\lambda(x) = 1/(b-x)$, para $a < x < b$. Puede comprobarse que la función de riesgo acumulado es

$$\Lambda(x) = \int_a^x \frac{1}{b-u}\, du = \log \frac{b-a}{b-x}, \quad a < x < b.$$

∎

A partir de lo anterior, es inmediato verificar la siguiente identidad.

Proposición 2.7 *Para tiempos de vida continuos,*

$$\Lambda(x) = -\log S(x), \quad \text{para } x > 0 \text{ tal que } S(x) > 0.$$

Demostración. Por la segunda fórmula de la Proposición 2.6,

$$S(x) = \exp\{-\int_0^x \lambda(u)\, du\} = \exp\{-\Lambda(x)\}.$$

Resolviendo para $\Lambda(x)$ se obtiene el resultado. Recuerde que log indica logaritmo natural. ∎

Para un modelo paramétrico cualquiera, la facilidad para encontrar la función de riesgo acumulado $\Lambda(x)$ usando la fórmula de la última proposición dependerá de la disponibilidad de una expresión sencilla o manejable para $S(x)$.

Función de riesgo discreta

A continuación estudiaremos la función de riesgo para tiempos de vida discretos. La definición e interpretación es similar al caso continuo.

Definición 2.6 *Sea X un tiempo de vida discreto con valores $0 < x_1 < x_2 < \cdots$, con función de probabilidad $f(x_i)$ y función de supervivencia $S(x_i)$. La función de riesgo asociada es*

$$\lambda(x_i) := \frac{f(x_i)}{S(x_{i-1})}, \quad \text{para } \ i = 1, 2, \ldots$$

Observe que debemos definir $x_0 := 0$, $S(0) := 1$ y que, además, el cociente

anterior tiene sentido sólo cuando $S(x_{i-1}) > 0$. La función de riesgo discreta $\lambda(x_i)$ también se puede escribir como

$$\lambda(x_i) = \frac{P(X = x_i)}{P(X \geqslant x_i)}, \quad \text{para } \ i = 1, 2, \ldots$$

Esta expresión tiene un aspecto más similar al caso continuo. La aparente diferencia en la definición de función de riesgo para los casos continuo y discreto se resuelve al recordar que si X es una variable aleatoria continua, entonces $S(x) = P(X > x) = P(X \geqslant x)$. Con esto se compatibilizan ambas definiciones.

En el caso discreto, a partir de la definición, es evidente que la función de riesgo $\lambda(x_i)$ representa la probabilidad de que un individuo fallezca a edad x_i dado que ha sobrevivido a edad x_{i-1}, de modo que el fallecimiento ocurrirá en alguna de las edades $x_i, x_{i+1}, \ldots$ Por simplicidad, a menudo supondremos que los valores de un tiempo de vida discreto son los números naturales $1, 2, \ldots$ o un subconjunto de éstos.

Ejemplo 2.9 (Tiempo de vida geométrico)
Sea X un tiempo de vida discreto con la distribución geométrica que aparece abajo, en donde $0 < p < 1$.

$$f(x) = \begin{cases} p(1-p)^{x-1} & \textit{si } \ x = 1, 2, \ldots, \\ 0 & \textit{en otro caso.} \end{cases}$$

Puede comprobarse que $S(x) = (1-p)^x$, para $x = 0, 1, \ldots$ Entonces X tiene función de riesgo constante pues, para cualquier $x = 1, 2, \ldots$

$$\lambda(x) = \frac{f(x)}{S(x-1)} = \frac{p(1-p)^{x-1}}{(1-p)^{x-1}} = p.$$

Es decir, a cualquier edad alcanzada x, la propensión o intensidad a morirse es constante igual a p. Esta es la misma situación que en el caso de tiempos de vida exponenciales pero ahora en el caso discreto: para tiempos de vida geométricos, la vejez no incrementa la mortalidad. Esta es otra forma de expresar la propiedad de pérdida de memoria de la distribución geométrica. Puede comprobarse que esta distribución es la única distribución discreta con soporte $\{1, 2, \ldots\}$ y con función de riesgo constante. Véase el Ejercicio 56.

■

Ejemplo 2.10 (Tiempo de vida uniforme discreto)
Sean $0 < x_1 < x_2 < x_3$ tres números y sea X un tiempo de vida discreto con distribución unif$\{x_1, x_2, x_3\}$. Es fácil comprobar que

$$\lambda(x) = \begin{cases} 1/3 & si \quad x = x_1, \\ 1/2 & si \quad x = x_2, \\ 1 & si \quad x = x_3. \end{cases}$$

Esta es una función creciente cuya gráfica se muestra en la Figura 2.4. Se observa que $\lambda(x_3) = 1$, es decir, la probabilidad de que una persona que ha sobrevivido a edad x_2 muera a edad x_3 es 1, pues nadie puede sobrepasar la edad máxima x_3. ∎

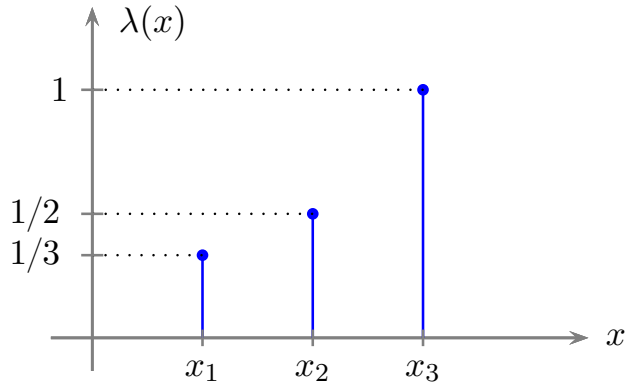

Figura 2.4: Función de riesgo discreta.

La relación entre la función de riesgo, la función de supervivencia y la función de probabilidad para un tiempo de vida discreto se establecen en los siguientes resultados.

Proposición 2.8 *Sea X un tiempo de vida discreto con valores $0 < x_1 < x_2, \ldots$ con función de supervivencia $S(x_i)$ y función de riesgo $\lambda(x_i)$. Entonces*

1. $P(X = x_i) = S(x_{i-1}) - S(x_i)$, *para* $i=1,2,\ldots$
2. $\lambda(x_i) = \dfrac{S(x_{i-1}) - S(x_i)}{S(x_{i-1})} = 1 - \dfrac{S(x_i)}{S(x_{i-1})}$, *para* $i=1,2,\ldots$
3. $S(x) = \displaystyle\prod_{x_i \leqslant x} \frac{S(x_i)}{S(x_{i-1})} = \prod_{x_i \leqslant x} [1 - \lambda(x_i)]$, *para* $x>0$.
4. $S(x_i) = [1 - \lambda(x_i)]\ S(x_{i-1})$, *para* $i=1,2,\ldots$

Demostración. Para $i = 1, 2, \ldots$

1. $P(X = x_i) = P(X > x_{i-1}) - P(X > x_i) = S(x_{i-1}) - S(x_i)$.

2. $\lambda(x_i) = \dfrac{f(x_i)}{S(x_{i-1})} = \dfrac{P(X = x_i)}{P(X > x_{i-1})} = \dfrac{S(x_{i-1}) - S(x_i)}{S(x_{i-1})} = 1 - \dfrac{S(x_i)}{S(x_{i-1})}$.

3. Sea $x > 0$ un valor fijo y sea x_u el valor más grande que puede tomar X tal que $x_u \leqslant x$. Entonces, por el resultado (2),

$$\begin{aligned}
\prod_{x_i \leqslant x} [1 - \lambda(x_i)] &= \prod_{x_i \leqslant x} \frac{S(x_i)}{S(x_{i-1})} \\
&= \frac{S(x_1)}{S(x_0)} \frac{S(x_2)}{S(x_1)} \frac{S(x_3)}{S(x_2)} \cdots \frac{S(x_u)}{S(x_{u-1})} \\
&= S(x_u) \\
&= S(x).
\end{aligned}$$

4. Por el resultado (3),

$$\begin{aligned} S(x_i) &= \prod_{j=1}^{i} [1 - \lambda(x_j)] \\ &= [1 - \lambda(x_i)] \prod_{j=1}^{i-1} [1 - \lambda(x_j)] \\ &= [1 - \lambda(x_i)]\, S(x_{i-1}). \end{aligned}$$

■

Observe que la relación (4) establece una relación recursiva para $S(x)$. Las identidades recién demostradas establecen, también en este caso discreto, que las funciones $S(x_i)$ y $\lambda(x_i)$ son equivalentes.

Función de riesgo acumulado (caso discreto)

La definición de la función de riesgo acumulado en el caso discreto es análoga al caso continuo. La revisaremos a continuación.

Definición 2.7 *Sea X un tiempo de vida discreto con valores $0 < x_1 < x_2 < \cdots$ y con función de riesgo $\lambda(x_i)$. La función de riesgo acumulado es*

$$\Lambda(x) := \sum_{x_i \leqslant x} \lambda(x_i), \quad \text{para } x > 0.$$

Es decir, $\Lambda(x_i) = \lambda(x_1) + \lambda(x_2) + \cdots + \lambda(x_i)$, para $i = 1, 2, \ldots$ En el caso continuo sabemos que se cumplen las siguientes relaciones:

$$\begin{aligned} \Lambda(x) &= -\log S(x), \\ S(x) &= \exp\{-\Lambda(x)\}. \end{aligned}$$

En el caso discreto esto no necesariamente es cierto. Por ejemplo, para el caso $X \sim \text{unif}\{x_1, x_2, x_3\}$ que se estudió en el Ejemplo 2.10, tenemos que

$\Lambda(x_2) = \lambda(x_1) + \lambda(x_2) = 1/3 + 1/2 = 5/6$, mientras que $S(x_2) = 1/3$. Por lo tanto, $S(x_2) \neq \exp\{-\Lambda(x_2)\}$.

Función de riesgo truncada

Veremos ahora la forma en la que cambia la función de riesgo bajo truncamientos. Utilizaremos los resultados vistos anteriormente acerca de la función de supervivencia.

Proposición 2.9 *Sea X un tiempo de vida continuo con función de riesgo $\lambda(x)$. Sean $0 \leqslant a < b$ dos constantes tales que $P(a < X \leqslant b) > 0$. La función de riesgo truncada al intervalo $(a, b]$ es*

$$\lambda(x \,|\, a < X \leqslant b) = \frac{f(x)}{S(x) - S(b)}, \quad a < x \leqslant b.$$

Demostración. Por definición, para $a < x \leqslant b$,

$$\begin{aligned} \lambda(x \,|\, a < X \leqslant b) &= \frac{f(x \,|\, a < X \leqslant b)}{S(x \,|\, a < X \leqslant b)} \\ &= \frac{f(x)}{F(b) - F(a)} \frac{S(a) - S(b)}{S(x) - S(b)} \\ &= \frac{f(x)}{S(x) - S(b)}. \end{aligned}$$

∎

Puede comprobarse que $\lambda(x \,|\, a < X \leqslant b)$ es la función de riesgo asociada a la función de supervivencia $S(x \,|\, a < X \leqslant b)$. A partir del resultado recién demostrado es inmediato ahora encontrar una fórmula para la función de riesgo acumulado truncada, en términos de la función de supervivencia.

Proposición 2.10 *Sea X un tiempo de vida continuo con función de riesgo $\lambda(x)$. Sean $0 \leqslant a < b$ dos constantes tales que $P(a < X \leqslant b) > 0$. La función de riesgo acumulado truncada al intervalo $(a, b]$ es*

$$\Lambda(x \,|\, a < X \leqslant b) = \log\left(\frac{S(a) - S(b)}{S(x) - S(b)}\right), \quad a < x \leqslant b.$$

Demostración. Por definición, para $a < x \leqslant b$,

$$\begin{aligned}
\Lambda(x \,|\, a < X \leqslant b) &= \int_0^x \lambda(u \,|\, a < X \leqslant b)\, du \\
&= \int_a^x \lambda(u \,|\, a < X \leqslant b)\, du \\
&= \int_a^x \frac{f(u)}{S(u) - S(b)}\, du \\
&= \int_a^x -\frac{d}{du} \log\left[S(u) - S(b)\right] du \\
&= -\log\left[S(u) - S(b)\right] \Big|_a^x \\
&= \log\left(\frac{S(a) - S(b)}{S(x) - S(b)}\right).
\end{aligned}$$

Alternativamente,

$$\begin{aligned}
\Lambda(x \,|\, a < X \leqslant b) &= -\log S(x \,|\, a < X \leqslant b) \\
&= \log\left(\frac{S(a) - S(b)}{S(x) - S(b)}\right).
\end{aligned}$$

∎

Recordemos que todas las fórmulas encontradas pueden reducirse a expresiones más simples cuando el intervalo de doble truncamiento $(a, b]$ es un truncamiento por abajo, (a, ∞), o bien un truncamiento por arriba, $(0, b]$.

Para el caso discreto se tienen las fórmulas que aparecen abajo y que son análogas a las demostradas en el caso continuo. Se deja como ejercicio la

comprobación de estas identidades.

$$\begin{aligned}\lambda(x_i \,|\, a < X \leqslant b) &= \frac{f(x_i)}{S(x_i) - S(b)}, \quad a < x_i \leqslant b,\\ \Lambda(x_i \,|\, a < X \leqslant b) &= \sum_{a < x_j \leqslant x_i} \frac{f(x_j)}{S(x_j) - S(b)}, \quad a < x_i \leqslant b.\end{aligned}$$

2.4. Función tiempo medio de vida restante

Se le llama tiempo de vida restante a un tiempo de vida truncado por abajo y medido a partir del momento del truncamiento. También se le llama tiempo de vida residual. Aquí tenemos la definición.

Definición 2.8 *Sea X un tiempo de vida con función de supervivencia $S(x)$. Sea $x \geqslant 0$ un tiempo fijo tal que $S(x) > 0$. A la variable aleatoria $T := X - x$, condicionada al evento $(X > x)$, se le llama tiempo de vida restante, o residual, a edad x.*

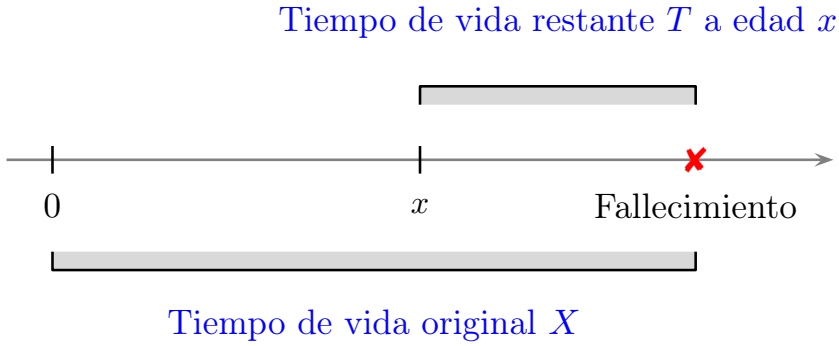

Figura 2.5: Tiempo de vida restante a edad x.

De este modo, para cada tiempo o edad $x \geqslant 0$, se puede definir el tiempo de vida T que le resta por vivir a una persona de edad x. El evento $(X > x)$

significa que el individuo se encuentra con vida al tiempo x. Véase la Figura 2.5. La definición anterior es válida tanto para tiempos de vida discretos con continuos. Observe que la variable aleatoria T es no negativa y coincide con el tiempo de vida original X cuando se toma $x = 0$. Observe también que en la notación usada para la variable T no se escribe explícitamente la edad de referencia x. Este tiempo fijo x aparece, generalmente, como un parámetro en la distribución de T. Esto se muestra en las siguientes fórmulas, en donde se establece la distribución de T.

Proposición 2.11 *Sea X un tiempo de vida discreto o continuo con función de probabilidad o densidad $f(x)$, función de supervivencia $S(x)$ y función de riesgo $\lambda(x)$. Sea $x \geqslant 0$ un tiempo fijo tal que $S(x) > 0$. La distribución del tiempo de vida restante $T = X - x$ satisface:*

1. $F_T(t) = \dfrac{S(x) - S(x+t)}{S(x)}$, *para* $t > 0$.
2. $f_T(t) = \dfrac{f(x+t)}{S(x)}$, *para* $t > 0$.
3. $S_T(t) = \dfrac{S(x+t)}{S(x)}$, *para* $t > 0$.
4. $\lambda_T(t) = \lambda(x+t)$, *para* $t > 0$.

Demostración.

1. Por definición,

$$\begin{aligned} F_T(t) &= P(X - x \leqslant t \,|\, X > x) \\ &= \frac{P(x < X \leqslant x + t)}{S(x)} \\ &= \frac{F(x+t) - F(x)}{S(x)} \\ &= \frac{S(x) - S(x+t)}{S(x)}, \quad \text{para } t > 0. \end{aligned}$$

2. En el caso continuo, derivando la fórmula anterior respecto de t,

$$f_T(t) = \frac{d}{dt} F_T(t) = \frac{f(x+t)}{S(x)}, \quad \text{para } t > 0.$$

En el caso discreto, para $t > 0$,

$$f_T(t) = P(X - x = t \,|\, X > x) = \frac{P(X = x+t)}{P(X > x)} = \frac{f(x+t)}{S(x)}.$$

3. Por definición,

$$\begin{aligned} S_T(t) &= P(T > t \,|\, X > x) \\ &= P(X - x > t \,|\, X > x) \\ &= \frac{P(X > x+t)}{P(X > x)} \\ &= \frac{S(x+t)}{S(x)}, \quad \text{para } t > 0. \end{aligned}$$

4. Por definición y por los resultados anteriores,

$$\begin{aligned} \lambda_T(t) &= \frac{f_T(t)}{S_T(t)} \\ &= \frac{f(x+t)}{S(x)} \frac{S(x)}{S(x+t)} \\ &= \lambda(x+t), \quad \text{para } t > 0. \end{aligned}$$

■

Recordemos que cuando no aparece ningún subíndice en alguna de las funciones que se muestran en el enunciado anterior, se entiende que se trata de funciones asociadas a la variable aleatoria inicial X. Observemos nuevamente que, cuando $x = 0$, el tiempo de vida restante T es el tiempo de vida completo X y las funciones anteriores se reducen a la funciones asociadas a X, escritas como dependientes del tiempo $t > 0$.

Ejemplo 2.11 (Tiempo de vida exponencial)
Para un tiempo de vida X con distribución exp(λ), la propiedad de pérdida de memoria implica que el tiempo de vida restante $T = X - x$ no depende del valor $x \geqslant 0$ y la distribución de T es la misma que la de X. En efecto, para $t > 0$,

$$\begin{aligned} P(T \leqslant t) &= P(X - x \leqslant t \,|\, X > x) \\ &= \frac{P(x < X \leqslant x + t)}{P(X > x)} \\ &= \frac{e^{-\lambda x} - e^{-\lambda(x+t)}}{e^{-\lambda x}} \\ &= 1 - e^{-\lambda t}. \end{aligned}$$

■

Para tiempos de vida discretos, conviene suponer que sus valores son los números naturales $1, 2, \ldots$, de este modo la Definición 2.8 y las fórmulas de la Proposición 2.11 permanecen válidos para valores $x = 0, 1, \ldots$ y $t = 1, 2, \ldots$ Veremos a continuación el ejemplo de la distribución geométrica que inicia en el valor 1. Encontraremos que T tiene la misma distribución que X. Este resultado es la versión discreta del caso exponencial del Ejemplo 2.11.

Ejemplo 2.12 (Tiempo de vida geométrico)
Sea X un tiempo de vida discreto con la distribución geométrica que aparece abajo, en donde $0 < p < 1$.

$$f(x) = \begin{cases} p(1-p)^{x-1} & si \;\; x = 1, 2, \ldots, \\ 0 & en\ otro\ caso. \end{cases}$$

Puede comprobarse que $S(x) = (1-p)^x$, equivalentemente, $F(x) = 1 - (1-p)^x$, para $x = 0, 1, \ldots$ Sea $x \geqslant 0$ un entero fijo. La distribución del tiempo

de vida restante a edad x es, para $t = 1, 2, \ldots$

$$\begin{aligned} P(T \leqslant t) &= P(X - x \leqslant t \,|\, X > x) \\ &= \frac{P(x < X \leqslant x + t)}{P(X > x)} \\ &= \frac{S(x) - S(x + t)}{S(x)} \\ &= \frac{(1 - p)^x - (1 - p)^{x+t}}{(1 - p)^x} \\ &= 1 - (1 - p)^t. \end{aligned}$$

Esta es la función de distribución del tiempo de vida original X, de modo que X y T tienen la misma distribución. ∎

En la sección de ejercicios se encuentran fórmulas generales para la esperanza y varianza del tiempo de vida restante. Con la información anterior, podemos ahora definir una cuarta función básica para un tiempo de vida.

Función tiempo promedio de vida restante

Definición 2.9 *Sea X un tiempo de vida con esperanza finita y sea $x \geqslant 0$ un tiempo o edad dentro del rango de valores de X tal que $P(X > x) > 0$. La función tiempo promedio de vida restante, o residual, a edad x es*

$$R(x) := E(X - x \,|\, X > x).$$

Es decir, $R(x)$ es la esperanza de la variable aleatoria $T = X - x$ definida antes. Esta función indica el tiempo promedio que le resta por vivir a una persona que ha alcanzado la edad x. Se entiende implícitamente que este tiempo promedio restante por vivir se mide a partir de la edad x. Observemos que $R(0) = E(X)$, es decir, a edad $x = 0$ el tiempo promedio de vida restante es el tiempo promedio de vida completo $E(X)$. La afirmación anterior se cumple para tiempos de vida X que inician desde el valor 0.

Cuando un tiempo de vida X inicia desde un valor $\gamma > 0$, (véase por ejemplo el modelo Pareto mencionado en la página 109), se cumple la relación $\gamma + R(\gamma) = E(X)$, cuando esta esperanza es finita.

Debido a su nombre en inglés *Mean Residual Lifetime*, a la función $R(x)$ también se le denota por $\text{MRL}(x)$. Esta función es importante pues puede demostrarse que caracteriza de manera única a la distribución de una variable aleatoria con esperanza finita. Sin embargo, puede no ser fácil encontrar una expresión para $R(x)$ a partir de la distribución del tiempo de vida. En términos de la función de supervivencia, la función $R(x)$ se expresa de la siguiente forma

$$R(x) = \frac{1}{S(x)} \int_x^\infty S(u)\, du, \quad x > 0 \text{ tal que } S(x) > 0.$$

Demostraremos esta fórmula más adelante. Como un ejemplo sencillo veamos el caso exponencial.

Ejemplo 2.13 (Tiempo de vida exponencial)
Considere que el tiempo de vida X tiene distribución exp(λ). Sea $x \geqslant 0$ un tiempo fijo. Recordando que $S(x) = e^{-\lambda x}$, para $x > 0$, se puede comprobar que $R(x) = 1/\lambda$. Esto no es ninguna sorpresa pues hemos demostrado antes que T y X tienen la misma distribución debido a la propiedad de pérdida de memoria. ∎

Cuando X es un tiempo de vida discreto con valores $0 < x_1 < x_2 < \cdots$, la función tiempo promedio de vida residual se puede escribir de la siguiente forma:

$$R(x_i) = E(X - x_i \,|\, X > x_i), \quad i = 1, 2, \ldots$$

Ejemplo 2.14 (Tiempo de vida geométrico)
Sea X un tiempo de vida discreto con la distribución geométrica como aparece abajo, en donde $0 < p < 1$. Como el tiempo de vida original X y el tiempo de vida restante T a cualquier edad x tienen la misma distribución, tenemos que $R(x) = E(X) = 1/p$, para $x = 0, 1, \ldots$

$$f(x) = \begin{cases} p(1-p)^{x-1} & si \;\; x = 1, 2, \ldots, \\ 0 & en \; otro \; caso. \end{cases}$$

■

En la Tabla 2.3 se muestran las expresiones de las funciones tiempo promedio de vida restante de los modelos paramétricos que se revisan en la Sección 3.1 "Distribuciones paramétricas".

Tiempo de vida	Función tiempo promedio de vida residual
unif(a, b)	$R(x) = 1/(b-x), \quad a < x < b$
exp(λ)	$R(x) = 1/\lambda, \quad x > 0$
Gompertz(b, c)	Fórmula general $R(x) = \frac{1}{S(x)} \int_x^\infty S(u)du, \quad x > 0$
Makeham(a, b, c)	Fórmula general $R(x) = \frac{1}{S(x)} \int_x^\infty S(u)du, \quad x > 0$
Weibull(α, λ)	Fórmula general $R(x) = \frac{1}{S(x)} \int_x^\infty S(u)du, \quad x > 0$
gama(γ, λ)	Fórmula general $R(x) = \frac{1}{S(x)} \int_x^\infty S(u)du, \quad x > 0$
lognormal(μ, σ^2)	Fórmula general $R(x) = \frac{1}{S(x)} \int_x^\infty S(u)du, \quad x > 0$
Pareto(α, γ)	$R(x) = x/(\alpha - 1), \quad x \geqslant \gamma$

Tabla 2.3: Algunas distribuciones y sus funciones tiempo promedio de vida restante.

Se puede encontrar mayor información sobre el tiempo promedio de vida residual $R(x)$ en el trabajo de F. Guess y F. Proschan [23]. En el trabajo citado se encuentran algunas referencias de artículos de investigación en donde se da respuesta al problema de determinar las características que debe cumplir una función general $R(x)$ para que sea la función tiempo promedio de vida residual de un tiempo de vida.

En la siguiente sección estudiaremos la relación que guarda $R(x)$ con las otras funciones básicas que hemos definido antes.

2.5. Equivalencia de las funciones básicas

En esta sección se demuestra que las cuatro funciones básicas que se han definido para un tiempo de vida X son equivalentes, es decir, a partir de cualquiera de ellas se pueden determinar, de manera única, a las otras funciones. Por simplicidad y por ser ya conocido, se ha omitido la equivalencia entre $f(x)$ y $F(x)$, y la equivalencia entre $\lambda(x)$ y $\Lambda(x)$. Además, se presenta sólo el caso continuo, en el Ejercicio 95 se encuentra la mayoría las fórmulas correspondientes al caso discreto.

Como recordatorio, se indican en el siguiente recuadro las definiciones de las funciones básicas en el caso continuo.

- Función de distribución $F(x) = P(X \leqslant x)$.
- Función de supervivencia $S(x) = P(X > x)$.
- Función de riesgo $\lambda(x) = f(x)/S(x)$.
- Función tiempo prom. de vida residual $R(x) = E(X - x \,|\, X > x)$.

Conocemos bien la equivalencia entre $f(x)$ y $F(x)$, así es que tomaremos a $f(x)$ como la función básica. Recordemos también que las funciones $\lambda(x)$ y $\Lambda(x)$ son equivalentes, tomaremos a $\lambda(x)$ como la función básica. Es también conveniente recordar que la determinación única se cumple en el sentido de que dos funciones son iguales cuando difieren en a lo sumo un número numerable de valores.

Iniciamos una serie de proposiciones que muestran fórmulas para calcular cada función básica a partir de cualquier otra.

Proposición 2.12 *Dada la función de densidad $f(x)$ de un tiempo de vida continuo, las otras funciones básicas quedan determinadas de manera única mediante las siguientes fórmulas:*

1. $S(x) = \displaystyle\int_x^\infty f(u)\,du, \quad x \geqslant 0.$
2. $\lambda(x) = \dfrac{f(x)}{\int_x^\infty f(u)\,du}, \quad x \geqslant 0.$
3. $R(x) = \dfrac{\int_x^\infty (u-x) f(u)\,du}{\int_x^\infty f(u)\,du}, \quad x \geqslant 0.$

Demostración.

1. Por definición, $S(x) = P(X > x) = \displaystyle\int_x^\infty f(u)\,du, \quad x \geqslant 0.$
2. Por definición y por el primer inciso,
$$\lambda(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{\int_x^\infty f(u)\,du}, \quad x \geqslant 0.$$
3. Por definición,
$$\begin{aligned} R(x) &= E(X - x \,|\, X > x) \\ &= \int_0^\infty (u-x)\, f(u \,|\, X > x)\, du \\ &= \int_x^\infty (u-x)\, \frac{f(u)}{P(X > x)}\, du \\ &= \frac{\int_x^\infty (u-x) f(u)\,du}{\int_x^\infty f(u)\,du}, \quad x \geqslant 0. \end{aligned}$$

■

Proposición 2.13 *Dada la función de supervivencia $S(x)$ de un tiempo de vida continuo, las otras funciones básicas quedan determinadas de manera única mediante las siguientes fórmulas:*

1. $f(x) = -\dfrac{d}{dx}\, S(x), \quad x \geqslant 0$ *tal que* $S(x) > 0$.
2. $\lambda(x) = -\dfrac{d}{dx} \log\, S(x), \quad x \geqslant 0$ *tal que* $S(x) > 0$.
3. $R(x) = \dfrac{1}{S(x)} \displaystyle\int_x^\infty S(u)\, du, \quad x \geqslant 0$ *tal que* $S(x) > 0$.

Demostración. En cada caso se desarrolla el lado derecho.

1. $-\dfrac{d}{dx} S(x) = -\dfrac{d}{dx}\left[1 - F(x)\right] = f(x), \quad x \geqslant 0.$
2. Por el inciso anterior,

$$-\frac{d}{dx} \log\, S(x) = -\frac{S'(x)}{S(x)} = \frac{f(x)}{S(x)} = \lambda(x), \quad x \geqslant 0.$$

3. Usaremos la expresión de $R(x)$ en términos de $f(x)$ demostrada en la proposición anterior. Por definición,

$$\begin{aligned}
\frac{1}{S(x)} \int_x^\infty S(u)\, du &= \frac{1}{S(x)} \int_x^\infty P(X > u)\, du \\
&= \frac{1}{S(x)} \int_x^\infty \int_u^\infty f(v)\, dv\, du \\
&= \frac{1}{S(x)} \int_x^\infty \int_x^v f(v)\, du\, dv \\
&= \frac{1}{S(x)} \int_x^\infty (v - x)\, f(v)\, dv \\
&= R(x), \quad x \geqslant 0.
\end{aligned}$$

■

Proposición 2.14 *Dada la función de riesgo $\lambda(x)$ de un tiempo de vida continuo, las otras funciones básicas quedan determinadas de manera única mediante las siguientes fórmulas:*

1. $f(x) = \lambda(x)\, \exp\{-\int_0^x \lambda(u)\, du\}, \quad x \geqslant 0.$
2. $S(x) = \exp\{-\int_0^x \lambda(u)\, du\}, \quad x \geqslant 0.$
3. $R(x) = \int_x^\infty \exp\{-\int_x^u \lambda(v)\, dv\}\, du, \quad x \geqslant 0.$

Demostración. En cada caso se desarrolla el lado derecho. Es conveniente recordar que $S(0) = 1$ y que $\log x$ denota el logaritmo natural.

1. Por definición,

$$\begin{aligned}
\lambda(x)\, \exp\{-\int_0^x \lambda(u)\, du\} &= \frac{f(x)}{S(x)}\, \exp\{-\int_0^x \frac{f(u)}{S(u)}\, du\} \\
&= \frac{f(x)}{S(x)}\, \exp\{\int_0^x \frac{S'(u)}{S(u)}\, du\} \\
&= \frac{f(x)}{S(x)}\, \exp\{\int_0^x \frac{d}{du}\, \log S(u)\, du\} \\
&= \frac{f(x)}{S(x)}\, \exp\{\log \frac{S(x)}{S(0)}\} \\
&= f(x), \quad x \geqslant 0.
\end{aligned}$$

2. El procedimiento es idéntico al inciso anterior. Por definición,

$$\begin{aligned}
\exp\{-\int_0^x \lambda(u)\,du\} &= \exp\{-\int_0^x \frac{f(u)}{S(u)}\,du\} \\
&= \exp\{\int_0^x \frac{S'(u)}{S(u)}\,du\} \\
&= \exp\{\int_0^x \frac{d}{du}\,\log S(u)\,du\} \\
&= \exp\{\log\frac{S(x)}{S(0)}\} \\
&= S(x), \quad x \geqslant 0.
\end{aligned}$$

3. El procedimiento es nuevamente similar. Usaremos además la expresión de $R(x)$ en términos de $S(x)$ demostrada antes. Por definición,

$$\begin{aligned}
\int_x^\infty \exp\{-\int_x^u \lambda(v)\,dv\}\,du &= \int_x^\infty \exp\{-\int_x^u \frac{f(v)}{S(v)}\,dv\}\,du \\
&= \int_x^\infty \exp\{\int_x^u \frac{S'(v)}{S(v)}\,dv\}\,du \\
&= \int_x^\infty \exp\{\int_x^u \frac{d}{dv}\,\log S(v)\,dv\}\,du \\
&= \int_x^\infty \exp\{\log\frac{S(u)}{S(x)}\}\,du \\
&= \int_x^\infty \frac{S(u)}{S(x)}\,du \\
&= R(x), \quad x \geqslant 0.
\end{aligned}$$

■

Proposición 2.15 *Dada la función tiempo promedio de vida residual $R(x)$ de un tiempo de vida continuo, las otras funciones básicas quedan determinadas de manera única mediante las siguientes fórmulas:*

1. $S(x) = \dfrac{R(0)}{R(x)} \exp\{-\displaystyle\int_0^x \frac{du}{R(u)}\}, \quad x \geqslant 0.$

2. $f(x) = \dfrac{R(0)(1 + R'(x))}{R^2(x)} \exp\{-\displaystyle\int_0^x \frac{du}{R(u)}\}, \quad x \geqslant 0.$

3. $\lambda(x) = \dfrac{1 + R'(x)}{R(x)}, \quad x \geqslant 0.$

Demostración.

1. Por lo demostrado antes, sabemos que

$$R(x) = \frac{1}{S(x)} \int_x^\infty S(u)\, du.$$

Derivando,

$$\begin{aligned} R'(x) &= -(S(x))^{-2} S'(x) \int_x^\infty S(u)\, du - 1 \\ &= -\frac{S'(x)}{S(x)} R(x) - 1. \end{aligned}$$

Es decir,

$$\frac{S'(x)}{S(x)} = -\frac{1 + R'(x)}{R(x)}.$$

Esta es una ecuación diferencial para $S(x)$ con condición inicial $S(0) = 1$ y que se puede escribir como

$$\frac{d}{dx} \log R(x)\, S(x) = -\frac{1}{R(x)}.$$

Integrando,

$$\int_0^x \frac{d}{du} \log R(u)\, S(u)\, du = -\int_0^x \frac{1}{R(u)}\, du.$$

Por lo tanto,

$$\log \frac{R(x)\,S(x)}{R(0)\,S(0)} = -\int_0^x \frac{1}{R(u)}\,du.$$

Despejando $S(x)$ se obtiene el resultado buscado,

$$S(x) = \frac{R(0)}{R(x)}\,\exp\{-\int_0^x \frac{du}{R(u)}\}, \quad x \geqslant 0.$$

2. Derivando la expresión encontrada en el inciso anterior, tenemos que

$$\begin{aligned} S'(x) &= -R(0)(R(x))^{-2}R'(x)\,\exp\{-\int_0^x \frac{du}{R(u)}\} \\ &\quad + \frac{R(0)}{R(x)}\,\exp\{-\int_0^x \frac{du}{R(u)}\}\,(-\frac{1}{R(x)}). \end{aligned}$$

Es decir,

$$f(x) = \frac{R(0)(1+R'(x))}{R^2(x)}\,\exp\{-\int_0^x \frac{du}{R(u)}\}, \quad x \geqslant 0.$$

3. Nuevamente iniciamos con la fórmula

$$R(x) = \frac{1}{S(x)}\,\int_x^\infty S(u)\,du.$$

Derivando,

$$\begin{aligned} R'(x) &= -(S(x))^{-2}\,S'(x)\,\int_x^\infty S(u)\,du - 1 \\ &= -\frac{S'(x)}{S(x)}\,R(x) - 1 \\ &= \lambda(x)\,R(x) - 1. \end{aligned}$$

Despejando $\lambda(x)$ se obtiene la expresión buscada.

■

En conclusión, un tiempo de vida se puede caracterizar matemáticamente por cualquiera de las cuatro funciones: $f(x)$, $S(x)$, $\lambda(x)$ ó $R(x)$. En vista de

este resultado, se puede establecer un diagrama cíclico para cualquier orden que se desee de las cuatro funciones básicas. Véase la Figura 2.6.

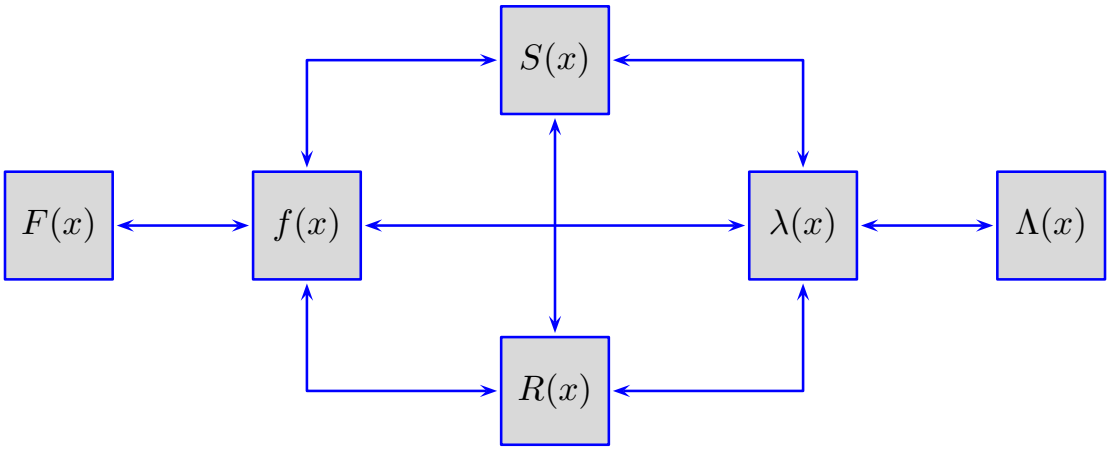

Figura 2.6: Funciones asociadas a un tiempo de vida.

Debe observarse que las demostraciones que hemos presentado y que llevan una función básica a otra no son únicas; existen varias maneras de demostrar las fórmulas enunciadas.

Pueden encontrarse otras exposiciones sobre las funciones básicas asociadas a los tiempos de vida en los trabajos de R. C. Elandt-Johnson y N. L. Johnson [19], E. T. Lee y J. W. Wang [37], X. Liu [42], D. F. Moore [44] ó A. Nag [45]. Debido a su importancia en la modelación, la mayoría de las fuentes que aparecen en la bibliografía proporcionan un tratamiento más detallado de la función de supervivencia $S(x)$ y la función de riesgo $\lambda(x)$.

2.6. Ejercicios

Función de distribución

22. Sea X un tiempo de vida con función de distribución $F(x)$ como aparece abajo, en donde la unidad de medición es de 1 año. Calcule la

probabilidad de que un individuo a edad x sobreviva un año adicional.

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ x/100 & \text{si } 0 \leqslant x \leqslant 100, \\ 1 & \text{si } x > 100. \end{cases}$$

23. Sea $0 \leqslant x < y$ dos números. Sea X un tiempo de vida con función de distribución continua $F(x)$. Exprese las siguientes probabilidades en términos de la función de distribución.

    a) $P(X > x)$.
    b) $P(x < X \leqslant y)$.
    c) $P(X \leqslant x \,|\, X \leqslant y)$.
    d) $P(X \leqslant y \,|\, X > x)$.
    e) $P(X > x \,|\, X \leqslant y)$.
    f) $P(X > y \,|\, X > x)$.
    g) $P(X > x + y \,|\, X > x)$, para $x \geqslant 0$, $y \geqslant 0$.

24. Sea X un tiempo de vida con función de distribución continua $F(x)$ y con esperanza finita. Demuestre que

$$E(X) = \int_0^\infty (1 - F(x))\, dx.$$

25. Sea X un tiempo de vida con función de distribución continua $F(x)$. Demuestre que

$$F(X) \sim \text{unif}(0, 1).$$

Función de supervivencia

26. Sean $0 \leqslant x < y$ dos números. Sea X un tiempo de vida con función de supervivencia continua $S(x)$. Exprese las siguientes probabilidades en términos de $S(x)$.

    a) $P(X > x)$.
    b) $P(x < X \leqslant y)$.
    c) $P(X \leqslant x \,|\, X \leqslant y)$.
    d) $P(X \leqslant y \,|\, X > x)$.
    e) $P(X > x \,|\, X \leqslant y)$.
    f) $P(X > y \,|\, X > x)$.
    g) $P(X > x + y \,|\, X > x)$, para $x \geqslant 0$, $y \geqslant 0$.

27. *Tiempo de vida uniforme.*
Suponga que un tiempo de vida X tiene distribución unif(a, b), en donde $0 < a < b < \infty$ son dos constantes. Encuentre y grafique la función de supervivencia $S(x)$, para $a < x < b$.

28. *Tiempo de vida geométrico.*
Sea X un tiempo de vida discreto con función de probabilidad $f(x)$ como aparece abajo, en donde $0 < p < 1$. Encuentre la función de supervivencia $S(x)$, para $x = 0, 1, \ldots$

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{si } \ x = 1, 2, \ldots \\ 0 & \text{en otro caso.} \end{cases}$$

29. *Tiempo de vida uniforme discreto.*
Sea X un tiempo de vida con distribución unif$\{x_1, \ldots, x_n\}$, en donde $0 < x_1 < \cdots < x_n$. Demuestre que

$$S(x) = \begin{cases} 1 & \text{si } \ x < x_1, \\ \dfrac{n-i}{n} & \text{si } \ x_i \leqslant x < x_{i+1}, \quad i = 1, \ldots, n-1, \\ 0 & \text{si } \ x \geqslant x_n. \end{cases}$$

30. Sea $a > 0$ una constante y suponga que un tiempo de vida continuo X tiene función de supervivencia $S(x)$ dada por

$$S(x) = \begin{cases} 1 - x^2/2a^2 & \text{si } \ 0 < x \leqslant a, \\ (2a-x)^2/2a^2 & \text{si } \ a < x < 2a, \\ 0 & \text{si } \ x \geqslant 2a. \end{cases}$$

   a) Grafique $S(x)$.

   b) Encuentre y grafique la función de densidad de X.

   c) Compruebe que la función encontrada en el inciso anterior es, efectivamente, una función de densidad.

31. *Propiedades de la función de supervivencia.*
Sea X un tiempo de vida con función de supervivencia $S(x)$. Demuestre que:

a) $\lim_{x\to\infty} S(x) = 0$.

b) $\lim_{x\searrow 0} S(x) = 1$.

c) Si $x_1 \leqslant x_2$ entonces $S(x_1) \geqslant S(x_2)$.

d) $S(x)$ es continua por la derecha, es decir, $S(x+) = S(x)$.

32. Demuestre que las siguientes funciones continuas son de supervivencia. Para ello, compruebe que las cuatro propiedades de la Proposición 2.1 se cumplen. Grafique, además, estas funciones.

a) $S(x) = \dfrac{1}{1+x}$, para $x > 0$.

b) $S(x) = e^{-x^2}$, para $x > 0$.

33. Demuestre que la siguiente función no es de supervivencia.

$$S(x) = \begin{cases} 1 & \text{si } x \leqslant 0, \\ \dfrac{4+x}{4+2x} & \text{si } x > 0. \end{cases}$$

34. *Combinación lineal convexa.*
Sean $S_1(x)$ y $S_2(x)$ dos funciones de supervivencia y sea $\lambda \in [0,1]$ una constante. Demuestre que la siguiente función es de supervivencia.

$$S(x) := \lambda\, S_1(x) + (1-\lambda)\, S_2(x).$$

35. Considere un tiempo de vida con función de supervivencia

$$S(x) = 1 - (x/10)^2, \quad \text{para } 0 \leqslant x \leqslant 10.$$

a) Grafique $S(x)$.

b) Encuentre y grafique la función de distribución.

c) Encuentre y grafique la función de densidad.

d) Encuentre el tiempo de vida promedio.

e) Encuentre la mediana del tiempo de vida.

36. *Esperanza y varianza.*
Sea X un tiempo de vida continuo con función de supervivencia $S(x)$ y con segundo momento finito. Demuestre que:

a) $E(X) = \displaystyle\int_0^\infty S(x)\,dx.$

b) $E(X^2) = 2\displaystyle\int_0^\infty x\,S(x)\,dx.$

c) $Var(X) = 2\displaystyle\int_0^\infty x\,S(x)\,dx - \Big(\int_0^\infty S(x)\,dx\Big)^2.$

37. *Esperanza y varianza.*
Sea X un tiempo de vida discreto con valores $1, 2, \ldots$, con función de supervivencia $S(x)$ y con segundo momento finito. Demuestre que:

a) $E(X) = \displaystyle\sum_{x=0}^{\infty} S(x).$

b) $E(X^2) = \displaystyle\sum_{x=1}^{\infty} (2x-1)\,S(x-1).$

c) $Var(X) = \displaystyle\sum_{x=1}^{\infty} (2x-1)\,S(x-1) - \Big(\sum_{x=0}^{\infty} S(x)\Big)^2.$

38. *Sistemas en serie.*
Considere un sistema de n componentes conformado mediante un arreglo en serie como el que se muestra en la Figura 2.7. Suponga que cada componente tiene un tiempo de vida independiente X_i, con función de supervivencia $S_i(x)$, para $i = 1, \ldots, n$. El tiempo de vida del sistema completo es la variable aleatoria

$$X := \text{mín}\,\{X_1, \ldots, X_n\}.$$

Figura 2.7: Componentes en serie.

a) Demuestre que la función de supervivencia del tiempo de vida del sistema es

$$S(x) = \prod_{i=1}^{n} S_i(x), \quad x > 0.$$

b) Demuestre que $S(x) \leqslant S_i(x)$, para cada $i = 1, \ldots, n$. Esta es otra manera de escribir el hecho de que el tiempo de vida del sistema completo es menor o igual al tiempo de vida de cualquiera de sus componentes. El sistema completo funciona si, y sólo si, todos los componentes funcionan.

c) Encuentre la función de supervivencia del tiempo de vida del sistema completo cuando el componente i tiene tiempo de vida $\exp(\lambda_i)$.

d) Bajo la hipótesis del inciso (c), encuentre $E(X)$.

e) Bajo la hipótesis del inciso (c), calcule la probabilidad de que el tiempo de vida X exceda a su valor esperado.

39. *Sistemas en paralelo.*
Considere un sistema de n componentes conformado mediante un arreglo en paralelo como el que se muestra en la Figura 2.8. Suponga que cada componente tiene un tiempo de vida independiente X_i, con función de supervivencia $S_i(x)$, para $i = 1, \ldots, n$. El tiempo de vida del sistema completo es la variable aleatoria

$$X := \max\{X_1, \ldots, X_n\}.$$

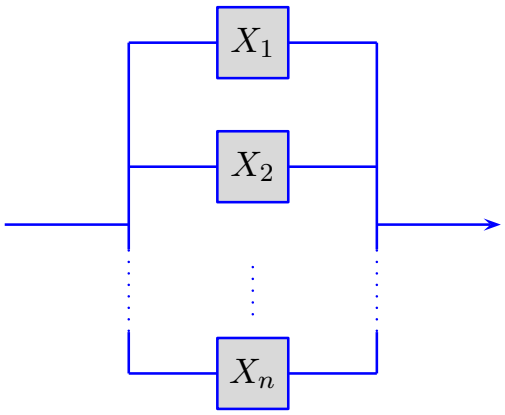

Figura 2.8: Componentes en paralelo.

a) Demuestre que la función de supervivencia del tiempo de vida del sistema es

$$S(x) = 1 - \prod_{i=1}^{n} (1 - S_i(x)).$$

b) Demuestre que $S(x) \geqslant S_i(x)$, para cada $i = 1, \ldots, n$. Esta es otra manera de escribir el hecho de que el tiempo de vida del sistema completo es mayor o igual al tiempo de vida de cualquiera de sus componentes. Una estructura en paralelo hace que el sistema sea más confiable pues la falla de algún componente no implica necesariamente la falla del sistema completo.

c) Encuentre la función de supervivencia del tiempo del vida del sistema completo cuando el componente i tiene tiempo de vida $\exp(\lambda_i)$.

d) Suponga que cada componente tiene tiempo de vida $\exp(\lambda)$. Demuestre que

$$E(X) = \frac{1}{\lambda}\,[1 + 1/2 + \cdots + 1/n].$$

40. Sea X un tiempo de vida continuo con función de supervivencia $S(x)$. Suponga que $S(x)$ es invertible para $x \in (0, \infty)$. Demuestre que

$$S(X) \sim \text{unif}(0, 1).$$

41. Sea X un tiempo de vida con función de supervivencia $S(x)$. Suponga que $S(x)$ es invertible para $x \in (0, \infty)$. Sea $\lambda > 0$ una constante. Demuestre que

$$-(1/\lambda) \log (S \circ X) \sim \exp(\lambda).$$

42. Sean $S_1(x)$ y $S_2(x)$ dos funciones de supervivencia. Determine si las siguientes funciones también son de supervivencia.

a) $S_1(x) \cdot S_2(x)$.
b) $S_1(x) + S_2(x)$.
c) $S_1(x) - S_2(x)$.
d) $[S_1(x)]^n, \quad n = 1, 2, \ldots$
e) $aS_1(x), \quad 0 < a < 1$.
f) $1 - S_1(x)$.

43. *Funciones de supervivencia truncadas.*
Sea X un tiempo de vida con función de supervivencia $S(x)$. Sean $0 \leqslant a < b$ dos constantes tales que $P(a < X \leqslant b) > 0$. Demuestre que las siguientes funciones son de supervivencia:

a) $S(x \,|\, X > a)$.
b) $S(x \,|\, X \leqslant b)$.
c) $S(x \,|\, a < X \leqslant b)$.

44. *Distribución exponencial truncada.*
Suponga que un tiempo de vida X se puede modelar mediante la distribución $\exp(\lambda)$ truncada al intervalo $(a, b]$, en donde $0 \leqslant a < b$. Encuentre $S(x \,|\, a < X \leqslant b)$.

45. Sea X un tiempo de vida con distribución $\exp(\lambda)$ y sean x y t dos cantidades positivas. Demuestre que:

a) $P(x < X \leqslant x + t) = e^{-\lambda x}(1 - e^{-\lambda t})$.
b) $P(x < X \leqslant x + t \,|\, X > x) = 1 - e^{-\lambda t}$.
c) $P(X > x + t \,|\, X > x) = e^{-\lambda t}$.

46. Sea X un tiempo de vida discreto con función de probabilidad $f(x)$ como aparece abajo, en donde $0 < p < 1$.

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{si } x = 1, 2, \ldots \\ 0 & \text{en otro caso.} \end{cases}$$

Demuestre que para tiempos $n = 0, 1, \ldots$ y $m \in \mathbb{N}$,

a) $P(n < X \leqslant n + m) = (1-p)^n \,[1 - (1-p)^m]$.
b) $P(n < X \leqslant n + m \,|\, X > n) = 1 - (1-p)^m$.
c) $P(X > n + m \,|\, X > n) = (1-p)^m$.

47. Sean X y Y dos tiempos de vida con distribución $\exp(\lambda_1)$ y $\exp(\lambda_2)$, respectivamente. Sean $S_1(x)$ y $S_2(x)$ sus correspondientes funciones de supervivencia. Demuestre que

$$S_1(x) \leqslant S_2(x) \text{ para todo } x \geqslant 0 \;\Leftrightarrow\; E(X) \geqslant E(Y).$$

48. *Otra caracterización de un tiempo de vida exponencial.*
Sea X un tiempo de vida continuo, con soporte $(0, \infty)$ y con función de supervivencia $S(x)$. Demuestre que

$$\begin{array}{lcl} S(x+y) = S(x)\, S(y) & \Leftrightarrow & \text{Existe } \lambda > 0 \text{ tal que } X \text{ tiene} \\ \text{para todo } x \geqslant 0,\ y \geqslant 0 & & \text{distribución } \exp(\lambda). \end{array}$$

Fórmula del producto

49. Sea X un tiempo de vida y sea x un tiempo tal que $P(X > x) > 0$.

    a) Demuestre que para cualquier $t > 0$ se cumple la relación que aparece abajo

    b) Interprete esta desigualdad cuando X representa el tiempo de vida de una persona.

$$P(x < X \leqslant x + t \,|\, X > x) \geqslant P(x < X \leqslant x + t).$$

Función de riesgo

50. Encuentre la función de riesgo $\lambda(x)$ de un tiempo de vida con función de densidad $f(x)$ que aparece abajo. Grafique con precisión $\lambda(x)$ y $f(x)$.

    a) $f(x) = 2x, \quad 0 < x < 1.$
    b) $f(x) = 2(1-x), \quad 0 < x < 1.$
    c) $f(x) = 3x^2, \quad 0 < x < 1.$
    d) $f(x) = 3(1-x)^2, \quad 0 < x < 1.$
    e) $f(x) = e^{-x}, \quad x > 0.$

51. Sean $a > 0$ y $b > 0$ dos constantes. Demuestre que las siguientes funciones son de riesgo y, en cada caso, encuentre la función de supervivencia $S(x)$ asociada.

    a) $\lambda(x) = a, \quad x > 0.$
    b) $\lambda(x) = ax + b, \quad x > 0.$
    c) $\lambda(x) = ax^2 + b, \quad x > 0.$
    d) $\lambda(x) = a(1+x)^2 + b, \quad x > 0.$
    e) $\lambda(x) = \dfrac{1}{1+x}, \quad x > 0.$

52. Sea X un tiempo de vida discreto con distribución unif$\{x_1, \ldots, x_n\}$, en donde $0 < x_1 < \cdots < x_n$. Demuestre que

$$\lambda(x) = \begin{cases} \dfrac{1}{n-i+1} & \text{si} \quad x = x_1, \ldots, x_n, \\ 0 & \text{en otro caso.} \end{cases}$$

53. Encuentre la función de riesgo $\lambda(x)$ asociada a la función de supervivencia $S(x)$ como aparece abajo. Grafique con precisión $\lambda(x)$.

    a) $S(x) = e^{-x^2/2}$, para $x > 0$.

    b) $S(x) = \dfrac{\alpha\, x^{\alpha-1}}{1+x^{\alpha}}$, para $x > 0$. ($\alpha > 0$ constante.)

    c) $S(x) = e^{-\alpha\, x^{\beta}}$, para $x > 0$. ($\alpha > 0,\ \beta > 0$ constantes.)

54. *Función de riesgo exponencial.*
    Sean $\alpha > 0$ y $\beta \geqslant 0$ dos constantes. La función de riesgo exponencial se define como

$$\lambda(x) := \alpha \exp\{\, \beta x \,\}, \quad \text{para } x > 0.$$

    Esta función no debe confundirse con la función de riesgo de la distribución exponencial. Esta última se obtiene en el caso $\beta = 0$, la cual es constante $\lambda(x) = \alpha$, para $x > 0$.

    a) Demuestre que $\lambda(x)$ es una función de riesgo.

    b) Demuestre que la función de supervivencia asociada es

$$S(x) = \exp\{\, \frac{\alpha}{\beta}(1 - e^{\beta x}) \,\}, \quad \text{para } x > 0.$$

    c) Demuestre el límite que aparece abajo. Como es de esperarse, este límite produce la función de supervivencia de la distribución $\exp(\alpha)$.

$$\lim_{\beta \to 0} S(x) = \exp\{-\alpha x\}, \quad \text{para } x > 0.$$

55. *Transformación lineal de una función de riesgo.*
    Sea $\lambda(x)$ una función de riesgo y sean $a > 0$ y $b \geqslant 0$ dos constantes. Demuestre que las siguientes funciones son de riesgo:

    a) $a\,\lambda(x)$.

    b) $\lambda(x) + b$.

    c) $a\,\lambda(x) + b$.

56. *Únicas distribuciones con funciones de riesgo constante.*

    a) Sea $\lambda > 0$ una constante. Demuestre que si X es un tiempo de vida continuo, con soporte $(0, \infty)$ y con función de riesgo constante $\lambda(x) = \lambda$, para $x > 0$, entonces X tiene distribución $\exp(\lambda)$.

    b) Sea $0 < p < 1$ una constante. Demuestre que si X es un tiempo de vida discreto, con soporte $\{1, 2, \ldots\}$ y con función de riesgo constante $\lambda(x) = p$, para $x = 1, 2, \ldots$, entonces X tiene la siguiente distribución geométrica

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{si } \ x = 1, 2, \ldots, \\ 0 & \text{en otro caso.} \end{cases}$$

57. *Sistemas en serie.*
    Sean $\lambda_1(x), \ldots, \lambda_n(x)$ las funciones de riesgo de n componentes de un sistema en serie como el que se muestra en la Figura 2.7 del Ejercicio 38.

    a) Demuestre que la función de riesgo del tiempo de vida del sistema completo es

$$\lambda(x) = \sum_{i=1}^{n} \lambda_i(x).$$

    b) Encuentre la función de riesgo $\lambda(x)$ del sistema completo cuando X_i tiene distribución $\exp(\lambda_i)$, para $i = 1, \ldots, n$.

58. *Sistemas en paralelo.*
    Considere el sistema n componentes en paralelo que aparece en la Figura 2.8 del Ejercicio 39. Suponga que los componentes $X_1, \ldots, X_n$ tienen tiempos de vida con idéntica función de distribución continua $F_0(x)$ y función de riesgo $\lambda_0(x)$. Demuestre que la función de riesgo del tiempo de vida del sistema completo es

$$\lambda(x) = (\lambda_0(x) + F_0(x))^n - (F_0(x))^n.$$

    Cuando los componentes tienen tiempos de vida con distribuciones distintas, la expresión para $\lambda(x)$ es más complicada.

59. Sea X un tiempo de vida con función de riesgo $\lambda(x)$ y sean $x \geqslant 0$ y $y \geqslant 0$ dos tiempos cualesquiera con $P(X > x) > 0$.

a) Demuestre que

$$P(X > x + y \,|\, X > x) = \exp\{-\int_x^{x+y} \lambda(u)\, du\}.$$

b) Demuestre que esta es una función decreciente en y.

c) Demuestre que esta es una función decreciente en $x \Leftrightarrow \lambda(x)$ es creciente.

60. *Cuantiles.*
Sea X un tiempo de vida continuo y sea $p \in (0, 1)$. A la edad $x_p > 0$ tal que $P(X > x_p) = p$ se le llama cuantil de orden p y es tal que la probabilidad de sobrevivir a dicha edad es el valor p. En particular, cuando $p = 0.5$ el valor x_p es la mediana, es decir, la probabilidad de que un individuo alcance esa edad es 0.5 . Encuentre la mediana de un tiempo de vida que sigue una distribución exp(λ).

61. *Tiempo de vida uniforme continuo.*
Sea X un tiempo de vida con distribución unif(a, b), en donde $0 \leqslant a < b$. Demuestre que:

$$a)\ \ S(x) = \begin{cases} 1 & \text{si} \ \ x \leqslant a, \\ \dfrac{b-x}{b-a} & \text{si} \ \ a < x < b, \\ 0 & \text{si} \ \ x \geqslant b. \end{cases}$$

$$b)\ \ \lambda(x) = \begin{cases} \dfrac{1}{b-x} & \text{si} \ \ a < x < b, \\ 0 & \text{en otro caso.} \end{cases}$$

62. Sea $a > 0$ una constante. Explique las razones por las cuales la función $\lambda(x)$ especificada abajo no es una función de riesgo.

$$\lambda(x) = e^{-ax} \quad \text{para} \ \ x \geqslant 0.$$

63. *Tiempo de vida Weibull.*
Sea X un tiempo de vida con distribución Weibull(α, λ), es decir, su función de densidad es

$$f(x) = \begin{cases} \alpha\lambda x^{\alpha-1} \exp\{-\lambda x^\alpha\} & \text{si} \ \ x > 0, \\ 0 & \text{en otro caso,} \end{cases}$$

en donde $\alpha > 0$ es el parámetro de forma y $\lambda > 0$ es el parámetro de escala. Demuestre que:

a) $S(x) = \exp\{-\lambda x^{\alpha}\}, \quad x > 0.$

b) $\lambda(x) = \alpha\lambda x^{\alpha-1}, \quad x > 0.$

c) $\lambda(x)$ es $\begin{cases} \text{creciente} & \text{si } \alpha > 1, \\ \text{decreciente} & \text{si } \alpha < 1, \\ \text{constante} & \text{si } \alpha = 1. \end{cases}$

64. *Distribución lineal-exponencial.*
Sea X un tiempo de vida con función de riesgo lineal $\lambda(x)$ como aparece abajo, en donde $a > 0$ y $b > 0$ son dos constantes. Encuentre $S(x)$ y $f(x)$.

$$\lambda(x) = ax + b, \quad \text{para } x > 0.$$

65. Sea X un tiempo de vida con función de riesgo $\lambda(x)$ como parece abajo, en donde $a > 0$ y $b > 0$ son dos constantes. Demuestre que $\lambda(x)$ es una función de riesgo y encuentre $S(x)$ y $f(x)$.

$$\lambda(x) = \frac{b}{a+x}, \quad \text{para } x > 0.$$

66. *Función de riesgo discreta truncada.*
Sea X un tiempo de vida discreto con valores $0 < x_1 < x_2 < \cdots$ y con función de riesgo $\lambda(x)$. Sean $0 \leqslant a < b$ dos constantes tales que $P(a < X \leqslant b) > 0$. Demuestre que:

a) $\lambda(x_i \,|\, a < X \leqslant b) = \dfrac{f(x_i)}{S(x_i) - S(b)}, \quad a < x_i \leqslant b.$

b) $\Lambda(x_i \,|\, a < X \leqslant b) = \displaystyle\sum_{a < x_j \leqslant x_i} \frac{f(x_j)}{S(x_j) - S(b)}, \quad a < x_i \leqslant b.$

67. *Distribución exponencial truncada.*
Suponga que un tiempo de vida X se puede modelar mediante la distribución $\exp(\lambda)$ truncada al intervalo $(a, b]$, en donde $0 \leqslant a < b$. Encuentre $\lambda(x \,|\, a < X \leqslant b)$. ¿Por qué esta función no es constante?

68. *Distribución geométrica truncada.*
Suponga que un tiempo de vida X se puede modelar mediante la distribución geo(p) truncada al intervalo $(n, m]$, en donde n y m son dos enteros tales que $0 \leqslant n < m$. Encuentre $\lambda(x \,|\, n < X \leqslant n)$. La distribución geo($p$) se especifica a continuación.

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{si } \ x = 1, 2, \ldots \\ 0 & \text{en otro caso.} \end{cases}$$

69. Sea X el tiempo continuo que transcurre entre el momento de la puesta en operación de un componente mecánico o electrónico hasta el momento de la primera falla. En sistemas en los que se requiere alta confiabilidad, estos componentes se deben reemplazar cuando se detecta la primera falla, o bien cuando transcurre un cierto tiempo fijo $t > 0$. Así, el tiempo de uso del componente es la variable aleatoria mixta

$$Y := \text{mín}\,\{X, t\}.$$

a) Encuentre $S_Y(y)$ en términos de $S_X(x)$.

b) Encuentre $\lambda_Y(y)$ en términos de $\lambda_X(x)$.

c) Demuestre que $E(Y) = \displaystyle\int_0^t S_X(x)\, dx$.

d) Demuestre que $E(Y) \leqslant E(X)$.

70. Sea X un tiempo de vida continuo con función de supervivencia $S(x)$ y función de riesgo acumulado $\Lambda(x)$. Demuestre las siguientes afirmaciones de manera sucesiva:

a) $\Lambda(x) = -\log\, S(x)$.

b) $S(X) \sim \text{unif}(0, 1)$.

c) $\Lambda(X) \sim \exp(\lambda)$, con $\lambda = 1$.

71. Sea $X_1, \ldots, X_n$ una muestra aleatoria de un tiempo de vida X con función de riesgo acumulado $\Lambda(x)$.

a) Demuestre que $\Lambda(X_{(1)}), \ldots, \Lambda(X_{(n)})$ es una muestra ordenada (no independiente) de la distribución $\exp(\lambda)$, con $\lambda = 1$.

b) Demuestre que

$$E(\Lambda(X_{(i)})) = \sum_{j=1}^{i} \frac{1}{n-j+1}, \quad i = 1, \ldots, n.$$

72. *Una aproximación.*
Sea X un tiempo de vida discreto con valores $0 < x_1 < x_2 < \cdots$ y con función de riesgo $\lambda(x_i)$. Use la aproximación (4.24) que aparece en la página 190 para demostrar que, cuando los valores $\lambda(x_i)$ son pequeños, se cumple la aproximación

$$\Lambda(x) \approx -\sum_{x_i \leqslant x} \log\left[1 - \lambda(x_i)\right], \quad \text{para } x > 0.$$

Función de riesgo acumulado

73. Sea X un tiempo de vida continuo con función de riesgo acumulado $\Lambda(x)$. Demuestre que:

a) $\Lambda(X) \sim \exp(1)$.

b) $E(\Lambda(X) \,|\, \Lambda(X) > x) = x + 1, \quad x \geqslant 0.$

Función tiempo medio de vida restante

74. Sea X un tiempo de vida continuo con esperanza finita, con función de densidad $f(x)$ y función tiempo promedio de vida residual $R(x)$. Demuestre que para valores de x tales que $S(x) > 0$:

a) $R(x) = \dfrac{1}{S(x)} \displaystyle\int_x^\infty u f(u)\, du - x.$

b) $R(x) = \dfrac{1}{S(x)} \displaystyle\int_x^\infty S(u)\, du.$

75. Encuentre las funciones $f(x)$, $S(x)$ y $\lambda(x)$ para un tiempo de vida con función tiempo promedio de vida restante:

a) $R(x) = x + 1, \quad x \geqslant 0.$

b) $R(x) = \dfrac{1}{x+1}, \quad x \geqslant 0.$

76. *Tiempo de vida exponencial.*
Sea X un tiempo de vida con distribución $\exp(\lambda)$ y sea $x \geqslant 0$ una constante. Sea T el tiempo de vida restante a edad x. Encuentre:

a) $F_T(t), \quad t > 0.$
b) $f_T(t), \quad t > 0.$
c) $S_T(t), \quad t > 0.$
d) $\lambda_T(t), \quad t > 0.$

77. *Tiempo de vida uniforme continuo.*
Sea X un tiempo de vida con distribución $\text{unif}(a, b)$, en donde $0 \leqslant a < b$. Sea T el tiempo de vida restante a edad x, en donde $a < x < b$. Encuentre:

a) $F_T(t), \quad 0 < t < b - x.$
b) $f_T(t), \quad 0 < t < b - x.$
c) $S_T(t), \quad 0 < t < b - x.$
d) $\lambda_T(t), \quad 0 < t < b - x.$

78. *Tiempo de vida restante discreto.*
Sea X un tiempo de vida discreto con valores $1, 2, \ldots$. Sea $x \geqslant 0$ un entero fijo y suponga que ocurre el evento $(X > x)$. Demuestre que se cumplen las fórmulas de la Proposición 2.11 de la página 61 para el tiempo de vida restante discreto $T = X - x$, es decir, demuestre que:

a) $F_T(t) = \dfrac{S(x) - S(x+t)}{S(x)}, \quad \text{para} \;\; t = 1, 2, \ldots$
b) $f_T(t) = \dfrac{f(x+t)}{S(x)}, \quad \text{para} \;\; t = 1, 2, \ldots$
c) $S_T(t) = \dfrac{S(x+t)}{S(x)}, \quad \text{para} \;\; t = 1, 2, \ldots$
d) $\lambda_T(t) = \lambda(x+t), \quad \text{para} \;\; t = 1, 2, \ldots$

79. *Tiempo de vida uniforme discreto.*
Sea X un tiempo de vida discreto con distribución $\text{unif}\{x_1, x_2, x_3\}$, en donde $0 < x_1 < x_2 < x_3$.

a) Encuentre el tiempo promedio de vida restante a edad x, $R(x) = E(X - x \,|\, X > x)$, para cada valor $x = 0, x_1, x_2$.
b) Grafique la función $R(x)$, $x = 0, x_1, x_2$.

80. *Tiempo de vida uniforme discreto.*
Sea X un tiempo de vida discreto con distribución unif$\{1, \ldots, n\}$, en donde $n \geqslant 1$.

   a) Encuentre el tiempo promedio de vida restante a edad x, $R(x) = E(X - x \,|\, X > x)$, para cada valor $x = 0, 1, \ldots, n-1$.
   b) Grafique la función $R(x)$, $x = 0, 1, \ldots, n-1$.

81. *Tiempo de vida uniforme continuo.*
Sea X un tiempo de vida con distribución unif(a, b), en donde $0 \leqslant a < b$. Demuestre que el tiempo promedio de vida restante a edad x tiene la expresión que aparece abajo. Grafique esta función e interprete los valores $R(a)$ y $R(b)$.

$$R(x) = \frac{b-x}{2}, \quad \text{para } a \leqslant x \leqslant b.$$

82. *Tiempo de vida geométrico.*
Sea X un tiempo de vida discreto con la distribución geométrica que aparece especificada abajo, en donde $0 < p < 1$.

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{si } \ x = 1, 2, \ldots, \\ 0 & \text{en otro caso.} \end{cases}$$

Sea $x \geqslant 0$ un entero y sea T el tiempo de vida restante a edad x. Encuentre:

   a) $F_T(t), \quad t = 1, 2, \ldots$
   b) $f_T(t), \quad t = 1, 2, \ldots$
   c) $S_T(t), \quad t = 1, 2, \ldots$
   d) $\lambda_T(t), \quad t = 1, 2, \ldots$

83. *Tiempo de vida exponencial.*
Sea X un tiempo de vida con distribución exp(λ). Demuestre que, para cualquier $x \geqslant 0$,

$$E(X \,|\, X > x) = \frac{1}{\lambda} + x.$$

84. Sea X un tiempo de vida continuo con función de riesgo acumulado $\Lambda(x)$ y función de riesgo $\lambda(x)$. Demuestre que

$$R(x) = \int_0^\infty e^{\Lambda(x) - \lambda(u+x)}\, du.$$

85. *Esperanza y varianza del tiempo de vida restante.*
Sea X un tiempo de vida continuo con esperanza y varianza finitas, con función de densidad $f(x)$, función de supervivencia $S(x)$ y función de riesgo $\lambda(x)$. Sea $T = X - x$ el tiempo de vida restante a edad $x \geqslant 0$. Demuestre que:

$$a)\ \ E(X - x \,|\, X > x) = \frac{1}{S(x)} \int_0^\infty t\, \lambda(x+t)\, S(x+t)\, dt.$$

$$b)\ \ \mathrm{Var}(X - x \,|\, X > x) = \frac{1}{S(x)} \int_0^\infty t^2\, f(x+t)\, dt - E^2(T).$$

Compruebe que, en el caso $x = 0$, las expresiones anteriores se reducen a $E(X)$ y $\mathrm{Var}(X)$, respectivamente.

86. *Expresión equivalente para la esperanza del tiempo de vida restante.*
Sea X un tiempo de vida continuo con función de supervivencia $S(x)$. Demuestre que el tiempo de vida media residual o restante $T = X - x$ a edad $x \geqslant 0$ satisface

$$E(X - x \,|\, X > x) = \int_x^\infty \frac{S(u)}{S(x)}\, du.$$

87. *Cuantiles del tiempo de vida restante.*
Sea $T = X - x$ el tiempo de vida restante a edad $x \geqslant 0$ de un tiempo de vida continuo X con función de supervivencia $S(x)$. A la edad x_p tal que

$$P(T > x_p \,|\, X > x) = p \tag{2.5}$$

se le llama cuantil de orden p del tiempo de vida restante T. La edad x_p es tal que la probabilidad de que un individuo alcance esa edad es p. Observe que la igualdad (2.5) se puede escribir como

$$\frac{S(x + x_p)}{S(x)} = p. \tag{2.6}$$

En particular, cuando $p = 0.5$ el valor x_p es la mediana, es decir, la probabilidad de que un individuo alcance esa edad es 0.5. Encuentre la mediana del tiempo de vida restante de un tiempo de vida que sigue una distribución $\exp(\lambda)$.

Equivalencia de las funciones básicas

88. Sea X un tiempo de vida continuo. Diga falso o verdadero:

    a) $f(x) = -\dfrac{d}{dx}S(x)$, para $x > 0$.

    b) $\lambda(x) = -\dfrac{d}{dx}\log S(x)$, para x tal que $S(x) > 0$.

    c) $\Lambda(x) = -\log S(x)$, para x tal que $S(x) > 0$.

    d) $R(x)$ es decreciente.

89. Sea X un tiempo de vida continuo con función de supervivencia $S(x)$ como aparece abajo. Grafique $S(x)$, compruebe que es una función de supervivencia y encuentre las cantidades indicadas.

$$S(x) = \frac{1}{10}\sqrt{100 - x}, \quad \text{para } 0 \leqslant x \leqslant 100.$$

    a) $f(50)$.
    b) $\lambda(30)$.
    c) $\Lambda(70)$.
    d) $E(X)$.
    e) $E(X^2)$.
    f) $\text{Var}(X)$.

90. Sean a y b dos constantes positivas. Suponga que una función de riesgo es de la forma $\lambda(x) = ax + b$, para $x \geqslant 0$.

    a) Encuentre y grafique $S(x)$.

    b) Encuentre y grafique $f(x)$.

    c) Encuentre la mediana de X.

91. Encuentre las otras funciones básicas asociadas a la función de supervivencia:

    a) $S(x) = \begin{cases} 1 - x & \text{si } 0 \leqslant x \leqslant 1, \\ 0 & \text{si } x > 1. \end{cases}$

    b) $S(x) = \begin{cases} 1 - x/2 & \text{si } 0 \leqslant x \leqslant 2, \\ 0 & \text{si } x > 2. \end{cases}$

$c)\ S(x) = \begin{cases} 1-2x & \text{si } \ 0 \leqslant x \leqslant 1/2, \\ 0 & \text{si } \ x > 1/2. \end{cases}$

$d)\ S(x) = \begin{cases} 1-x/a & \text{si } \ 0 \leqslant x \leqslant a, \quad (a \text{ constante}) \\ 0 & \text{si } \ x > a. \end{cases}$

$e)\ S(x) = \begin{cases} 1-(x/10)^2 & \text{si } \ 0 \leqslant x \leqslant 10, \\ 0 & \text{si } \ x > 10. \end{cases}$

92. *Tiempo de vida uniforme.*
Considere un tiempo de vida con distribución unif$(0, \omega)$, en donde $\omega > 0$ es la edad máxima posible. Demuestre que para $0 < x \leqslant \omega$,

$a)\ f(x) = 1/\omega.$
$b)\ F(x) = x/\omega.$
$c)\ S(x) = (\omega - x)/\omega.$
$d)\ \lambda(x) = 1/(\omega - x).$
$e)\ R(x) = (w - x)/2.$

93. *Transformación lineal del tiempo.*
Sea X un tiempo de vida continuo con funciones asociadas $f_X(x)$, $F_X(x)$, $S_X(x)$, $\lambda_X(x)$, $\Lambda_X(x)$ y $R_X(x)$. Sean $a > 0$ y $b > 0$ dos constantes. Encuentre las funciones asociadas de los siguientes tiempos de vida en términos de las funciones de X.

$a)\ aX.$ (Cambio de escala)
$b)\ X + b.$ (Desplazamiento)
$c)\ aX + b.$

94. Sea X un tiempo de vida continuo con funciones básicas $f_X(x)$, $F_X(x)$, $S_X(x)$, $\lambda_X(x)$ y $R_X(x)$. Sean $a > 0$ y $b > 0$ dos constantes. Exprese cada función básica del tiempo de vida Y, en términos de la correspondiente función básica de X, cuando su función de riesgo es:

$a)\ \lambda_Y(x) = a\,\lambda_X(x).$
$b)\ \lambda_Y(x) = \lambda_X(x) + b.$
$c)\ \lambda_Y(x) = a\,\lambda_X(x) + b.$

95. *Equivalencia de las funciones básicas para tiempos de vida discretos.* Sea X un tiempo de vida discreto con valores $0 < x_1 < x_2 < \cdots$ Definamos $x_0 := 0$. La funciones básicas que caracterizan la distribución de X son las siguientes: para $i = 1, 2, \ldots$

- Función de probabilidad $f(x_i) := P(X = x_i)$.
- Función de supervivencia $S(x_i) := P(X > x_i)$.
- Función de riesgo $\lambda(x_i) := f(x_i)/S(x_{i-1})$.
- Función tiempo promedio de vida residual $R(x_i) := E(X - x_i \,|\, X > x_i)$.

Demuestre que se cumplen las siguientes fórmulas:

a) Dada la función de probabilidad $f(x_i)$,

a.1) $S(x_i) = \sum_{j=i+1}^{\infty} f(x_j)$.

a.2) $\lambda(x_i) = f(x_i) / \sum_{j=i}^{\infty} f(x_j)$.

a.3) $R(x_i) = \sum_{j=i+1}^{\infty} x_j \, f(x_j) / \sum_{j=i+1}^{\infty} f(x_j) - x_i$.

b) Dada la función de supervivencia $S(x_i)$,

b.1) $f(x_i) = S(x_{i-1}) - S(x_i)$.

b.2) $\lambda(x_i) = 1 - S(x_i)/S(x_{i-1})$.

b.3) $R(x_i) = \dfrac{1}{S(x_i)} \sum_{j=i+1}^{\infty} x_j \, [S(x_{j-1}) - S(x_j)] - x_i$.

c) Dada la función de riesgo $\lambda(x_i)$,

c.1) $f(x_i) = \lambda(x_i) \prod_{j=1}^{i-1} (1 - \lambda(x_j))$.

$$c.2)\ \ S(x_i) = \sum_{j=i+1}^{\infty} \lambda(x_j) \cdot \sum_{k=1}^{j-1} (1 - \lambda(x_k)).$$

$$c.3)\ \ R(x_i) = \frac{\sum_{j=i}^{\infty} x_j\, \lambda(x_j) \prod_{k=1}^{j-1}(1 - \lambda(x_k))}{\sum_{j=i}^{\infty} \lambda(x_j) \prod_{k=1}^{j-1}(1 - \lambda(x_k))} - x_i.$$

d) A partir de $R(x_i)$, las expresiones para $f(x_i)$, $S(x_i)$ y $\lambda(x_i)$ son más complicadas y las omitimos de este ejercicio.

96. *Tiempo de vida exponencial trasladado.*
Sea $a > 0$ una constante. Sea X un tiempo de vida con función de densidad como aparece abajo, en donde $\lambda > 0$.

$$f(x) = \lambda\, e^{-\lambda(x-a)}, \quad \text{para} \ \ x > a.$$

Demuestre que:

a) $F(x) = 1 - e^{-\lambda(x-a)}, \quad x > a.$

b) $S(x) = e^{-\lambda(x-a)}, \quad x > a.$

c) $\lambda(x) = \lambda, \quad x > a.$

d) $\Lambda(x) = \lambda x, \quad x > a.$

e) $R(x) = 1/\lambda, \quad x > a.$

97. *Transformación lineal de un tiempo de vida exponencial.*
Sea $X \sim \exp(\lambda)$ y defina $Y = aX + b$, en donde $a > 0$ y $b > 0$ son constantes. Encuentre las funciones $f(y)$, $F(y)$, $S(y)$, $\lambda(y)$, $\Lambda(y)$ y $R(y)$ asociadas al tiempo de vida Y.

98. *Distribución exponencial truncada.*
Sea X un tiempo de vida con distribución $\exp(\lambda)$. Sean a y b dos constantes tales que $0 \leqslant a < b \leqslant \infty$. Demuestre que:

$$a)\ \ F(x \,|\, a < X \leqslant b) = \begin{cases} 0 & \text{si} \ \ x \leqslant a, \\ \dfrac{1 - e^{-\lambda(x-a)}}{1 - e^{-\lambda(b-a)}} & \text{si} \ \ a < x \leqslant b, \\ 1 & \text{si} \ \ x > b. \end{cases}$$

$$b)\ f(x \,|\, a < X \leqslant b) = \begin{cases} \dfrac{\lambda\, e^{-\lambda(x-a)}}{1 - e^{-\lambda(b-a)}} & \text{si } \ a < x < b, \\ 0 & \text{en otro caso.} \end{cases}$$

$$c)\ S(x \,|\, a < X \leqslant b) = \begin{cases} 1 & \text{si } \ x \leqslant a, \\ \dfrac{e^{-\lambda x} - e^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}} & \text{si } \ a < x \leqslant b, \\ 0 & \text{si } \ x > b. \end{cases}$$

$$d)\ S(x \,|\, X \leqslant b) = \begin{cases} 1 & \text{si } \ x \leqslant 0, \\ \dfrac{e^{-\lambda x} - e^{-\lambda b}}{1 - e^{-\lambda b}} & \text{si } \ 0 < x \leqslant b, \\ 0 & \text{si } \ x > b. \end{cases}$$

$$e)\ S(x \,|\, X > a) = \begin{cases} 1 & \text{si } \ x \leqslant a, \\ e^{-\lambda(x-a)} & \text{si } \ x > a. \end{cases}$$

$$f)\ \lambda(x \,|\, a < X \leqslant b) = \frac{\lambda}{1 - e^{-\lambda(b-x)}}, \quad \text{para } a < x < b.$$

$$g)\ \Lambda(x \,|\, a < X \leqslant b) = \log\left(\frac{e^{\lambda(b-a)} - 1}{e^{\lambda(b-x)} - 1}\right), \quad \text{para } a < x < b.$$

Estas fórmulas se reducen a expresiones más sencillas para el caso de truncamiento por abajo, $b = \infty$, y para el truncamiento por arriba, $a = 0$.

99. Sea X un tiempo de vida con función de densidad $f(x)$ como aparece abajo, en donde $\theta > -2$ es un parámetro.

$$f(x) = (\theta + 2)\, e^{-(\theta+2)x}, \quad \text{para } \ x > 0.$$

Demuestre que:

a) $F(x) = 1 - e^{-(\theta+2)x}$, para $x > 0$.

b) $S(x) = e^{-(\theta+2)x}$, para $x > 0$.

c) $\lambda(x) = \theta + 2$, para $x > 0$.

d) $\Lambda(x) = (\theta + 2)x$, para $x > 0$.

e) $R(x) = 1/(\theta + 2)$, para $x > 0$.

100. *Tiempo de vida geométrico trasladado.*
Sea $k \geqslant 1$ un entero. Sea X un tiempo de vida discreto con función de probabilidad como aparece abajo, en donde $0 < p < 1$.

$$f(x) = \begin{cases} p(1-p)^{x-k} & \text{si } x = k, k+1, \ldots \\ 0 & \text{en otro caso.} \end{cases}$$

Demuestre que:

a) $F(x) = 1 - (1-p)^{x+1-k}$, $x = k, k+1, \ldots$
b) $S(x) = (1-p)^{x+1-k}$, $x = k, k+1, \ldots$
c) $\lambda(x) = p$, $x = k, k+1, \ldots$
d) $R(x) = 1/p$, $x = k, k+1, \ldots$

101. *Distribución normal truncada.*
Suponga que un tiempo de vida X se puede modelar mediante la distribución $\mathrm{N}(\mu, \sigma^2)$ truncada al intervalo (a, ∞), con $a \geqslant 0$. Encuentre:

a) $f(x \mid X > a)$.
b) $E(X \mid X > a)$.
c) $\mathrm{Var}(X \mid X > a)$.
d) $S(x \mid X > a)$.
e) $\lambda(x \mid X > a)$.

Capítulo 3

Modelos paramétricos y funciones de verosimilitud

En este capítulo se revisan algunas distribuciones de probabilidad continuas que se pueden usar como modelos teóricos de tiempos de vida y que dependen de algún parámetro desconocido. Se aplica, además, el método de máxima verosimilitud para la estimación de parámetros bajo la presencia de censura. Debe advertirse, sin embargo, que esta perspectiva de ajustar una distribución conocida a un conjunto de datos de supervivencia no es siempre la más adecuada, pues un tiempo de vida cualquiera no necesariamente sigue alguno de los modelos paramétricos. En contraparte, en el siguiente capítulo consideraremos métodos no paramétricos para la estimación de la distribución de un tiempo de vida. Se puede encontrar mayor información sobre las distintas distribuciones de probabilidad aplicadas a tiempos de vida en el capítulo 33 del libro de N. L. Jonhson *et al* [29]. Véase también el capítulo 6 del libro de E. T. Lee y J. W. Wang [37].

3.1. Distribuciones paramétricas

Recordaremos a continuación algunas distribuciones continuas que pueden ser utilizadas para modelar tiempos de vida. En ocasiones algunas de estas distribuciones surgen cuando se efectúa alguna transformación sobre la función de supervivencia, por ejemplo, $\log S(x)$.

Cuando la sencillez de las fórmulas lo permita, mencionaremos algunas características numéricas de estas distribuciones. En particular, encontrar la expresión de todas las funciones básicas que estudiamos antes: $f(x)$, $S(x)$, $\lambda(x)$ y $R(x)$, puede ser un problema complicado en algunos casos. Aquí radica la importancia de la equivalencia de estas funciones, pues es suficiente especificar una de ellas para identificar de manera única a la distribución.

Es bien conocido que las distribuciones de probabilidad pueden parametrizarse de varias maneras. El lector debe tomar esto en cuenta si desea hacer uso de algún programa de cómputo para el tratamiento de alguna distribución de probabilidad, o bien comparar los resultados aquí presentados con otras fuentes bibliográficas.

Distribución uniforme continua

Sean $a < b$ dos constantes. La variable aleatoria continua X tiene distribución unif(a, b) si su función de densidad es

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } \ a < x < b, \\ 0 & \text{en otro caso.} \end{cases}$$

Se puede comprobar que $E(X) = (a+b)/2$ y $\text{Var}(X) = (b-a)^2/12$. Si se desea que esta variable aleatoria represente un tiempo de vida se requiere la condición $a \geqslant 0$. En este modelo los posibles tiempos de fallecimiento se encuentran en el intervalo (a, b). Las otras funciones básicas son:

1. $F(x) = (x-a)/(b-a)$, para $a < x < b$.
2. $S(x) = (b-x)/(b-a)$, para $a < x < b$.
3. $\lambda(x) = 1/(b-x)$, para $a < x < b$.
4. $\Lambda(x) = \log(b-a)/(b-x)$, para $a < x < b$.
5. $R(x) = (b-x)/2$, para $a < x < b$.

En particular, cuando $a = 0$ y la edad máxima b es denotada por la letra ω, esta distribución se puede caracterizar por cualquiera de las siguientes funciones básicas:

1. $f(x) = 1/\omega$, para $0 < x < \omega$.
2. $F(x) = x/\omega$, para $0 < x < \omega$.
3. $S(x) = (\omega - x)/\omega$, para $0 < x < \omega$.
4. $\lambda(x) = 1/(\omega - x)$, para $0 < x < \omega$.
5. $\Lambda(x) = \log(\omega/(\omega - x))$, para $0 < x < \omega$.
6. $R(x) = (w - x)/2$, para $0 < x < \omega$.

Las gráficas de las funciones de supervivencia y de riesgo de la distribución unif$(0, \omega)$ se muestran en la Figura 3.1. Observe que la probabilidad de supervivencia $S(x)$ decae de manera lineal, y que la función de riesgo $\lambda(x)$ se dispara a infinito instantes antes del valor ω indicando que nadie puede rebasar esa edad.

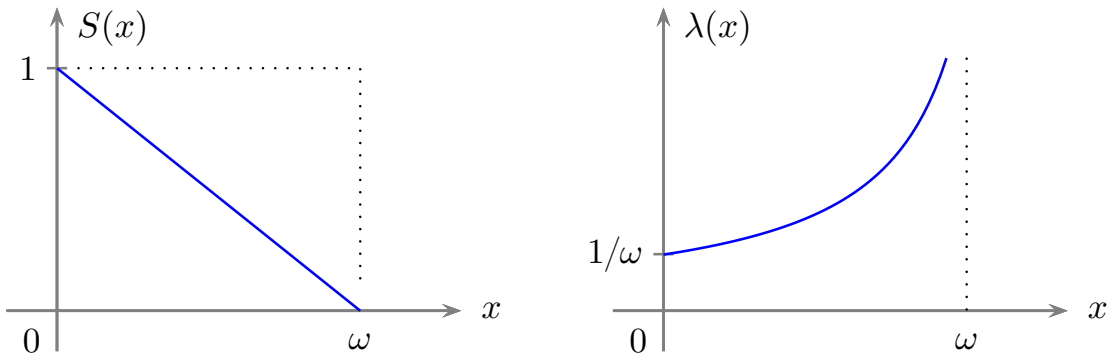

Figura 3.1: Funciones de supervivencia y de riesgo para un tiempo de vida unif $(0, \omega)$.

Distribución exponencial

Sea $\lambda > 0$ una constante. La variable aleatoria continua X tiene distribución exp(λ) si su función de densidad es

$$f(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{si } \ x > 0, \\ 0 & \text{en otro caso.} \end{cases}$$

En este modelo los tiempos de vida X toman valores en el intervalo $(0, \infty)$ y se puede comprobar que $E(X) = 1/\lambda$ y $\text{Var}(X) = 1/\lambda^2$. Las otras funciones básicas son:

1. $F(x) = 1 - e^{-\lambda x}$, para $x > 0$.
2. $S(x) = e^{-\lambda x}$, para $x > 0$.
3. $\lambda(x) = \lambda$, para $x > 0$.
4. $\Lambda(x) = \lambda x$, para $x > 0$.
5. $R(x) = 1/\lambda$, para $x > 0$.

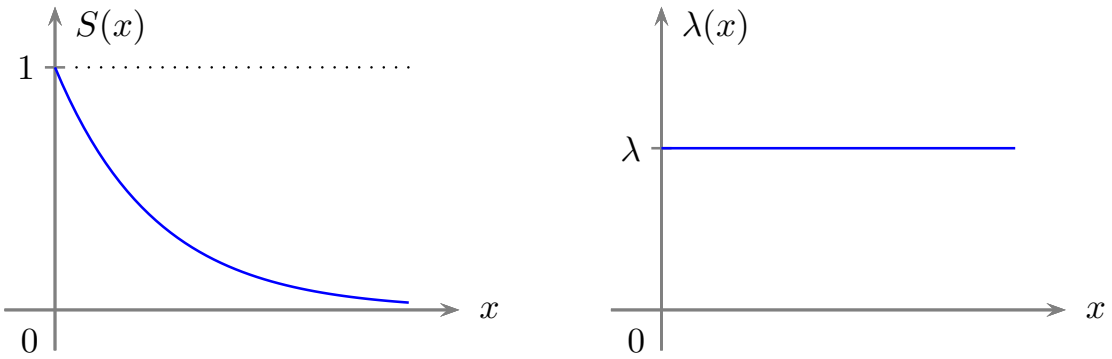

Figura 3.2: Funciones de supervivencia y de riesgo para un tiempo de vida exp(λ).

Dentro del conjunto de todas las distribuciones univariadas continuas, la distribución exponencial es la única que tiene fuerza de mortalidad constante y, justamente, se le refiere con esa característica: la distribución continua con fuerza de mortalidad constante. Esta distribución se ha usado con frecuencia para modelar tiempos de vida de componentes mecánicos o electrónicos. Las gráficas de las funciones de supervivencia y de riesgo de la distribución exponencial se muestran en la Figura 3.2.

Debe advertirse que, tomando $\lambda = 1/\theta > 0$, la distribución exponencial definida arriba también puede parametrizarse de la siguiente forma

$$f(x) = \frac{1}{\theta}\, e^{-x/\theta}, \quad \text{si} \quad x > 0.$$

Esta manera de escribir la densidad exponencial puede aparecer tanto en otras fuentes bibliográficas como en los programas de cómputo en los que se encuentra implementado el cálculo de probabilidades o la simulación de valores de esta distribución.

Distribución Gompertz[1]

A esta distribución se le denomina de tipo I dentro de la clase de las distribuciones de valor extremo. Su definición es más sencilla a partir de la especificación de su función de riesgo. Sean $b > 0$ y $c > 0$ dos constantes. La variable aleatoria continua X tiene distribución Gompertz(b, c) si su función de riesgo es

$$\lambda(x) = b\, e^{cx}, \quad \text{para} \quad x > 0.$$

En este modelo los tiempos de vida X toman valores en el intervalo $(0, \infty)$. La esperanza, varianza y momentos de esta distribución no tienen una expresión sencilla y omitiremos su escritura. Puede comprobarse que las otras funciones básicas son:

1. $F(x) = 1 - \exp\{\,(b/c)\,(1 - e^{cx})\,\}, \quad \text{para} \quad x > 0.$
2. $S(x) = \exp\{\,(b/c)\,(1 - e^{cx})\,\}, \quad \text{para} \quad x > 0.$
3. $f(x) = \lambda(x)\, S(x) = b \exp\{\,cx + (b/c)\,(1 - e^{cx})\,\}, \quad \text{para} \quad x > 0.$
4. $\Lambda(x) = (b/c)\,(e^{cx} - 1), \quad \text{para} \quad x > 0.$
5. $R(x) = \dfrac{1}{S(x)} \displaystyle\int_x^\infty S(u)\, du, \quad \text{para} \quad x > 0.$

Sólo se ha indicado una manera general de calcular $R(x)$. Las gráficas de las funciones de supervivencia y de riesgo de la distribución Gompertz se

[1] Benjamin Gompertz (1779–1865), actuario y matemático inglés.

muestran en la Figura 3.3 para $b = c = 1$.

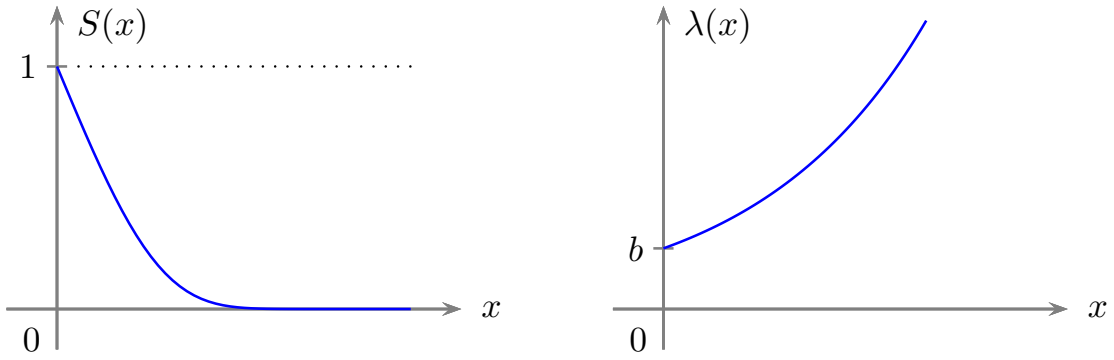

Figura 3.3: Funciones de supervivencia y de riesgo
para un tiempo de vida Gompertz(b, c).

Se puede comprobar que cuando $c \to 0$, la densidad Gompertz(b, c) converge a la densidad exp(b).

Distribución Makeham[2]

Esta distribución es una extensión de la distribución Gompertz. También en este caso su definición es más sencilla a partir de la especificación de su función de riesgo. Sean $b > 0$, $c > 0$ y $a > -b$ tres constantes. La variable aleatoria continua X tiene distribución Makeham(a, b, c) si su función de riesgo es

$$\lambda(x) = a + b\, e^{cx}, \quad \text{para} \quad x > 0.$$

Cuando $a = 0$ se obtiene la distribución Gompertz(b, c). En este modelo los tiempos de vida X toman valores en el intervalo $(0, \infty)$. La esperanza, varianza y momentos de esta distribución no tienen una expresión sencilla y se omite su escritura. Puede comprobarse que las otras funciones básicas son:

[2]William Matthew Makeham (1826–1891) actuario y matemático inglés.

1. $S(x) = \exp\{\,(b/c)\,(1-e^{cx}) - ax\,\}, \quad \text{para} \quad x > 0.$
2. $F(x) = 1 - \exp\{\,(b/c)\,(1-e^{cx}) - ax\,\}, \quad \text{para} \quad x > 0.$
3. $f(x) = \lambda(x)\,S(x) = (a+b\,e^{cx})\exp\{\,(b/c)\,(1-e^{cx}) - ax\,\}, \quad \text{para} \quad x > 0.$
4. $\Lambda(x) = ax + (b/c)\,(e^{cx}-1), \quad \text{para} \quad x > 0.$
5. $R(x) = \dfrac{1}{S(x)} \displaystyle\int_x^\infty S(u)\,du, \quad \text{para} \quad x > 0.$

Sólo se ha indicado una manera general de calcular $R(x)$. Las gráficas de las funciones de supervivencia y de riesgo de la distribución Makeham(a, b, c) se muestran en la Figura 3.4 para $a = 2$ y $b = c = 1$.

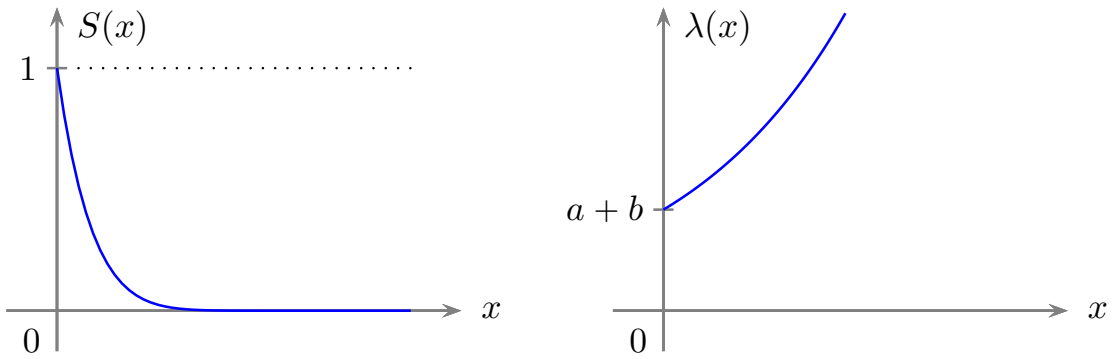

Figura 3.4: Funciones de supervivencia y de riesgo
para un tiempo de vida Makeham(a, b, c).

Distribución Weibull[3]

Sean $\alpha > 0$ y $\lambda > 0$ dos constantes. La variable aleatoria continua X tiene distribución Weibull(α, λ) si su función de densidad es

$$f(x) = \alpha\,\lambda\,x^{\alpha-1}\exp\{-\lambda x^{\alpha}\}, \quad \text{para} \quad x > 0.$$

[3]Waloddi Weibull (1887–1979) matemático e ingeniero sueco.

En este modelo los tiempos de vida X toman valores en el intervalo $(0, \infty)$. Observe que esta función se reduce a la función de densidad $\exp(\lambda)$ cuando $\alpha = 1$. Se puede comprobar que la esperanza y varianza de la distribución Weibull(α, λ) son:

$$\begin{aligned} E(X) &= \lambda^{-1/\alpha}\, \Gamma(1 + 1/\alpha), \\ \text{Var}(X) &= \lambda^{-1/\alpha}\, [\, \Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)\,], \end{aligned}$$

en donde $\Gamma(x)$ es la función gama definida como la integral que aparece abajo para valores de x en donde la integral es finita,

$$\Gamma(x) = \int_0^\infty x^{t-1}\, e^{-t}\, dt.$$

Puede comprobarse que las otras funciones básicas de la distribución Weibull son:

1. $F(x) = 1 - \exp\{-\lambda x^\alpha\}$, para $x > 0$.
2. $S(x) = \exp\{-\lambda x^\alpha\}$, para $x > 0$.
3. $\lambda(x) = f(x)/S(x) = \alpha\, \lambda\, x^{\alpha-1}$, para $x > 0$.
4. $\Lambda(x) = \lambda\, x^\alpha$, para $x > 0$
5. $R(x) = \exp\{\lambda x^\alpha\} \int_x^\infty \alpha\, \lambda\, u^\alpha \exp\{-\lambda u^\alpha\}\, du - x$, para $x > 0$.

Véase el Ejercicio 110 en la página 129 para conocer otras formas equivalentes de escribir la densidad Weibull al modificar la forma de escribir sus parámetros. Debe observarse que otras fuentes bibliográficas y los diversos programas de cómputo pueden hacer uso de cualquiera de estas otras representaciones.

Distribución gama

Sean $\gamma > 0$ y $\lambda > 0$ dos constantes. La variable aleatoria continua X tiene distribución gama(γ, λ) si su función de densidad es

$$f(x) = \frac{(\lambda x)^{\gamma-1}}{\Gamma(\gamma)} \lambda e^{-\lambda x}, \quad \text{para} \quad x > 0.$$

En este modelo los tiempos de vida X toman valores en el intervalo $(0, \infty)$. Cuando $\gamma = 1$ se obtiene la distribución $\exp(\lambda)$. Además, cuando el parámetro γ es un entero positivo n, a la distribución gama también se le llama distribución Erlang y se le denota por $\text{Erlang}(n, \lambda)$. A la distribución $\text{gama}(\nu/2, 1/2)$ en donde $\nu > 0$ se le llama distribución ji-cuadrada, se le denota por $\chi^2(\nu)$ y su función de densidad es

$$f(x) = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)}\, x^{\nu/2-1}\, e^{-x/2}, \quad \text{para} \quad x > 0.$$

Regresando al caso general de la distribución gama, se puede comprobar que $E(X) = \gamma/\lambda$ y $\text{Var}(X) = \gamma/\lambda^2$. Puede comprobarse que las otras funciones básicas son:

1. $F(x) = \dfrac{1}{\Gamma(\gamma)} \displaystyle\int_0^{\lambda x} u^{\gamma-1} e^{-u}\, du, \quad \text{para} \quad x > 0.$

2. $S(x) = \dfrac{1}{\Gamma(\gamma)} \displaystyle\int_{\lambda x}^{\infty} u^{\gamma-1} e^{-u}\, du, \quad \text{para} \quad x > 0.$

3. $\lambda(x) = \dfrac{f(x)}{S(x)} = \dfrac{(\lambda x)^{\gamma-1}\lambda e^{-\lambda x}}{\displaystyle\int_{\lambda x}^{\infty} u^{\gamma-1} e^{-u}\, du} = \dfrac{\lambda^{\gamma}}{x\displaystyle\int_0^{\infty} (u+\lambda)^{\gamma-1} e^{-ux}\, du},$
   para $x > 0$.

4. $R(x) = \dfrac{1}{\lambda} \dfrac{\displaystyle\int_{\lambda x}^{\infty} u^{\gamma} e^{-u}\, du}{\displaystyle\int_{\lambda x}^{\infty} u^{\gamma-1} e^{-u}\, du} - x, \quad \text{para} \quad x > 0.$

Por su complejidad, se omite la expresión para $\Lambda(x)$. A la integral que aparece en la expresión de $S(x)$ se le llama función gama incompleta y no es posible calcularla de manera exacta. Se han elaborado tablas (véase el manual de M. Abramowitz e I. A. Stegun [1]) y programas de cómputo para determinar sus valores aproximados mediante integración numérica o simulación. Las funciones $F(x)$ y $\lambda(x)$ están expresadas nuevamente en términos de una función gama incompleta.

Distribución lognormal

Sean μ y $\sigma^2 > 0$ dos constantes. La variable aleatoria continua X tiene distribución lognormal(μ, σ^2) si su función de densidad es

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\{-\frac{(\log x - \mu)^2}{2\sigma^2}\}, \quad \text{para} \quad x > 0.$$

Esta distribución surge de la siguiente manera: si Y tiene distribución $\mathrm{N}(\mu, \sigma^2)$, entonces la variable $X := e^Y$ tiene distribución lognormal con los parámetros indicados. Claramente, en este modelo los tiempos de vida X toman valores en el intervalo $(0, \infty)$. Se puede comprobar que la esperanza y la varianza de esta distribución son:

$$\begin{aligned} E(X) &= \exp\{\mu + \sigma^2/2\}, \\ \mathrm{Var}(X) &= \exp\{2\mu + 2\sigma^2\} - \exp\{2\mu + \sigma^2\}. \end{aligned}$$

Denotando por $\Phi(x)$ a la función de acumulación de la distribución normal estándar, puede comprobarse que las otras funciones básicas de la distribución lognormal son:

1. $F(x) = \Phi(\frac{\log x - \mu}{\sigma}), \quad \text{para} \quad x > 0.$

2. $S(x) = 1 - \Phi(\frac{\log x - \mu}{\sigma}), \quad \text{para} \quad x > 0.$

3. $\lambda(x) = \frac{f(x)}{S(x)} = \frac{1}{x\sqrt{2\pi\sigma^2}} \frac{\exp\{-\frac{(\log x - \mu)^2}{2\sigma^2}\}}{(1 - \Phi(\frac{\log x - \mu}{\sigma}))}, \quad \text{para} \quad x > 0.$

4. $R(x) = \frac{1}{S(x)} \int_x^\infty S(u)\, du, \quad \text{para} \quad x > 0.$

Sólo se ha indicado una manera general de calcular $R(x)$ y se ha omitido la expresión para $\Lambda(x)$.

Distribución Pareto[4]

Sean $\alpha > 0$ y $\gamma > 0$ dos constantes. La variable aleatoria continua X tiene una distribución Pareto(α, γ) tipo I cuando su función de densidad es

$$f(x) = \frac{\alpha}{x}\left(\frac{\gamma}{x}\right)^{\alpha}, \quad \text{para} \quad x \geqslant \gamma.$$

Observe que un tiempo de vida que sigue este modelo toma valores en el intervalo (γ, ∞). Puede demostrarse que $E(X) = \alpha\gamma/(\alpha - 1)$, para $\alpha > 1$, y $\text{Var}(X) = \alpha\gamma^2/[(\alpha-1)^2(\alpha-2)]$, para $\alpha > 2$. Las otras funciones básicas de la distribución Pareto son:

1. $F(x) = 1 - \left(\frac{\gamma}{x}\right)^{\alpha}$, para $x \geqslant \gamma$.
2. $S(x) = \left(\frac{\gamma}{x}\right)^{\alpha}$, para $x \geqslant \gamma$.
3. $\lambda(x) = \frac{\alpha}{x}$, para $x \geqslant \gamma$.
4. $\Lambda(x) = \alpha \log(x/\gamma)$, para $x \geqslant \gamma$.
5. $R(x) = \frac{x}{\alpha - 1}$, para $x \geqslant \gamma$.

§

Con esto concluimos una breve lista de algunas distribuciones continuas de probabilidad que pueden usarse como modelos paramétricos en el análisis de supervivencia. No se han incluido distribuciones discretas, las cuales pueden también ser usadas para modelar tiempos de vida. Se puede encontrar mayor información sobre las distintas distribuciones univariadas, continuas y discretas, en el libro de N. Balakrishnan y V. B. Nevzorov [5], en C. Forbes *et al* [21], o en el libro de N. L. Johnson *et al* [29]. Véase también el capítulo 6 del libro E. T. Lee y J. W. Wang [37].

[4] Vilfredo Pareto (1826–1891) ingeniero civil, economista y sociólogo italiano.

3.2. Función de verosimilitud para datos censurados

Sea X un tiempo de vida para el cual se ha adoptado una distribución dependiente de un parámetro o vector de parámetros desconocidos θ. En general, en el análisis de supervivencia no se adopta esta perspectiva paramétrica para estimar la distribución de un tiempo de vida, sin embargo, tal perspectiva de ajustar un modelo paramétrico es una posibilidad y en las siguientes secciones se revisará la forma de llevar a cabo las estimaciones. Así, se desea estimar al parámetro θ por el método de máxima verosimilitud a partir de una serie de observaciones posiblemente censuradas. El problema es encontrar la función de verosimilitud para después tratar de descubrir el valor de θ que maximiza esa función.

Supongamos que $f(x)$ es la función de densidad de un tiempo de vida en estudio. Cuando todas las observaciones $x_1, \ldots, x_n$ son tiempos de vida completos, es decir, ninguna observación sufrió censura, y cuando estas observaciones fueron obtenidas de manera independiente, la función de verosimilitud es

$$L(\theta) = f(x_1)\, f(x_2) \cdots f(x_n).$$

Es decir, cada observación completa x_i contribuye a la función de verosimilitud a través del factor $f(x_i)$. En cambio, si la observación x_i ha sido censurada, el factor es diferente y el problema es encontrarlo. Para ello supondremos las siguientes hipótesis:

a) Los tiempos de censura son independientes de los tiempos de vida.

b) La distribución de los tiempos de censura no dependen del parámetro a estimar.

Tenemos entonces los siguientes tres casos:

- Censura por la derecha. Sea $C > 0$ la variable aleatoria que representa el tiempo de censura por la derecha. La probabilidad de que una observación de un tiempo de vida X presente censura por la derecha en

un valor dado c es

$$\begin{aligned} P(X > C, C = c) &= \int_c^\infty f_{X,C}(u,c)\,du \\ &= \int_c^\infty f_X(u)\,f_C(c)\,du \\ &= P(X > c)\,f_C(c) \\ &\propto S(c). \end{aligned}$$

El símbolo $\propto$ significa "proporcional a", de modo que consideramos al factor $f_C(c)$ como una constante que no depende de θ y, de esta manera, se da preponderancia a la expresión $S(c)$, en la cual, en general, aparece θ.

- Censura por la izquierda. Sea $C > 0$ la variable aleatoria que representa el tiempo de censura por la izquierda. La probabilidad de que una observación del tiempo de vida X presente censura por la izquierda en un valor dado c es

$$\begin{aligned} P(X < C, C = c) &= \int_0^c f_{X,C}(u,c)\,du \\ &= \int_0^c f_X(u)\,f_C(c)\,du \\ &= P(X < c)\,f_C(c) \\ &\propto 1 - S(c). \end{aligned}$$

- Censura por intervalo. Suponga que (A, B) denota el intervalo aleatorio de censura. La probabilidad de que una observación del tiempo de vida X presente censura por un intervalo dado (a, b) es

$$\begin{aligned} \int_a^b f_{X,A,B}(u,a,b)\,du &= P(a < X < b) \cdot f_{A,B}(a,b) \\ &\propto F(b) - F(a) \\ &= S(a) - S(b). \end{aligned}$$

Sean $x_1, \ldots, x_n$ los datos de supervivencia de n individuos. Cada uno de estos valores pertenece a uno, y sólo uno, de los siguientes conjuntos:

$$
\begin{aligned}
D &= \text{“Conjunto de observaciones exactas”.} \\
&\quad (D = \textit{Death}) \qquad (3.1)
\end{aligned}
$$

$$
\begin{aligned}
R &= \text{“Conjunto de observaciones censuradas por la derecha”.} \\
&\quad (R = \textit{Right})
\end{aligned}
$$

$$
\begin{aligned}
L &= \text{“Conjunto de observaciones censuradas por la izquierda”.} \\
&\quad (L = \textit{Left})
\end{aligned}
$$

$$
\begin{aligned}
I &= \text{“Conjunto de observaciones censuradas por intervalo”.} \\
&\quad (I = \textit{Interval})
\end{aligned}
$$

El uso de estas letras para denotar a los conjuntos indicados tiene sentido al recordar la traducción de los términos en inglés: *Death*, *Right*, *Left* e *Interval*.

Como estamos suponiendo que cada observación posiblemente está sujeta a alguno de los tres tipos de censura mencionados, es importante notar que el término x_i representa el mismo dato cuando no hay censura, o bien el valor c_i cuando existe censura por la izquierda o por la derecha al tiempo c_i, o bien el intervalo (a_i, b_i) cuando el dato ha sufrido censura por este intervalo. Es conveniente tener en mente esta notación implícita para entender mejor la fórmula (3.2) que aparece abajo para la verosimilitud de la muestra.

Considerando de manera conjunta todos los datos y suponiendo independencia entre las observaciones, la función de verosimilitud de los datos de supervivencia cuando se adopta un modelo paramétrico particular se puede escribir como se establece en el siguiente recuadro.

Proposición 3.1 *Sea X un tiempo de vida para el que se asume un modelo paramétrico dependiente de θ, con función de densidad $f(x)$ y función de supervivencia $S(x)$. Sean $x_1, \ldots, x_n$ observaciones independientes de X que pueden estar censurados de tal forma que cada uno de ellos pertenece a uno de los conjuntos: D, R, L e I especificados en* (3.1). *Para cada observación x_i y según sea el caso, sea c_i el tiempo de censura por la izquierda, o por la derecha, y sea (a_i, b_i) el posible intervalo de censura. Entonces la función de verosimilitud de la muestra es*

$$\begin{aligned} L(\theta) \quad \propto \quad & [\prod_{x_i \in D} f(x_i)\,] \cdot [\prod_{x_i \in R} S(c_i)\,] \\ & \cdot [\prod_{x_i \in L} (1 - S(c_i))\,] \cdot [\prod_{x_i \in I} (S(a_i) - S(b_i))\,]. \end{aligned} \tag{3.2}$$

De la expresión general (3.2), es evidente que es conveniente contar con una expresión sencilla o manejable para la función de densidad $f(x)$ y para la función de supervivencia $S(x)$ de la distribución en estudio. Como se mencionó en la sección anterior, esto no necesariamente se cumple para un modelo paramétrico dado. Claramente la fórmula general (3.2) puede contener un menor número de términos en el caso cuando los datos presenten sólo uno o dos de los tres tipos de censura considerados. Es claro que para muestras que no presenten ningún tipo de censura, la función de verosimilitud (3.2) se reduce a la expresión conocida $L(\theta) = f(x_1)\, f(x_2) \cdots f(x_n)$.

En la siguiente sección se analizará el caso particular cuando el tipo de censura que se presenta sólo es por la derecha.

3.3. Función de verosimilitud para datos censurados por la derecha

En esta sección se encontrará la función de verosimilitud de un conjunto de datos bajo varios tipos de censura por la derecha, y suponiendo que se ha adoptado un modelo paramétrico, el cual depende de un parámetro desconocido θ.

Función de verosimilitud bajo censura tipo I

En el caso cuando los datos sólo pueden presentar censura por la derecha, éstos se pueden escribir de la forma

$$(x_1, \delta_1), (x_2, \delta_2), \ldots, (x_n, \delta_n),$$

en donde δ_i toma el valor 1 cuando el dato x_i no está censurado (δ=*Death*), y toma el valor 0 cuando el dato x_i está censurado. En este caso, la función de verosimilitud (3.2) se reduce al siguiente producto.

Proposición 3.2 *Sea X un tiempo de vida para el que se asume un modelo paramétrico dependiente de θ, con función de densidad $f(x)$ y función de supervivencia $S(x)$. Sean $x_1, \ldots, x_n$ observaciones independientes de X y suponga que los datos pueden presentar censura por la derecha y se les escribe como $(x_1, \delta_1), \ldots, (x_n, \delta_n)$, en donde la variable δ_i toma el valor 0 cuando el dato x_i presenta censura y toma el valor 1 cuando el dato x_i no presenta censura. Entonces la función de verosimilitud de la muestra es*

$$L(\theta) = \prod_{i=1}^{n} [f(x_i)]^{\delta_i} [S(x_i)]^{1-\delta_i}. \tag{3.3}$$

La fórmula (3.3) se obtiene de la fórmula general (3.2) demostrada antes. Observe que cuando $\delta_i = 1$, es decir, cuando no hay censura en el dato x_i, el factor que aparece en (3.3) es $f(x_i)$, y cuando $\delta_i = 0$, es decir, cuando hay censura en el dato x_i, el factor es $S(x_i)$. Por otro lado, substituyendo la identidad $f(x) = \lambda(x)\, S(x)$ en (3.3) se obtiene la expresión equivalente

$$L(\theta) = \prod_{i=1}^{n} [\lambda(x_i)]^{\delta_i} S(x_i). \tag{3.4}$$

En general, la expresión para $L(\theta)$ puede ser complicada como función de θ y encontrar el punto $\hat{\theta}$ en donde se alcanza el máximo puede ser un problema difícil. El ejemplo que se presenta a continuación es el caso exponencial y allí los cálculos no son complicados.

Ejemplo 3.1 (Tiempo de vida exponencial)
Sea $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ un conjunto de observaciones, posiblemente censuradas por la derecha, de un tiempo de vida con distribución exp(λ), en donde λ es desconocido. Por la fórmula (3.3)*, la función de verosimilitud de los datos es*

$$L(\lambda) \propto \prod_{i=1}^{n} [\lambda\, e^{-\lambda x_i}]^{\delta_i}\, [e^{-\lambda x_i}]^{1-\delta_i} \;=\; \lambda^r\, e^{-\lambda x},$$

en donde

$$r \;=\; \sum_{i=1}^{n} \delta_i \;=\; \text{“Número de fallecimientos observados.”}$$

$$x \;=\; \sum_{i=1}^{n} x_i \;=\; \text{“Suma de todos los tiempos registrados.”}$$

Se busca maximizar $L(\lambda)$. La derivada de $L(\lambda)$ es

$$\frac{d}{d\lambda} L(\lambda) \quad \propto \quad r\, \lambda^{r-1}\, e^{-\lambda x} - \lambda^r\, x\, e^{-\lambda x}.$$

Puede comprobarse que $L(\lambda)$ tiene un máximo cuando su derivada es cero y esto ocurre cuando $\lambda = r/x$. Así, la estimación máximo verosímil para λ es

$$\hat{\lambda} = \frac{r}{x} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}.$$

El estimador como variable aleatoria se puede escribir como

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \Delta_i}{\sum_{i=1}^{n} X_i},$$

en donde $\Delta_i \sim Ber(p_i)$, con $p_i = P(X_i < c_i)$. Los cálculos anteriores se pudieron hacer, en buena medida, debido a la expresión compacta para $S(x)$.

Observemos que, cuando todas las observaciones son completas, es decir, ninguna observación presenta censura por la derecha, tenemos que $\sum_{i=1}^{n} \delta_i = n$ *y, por lo tanto, el estimador se reduce a*

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

Este es el estimador máximo verosímil para el parámetro λ *de la distribución* $exp(\lambda)$ *en el caso de datos no censurados.* ∎

En general, la estimación de parámetros por máxima verosimilitud para datos censurados es más complicada que en el caso de datos provenientes de observaciones completas. A continuación se presenta el caso de la distribución Weibull, en donde se tienen dos parámetros y sólo se plantean las ecuaciones que determinan las estimaciones de esos parámetros. Para resolver estas ecuaciones es necesario usar algún método de aproximación. Véase el artículo de C. A. Clifford [13] para mayor información sobre la estimación de parámetros para la distribución Weibull.

Ejemplo 3.2 (Tiempo de vida Weibull)
Consideremos nuevamente que se cuenta con un conjunto de n *observaciones* $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ *con posible censura por la derecha. Supondremos ahora que estos datos provienen del modelo* Weibull(α, λ)*, en donde ambos parámetros,* $\alpha > 0$ *y* $\lambda > 0$*, son desconocidos. Por la fórmula* (3.3) *y recordando las expresiones para* $f(x)$ *y* $S(x)$ *de esta distribución, la función de verosimilitud de los datos es*

$$L(\alpha, \lambda) \propto \prod_{i=1}^{n} [\alpha\lambda\, x_i^{\alpha-1}\, e^{-\lambda x_i^{\alpha}}]^{\delta_i}\, [e^{-\lambda x_i^{\alpha}}]^{1-\delta_i} \;=\; \prod_{i=1}^{n} [\alpha\lambda\, x_i^{\alpha-1}]^{\delta_i}\, [e^{-\lambda x_i^{\alpha}}].$$

Como antes, denotaremos por r *al número de fallecimientos observados, esto es,* $r = \delta_1 + \cdots + \delta_n$*. Tomando logaritmo de la función de verosimilitud y simplificando se llega a*

$$\log L(\alpha, \lambda) = r \log \alpha + r \log \lambda + (\alpha - 1) \sum_{i=1}^{n} \delta_i \log x_i - \lambda \sum_{i=1}^{n} x_i^{\alpha}.$$

Las derivadas respecto de α y λ igualadas a cero son:

$$\frac{\partial}{\partial\alpha}\log L(\alpha,\lambda) = \frac{r}{\alpha}+\sum_{i=1}^{n}\delta_i\log x_i-\lambda\sum_{i=1}^{n}x_i^{\alpha}\log x_i = 0, \quad (3.5)$$

$$\frac{\partial}{\partial\lambda}\log L(\alpha,\lambda) = \frac{r}{\lambda}-\sum_{i=1}^{n}x_i^{\alpha} = 0. \quad (3.6)$$

Resolver simultáneamente estas dos ecuaciones para α y λ no es sencillo y se deben emplear métodos numéricos para obtener alguna aproximación de la solución. Dada esta dificultad técnica, dejaremos en este punto el desarrollo del ejemplo. Cuando los datos son completos ($\delta_1 = \cdots = \delta_n = 1$), las ecuaciones (3.5) *y* (3.6) *se reducen a las ecuaciones conocidas que determinan las estimaciones por máxima verosimilitud en el caso cuando no hay censura, las cuales son*

$$\frac{n}{\alpha}+\sum_{i=1}^{n}\log x_i-\lambda\sum_{i=1}^{n}x_i^{\alpha}\log\alpha = 0, \quad (3.7)$$

$$\frac{n}{\lambda}-\sum_{i=1}^{n}x_i^{\alpha} = 0. \quad (3.8)$$

Regresando al caso cuando hay censura, observemos que de la ecuación (3.6) *se obtiene la siguiente expresión para $\hat{\lambda}$, la cual es semejante a la que se obtuvo antes para este parámetro en el caso exponencial ($\alpha = 1$).*

$$\hat{\lambda} = \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n}x_i^{\hat{\alpha}}}.$$

■

En la sección de ejercicios aparecen otros ejemplos paramétricos que pueden estudiarse en el caso de censura tipo I. La función de verosimilitud (3.3) puede usarse cuando se cuente con expresiones para $f(x)$ y $S(x)$, o bien $\lambda(x)$ y $S(x)$ según (3.4).

Función de verosimilitud bajo censura tipo II

Sea $x_1, \ldots, x_n$ una colección de observaciones de un tiempo de vida con función de densidad $f(x, \theta)$, dependiente de un parámetro desconocido θ. Suponga que las observaciones se obtienen a través de un esquema de censura por la derecha tipo II, es decir, sólo se registran los r tiempos de vida más pequeños y el resto se censuran por la derecha, $1 \leqslant r \leqslant n$. En la siguiente proposición se muestra una expresión para la función de verosimilitud de la muestra.

Proposición 3.3 *Sea X un tiempo de vida para el que se asume un modelo paramétrico dependiente de θ, con función de densidad $f(x)$ y función de supervivencia $S(x)$. Si $x_1, \ldots, x_n$ son observaciones independientes de X bajo censura por la derecha tipo II en donde sólo se registran los r tiempos más pequeños, entonces la función de verosimilitud de la muestra es*

$$L(\theta) = \frac{n!}{(n-r)!} \Big[\prod_{i=1}^{r} f(x_{(i)})\Big] \big[S(x_{(r)})\big]^{n-r}. \tag{3.9}$$

Demostración. Para cada $i = 1, \ldots, r$, el dato $x_{(i)}$ es observado (no censurado) y su verosimilitud es $f(x_{(i)})$. En cambio, para $i = r+1, \ldots, n$, el dato $x_{(i)}$ no es observado (pues es censurado) y la probabilidad asociada es $P(X_i > x_{(r)}) = S(x_{(r)})$. Es evidente que no se cumple la hipótesis de independencia entre las observaciones pues éstas son estadísticas de orden. La verosimilitud de las observaciones es

$$L(\theta) = \int_{x_{(r)}}^{\infty} \cdots \int_{x_{(r)}}^{\infty} f(x_{(1)}, \ldots, x_{(r)}, u_1, \ldots, u_{n-r})\, du_1 \cdots du_{n-r},$$

en donde f es la densidad conjunta de $X_{(1)}, \ldots, X_{(n)}$. La primera observación puede ser cualquiera de los n tiempos de vida, la segunda observación puede obtenerse de cualquiera de los $n-1$ tiempos restantes, y así sucesivamente, hasta la r-ésima observación, para la cual se tienen $n-r+1$ posibilidades.

El resto de los tiempos de vida son los tiempos censurados. Entonces

$$\begin{aligned}
&f(x_{(1)}, \ldots, x_{(r)}, u_1, \ldots, u_{n-r}) \\
&\quad = n(n-1)\cdots(n-r+1) f_{X_1,\ldots,X_n}(x_{(1)}, \ldots, x_{(r)}, u_1, \ldots, u_{n-r}) \\
&\quad = \frac{n!}{(n-r)!} f_{X_1,\ldots,X_n}(x_{(1)}, \ldots, x_{(r)}, u_1, \ldots, u_{n-r}) \\
&\quad = \frac{n!}{(n-r)!} f(x_{(1)}) \cdots f(x_{(r)}) \cdot f(u_1) \cdots f(u_{n-r}),
\end{aligned}$$

en donde ahora se cumple la hipótesis de independencia y la función f que aparece en la última expresión es la densidad univariada inicial. Entonces

$$\begin{aligned}
L(\theta) &= \int_{x_{(r)}}^{\infty} \cdots \int_{x_{(r)}}^{\infty} \frac{n!}{(n-r)!} f(x_{(1)}) \cdots f(x_{(r)})\, f(u_1) \cdots f(u_{n-r})\, du_1 \cdots du_{n-r} \\
&= \frac{n!}{(n-r)!} \, [\prod_{i=1}^{r} f(x_{(i)})\,] \, [\, S(x_{(r)})\,]^{n-r}.
\end{aligned}$$

■

Observe que la función de verosimilitud (3.9) se reduce a la verosimilitud usual $L(\theta) = f(x_1) \cdots f(x_n)$ cuando $r = n$, es decir, cuando en realidad no hay censura. La facilidad en la maximización de la función de verosimilitud (3.9) dependerá de las expresiones con las que se cuenten para $f(x)$ y $S(x)$ del modelo paramétrico particular. En el caso exponencial los cálculos son sencillos y se muestran a continuación.

Ejemplo 3.3 (Tiempo de vida exponencial)
Suponiendo el modelo exp(λ) para un conjunto de observaciones $x_1, \ldots, x_n$ que están sujetas a censura por la derecha tipo II, la función de verosimilitud (3.9) *es*

$$\begin{aligned}
L(\lambda) &\propto f(x_{(1)}) \cdots f(x_{(r)}) \cdot [S(x_{(r)})]^{n-r} \\
&= [\prod_{i=1}^{r} \lambda e^{-\lambda x_{(i)}}\,] \cdot [e^{-\lambda x_{(r)}}]^{n-r} \\
&= \lambda^r \cdot \exp\{-\lambda(\sum_{i=1}^{r} x_{(i)} + (n-r)\, x_{(r)})\}.
\end{aligned}$$

Derivando respecto de λ e igualando a cero se encuentra que

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{r} x_{(i)} + (n-r)\, x_{(r)}}. \tag{3.10}$$

Verificando que $L''(\hat{\lambda}) < 0$, puede comprobarse que la función de verosimilitud tiene un máximo en el punto $\hat{\lambda}$. Observemos que cuando $r = n$, la estimación $\hat{\lambda}$ se reduce a la expresión $1/\bar{x}$. Esta es la estimación máximo verosímil que se conoce cuando todos los datos son observados.

Por otro lado, para la variable aleatoria $Y := \sum_{i=1}^{r} X_{(i)} + (n-r)\, X_{(r)}$, puede comprobarse que $2\lambda Y$ tiene distribución $\chi^2(2r)$. Véase el texto de P. J. Smith [47]. Esto permite encontrar intervalos de confianza y llevar a cabo pruebas de hipótesis para el parámetro λ a partir de datos bajo el tipo de censura indicado. ∎

A continuación presentamos el caso de la distribución Weibull para datos con censura por la derecha tipo II. Encontraremos nuevamente la dificultad técnica de resolver un sistema de dos ecuaciones simultáneas. Dejaremos el desarrollo de este caso hasta el planteamiento de dichas ecuaciones.

Ejemplo 3.4 (Tiempo de vida Weibull)
Suponga que se acepta el modelo Weibull(α, λ) *para una colección de observaciones $x_1, \ldots, x_n$ que están sujetas a censura por la derecha tipo II. La función de verosimilitud* (3.9) *es*

$$\begin{aligned} L(\alpha, \lambda) &= \frac{n!}{(n-r)!}\, f(x_{(1)}) \cdots f(x_{(r)}) \cdot [S(x_{(r)})]^{n-r} \\ &= \frac{n!}{(n-r)!}\, \Big[\prod_{i=1}^{r} \alpha\lambda x_{(i)}^{\alpha-1} e^{-\lambda x_{(i)}^{\alpha}} \Big] \cdot [e^{-\lambda x_{(r)}^{\alpha}}]^{n-r}. \end{aligned}$$

Tomando logaritmo,

$$\begin{aligned} \log L(\alpha, \lambda) &= \log \frac{n!}{(n-r)!} + r \log \alpha + r \log \lambda \\ &\quad + \sum_{i=1}^{r} (\alpha - 1) \log x_{(i)} - \lambda \sum_{i=1}^{r} x_{(i)}^{\alpha} - \lambda (n-r) x_{(r)}^{\alpha}. \end{aligned}$$

Derivando respecto de α y λ, e igualando a cero se encuentra el sistema de ecuaciones

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log L(\alpha,\lambda) &= \frac{r}{\alpha}+\sum_{i=1}^{r}\log x_{(i)}-\lambda\sum_{i=1}^{r}x_{(i)}^{\alpha}\log x_{(i)}\\ &\quad -\lambda(n-r)x_{(r)}^{\alpha}\log x_{(r)} = 0, \qquad (3.11)\\ \frac{\partial}{\partial\lambda}\log L(\alpha,\lambda) &= \frac{r}{\lambda}-\sum_{i=1}^{r}x_{(i)}^{\alpha}-(n-r)x_{(r)}^{\alpha} = 0. \qquad (3.12)\end{aligned}$$

En el caso cuando no hay censura ($r=n$), es inmediato comprobar que las ecuaciones (3.11) *y* (3.12) *se reducen a las presentadas antes en* (3.7) *y* (3.8). *Por otro lado, cuando $\alpha=1$, la ecuación* (3.12) *produce el mismo estimador que se había encontrado en* (3.10), *correspondiente a la distribución exp(λ), es decir,*

$$\hat{\lambda}=\frac{r}{\sum_{i=1}^{r}x_{(i)}+(n-r)\,x_{(r)}}.$$

Más particularmente, cuando no hay datos censurados ($r=n$), se obtiene la estimación $\hat{\lambda}=1/\sum_{i=1}^{n}x_i$. ∎

Función de verosimilitud bajo censura tipo I aleatoria

Sea $X_1,\ldots,X_n$ una muestra aleatoria de un tiempo de vida continuo X con distribución dependiente de un parámetro θ, con función de densidad $f(x)$ y función de supervivencia $S(x)$. Suponga que cada elemento X_i de la muestra aleatoria posee un tiempo de censura por la derecha que es aleatorio y está dado por la variable continua $T_i>0$, cuya función de densidad es $f_{T_i}(t)$ y su función de supervivencia es $S_{T_i}(t)$. Supondremos que las variables aleatorias X_i y T_i son independientes.

La función δ_i que indica si el dato X_i no presenta censura es la variable aleatoria

$$\delta_i=\begin{cases}1 & \text{si } X_i\leqslant T_i,\\ 0 & \text{si } X_i>T_i.\end{cases}$$

De esta manera, $\delta_i = 1$ cuando no hay censura, es decir, hay un fallecimiento ($\delta =$*death*), y $\delta_i = 0$ cuando hay censura. Notemos que el valor observado se puede escribir como la variable aleatoria $Z_i = \text{mín}\{X_i, T_i\}$, para $i = 1, \ldots, n$. Bajo este esquema de censura aleatoria por la derecha, la función de verosimilitud de la muestra es la siguiente.

Proposición 3.4 *Sea $X_1, \ldots, X_n$ una muestra aleatoria de un tiempo de vida X para el cual se asume un modelo paramétrico dependiente de θ, con función de densidad continua $f(x)$ y función de supervivencia $S(x)$. Suponga que cada elemento X_i de la muestra puede ser censurado por un tiempo aleatorio $T_i > 0$, independiente de X_i y con función de densidad $f_{T_i}(t)$ y función de supervivencia $S_{T_i}(t)$. Sea $z_1, \ldots, z_n$ el conjunto de observaciones de las variables $Z_i = \text{mín}\{X_i, T_i\}$, para $i = 1, \ldots, n$. Entonces la función de verosimilitud de la muestra observada es*

$$L(\theta) = \prod_{i=1}^{n} [S(z_i)\, f_{T_i}(z_i)]^{1-\delta_i} \cdot [S_{T_i}(z_i)\, f(z_i)]^{\delta_i}. \tag{3.13}$$

Demostración. Para cualquier $z > 0$,

$$\begin{aligned} P(Z_i < z, \delta_i = 0) &= P(T_i < z, X_i > T_i) \\ &= \int_0^\infty P(T_i < z, X_i > T_i \,|\, T_i = t)\, f_{T_i}(t)\, dt \\ &= \int_0^\infty P(t < z, X_i > t)\, f_{T_i}(t)\, dt \\ &= \int_0^z S(t)\, f_{T_i}(t)\, dt. \end{aligned}$$

Derivando respecto de z,

$$f_{Z_i,\delta_i}(z, 0) = S(z)\, f_{T_i}(z). \tag{3.14}$$

Hemos usado aquí la hipótesis de que el integrando es una función continua. A continuación haremos el mismo cálculo bajo la condición $\delta = 1$ y ahora

conviene condicionar respecto del valor de X_i,

$$\begin{aligned} P(Z_i < z, \delta_i = 1) &= P(X_i < z, X_i \leqslant T_i) \\ &= \int_0^\infty P(X_i < z, X_i \leqslant T_i \,|\, X_i = x)\, f(x)\, dx \\ &= \int_0^\infty P(x < z, x \leqslant T_i)\, f(x)\, dx \\ &= \int_0^z S_{T_i}(x)\, f(x)\, dx. \end{aligned}$$

Derivando respecto de z,

$$f_{Z_i,\delta_i}(z, 1) = S_{T_i}(z)\, f(z). \tag{3.15}$$

Las fórmulas (3.14) y (3.15) se pueden escribir en una sola expresión de la siguiente forma

$$f(z, \delta) = [S(z)\, f_{T_i}(z)]^{1-\delta_i} \cdot [S_{T_i}(z)\, f(z)]^{\delta_i}.$$

Por la hipótesis de independencia, la función de verosimilitud de la muestra $z_1, \ldots, z_n$ es el producto

$$L(\theta) = \prod_{i=1}^{n} [S(z_i)\, f_{T_i}(z_i)]^{1-\delta_i} \cdot [S_{T_i}(z_i)\, f(z_i)]^{\delta_i}.$$

■

Es de utilidad observar que cuando la distribución de los tiempos aleatorios de censura T_i no dependen del parámetro θ, los factores $f_{T_i}(z_i)$ y $S_{T_i}(z_i)$ son constantes respecto de θ, de modo que la fórmula (3.13) se puede escribir en forma reducida como

$$L(\theta) \propto \prod_{i=1}^{n} [f(z_i)]^{\delta_i}\ [S(z_i)]^{1-\delta_i}. \tag{3.16}$$

Cuando los tiempos aleatorios de censura son las constantes $t_1, \ldots, t_n$, tenemos que $z_i = \text{mín}\,\{x_i, t_i\}$ y distinguimos dos casos:

- Cuando no hay censura ($\delta_i = 1$), $z_i = x_i$ y el factor distinto de 1 en (3.13) es $S_{T_i}(z_i)\, f(z_i) \propto f(z_i)$.

- Cuando hay censura ($\delta_i = 0$), $z_i = t_i$ y el factor distinto de 1 en (3.13) es $S(z_i)\, f_{T_i}(z_i) \propto S(z_i)$.

Así, la función de verosimilitud recién demostrada (3.13) toma la forma de la función de verosimilitud (3.3) encontrada antes en la página 114, cuando los tiempos de censura por la derecha son cantidades fijas.

Como antes, las funciones de verosimilitud (3.13) ó (3.16), podrán ser de fácil manejo dependiendo de la complejidad de las expresiones para $f(x)$ y $S(x)$. El caso exponencial es sencillo y se muestra a continuación.

Ejemplo 3.5 (Tiempo de vida exponencial)
Sea $X_1, \ldots, X_n$ una muestra aleatoria de la distribución exp(λ) sujeta a censura aleatoria por la derecha dada por los tiempos aleatorios $T_1, \ldots, T_n$, respectivamente. Recordemos que las observaciones se expresan como $Z_i = \min\{X_i, T_i\}$ y la presencia de censura se escribe a través de las funciones indicadoras

$$\delta_i = \begin{cases} 1 & si\ X_i \leqslant T_i, \\ 0 & si\ X_i > T_i. \end{cases}$$

De este modo, los datos se pueden escribir como $(z_1, \delta_1), \ldots, (z_n, \delta_n)$. Es conveniente también definir las cantidades $r = \delta_1 + \cdots + \delta_n$ y $y = z_1 + \cdots + z_n$. Suponiendo entonces que la distribución de los tiempos de censura T_i no dependen del parámetro λ, la expresión de la función de verosimilitud (3.13) *es*

$$\begin{aligned} L(\lambda) &\propto \prod_{i=1}^{n} [f(z_i)]^{\delta_i}\ [S(z_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda^{\delta_i}\, e^{-\lambda z_i \delta_i}\, e^{-\lambda z_i (1-\delta_i)} \\ &= \lambda^r \cdot e^{-\lambda y}. \end{aligned}$$

Derivando respecto de λ e igualando a cero se encuentra la estimación $\hat{\lambda} = r/y = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} z_i$. Cuando no hay censura ($\delta_1 = 1, \ldots, \delta_n = 1$), se tiene que $r = n$ y la estimación anterior arroja el valor $\hat{\lambda} = 1/\bar{x}$.

Como variable aleatoria, el estimador se puede escribir como $\hat{\lambda} = R/Y$. *En este caso, la distribución de* $\hat{\lambda}$ *no es fácil de obtener como ocurrió en el caso de la censura tipo II para este modelo.* ■

3.4. Función de verosimilitud para datos truncados

Supongamos nuevamente que hemos adoptado un modelo paramétrico para un tiempo de vida X. En esta breve sección se explica la forma de encontrar la verosimilitud de una muestra que presenta algún tipo de truncamiento.

- Truncamiento por la izquierda

  Si una observación x está truncada por la izquierda por el valor $t > 0$, la verosimilitud de la observación es la distribución condicionada al evento $(X > t)$, es decir,

  $$f(x \mid X > t) = \frac{f(x)}{S(t)}, \quad x > t.$$

  Si tenemos una colección de observaciones $x_1, \ldots, x_n$ que fueron obtenidas de manera independiente, y que presentan truncamiento por la izquierda por los valores $t_1, \ldots, t_n$, respectivamente, entonces la función de verosimilitud completa es

  $$L(\theta) = \prod_{i=1}^{n} \frac{f(x_i)}{S(t_i)}, \quad x_i > t_i. \tag{3.17}$$

- Truncamiento por la derecha

  En el caso de truncamiento por la derecha por el valor $t > 0$, la verosimilitud de una observación x es

  $$f(x \mid X < t) = \frac{f(x)}{1 - S(t)}, \quad x < t.$$

  De modo que la función de verosimilitud de una muestra $x_1, \ldots, x_n$ con truncamiento por la derecha por los valores $t_1, \ldots, t_n$, respectivamente, es

  $$L(\theta) = \prod_{i=1}^{n} \frac{f(x_i)}{1 - S(t_i)}, \quad x_i < t_i. \tag{3.18}$$

Las expresiones (3.17) y (3.18) se simplifican cuando los tiempos de truncamiento t_i son un mismo valor t para todos los datos. Consideraremos esta situación en el siguiente ejemplo.

Ejemplo 3.6 (Tiempo de vida exponencial)
Sea $x_1, \ldots, x_n$ una muestra de la distribución exp(λ) truncada por la izquierda por un valor fijo $t \geqslant 0$. Supondremos que el parámetro λ es desconocido y que se desea estimar por máxima verosimilitud. La función de verosimilitud (3.17) *es*

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \frac{f(x_i)}{S(t)}, \quad x_i > t \\ &= \prod_{i=1}^{n} \frac{\lambda \exp\{-\lambda x_i\}}{\exp\{-\lambda t\}} \\ &= \lambda^n \exp\{\lambda n(t - \bar{x})\}. \end{aligned}$$

Tomando logaritmo, derivando respecto de λ e igualando a cero se encuentra la estimación $\hat{\lambda}$ que aparece abajo. Observe que cuando $t = 0$, es decir, cuando no hay truncamiento, se obtiene el estimador conocido $\hat{\lambda} = 1/\bar{x}$.

$$\hat{\lambda} = \frac{1}{\bar{x} - t}.$$

■

Un poco más generalmente, también pueden considerarse escenarios en donde los datos presentan censura y truncamiento al mismo tiempo, y se desea encontrar la función de verosimilitud para estimar un parámetro. No trataremos esa situación en este trabajo. En los libros de R. C. Elandt-Johnson y N. L. Johnson [19], E. T. Lee y J. W. Wang [37] ó P. J. Smith[47], pueden encontrarse otras exposiciones del tema de funciones de verosimiltud para datos censurados.

3.5. Ejercicios

Distribuciones paramétricas

102. Sea $x_1, \ldots, x_n$ una muestra de una distribución continua con función de supervivencia $S(x; \theta)$, dependiente de un parámetro θ. Demuestre

que el estimador máximo verosímil para θ es la solución de la ecuación

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta}\frac{\partial}{\partial x}S(x_i;\theta)}{\frac{\partial}{\partial x}S(x_i;\theta)} = 0. \tag{3.19}$$

103. *Tiempo de vida uniforme.*
Sea X un tiempo de vida con distribución unif$(0,\omega)$, en donde $\omega > 0$ es una constante. Demuestre las siguientes expresiones para las funciones básicas de X. Grafique estas funciones.

   a) $F(x) = x/\omega$, para $0 < x < \omega$.
   b) $S(x) = (\omega - x)/\omega$, para $0 < x < \omega$.
   c) $\lambda(x) = 1/(\omega - x)$, para $0 < x < \omega$.
   d) $\Lambda(x) = \log(\omega/(\omega - u))$, para $0 < x < \omega$.
   e) $R(x) = (w - x)/2$, para $0 < x < \omega$.

104. *Unicidad de la distribución con función tiempo promedio de vida residual constante.*
Sea X una variable aleatoria cuya distribución tiene soporte el intervalo $(0,\infty)$ y con esperanza finita. Suponga que su función tiempo promedio de vida residual es constante $R(x) = 1/\lambda > 0$, para $x > 0$. Demuestre que X tiene distribución exp(λ).

105. *Distribución Gompertz.*
Se ha definido la densidad Gompertz(b,c) como

$$f(x) = b\exp\{\, cx + \frac{b}{c}\,(1 - e^{cx})\,\}, \quad \text{para } \ x > 0,$$

en donde $b > 0$ y $c > 0$ son dos parámetros.

   a) Demuestre que cuando $c \searrow 0$, la densidad Gompertz(b,c) converge a la densidad exp(b).
   b) Tomando $b = \lambda > 0$ y $c = \log\varphi > 0$, compruebe que la densidad Gompertz(b,c) también se puede escribir como

$$f(x) = \lambda\varphi^x \exp\{-\frac{\lambda}{\log\varphi}(\varphi^x - 1)\}, \quad \text{para } \ x > 0,$$

   en donde ahora los parámetros son $\lambda > 0$ y $\varphi > 1$.

106. *Tiempo de vida Gompertz.*
Sea X un tiempo de vida con distribución Gompertz(b, c). Demuestre las siguientes expresiones para las funciones básicas de X.

a) $F(x) = 1 - \exp\{\frac{b}{c}(1 - e^{cx})\}$, para $x > 0$.

b) $S(x) = \exp\{\frac{b}{c}(1 - e^{cx})\}$, para $x > 0$.

c) $f(x) = \lambda(x)\, S(x) = b \exp\{cx + \frac{b}{c}(1 - e^{cx})\}$, para $x > 0$.

d) $\Lambda(x) = \frac{b}{c}(e^{cx} - 1)$.

e) $R(x) = \dfrac{1}{S(x)} \displaystyle\int_x^\infty S(u)\, du$, para $x > 0$.

107. *Relación Weibull-exponencial.*
Sean $\alpha > 0$ y $\lambda > 0$ dos constantes. Demuestre que

$$X \sim \text{Weibull}(\alpha, \lambda) \;\Leftrightarrow\; \lambda X^\alpha \sim \exp(1).$$

108. *Tiempo de vida Weibull.*
Suponga que un tiempo de vida X sigue una distribución Weibull(α, λ) con función de densidad

$$f(x) = \alpha\, \lambda\, x^{\alpha-1} \exp\{-\lambda\, x^\alpha\}, \quad \text{para} \;\; x > 0.$$

Demuestre que:

a) $F(x) = 1 - \exp\{-\lambda x^\alpha\}$, para $x > 0$.

b) $S(x) = \exp\{-\lambda x^\alpha\}$, para $x > 0$.

c) $\lambda(x) = \alpha\, \lambda\, x^{\alpha-1}$, para $x > 0$.

d) $\Lambda(x) = \lambda\, x^\alpha$, para $x > 0$.

e) $R(x) = \exp(\lambda x^\alpha) \displaystyle\int_x^\infty \alpha\, \lambda\, u^\alpha \exp(-\lambda u^\alpha)\, du - x$, para $x > 0$.

109. *Tiempo de vida Weibull.*
Sea X un tiempo de vida que sigue una distribución Weibull(α, λ). Grafique $\lambda(x)$ y demuestre que esta función es:

a) Creciente cuando $\alpha > 1$.

b) Decreciente cuando $\alpha < 1$.

c) Constante cuando $\alpha = 1$.

110. *Otras parametrizaciones de la distribución Weibull.*
Sean $\alpha > 0$ y $\lambda > 0$ dos parámetros. La densidad Weibull(α, λ) se define como la función $f(x)$ que aparece abajo. En el caso $\alpha = 1$ se obtiene la distribución exp(λ).

$$f(x) = \alpha\,\lambda\,x^{\alpha-1}\,\exp\{-\lambda\,x^{\alpha}\}, \quad \text{para} \;\; x > 0.$$

a) Tomando $\lambda = \theta^{\alpha}$, compruebe que

$$f(x) = \alpha\,\theta(\theta x)^{\alpha-1}\,\exp\{-(\theta x)^{\alpha}\}, \quad \text{para} \;\; x > 0,$$

b) Tomando $\lambda = 1/\theta$, compruebe que

$$f(x) = (\alpha/\theta)\,x^{\alpha-1}\,\exp\{-x^{\alpha}/\theta\}, \quad \text{para} \;\; x > 0,$$

c) Tomando $\lambda = (1/\theta)^{\alpha}$, compruebe que

$$f(x) = \frac{\alpha}{\theta}\,(x/\theta)^{\alpha-1}\,\exp\{-(x/\theta)^{\alpha}\}, \quad \text{para} \;\; x > 0,$$

111. *Distribución de valor extremo.*
Sea X una variable aleatoria con distribución Weibull(α, λ). A la distribución de $Y = \log X$ se le llama distribución de valor extremo y se le denota por EV(α, λ). Este no es un tiempo de vida pues sus valores son $(-\infty, \infty)$.

a) Encuentre $S_Y(y) = P(Y > y)$.

b) Encuentre $f_Y(y)$.

112. *Tiempo de vida Pareto.*
Suponga que un tiempo de vida X sigue una distribución Pareto(α, γ) con función de densidad

$$f(x) = \frac{\alpha}{x}\,\left(\frac{\gamma}{x}\right)^{\alpha}, \quad \text{para} \;\; x \geqslant \gamma.$$

Demuestre que:

a) $F(x) = 1 - \left(\dfrac{\gamma}{x}\right)^{\alpha}$, para $x \geqslant \gamma$.

b) $S(x) = \left(\dfrac{\gamma}{x}\right)^{\alpha}$, para $x \geqslant \gamma$.

c) $\lambda(x) = \dfrac{\alpha}{x}$, para $x \geqslant \gamma$.

d) $\Lambda(x) = \alpha \log(x/\gamma)$, para $x \geqslant \gamma$.

e) $R(x) = \dfrac{x}{\alpha - 1}$, para $x \geqslant \gamma$.

113. *Distribución Pareto.*
Sea X una variable aleatoria con distribución Pareto(α, γ), es decir, su función de densidad es

$$f(x) = \frac{\alpha}{x}\left(\frac{\gamma}{x}\right)^{\alpha}, \quad \text{para} \quad x \geqslant \gamma.$$

a) Demuestre que $E(X) = \alpha\gamma/(\alpha - 1)$ cuando $\alpha > 1$.

b) Demuestre que $\gamma + R(\gamma) = E(X)$ cuando $\alpha > 1$.

c) Suponga que los parámetros son tales que $\gamma < \alpha$ y sea $t > 0$ fijo tal que $\gamma < t < \alpha$. Demuestre que $X \mid (X > t)$ tiene distribución Pareto$(\alpha - t, \gamma)$.

114. *Tiempo de vida Rayleigh.*
Suponga que un tiempo de vida X sigue una distribución Rayleigh(θ), es decir, función de densidad es

$$f(x) = \frac{2x}{\theta}\, e^{-x^2/\theta}, \quad \text{para} \quad x > 0,$$

en donde $\theta > 0$ es un parámetro. Demuestre que:

a) $F(x) = 1 - e^{-x^2/\theta}$, para $x > 0$.

b) $S(x) = e^{-x^2/\theta}$, para $x > 0$.

c) $\lambda(x) = 2x/\theta$, para $x > 0$.

d) $\Lambda(x) = x^2/\theta$, para $x > 0$.

e) $R(x) = e^{x^2/\theta} \displaystyle\int_x^{\infty} e^{-u^2/\theta}\, du$, para $x > 0$.

Función de verosimilitud para datos censurados

115. Sea X un tiempo de vida continuo con función de distribución $F(x)$ y función de supervivencia $S(x)$. Sea $C > 0$ otra variable aleatoria continua con función de densidad $f_C(c)$. Demuestre que:

a) $P(X > C) = \int_0^\infty S(c)\, f_C(c)\, dc.$

b) $P(X > C) = \int_0^\infty F(c)\, f_C(c)\, dc.$

Función de verosimilitud bajo censura tipo I

116. Para la estimación del parámetro θ en el caso de censura por la derecha tipo I, en la Proposición 3.2 se establece que la función de verosimilitud es

$$L(\theta) = \prod_{i=1}^{n} [f(x_i)]^{\delta_i} [S(x_i)]^{1-\delta_i}.$$

a) Demuestre que se cumple la igualdad que aparece abajo, en donde $\lambda(x)$ es la función de riesgo y $\Lambda(x)$ es la función de riesgo acumulado.

$$\log L(\theta) = \sum_{i=1}^{n} \delta_i \log \lambda(x_i) - \sum_{i=1}^{n} \Lambda(x_i).$$

b) Demuestre que para datos completos ($\delta_1 = 1, \ldots, \delta_n = 1$), la expresión del inciso anterior se reduce al logaritmo de la función de verosimilitud usual, es decir,

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i).$$

117. Sean $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ observaciones con posible censura por la derecha de un tiempo de vida X con función de supervivencia $S(x; \theta)$, dependiente de un parámetro desconocido θ.

a) Demuestre que el estimador máximo verosímil para θ es la solución de la ecuación

$$\sum_{i=1}^{n} \delta_i \frac{\frac{\partial}{\partial\theta}\frac{\partial}{\partial x}S(x_i;\theta)}{\frac{\partial}{\partial x}S(x_i;\theta)} + \sum_{i=1}^{n}(1-\delta_i)\frac{\frac{\partial}{\partial\theta}S(x_i;\theta)}{S(x_i;\theta)} = 0.$$

b) Demuestre que para datos completos ($\delta_1 = 1, \ldots, \delta_n = 1$), la expresión del inciso anterior se reduce a la ecuación (3.19) que aparece en el Ejercicio 102.

118. *Tiempo de vida Rayleigh.*
Sean $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ observaciones con posible censura por la derecha de un tiempo de vida X con distribución Rayleigh(θ), es decir, con función de densidad $f(x)$ como aparece abajo, en donde $\theta > 0$ es desconocido.

$$f(x) = \frac{2x}{\theta}\, e^{-x^2/\theta}, \quad \text{si} \;\; x > 0,$$

a) Demuestre que el estimador máximo verosímil para θ es

$$\hat{\theta} = (\sum_{i=1}^{n} x_i^2) \,/\, (\sum_{i=1}^{n} \delta_i).$$

b) Encuentre $\hat{\theta}$ cuando ninguna de las observaciones está censurada.

119. *Tiempo de vida Gompertz.*
Sean $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ observaciones con posible censura por la derecha de un tiempo de vida X con distribución Gompertz(b, c), es decir, con función de densidad $f(x)$ como aparece abajo, en donde los parámetros $b > 0$ y $c > 0$ son desconocidos.

$$f(x) = b \exp\{\, cx + \frac{b}{c}\,(1 - e^{cx})\,\}, \quad \text{para} \;\; x > 0.$$

a) Demuestre que los estimadores por máxima verosimilitud $\hat{b}$ y $\hat{c}$ son la solución al sistema de ecuaciones

$$\begin{aligned} c\sum_{i=1}^{n}\delta_i + b\sum_{i=1}^{n}(1-e^{cx_i}) &= 0, \\ c^2\sum_{i=1}^{n}\delta_i x_i - bc\sum_{i=1}^{n} x_i e^{cx_i} - b\sum_{i=1}^{n}(1-e^{cx_i}) &= 0. \end{aligned}$$

b) Demuestre que

$$\lim_{c \searrow 0} \hat{b} = (\sum_{i=1}^{n} \delta_i) \,/\, (\sum_{i=1}^{n} x_i).$$

Este es el estimador por máxima verosimilitud para datos censurados por la derecha para tiempos de vida con distribución $\exp(b)$. Recordemos que cuando $c \searrow 0$, la densidad Gompertz(b, c) converge a la densidad $\exp(b)$.

120. *Tiempo de vida Pareto.*
Sean $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ observaciones con posible censura por la derecha de un tiempo de vida X con distribución Pareto(α, γ), es decir, con función de densidad $f(x)$ como aparece abajo, en donde los parámetros $\alpha > 0$ y $\gamma > 0$ son desconocidos.

$$f(x) = \frac{\alpha}{x} \left(\frac{\gamma}{x}\right)^{\alpha}, \quad \text{para} \;\; x \geqslant \gamma.$$

a) Demuestre que los estimadores máximo verosímiles para α y γ son

$$\begin{aligned} \hat{\gamma} &= x_{(1)}, \\ \hat{\alpha} &= (\sum_{i=1}^{n} \delta_i) \,/\, (\sum_{i=1}^{n} \log (x_i / x_{(1)})). \end{aligned}$$

b) Encuentre $\hat{\alpha}$ cuando ninguna de las observaciones está censurada.

121. Sean $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ observaciones con posible censura por la derecha de un tiempo de vida X con función de densidad $f(x)$ como aparece abajo, en donde $\theta > -2$ es un parámetro desconocido.

$$f(x) = (\theta + 2)\, e^{-(\theta+2)x}, \quad \text{si} \;\; x > 0.$$

Demuestre que el estimador máximo verosímil para θ es

$$\hat{\theta} = (\sum_{i=1}^{n} \delta_i) \,/\, (\sum_{i=1}^{n} x_i) - 2.$$

Función de verosimilitud bajo censura tipo II

122. Para la estimación del parámetro θ en el caso de censura por la derecha tipo II, en la Proposición 3.3 se establece que la función de verosimilitud es

$$L(\theta) = \frac{n!}{(n-r)!}\,[\prod_{i=1}^{r} f(x_i)\,]\,[\,S(x_{(r)})\,]^{n-r}.$$

a) Demuestre que se cumple la igualdad que aparece abajo, en donde $\lambda(x)$ es la función de riesgo y $\Lambda(x)$ es la función de riesgo acumulado.

$$\log L(\theta) = \log \frac{n!}{(n-r)!} + \sum_{i=1}^{r} \log \lambda(x_i) - \sum_{i=1}^{r} \Lambda(x_i) - (n-r)\Lambda(x_{(r)}).$$

b) Demuestre que para datos completos ($r = n$), la expresión del inciso anterior se reduce al logaritmo de la función de verosimilitud usual, es decir,

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i).$$

123. Sean $x_1, \ldots, x_n$ observaciones con posible censura por la derecha tipo II de un tiempo de vida continuo X con función de supervivencia $S(x;\theta)$, dependiente de un parámetro desconocido θ.

a) Demuestre que el estimador máximo verosímil para θ es la solución de la ecuación

$$\sum_{i=1}^{r} \frac{\frac{\partial}{\partial\theta}\frac{\partial}{\partial x}S(x_i;\theta)}{\frac{\partial}{\partial x}S(x_i;\theta)} + (n-r)\frac{\frac{\partial}{\partial\theta}S(x_{(r)};\theta)}{S(x_{(r)};\theta)} = 0.$$

b) Demuestre que para datos completos ($r = n$), la expresión del inciso anterior se reduce a la ecuación (3.19) que aparece en el Ejercicio 102.

124. *Tiempo de vida Rayleigh.*
Sean $x_1, \ldots, x_n$ observaciones con posible censura por la derecha tipo II de un tiempo de vida X con distribución Rayleigh(θ), es decir, con

función de densidad $f(x)$ como aparece abajo, en donde $\theta > 0$ es desconocido.

$$f(x) = \frac{2x}{\theta}\, e^{-x^2/\theta}, \quad \text{si} \;\; x > 0,$$

a) Demuestre que el estimador máximo verosímil para θ es

$$\hat{\theta} = \frac{1}{r}\sum_{i=1}^{r} x_i^2 + \frac{n-r}{r}\, x_{(r)}^2.$$

b) Encuentre $\hat{\theta}$ cuando ninguna de las observaciones está censurada.

125. *Tiempo de vida Gompertz.*
Sean $x_1, \ldots, x_n$ observaciones con posible censura por la derecha tipo II de un tiempo de vida X con distribución Gompertz(b, c), es decir, con función de densidad $f(x)$ como aparece abajo, en donde los parámetros $b > 0$ y $c > 0$ son desconocidos.

$$f(x) = b \exp\{\, cx + \frac{b}{c}\,(1 - e^{cx})\,\}, \quad \text{para} \;\; x > 0.$$

a) Demuestre que los estimadores por máxima verosimilitud $\hat{b}$ y $\hat{c}$ son la solución al sistema de ecuaciones

$$\begin{aligned}
&\frac{r}{b} + \frac{1}{c}\sum_{i=1}^{r}(1 - e^{cx_{(i)}}) + (n-r)\frac{1}{c}(1 - e^{cx_{(r)}}) \;\; = \;\; 0, \\
&\sum_{i=1}^{r} x_{(i)} - \frac{b}{c^2}\sum_{i=1}^{r}(1 - e^{cx_{(i)}}) + \frac{b}{c}\sum_{i=1}^{r} x_{(i)} e^{cx_{(i)}} \\
&\qquad -(n-r)\frac{b}{c^2}(1 - e^{cx_{(r)}}) - (n-r)\frac{b}{c} x_{(r)} e^{cx_{(r)}} \;\; = \;\; 0.
\end{aligned}$$

b) Demuestre que

$$\lim_{c \searrow 0} \hat{b} = \frac{r}{\displaystyle\sum_{i=1}^{r} x_{(i)} + (n-r)\, x_{(r)}}.$$

Este es el estimador por máxima verosimilitud para datos censurados por la derecha tipo II para tiempos de vida con distribución

$\exp(b)$, el cual se presentó antes en la ecuación (3.10). Recordemos nuevamente que, cuando $c \searrow 0$, la densidad Gompertz(b, c) converge a la densidad $\exp(b)$. Más particularmente, cuando no hay datos censurados, es decir, cuando $r = n$, se cumple que $\lim_{c \searrow 0} \hat{b} = 1/\bar{x}$.

126. *Tiempo de vida Pareto.*
Sean $x_1, \ldots, x_n$ observaciones con posible censura por la derecha tipo II de un tiempo de vida X con distribución Pareto(α, γ), es decir, con función de densidad $f(x)$ como aparece abajo, en donde los parámetros $\alpha > 0$ y $\gamma > 0$ son desconocidos.

$$f(x) = \frac{\alpha}{x} \left(\frac{\gamma}{x}\right)^{\alpha}, \quad \text{para} \;\; x \geqslant \gamma.$$

a) Demuestre que los estimadores máximo verosímiles para α y γ son

$$\begin{aligned} \hat{\gamma} &= x_{(1)}, \\ \hat{\alpha} &= \frac{r}{\sum_{i=1}^{r} \log\left(x_{(i)}/x_{(1)}\right) + (n-r) \log\left(x_{(r)}/x_{(1)}\right)}. \end{aligned}$$

b) Encuentre $\hat{\alpha}$ cuando ninguna de las observaciones está censurada.

127. Sean $x_1, \ldots, x_n$ observaciones con posible censura por la derecha tipo II de un tiempo de vida X con función de densidad $f(x)$ como aparece abajo, en donde $\theta > -1$ es un parámetro desconocido.

$$f(x) = (\theta + 1)\, e^{-(\theta+1)x}, \quad \text{si} \;\; x > 0.$$

a) Demuestre que el estimador máximo verosímil para θ es

$$\hat{\theta} = \frac{r}{\sum_{i=1}^{r} x_{(i)} + (n-r)\, x_{(r)}} - 1.$$

b) Encuentre $\hat{\theta}$ cuando ninguna de las observaciones está censurada.

Función de verosimilitud bajo censura tipo I aleatoria

128. Considere el caso de censura tipo I aleatoria en donde los tiempos de vida X_i tienen distribución $\exp(\lambda)$, y los tiempos de censura T_i tienen distribución $\exp(\theta)$, para $i = 1, \ldots, n$. Los tiempos observables son $Z_i = \min\{X_i, T_i\}$ y las variables que indican la presencia o ausencia de censura son

$$\delta_i = \begin{cases} 1 & \text{si } X_i \leqslant T_i, \\ 0 & \text{si } X_i > T_i. \end{cases}$$

Demuestre que:

a) $Z_i \sim \exp(\lambda + \theta)$.

b) $\delta_i \sim \text{Ber}(\lambda/(\lambda + \theta))$.

c) Z_i y δ_i son independientes.

Se demostró antes que el estimador máximo verosímil para λ es

$$\hat{\lambda} = (\sum_{i=1}^{n} \delta_i) \,/\, (\sum_{i=1}^{n} Z_i).$$

Usando los resultados anteriores, demuestre que:

d) $E(\hat{\lambda}) = \dfrac{n}{n-1}\,\lambda, \quad n \geqslant 2.$

e) $\text{Var}(\hat{\lambda}) = \dfrac{n\lambda\theta + n^2\lambda^2}{(n-1)(n-2)} - \dfrac{n^2\lambda^2}{(n-1)^2}, \quad n \geqslant 3.$

Función de verosimilitud para datos truncados

129. *Distribución exponencial.*
Sean $x_1, \ldots, x_n$ observaciones independientes de la distribución $\exp(\lambda)$ truncadas por la derecha por el mismo valor fijo y conocido $t > 0$, en donde λ es desconocido.

a) Demuestre que la distribución de X truncada por la derecha por el valor t es

$$f(x \,|\, X \leqslant t) = \frac{\lambda\, e^{-\lambda x}}{1 - e^{-\lambda t}}, \quad \text{para } \ 0 < x < t.$$

b) Encuentre la función de la verosimilitud de la muestra y compruebe que el estimador por máxima verosimilitud para λ es la solución de la ecuación

$$1 - e^{-\lambda t} = \lambda^2\, e^{-\lambda t} + \lambda \bar{x}(1 - e^{-\lambda t}).$$

c) Compruebe que $\hat{\lambda} = 1/\bar{x}$ cuando $t \to \infty$.

130. *Distribución exponencial.*
Sea X un tiempo de vida con distribución $\exp(\lambda)$.

a) Demuestre que la distribución de X truncada por la izquierda por el valor fijo $t \geqslant 0$ es

$$f(x \,|\, X > t) = \lambda\, e^{-\lambda(x-t)}, \quad \text{para} \;\; x > t.$$

b) Sean $x_1, \ldots, x_n$ observaciones independientes de la distribución $\exp(\lambda)$ truncadas por la izquierda por el valor $t > 0$, en donde λ y t son parámetros desconocidos. Demuestre que los estimadores por máxima verosimilitud para λ y t son

$$\begin{aligned} \hat{\lambda} &= \frac{1}{\bar{x} - x_{(1)}}\,, \\ \hat{t} &= x_{(1)}. \end{aligned}$$

131. *Distribución Pareto.*
Sea X un tiempo de vida con distribución Pareto(α, γ), es decir, su función de supervivencia es

$$S(x) = (\gamma/x)^{\alpha}, \quad \text{para} \;\; x \geqslant \gamma,$$

en donde $\alpha > 0$ y $\gamma > 0$ son dos parámetros desconocidos. Sean $(x_i, t_i), \ldots, (x_n, t_n)$ datos truncados por la izquierda de esta distribución, en donde $X = x_i$ es la observación sujeta a la condición $X > t_i$. Demuestre que los estimadores por máxima verosimilitud para α y γ son

$$\begin{aligned} \hat{\alpha} &= \frac{n}{\displaystyle\sum_{i=1}^{n} \log\left(x_i/x_{(1)}\right)}, \\ \hat{\gamma} &= x_{(1)}. \end{aligned}$$

Capítulo 4

Modelos no paramétricos

Este capítulo contiene una introducción al amplio tema de la aplicación de métodos no paramétricos al análisis de datos de supervivencia. Se revisan algunos métodos para estimar la distribución de un tiempo de vida a partir de un conjunto de observaciones con posible censura por la derecha, bajo determinados contextos y sin recurrir a modelos paramétricos.

4.1. Tablas de mortalidad

Las tablas de mortalidad constituyen una base fundamental de las ciencias actuariales. En esta sección recordaremos brevemente este concepto y definiremos algunas funciones básicas que surgen a partir de la información de estas tablas. Estudiaremos, además, algunas relaciones entre estas funciones. En particular, nos interesa estimar la función de supervivencia de un tiempo de vida a partir de la información contenida en una tabla de mortalidad.

Recordemos que se define una cohorte como un grupo de personas que comparten una característica en común y que se estudian durante un periodo de tiempo dado. Por ejemplo, un conjunto de personas que han nacido en un año particular conforman una cohorte. Una cohorte puede también estar integrada por aquel conjunto de personas a las que se le ha diagnosticado una cierta enfermedad.

Supondremos que tenemos una cohorte compuesta por ℓ_0 individuos en don-

de no hay migración. Se trata de una población cerrada en donde nadie ingresa y nadie sale, y en donde cada individuo eventualmente muere. El número entero ℓ_0 es llamado radix y, convencionalmente, toma el valor $10^5 = 100,000$ o $10^6 = 1,000,000$, aunque en realidad cualquier valor entero positivo es factible.

Definición 4.1 *Una tabla de mortalidad es un arreglo tabular en donde se registran las distintas edades que pueden tomar los individuos de una población y el número de sobrevivientes a esas edades.*

Un ejemplo de una tabla mínima de mortalidad se muestra en la Tabla 4.1. En esta tabla se ha registrado el número de personas vivas ℓ_x para diferentes edades enteras x, y en donde la población inicial consta de $100,000$ individuos. Nos interesa aquí llevar el registro de la mortalidad general para este conjunto de personas. Usualmente las edades para poblaciones humanas y medidas en años cumplidos son $0, 1, 2, \ldots, 120$. La edad máxima para cualquier individuo es denotada por la letra ω, en consecuencia, $\ell_{w+1} = 0$. En este caso hemos supuesto $\omega = 120$ años. Si una tabla de mortalidad inicia en la edad α, al radix se le denota por ℓ_α.

x	ℓ_x
0	100,000
1	98,287
2	97,604
3	96,844
4	96,003
$\vdots$	$\vdots$
120	1

Tabla 4.1: Tabla mínima de mortalidad.

Observe que en una tabla de mortalidad sólo se registra el número de perso-

nas vivas a cada edad y las posibles censuras por la derecha son consideradas como fallecimientos pues son individuos que ya no son reportados como vivos a la siguiente edad. Además, al considerarse que cada edad toma un valor entero (número de años cumplidos), los tiempos de vida están censurados por intervalo. Regularmente en una tabla de mortalidad también aparecen otras funciones que se construyen a partir de los valores de ℓ_x. Estas funciones son análogas a las definidas en la modelación probabilista dada por la función de supervivencia. Supondremos que la edad x de un individuo toma valores en el intervalo continuo $[0, \omega]$, en donde la unidad de medición es un año. Al número entero más grande que es menor o igual a x se le denota por $\lfloor x \rfloor$, esta es la edad en años cumplidos de un individuo de edad exacta x.

Estadísticas básicas

A partir de los valores de ℓ_x en una tabla de mortalidad, se pueden definir varias funciones que son de interés. Estas funciones tienen como principal variable dependiente la edad entera $x \geqslant 0$ de un individuo. Empezaremos reiterando la definición de la función principal ℓ_x.

- **Función número de personas vivas**

  Para cada valor entero $x \geqslant 0$, la función

  $$x \mapsto \ell_x$$

  indica el número de personas vivas a edad x. El valor inicial es el radix ℓ_0 y la función decrece a cero al paso del tiempo: $\ell_0 \geqslant \ell_1 \geqslant \ell_2 \geqslant \cdots \geqslant \ell_\omega \geqslant \ell_{\omega+1} = 0$. Estas cantidades constituyen el elemento esencial para construir otras cantidades de interés que pueden aparecer en una tabla de mortalidad.

- **Función número de fallecidos**

  Para cada valor entero $x \geqslant 0$, se define

  $$d_x := \ell_x - \ell_{x+1}.$$

Este es el número de personas de edad x que mueren antes de cumplir la edad $x+1$. La letra d proviene del término en inglés *death*. En particular, $d_\omega = \ell_\omega - \ell_{\omega+1} = \ell_\omega$. Más generalmente, para $n \geqslant 1$ entero se define

$$ {}_nd_x := \ell_x - \ell_{x+n}. $$

Esta cantidad es el número de personas de edad x que mueren antes de cumplir la edad $x+n$. Cuando $n = 1$ se obtiene ${}_1d_x$, pero se omite el subíndice izquierdo para recuperar la notación d_x.

- **Función probabilidad de fallecer**

  Para cada valor entero $x \geqslant 0$ tal que $\ell_x > 0$, se define

$$ q_x := \frac{d_x}{\ell_x} = \frac{\ell_x - \ell_{x+1}}{\ell_x}. $$

  Esta es la probabilidad de que una persona de edad x fallezca antes de cumplir la edad $x+1$. En otros términos, esta es la probabilidad condicional de que una persona fallezca a la edad x dado que ha cumplido esa edad. En particular, $q_\omega = d_\omega/\ell_\omega = 1$, suponiendo $\ell_\omega > 0$. En cambio, la probabilidad (no condicional) de que una persona fallezca a la edad x es $d_x/\ell_0 = (\ell_x - \ell_{x+1})/\ell_0$. Más generalmente, para $n \geqslant 1$ entero se define

$$ {}_nq_x := \frac{{}_nd_x}{\ell_x} = \frac{\ell_x - \ell_{x+n}}{\ell_x}. $$

  Esta es la probabilidad de que una persona de edad x muera antes de cumplir la edad $x+n$. Cuando $n = 1$ se obtiene ${}_1q_x$, pero se omite el subíndice izquierdo para recuperar la notación q_x. Todavía más generalmente, para cada edad x y para valores $n = 0, 1, \ldots$ y $m = 1, 2, \ldots$ se define

$$ {}_{n|m}q_x := \frac{{}_mq_{x+n}}{\ell_x} = \frac{\ell_{x+n} - \ell_{x+n+m}}{\ell_x}. $$

  Esta es la probabilidad de que un persona de edad x fallezca entre las edades $x+n$ y $x+n+m-1$, inclusive.

- **Función probabilidad de sobrevivir**

  Para cada valor entero $x \geqslant 0$ tal que $\ell_x > 0$, se define

$$p_x := \frac{\ell_{x+1}}{\ell_x}.$$

  Esta es la probabilidad de que una persona de edad x alcance la edad $x+1$. Claramente se cumple la identidad $p_x = 1 - q_x$. En particular, $p_\omega = \ell_{\omega+1}/\ell_\omega = 0$, suponiendo $\ell_\omega > 0$. Más generalmente, para $n \geqslant 1$ entero se define

$$ {}_np_x := \frac{\ell_{x+n}}{\ell_x}.$$

  Esta es la probabilidad de que una persona de edad x alcance la edad $x+n$. Es inmediato comprobar que ${}_np_x = 1 - {}_nq_x$. Cuando $n = 1$ se obtiene ${}_1p_x$, pero se omite el subíndice izquierdo para recuperar la notación p_x.

Muchas otras cantidades de interés se pueden definir y mostrar en una tabla de mortalidad. Por ejemplo, si se desean incluir las estadísticas d_x, q_x y p_x en el arreglo tabular, se obtiene el arreglo de la Tabla 4.2.

x	ℓ_x	d_x	q_x	p_x
0	100,000	1713	0.1713	0.98287
1	98,287	683	0.00694	0.99306
2	97,604	760	0.00778	0.99222
3	96,844	841	0.00868	0.99132
4	96,003	–	–	–
⋮	⋮	⋮	⋮	⋮
120	1	1	1	0

Tabla 4.2: Tabla de mortalidad.

A continuación presentaremos algunos resultados generales que pueden comprobarse a partir de las definiciones anteriores.

Proposición 4.1 *Para cada edad entera x y para cualquier número entero $n \geqslant 1$ se cumple que:*

1. ${}_nq_x = 1 - {}_np_x$.
2. $\ell_x = \ell_0\, p_0\, p_1 \cdots p_{x-1}$.
3. ${}_np_x = (1-q_x)(1-q_{x+1})\cdots(1-q_{x+n-1})$. *(Fórmula del producto)*

Demostración.

1. Por definición,
$$ {}_nq_x = \frac{{}_nd_x}{\ell_x} = \frac{\ell_x - \ell_{x+n}}{\ell_x} = 1 - \frac{\ell_{x+n}}{\ell_x} = 1 - {}_np_x. $$
2. Se desarrolla el lado derecho,
$$ \ell_0 \cdot p_0\, p_1 \cdots p_{x-1} = \ell_0 \cdot \frac{\ell_1}{\ell_0}\,\frac{\ell_2}{\ell_1} \cdots \frac{\ell_x}{\ell_{x-1}} = \ell_x. $$
3. Por definición,
$$ \begin{aligned} {}_np_x &= \frac{\ell_{x+n}}{\ell_x} \\ &= \frac{\ell_{x+1}}{\ell_x}\,\frac{\ell_{x+2}}{\ell_{x+1}} \cdots \frac{\ell_{x+n}}{\ell_{x+n-1}} \\ &= p_x\, p_{x+1} \cdots p_{x+n-1} \\ &= (1-q_x)(1-q_{x+1})\cdots(1-q_{x+n-1}). \end{aligned} $$

■

La identidad (3) es la fórmula del producto que aparece en la ecuación (2.1) en la página 45.

Funciones básicas

Se pueden definir también las cuatro funciones básicas $F(x)$, $S(x)$, $\lambda(x)$ y $R(x)$ que se estudiaron en el Capítulo 2, pero ahora en el contexto de una

tabla de mortalidad. Se revisan a continuación estas funciones. Aunque los tiempos de vida pueden ser continuos, las observaciones del número de sobrevivientes se llevan a cabo en tiempos periódicos. Esto hace que las funciones básicas correspondan a tiempos de vida discretos.

Consideraremos que la partición del tiempo está dada por los intervalos:

$$(0,1],(1,2],\ldots,(\omega-1,\omega],(\omega,\omega+1),$$

en donde a $(x-1,x]$ se le denomina el intervalo x, para $x=1,\ldots,\omega+1$. Observe que x es el extremo derecho del intervalo. Denotaremos por X a la variable aleatoria que indica el número del intervalo en el que fallece un individuo. Esta variable es estrictamente positiva y tiene como posibles valores $1,2,\ldots,\omega+1$. En cambio, la edad en años cumplidos de un individuo al fallecer puede tomar el valor 0. No debe confundirse a X con la edad de una persona al fallecer.

- **Función de distribución**

  La función de distribución en el modelo tabular es

  $$F(x) := P(X \leqslant x) = \frac{\ell_0 - \ell_x}{\ell_0}, \quad \text{para} \quad x = 0,1,\ldots,\omega+1.$$

  en donde ℓ_x es el número de personas con vida al inicio del intervalo $(x,x+1]$ y, por lo tanto, $\ell_0-\ell_x$ es el número de personas que fallecieron en alguno de los intervalos $1,\ldots,x$. Observe que $F(0)=0$ y $F(\omega+1)=1$. La correspondiente función de probabilidad es

  $$\begin{aligned} f(x) &:= F(x)-F(x-1) \\ &= \frac{\ell_{x-1}-\ell_x}{\ell_0}, \quad \text{para} \quad x=1,2,\ldots,\omega+1. \end{aligned}$$

  El número $f(x)$ es la probabilidad de que un individuo fallezca en el intervalo x, es decir, en $(x-1,x]$.

- **Función de supervivencia** Para cada valor entero de $x \geqslant 0$ se define la función de supervivencia en el modelo tabular como

  $$S(x) := P(X > x) = \frac{\ell_x}{\ell_0} = {}_x p_0, \quad \text{para} \quad x = 0,1,\ldots$$

Esta es la probabilidad de que una persona sobreviva al intervalo $(x-1, x]$. La definición de $S(x)$ que aparece arriba se puede escribir como

$$\ell_x = \ell_0\, S(x), \quad \text{para} \;\; x = 0, 1, \ldots$$

Observe que $S(0) = 1$ y $S(\omega + 1) = 0$. En la Tabla 4.3 se muestra una tabla de mortalidad en donde se ha incluido la función $S(x)$. Se muestran, además, los intervalos de valores en donde la función de supervivencia es constante. Los intervalos son de la forma indicada pues de esa manera la función $S(x)$ es continua por la derecha.

x	ℓ_x	$S(x) = \ell_x/\ell_0$	Intervalo para $S(x)$
0	ℓ_0	$S(0)$	$[0, 1)$
1	ℓ_1	$S(1)$	$[1, 2)$
2	ℓ_2	$S(2)$	$[2, 3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Tabla 4.3: Tabla de mortalidad con $S(x) = P(X > x)$ incorporado.

- **Función de riesgo**

  Para cada valor entero $x \geqslant 1$ tal que $S(x-1) > 0$, se define la función de riesgo $\lambda(x)$ en el modelo tabular como aparece abajo. Observe nuevamente que se trata de una función de riesgo discreta.

$$\begin{aligned} \lambda(x) &= \frac{f(x)}{S(x-1)} \\ &= \frac{(\ell_{x-1} - \ell_x)/\ell_0}{\ell_{x-1}/\ell_0} \\ &= \frac{d_{x-1}}{\ell_{x-1}} \\ &= q_{x-1}, \quad \text{para} \;\; x = 1, 2, \ldots, \omega + 1. \end{aligned}$$

  Es decir, $\lambda(x)$ es la probabilidad de que una persona de edad $x-1$ fallezca antes de cumplir la edad x. En particular, $\lambda(\omega + 1) = q_\omega = 1$, lo cual indica que nadie alcanza la edad $\omega + 1$.

- **Función tiempo medio de vida restante**

  Siguiendo la definición que habíamos dado antes para esta función en el caso de tiempos de vida discretos, tenemos que, para valores enteros $x \geqslant 0$ tales que $S(x) > 0$,

$$\begin{aligned} R(x) &= E(X - x \,|\, X > x) \\ &= \sum_{u=x+1}^{\omega+1} u\, f(u)/S(x) - x \\ &= \sum_{u=x+1}^{\omega+1} u\, \frac{(\ell_{u-1} - \ell_u)/\ell_0}{\ell_x/\ell_0} - x \\ &= \frac{1}{\ell_x} \sum_{u=x+1}^{\omega+1} u\,(\ell_{u-1} - \ell_u) - x, \end{aligned}$$

  en donde, recordemos, ω es la edad máxima en la tabla de mortalidad. Puede comprobarse que $R(0) = E(X)$.

Se puede encontrar mayor información sobre tablas de mortalidad en J. Leguina [38], G. Tutz y M. Schmid [51], por ejemplo. Véase también el capítulo 4 del libro de R. J. Cunningham, T. N. Herzog y R. L. London [17].

4.2. Método actuarial

Supondremos que tenemos el registro de los tiempos de vida $x_1, \ldots, x_n$ de n individuos, algunos de los cuales pueden estar censurados por la derecha. En esta sección veremos una forma no paramétrica de estimar la función de supervivencia $S(x)$ cuando los datos están agrupados en subintervalos como en una tabla de mortalidad.

Sea $0 = a_0 < a_1 < a_2 < \cdots < a_k < a_{k+1} = \infty$ una partición finita arbitraria del intervalo $(0, \infty)$, en donde a_k es un valor mayor o igual al tiempo de vida observado más grande. De esta manera, cada una de las observaciones pertenece a uno y sólo uno de los k subintervalos de tiempo $(0, a_1]$, $(a_1, a_2]$, $\ldots$, $(a_{k-1}, a_k]$. En la Figura 4.1 se muestra la situación cuando $k = 4$. Las constantes $a_1, a_2, \ldots, a_k$ pueden ser, por ejemplo, las

edades en años cumplidos de los individuos, es decir, $a_0 = 0, a_1 = 1, a_2 = 2, \ldots, a_k = \omega$, como en una tabla de mortalidad.

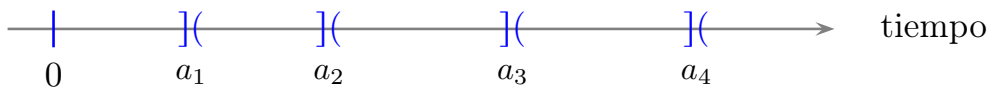

Figura 4.1: Subintervalos de tiempo.

Supondremos también que la tabla de mortalidad nos provee de la siguiente información para cada subintervalo.

Definición 4.2 *Para $j = 1, 2, \ldots, k$, se definen las cantidades:*

n_j = *"Número de personas en observación al inicio del intervalo* $(a_{j-1}, a_j]$*."*

d_j = *"Número de personas fallecidas en el intervalo* $(a_{j-1}, a_j]$*."*

w_j = *"Número de personas censuradas en el intervalo* $(a_{j-1}, a_j]$*."*

Al número n_j se le llama número en riesgo. El nombre completo es "número de personas en riesgo de morir" en el intervalo $(a_{j-1}, a_j]$. Claramente estos números son decrecientes, es decir, $n = n_1 \geqslant n_2 \geqslant \cdots \geqslant n_k \geqslant 0$, pues en cada intervalo pueden ocurrir fallecimientos o censuras. Por otro lado, la notación para d_j proviene del término *death*, y esta es la cantidad de personas que mueren en el intervalo en cuestión. Finalmente, w_j denota el número de tiempos de vida censurados en el intervalo en estudio. Cuando en una tabla de mortalidad no se proporciona la información de los datos censurados, definimos $w_j = 0$, de modo que los posibles datos censurados son considerados como fallecimientos y la mortalidad se sobreestima.

En general, se cumple la siguiente relación: para $j = 2, \ldots, k$,

$$n_j = n_{j-1} - d_{j-1} - w_{j-1}.$$

Es decir, el número en riesgo para el intervalo $(a_{j-1}, a_j]$ es el número de personas en riesgo de morir en el intervalo anterior, menos los fallecidos en ese intervalo, menos los censurados.

Definición 4.3 *La probabilidad condicional de que un individuo que inicia con vida el intervalo $(a_{j-1}, a_j]$ sobreviva al final de ese intervalo es*

$$p_j = P(X > a_j \,|\, X > a_{j-1}) = \frac{S(a_j)}{S(a_{j-1})}.$$

Con esta notación podemos ahora expresar la función de supervivencia mediante el siguiente producto: para $j = 1, 2, \ldots, k$,

$$S(a_j) = \frac{S(a_1)}{S(a_0)}\,\frac{S(a_2)}{S(a_1)} \cdots \frac{S(a_j)}{S(a_{j-1})} = p_1\, p_2 \cdots p_j.$$

Esta es otra forma de escribir la fórmula producto mencionada en la Proposición 2.5 de la página 45. La expresión anterior sugiere una forma de estimar $S(x)$ para las edades $x = a_1, a_2, \ldots, a_k$, a partir de alguna estimación que se tenga para las probabilidades condicionales p_j. Esto es lo que haremos a continuación.

Existe cierta justificación para definir el estimador $\hat{p}_j = 1 - d_j/n_j$. Sin embargo, esta expresión no toma en cuenta los posible individuos censurados en el intervalo en estudio. Más precisamente, esta fórmula indica de manera implícita que los individuos expuestos al riesgo son n_j, colocando los tiempos censurados en el intervalo $(a_{j-1}, a_j]$ hasta el tiempo final a_j, pues presupone que todos los individuos n_j estuvieron expuestos al riesgo de morir durante todo el subintervalo en cuestión. En realidad, los individuos censurados sólo se expusieron al riesgo de morir hasta el momento desconocido de su censura. Para tomar en cuenta esta consideración, se propone cambiar el denominador n_j por el siguiente promedio.

Definición 4.4 *El número efectivo de individuos en riesgo de morir durante el intervalo* $(a_{j-1}, a_j]$ *es*

$$n'_j = \frac{(n_j - w_j) + n_j}{2} = n_j - \frac{1}{2}w_j.$$

Este es el promedio aritmético de las siguientes dos cantidades: $n_j - w_j$, que es el número en riesgo cuando los tiempos censurados ocurren al inicio del intervalo, y n_j, que es el número en riesgo cuando los tiempos censurados ocurren al final del intervalo. Observemos que este promedio no es necesariamente un número entero, lo será únicamente cuando el número de tiempos censurados w_j es un número par. Además, es evidente que $n_j = n'_j$ cuando $w_j = 0$, es decir, en aquellos intervalos en donde no hay datos censurados. En general, se cumple que $n'_j \leqslant n_j$.

Se propone entonces como estimador para la probabilidad p_j al número

$$\hat{p}_j = 1 - \frac{d_j}{n'_j},$$

y se construye así el siguiente estimador para la función de supervivencia.

Definición 4.5 (Método actuarial) *A partir de un conjunto de datos de supervivencia con posible censura por la derecha y agrupados en los intervalos* $(0, a_1]$, $(a_1, a_2]$, $\ldots$, (a_k, ∞), *el estimador por el método actuarial para la función de supervivencia* $S(x)$ *es*

$$\begin{aligned} \hat{S}(a_0) &:= \hat{S}(0) = 1, \\ \hat{S}(a_j) &:= \hat{p}_1\, \hat{p}_2 \cdots \hat{p}_j \\ &= (1 - \frac{d_j}{n'_j})\, \hat{S}(a_{j-1}), \qquad j = 1, \ldots, k. \end{aligned} \tag{4.1}$$

Observe que lo anterior define a $\hat{S}(x)$ sólo en los puntos $x = a_1, \ldots, a_k$. Sin

embargo, la definición puede extenderse a cualquier valor continuo $x \geqslant 0$ de la siguiente forma

$$\hat{S}(x) = \begin{cases} 1 & \text{si } 0 \leqslant x < a_1, \\ \hat{S}(a_1) & \text{si } a_1 \leqslant x < a_2, \\ \hat{S}(a_2) & \text{si } a_2 \leqslant x < a_3, \\ \vdots & \vdots \\ \hat{S}(a_{k-1}) & \text{si } a_{k-1} \leqslant x < a_k, \\ 0 & \text{si } x \geqslant a_k. \end{cases}$$

Así es como se puede construir una función de supervivencia estimada $\hat{S}(x)$, para cualquier $x \geqslant 0$, la cual es constante en los intervalos $[a_j, a_{j+1})$, para $j = 0, 1, \ldots, k-1$.

Nota técnica 1. A diferencia de lo considerado antes, observe que los intervalos sobre los que se define $\hat{S}(x)$ ahora son cerrados por la izquierda y abiertos por la derecha. Esta modificación permite garantizar que la función $\hat{S}(x)$ sea continua por la derecha. Así la función estimada $\hat{S}(x)$ presenta, posiblemente, decrementos en los valores $a_1, \ldots, a_k$. Recordemos que estos números son arbitrarios y son aquellos a través de los cuales se define la partición del intervalo de tiempo $(0, \infty)$.

Nota técnica 2. Toda observación censurada en el último intervalo $(a_{k-1}, a_k]$ es considerada como un fallecimiento. De esta manera, $d_k = n_k = n'_k$ y el factor $1 - d_j/n'_j$ se anula. Esto tiene como consecuencia que $S(a_k) = 0$, es decir, la función de supervivencia estimada efectivamente toma el valor 0 a partir de a_k en adelante.

Ejemplo 4.1 *Considere el conjunto de datos de supervivencia que se muestra en la Tabla 4.4, agrupados en $k = 5$ subintervalos. Para cada subintervalo se cuenta con el número en riesgo n_j, el número de fallecimientos d_j y el número de individuos censurados por la derecha w_j. Con esta información se calcula el factor $1 - d_j/n'_j$ que aparece en la penúltima columna de la tabla y en la última columna tenemos los valores $\hat{S}(a_j)$, para $j = 1, 2, \ldots, 5$. Por*

brevedad, las cantidades son mostradas sólo con 4 dígitos decimales.

j	a_j	Intervalo de análisis	n_j	d_j	w_j	$\hat{p}_j = 1 - d_j/n'_j$	$\hat{S}(a_{j-1})$	Intervalo para $\hat{S}(x)$
1	1	$(0, 1]$	25	5	0	0.8000	1	$[0, 1)$
2	2	$(1, 2]$	20	3	1	0.8461	0.8000	$[1, 2)$
3	3	$(2, 3]$	16	6	0	0.6250	0.6769	$[2, 3)$
4	4	$(3, 4]$	10	5	2	0.4444	0.4230	$[3, 4)$
5	5	$(4, 5]$	3	2	1	0	0.1880	$[4, 5)$
–	–	$(5, \infty)$	0	–	–	–	0	$[5, \infty)$

Tabla 4.4: Datos de supervivencia del Ejemplo 4.1 .

Este ejemplo permite ilustrar varias cuestiones técnicas acerca del método actuarial:

- *Observe que los intervalos que aparecen en la tercera columna de la Tabla 4.4 son los utilizados para llevar a cabo el análisis del cálculo de los factores (penúltima columna), mientras que para especificar la definición de $\hat{S}(x)$, estos intervalos deben ser, contrariamente a como aquí aparecen, cerrados por la izquierda y abiertos por la derecha.*

- *Cada factor $1 - d_j/n'_j$ (penúltima columna) se debe multiplicar por el valor de $\hat{S}(a_j)$ que aparece en el mismo renglón para producir el valor de $\hat{S}(a_{j+1})$ que aparece en el renglón inferior inmediato.*

- *Para el último intervalo $(4, 5]$, tenemos que $n_5 = 3$ individuos ingresan con vida a ese intervalo, de los cuales $d_5 = 2$ individuos fallecen y $w_5 = 1$ individuo está censurado. Como se explicó antes, se deben modificar los datos y definir $d_5 = 3$ y $w_5 = 0$ para que el último factor $1 - d_5/n'_5$ sea cero, tal y como aparece en la tabla.*

- *El procedimiento que se muestra en la Tabla 4.4 puede ser implementado con facilidad en una computadora usando una hoja de cálculo.*

■

El estimador actuarial como variable aleatoria

El estimador por el método actuarial recién definido en (4.1) se puede escribir como una variable aleatoria de la siguiente manera. Se considera nuevamente una partición finita fija $0 = a_0 < a_1 < a_2 < \cdots < a_k < a_{k+1} = \infty$ del intervalo $(0, \infty)$. Se puede definir a N_j como el número aleatorio de individuos que inician con vida al intervalo $(a_{j-1}, a_j]$, para $j = 1, \ldots, k$. Observe que estas variables aleatorias no son independientes pues debe cumplirse que

$$n = N_1 \geqslant N_2 \geqslant \cdots \geqslant N_k.$$

Se puede definir la variable adicional N_{k+1} como el número de individuos que ingresan con vida al intervalo (a_k, ∞). Los términos D_j, N'_j y W_j tienen los mismos significados de antes, pero ahora no son números observados sino variables aleatorias y se debe cumplir que $W_j \leqslant N_j$ y, como antes, $N'_j = N_j - W_j/2$. El estimador para la probabilidad de supervivencia del intervalo $(a_{j-1}, a_j]$ es la variable aleatoria $\hat{p}_j = D_j/N'_j$, para $j = 1, \ldots, k$. El estimador para $S(x)$ es

$$\hat{S}(a_j) = \hat{p}_1\, \hat{p}_2 \cdots \hat{p}_j = (1 - \frac{D_j}{N'_j})\, S(a_{j-1}), \tag{4.2}$$

para $j = 1, \ldots, k$. Así es como tenemos una sucesión de variables aleatorias

$$\hat{S}(a_1) \geqslant \hat{S}(a_2) \geqslant \cdots \geqslant \hat{S}(a_k).$$

Bajo esta perspectiva, es natural preguntarse por las características numéricas de estas variables aleatorias. A continuación veremos algunos de estos resultados.

Proposición 4.2 *Sea $j \in \{1, 2, \ldots, k\}$ y suponga que el evento $(N'_j = n'_j)$ ocurre, en donde $n'_j \geqslant 1$ es un entero. Si los tiempos de vida son independientes, entonces*

1. $D_j \,|\, (N'_j = n'_j) \sim bin(n'_j, 1 - p_j)$.
2. $E(\hat{p}_j \,|\, N'_j = n'_j) = p_j$.

Demostración.

1. Cada individuo en riesgo de morir durante el intervalo $(a_{j-1}, a_j]$ tiene probabilidad de fallecer en ese intervalo el número desconocido $1 - p_j$. Suponiendo independencia entre los individuos y cuando n'_j individuos inician con vida el intervalo $(a_{j-1}, a_j]$, el total de muertes D_j en dicho intervalo tiene distribución $\text{bin}(n_j, 1 - p_j)$.

2. Este resultado es consecuencia de la definición de $\hat{p}_j$ y del inciso anterior,
$$E(\hat{p}_j \mid N'_j = n'_j) = E(1 - \frac{D_j}{N'_j} \mid N'_j = n'_j) = p_j.$$

■

Observe que en las dos afirmaciones de la proposición anterior se necesita como condición la ocurrencia del evento $(N'_j = n'_j)$, el cual especifica el número efectivo de individuos en riesgo de morir en el intervalo $(a_{j-1}, a_j]$. Sin esta información es complicado afirmar algo acerca de lo que ocurre en dicho intervalo. Existen, sin embargo, algunas aproximaciones que no incluyen esta condición; enunciaremos éstas a continuación. Su demostración se puede encontrar en el texto de D. London [39].

Proposición 4.3 *El estimador* $\hat{S}(a_j) = \hat{p}_1\, \hat{p}_2 \cdots \hat{p}_j$ *definido en* (4.2)*, para* $j \in \{1, 2, \ldots, k\}$*, satisface:*

1. $E(\hat{p}_j) \approx p_j$.

2. $E(\hat{S}(a_j)) \approx S(a_j)$.

3. $Var(\hat{S}(a_j) \mid N'_1 = n'_1, \ldots, N'_j = n'_j) \approx S^2(a_j) \cdot \sum_{i=1}^{j} \frac{1 - p_i}{p_i\, n'_i}$.

Las dos primera afirmaciones establecen que los estimadores $\hat{p}_j$ y $\hat{S}(a_j)$ por el método actuarial no son insesgados. Lo son sólo de manera aproximada. A la última aproximación se le conoce como fórmula de Greenwood. Se le puede usar para encontrar intervalos de confianza aproximados para los valores desconocidos $S(a_j)$.

Con esto concluimos nuestra breve introducción a algunos conceptos del análisis de supervivencia usando tablas de mortalidad. Existe una teoría bastante desarrollada de resultados, conceptos y estimaciones de modelos de supervivencia basados en datos de tablas de mortalidad. Una exposición muy completa sobre estos temas puede encontrarse en el libro de D. London [39]. Otras referencias generales sobre el tema de tablas de mortalidad son, por ejemplo, N. L. Jr. Bowers *et al* [9], D. C. M. Dickson *et al* [18], N. Keyfitz y H. Caswell [32], J. Leguina [38], y S. D. Promislow [46].

4.3. Interpolación en tablas de mortalidad

Para cada edad en años cumplidos x se busca definir la función $s \mapsto \ell_{x+s}$ para valores de s en el intervalo $[0, 1]$, en donde el valor inicial ℓ_x y el valor final ℓ_{x+1} son provistos por una tabla de mortalidad. En esta sección veremos tres maneras en las que esta interpolación puede efectuarse. De esta forma se puede construir una función $x \mapsto \ell_x$ definida ahora en el intervalo continuo $[0, \omega]$.

Interpolación lineal

En este tipo de interpolación se establece que la función $s \mapsto \ell_{x+s}$, para $0 \leqslant s \leqslant 1$, se define por la línea recta que une los valores ℓ_x y ℓ_{x+1}. Véase la gráfica de la izquierda de la Figura 4.2. Para cada entero x y para $0 \leqslant s \leqslant 1$, se define

$$\ell_{x+s} := s \cdot \ell_{x+1} + (1 - s) \cdot \ell_x. \tag{4.3}$$

De este modo se postula que la población decrece linealmente de ℓ_x a ℓ_{x+1} en una unidad de tiempo. Esta función se puede escribir de las siguientes formas equivalentes:

$$\begin{aligned} \ell_{x+s} &= \ell_x - s \cdot d_x \\ &= \ell_x - s \cdot (\ell_x - \ell_{x+1}) \\ &= s \cdot \ell_{x+1} + (1 - s) \cdot \ell_x. \end{aligned} \tag{4.4}$$

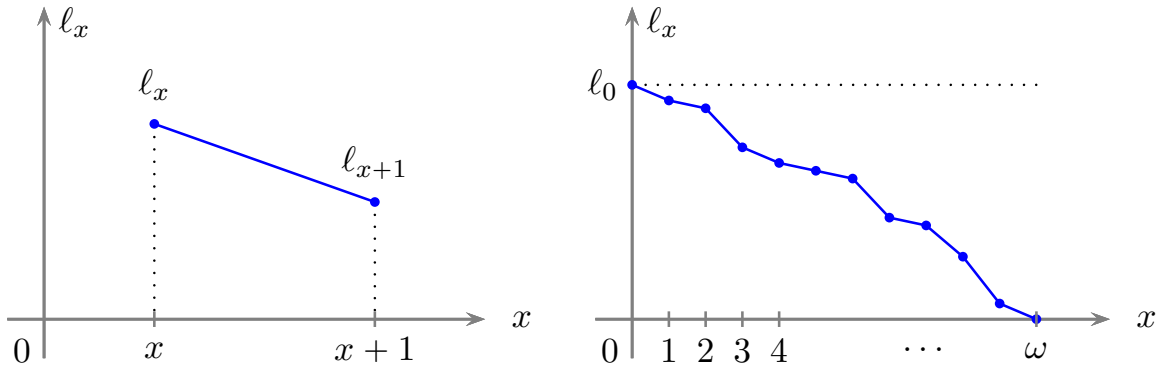

Figura 4.2: Interpolación lineal.

La gráfica de ℓ_x definida ahora para cualquier valor continuo $0 \leqslant x \leqslant \omega$ es la de una función continua, decreciente y lineal por pedazos. Véase la gráfica de la derecha de la Figura 4.2.

Bajo la interpolación lineal, se cumplen las identidades que aparecen en la siguiente proposición. Todas estas identidades se obtienen al substituir el valor de ℓ_{x+s} de acuerdo a la definición (4.3) ó (4.4). Se deja como ejercicio la verificación de estas igualdades.

Proposición 4.4 *Bajo la interpolación lineal* (4.3) *se cumplen las siguientes fórmulas: para* $x \geqslant 0$ *entero y para* $0 \leqslant s \leqslant 1$,

1. ${}_sp_x = 1 - s \cdot \dfrac{d_x}{\ell_x}$.
2. ${}_sq_x = s \cdot q_x$.
3. ${}_{(1-s)}p_{x+s} = \dfrac{p_x}{1 - s \cdot q_x}$.
4. ${}_{(1-s)}q_{x+s} = \dfrac{(1-s) \cdot q_x}{1 - s \cdot q_x}$.

Se ha mencionado que cuando se interpolan de manera lineal los valores ℓ_x de una tabla de mortalidad se produce una función continua decreciente

$\ell_x : [0, \infty) \to [\ell_0, 0]$. Esto induce un tiempo de vida continuo X para el cual se pueden calcular las funciones básicas. Estas fórmulas se muestran en el siguiente enunciado y su demostración se deja como ejercicio. Como se trata de un tiempo de vida continuo, las definiciones de estas funciones son las correspondientes a tiempos continuos.

Proposición 4.5 *El tiempo de vida continuo X inducido por la interpolación lineal $\ell_{x+s} = s \cdot \ell_{x+1} + (1-s) \cdot \ell_x$, tiene las siguientes funciones básicas: para $x = 0, 1, \ldots$*

1. $F(x+s) = (1-s)F(x) + sF(x+1), \quad 0 \leqslant s \leqslant 1.$

2. $f(x+s) = F(x+1) - F(x), \quad 0 < s < 1.$

3. $S(x+s) = (1-s)S(x) + sS(x+1), \quad 0 \leqslant s \leqslant 1.$

4. $\lambda(x+s) = \dfrac{F(x+1) - F(x)}{(1-s)S(x) + sS(x+1)}.$

La función tiempo promedio de vida restante $R(x+s)$ es más complicada y se ha omitido de la lista anterior.

Interpolación exponencial

Esta es una segunda manera de interpolar los valores entre ℓ_x y ℓ_{x+1}. Para cada entero $x \geqslant 0$ y para $0 \leqslant s \leqslant 1$, se define

$$\ell_{x+s} := \ell_x \left(\frac{\ell_{x+1}}{\ell_x} \right)^s = (\ell_{x+1})^s \, (\ell_x)^{1-s}. \tag{4.5}$$

Esta expresión es una función exponencial y su gráfica es la curva convexa que se muestra en la parte izquierda de la Figura 4.3. De esta manera, el valor inicial ℓ_x decae exponencialmente hacia el valor final ℓ_{x+1}.

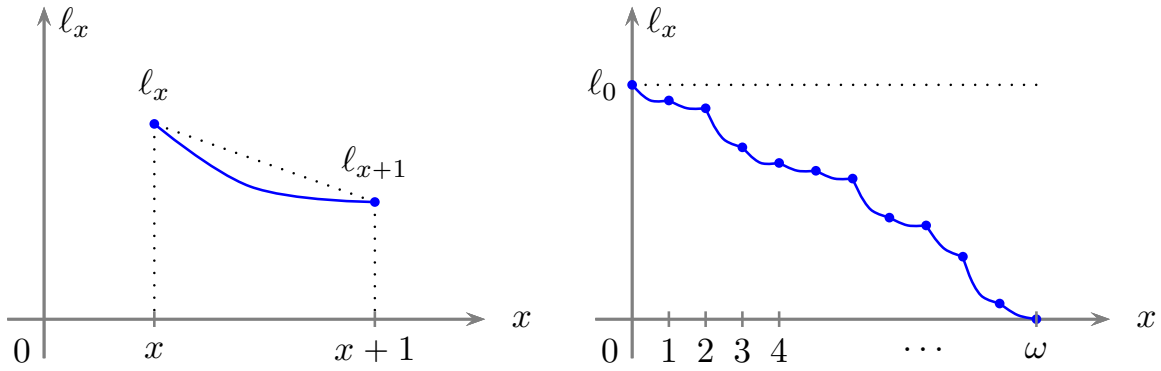

Figura 4.3: Interpolación exponencial.

La gráfica de ℓ_x definida ahora para cualquier valor continuo $0 \leqslant x \leqslant \omega$ es la de una función continua, decreciente y convexa por pedazos. Véase la gráfica de la derecha de la Figura 4.2.

Bajo la interpolación exponencial, se cumplen las fórmulas que aparecen a continuación. Su verificación no es difícil y se deja como ejercicio.

Proposición 4.6 *Bajo la interpolación exponencial* (4.5) *se cumple las siguientes fórmulas: para* $x \geqslant 0$ *entero y para* $0 \leqslant s \leqslant 1$,

1. ${}_sp_x = (p_x)^s$.
2. ${}_sq_x = 1 - (1 - q_x)^s$.
3. ${}_{(1-s)}p_{x+s} = (p_x)^{1-s}$.
4. ${}_{(1-s)}q_{x+s} = 1 - (1 - q_x)^{1-s}$.

Interpolación hiperbólica

Esta es otra manera alternativa de interpolar los valores entre ℓ_x y ℓ_{x+1}. Para cada entero $x \geqslant 0$ y para $0 \leqslant s \leqslant 1$, se define

$$\ell_{x+s} := \left(\frac{1}{\ell_x} + s \cdot \left[\frac{1}{\ell_{x+1}} - \frac{1}{\ell_x}\right]\right)^{-1}. \tag{4.6}$$

Equivalentemente,

$$\begin{aligned}\frac{1}{\ell_{x+s}} &= \frac{1}{\ell_x} + s \cdot \left(\frac{1}{\ell_{x+1}} - \frac{1}{\ell_x}\right) \\ &= s \cdot \left(\frac{1}{\ell_{x+1}}\right) + (1-s) \cdot \left(\frac{1}{\ell_x}\right). \end{aligned} \tag{4.7}$$

Esto corresponde a una interpolación lineal de los recíprocos de los valores ℓ_x y ℓ_{x+1}. Puede verificarse que la fórmula para ℓ_{x+s} que aparece en (4.6) produce el valor ℓ_x cuando $s = 0$ y es igual a ℓ_{x+1} cuando $s = 1$. La gráfica de la interpolación hiperbólica es similar al caso de la interpolación exponencial que se muestra en la Figura 4.3.

Bajo la interpolación hiperbólica se cumplen las fórmulas que aparecen abajo. Su verificación se deja como ejercicio.

Proposición 4.7 *Bajo la interpolación hiperbólica* (4.6) *se cumplen las siguientes fórmulas: para* $x \geqslant 0$ *entero y para* $0 \leqslant s \leqslant 1$,

1. ${}_sp_x = \dfrac{s \cdot q_x}{1 - (1-s) \cdot q_x}$.
2. ${}_sq_x = \dfrac{1 - q_x}{1 - (1-s) \cdot q_x}$.
3. ${}_{(1-s)}p_{x+s} = 1 - (1-s) \cdot q_x$.
4. ${}_{(1-s)}q_{x+s} = (1-s) \cdot q_x$.

§

No es difícil proponer otros métodos de interpolación entre ℓ_x y ℓ_{x+1} adicionales a los mencionados y que no producen necesariamente funciones continuas. Se puede definir, por ejemplo, una función $x \mapsto \ell_x$ que sea constante por pedazos y que sea decreciente en forma de escalera desde el valor ℓ_x hasta el valor ℓ_{x+1} para cada $x \geqslant 0$ entero. También debe advertirse que se pueden definir y estudiar otras funciones de interés basadas en los datos básicos $\ell_0, \ell_1, \ldots, \ell_\omega$ de una tabla de mortalidad (véase el libro de D. London [39]), y encontrar fórmulas del comportamiento de estas funciones bajo alguno de los métodos de interpolación.

Estimador para $S(x)$ a partir de una tabla de mortalidad

Para valores continuos de x en el intervalo $[0, \omega]$, sea $\hat{\ell}_x$ la interpolación mediante algún método de los datos $\ell_0, \ell_1, \ldots, \ell_\omega$ provenientes de una tabla de mortalidad. Esta función no es necesariamente diferenciable en los puntos $x = 0, 1, \ldots, \omega$, aunque claramente se cumple que $\hat{\ell}_x = \ell_x$ en los puntos indicados. Se puede ahora definir la estimación no paramétrica $\hat{S}(x)$ para la función de supervivencia de la siguiente forma.

Definición 4.6 (Estimador actuarial) *A partir de los datos $\ell_0, \ell_1, \ldots, \ell_\omega$ de una tabla de mortalidad y utilizando algún método de interpolación para definir $\hat{\ell}_x$ para cualquier valor x en el intervalo continuo $[0, \omega]$, el estimador actuarial para $S(x)$ se define como*

$$\hat{S}(x) = \frac{\hat{\ell}_x}{\ell_0}, \quad \text{para } x \in [0, \omega].$$

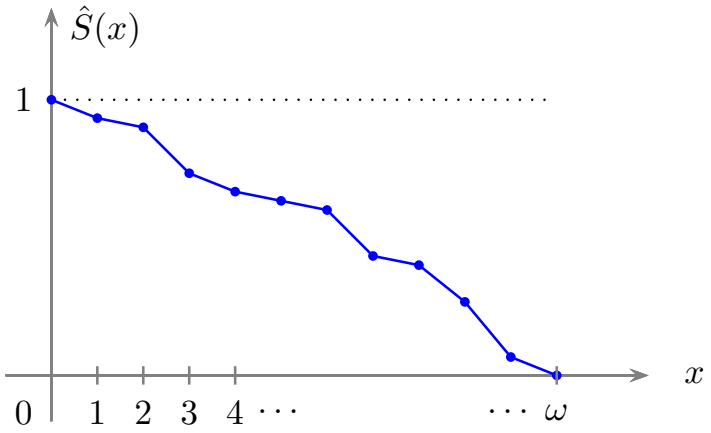

Figura 4.4: Estimación actuarial $\hat{S}(x) = \hat{\ell}_x/\ell_0$.

De esta forma el tiempo de vida en estudio y con datos registrados en una tabla de mortalidad queda caracterizado por la función de supervivencia

aproximada $\hat{S}(x)$. A partir de esta función se puede calcular de manera aproximada cualquier característica numérica del tiempo de vida. Una gráfica de $\hat{S}(x)$ se muestra en la Figura 4.4 en el caso de interpolación lineal. Se trata de una función que inicia en el valor $\hat{S}(0) = 1$, es continua, lineal por pedazos, decreciente y su último valor es $\hat{S}(\omega) = 0$.

4.4. Función de supervivencia empírica

Esta es una función que se construye a partir de un conjunto de observaciones no censuradas de una variable aleatoria cualquiera, en particular, un tiempo de vida. Definiremos primero esta función para datos numéricos y después consideraremos muestras aleatorias.

Función de supervivencia empírica para datos numéricos

Sea $x_1, \ldots, x_n$ una colección de n observaciones no censuradas de una variable aleatoria cualquiera X. Antes de definir la función de supervivencia, definiremos primero la función de distribución empírica de la siguiente forma

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(x_i), \quad -\infty < x < \infty, \tag{4.8}$$

en donde la función indicadora de un conjunto A de números reales, denotada por $1_A(x)$, toma el valor 1 cuando $x \in A$, y toma el valor 0 cuando $x \notin A$. Así, para cada valor real x, la función $F_n(x)$ es la proporción de observaciones que toman un valor a la izquierda de x respecto al número total de observaciones.

No es difícil comprobar que $F_n(x)$ es una función de distribución y su gráfica tiene las características de la función de distribución de una variable aleatoria discreta, es decir, tiene saltos hacia arriba en cada una de las observaciones x_i. Puede comprobarse que cuando se tienen n datos distintos y ordenados $x_1 < x_2 < \cdots < x_n$, los saltos de la función de distribución empírica son

$$F(x_i) - F(x_{i-1}) = \frac{1}{n}, \quad \text{para } i = 1, 2, \ldots, n,$$

en donde x_0 es cualquier valor a la izquierda del primer dato x_1. En el lado izquierdo de la Figura 4.5 se muestra un ejemplo gráfico de la función de distribución empírica $F_n(x)$ para un conjunto de 4 observaciones x_1, x_2, x_3, x_4 distintas, las cuales se escriben como $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)}$ cuando se ordenan de menor a mayor.

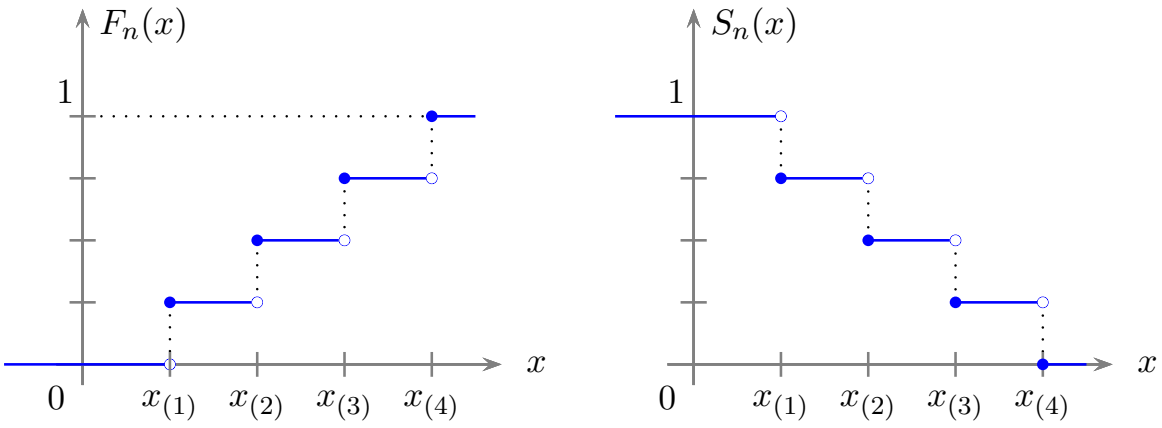

Figura 4.5: Función de distribución empírica $F_n(x)$ y su correspondiente función de supervivencia empírica $S_n(x)$.

La función $F_n(x)$ es una aproximación o estimación de la función de distribución de la variable aleatoria X. Mientras mayor sea el número de datos n, se espera que la función de distribución empírica se parezca cada vez más a la función de distribución $F(x)$ de la variable aleatoria. Definiremos ahora la función de supervivencia que es de nuestro interés.

Definición 4.7 *La función de supervivencia empírica de un conjunto de n observaciones no censuradas $x_1, \ldots, x_n$ es*

$$S_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{(x,\infty)}(x_i), \quad -\infty < x < \infty. \tag{4.9}$$

Esta definición es válida para cualesquiera observaciones numéricas $x_1, \ldots, x_n$ completamente especificadas. Cuando estas observaciones provienen de tiempos de vida no censurados, serán estrictamente positivas y esa es la situación que nos interesa. Claramente se cumple que $S_n(x) + F_n(x) = 1$. Para cualquier valor real de x, la función $S_n(x)$ indica la proporción de observaciones que son mayores al valor x respecto al número total de observaciones. Esta función tiene saltos hacia abajo en cada una de las observaciones x_i. En la parte derecha de la Figura 4.5 aparece un ejemplo gráfico de $S_n(x)$ para un conjunto de 4 observaciones x_1, x_2, x_3, x_4 distintas.

En particular, cuando se tienen n datos distintos y ordenados $x_1 < x_2 < \cdots < x_n$, los saltos de la función de supervivencia empírica son

$$S(x_i) - S(x_{i+1}) = \frac{1}{n}, \quad \text{para } i = 1, 2, \ldots, n,$$

en donde x_{n+1} es cualquier valor a la derecha del último dato x_n.

En el paquete estadístico `R` se puede obtener la función de distribución empírica con suma facilidad usando la función `ecdf()`, en donde el acrónimo `ecdf` significa *empirical cumulative distribution function*. Esto se muestra en el siguiente recuadro.

R

```
# Función de distribución empírica
x <- c(1,2,3,4)
plot(ecdf(x), axes=FALSE)
axis(1); axis(2)
```

El resultado de este código se muestra en el lado izquierdo de la Figura 4.6. La correspondiente función de supervivencia empírica se puede obtener usando el código que aparece abajo y que produce la gráfica que se muestra en el lado derecho de la Figura 4.6.

R

```
# Función de supervivencia empírica
library(survival)
x <- c(1,2,3,4)
plot(survfit(Surv(x)~1), axes=FALSE, conf.int=FALSE)
axis(1); axis(2)
```

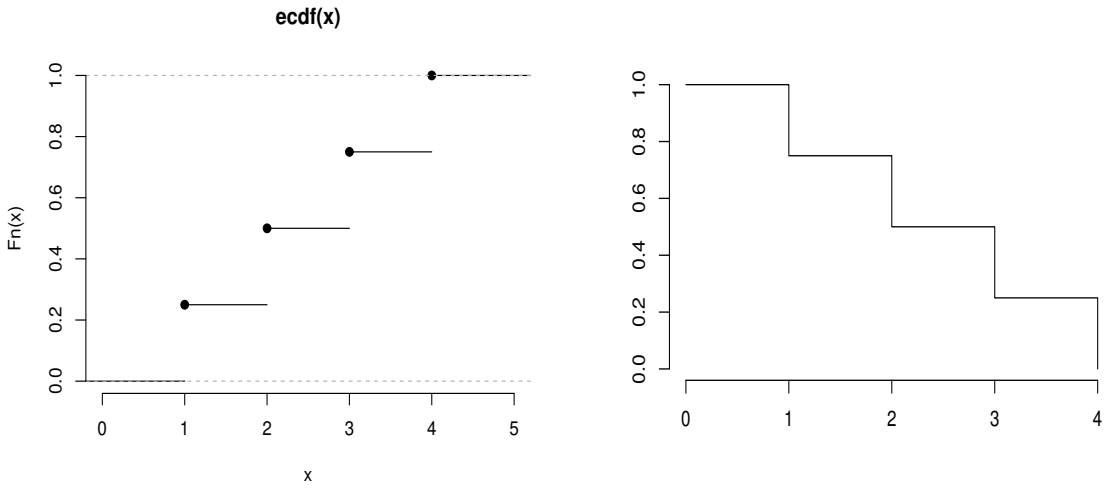

Figura 4.6: Gráficas de $F_n(x)$ y $S_n(x)$ producidas en el paquete estadístico R.

Función de supervivencia empírica para muestras aleatorias

En lugar de muestras numéricas particulares $x_1, \ldots, x_n$, consideraremos ahora colecciones de variables aleatorias que son independientes e idénticamente distribuidas. Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño n de una variable aleatoria X con función de distribución $F(x)$. Antes de definir a la función de supervivencia, definiremos primero a la función de distribución empírica, ahora como una variable aleatoria, de la siguiente forma

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(X_i), \quad -\infty < x < \infty. \tag{4.10}$$

Para cada valor real de x, la variable aleatoria $\hat{F}_n(x)$ resulta ser un estimador insesgado para $F(x)$. Si $x_1, \ldots, x_n$ son valores particulares de la muestra aleatoria $X_1, \ldots, X_n$, la expresión (4.10) se calcula como aparece en (4.8). Usando la ley de los grandes números puede comprobarse que $\hat{F}_n(x)$ converge a $F(x)$ para cada x fijo. Más aún, el teorema de Glivenko-Cantelli (véase el libro de A. Gut [25]) establece que, con probabilidad 1,

$$\lim_{n\to\infty} (\sup_x |\hat{F}_n(x) - F(x)|) = 0.$$

Es decir, $\hat{F}_n(x)$ converge uniformemente a $F(x)$ cuando $n \to \infty$. Ahora definiremos la función de supervivencia que nos interesa.

Definición 4.8 *La función de supervivencia empírica de una muestra aleatoria $X_1, \ldots, X_n$ es la variable aleatoria*

$$\hat{S}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{(x,\infty)}(X_i), \quad -\infty < x < \infty. \tag{4.11}$$

Esta definición se aplica para cualquier modelo o variable aleatoria X. El caso que nos interesa ocurre cuando $X \geqslant 0$ y, por lo tanto, puede representar un tiempo de vida. La función (4.11) es la proporción aleatoria de elementos de la muestra que toman un valor a la derecha del valor x respecto al tamaño n de la muestra. Puede comprobarse que $\hat{S}_n(x)$ es un estimador insesgado para la función de supervivencia de la variable aleatoria X. Además, es evidente que se cumple la identidad $\hat{F}_n(x) + \hat{S}_n(x) = 1$.

En la Figura 4.7 se muestra una función de supervivencia empírica cuando el número de datos n es grande. Observe que la función es decreciente y tiene pequeños saltos hacia abajo en múltiples ocasiones hasta llegar al valor 0. Ocurre un salto hacia abajo de magnitud $1/n$ cuando se observa un fallecimiento. En general, el tamaño del salto es k/n cuando ocurren k fallecimientos a un mismo tiempo. Estamos suponiendo aquí que no hay censura en los datos.

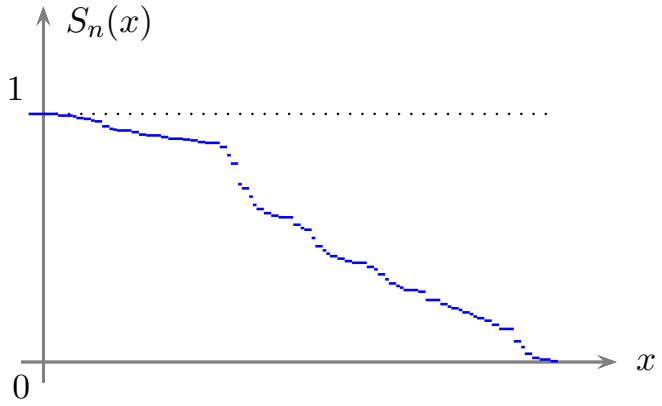

Figura 4.7: Ejemplo gráfico de una función de supervivencia empírica.

En el caso de datos con censura por la derecha y considerando una censura como una defunción, se puede calcular la función de supervivencia (4.11), sin embargo se estaría sobreestimando la verdadera mortalidad.

Proposición 4.8 *Sea $X_1, \ldots, X_n$ una muestra aleatoria de un tiempo de vida X con función de supervivencia $S(x)$. Sea $\hat{S}_n(x)$ la función de supervivencia empírica. Entonces, para $x \geqslant 0$,*

1. $E(\hat{S}_n(x)) = S(x)$. *(Insesgamiento)*
2. $Var(\hat{S}_n(x)) = \dfrac{1}{n} S(x)\,(1 - S(x))$.

Demostración. Es claro que cada sumando de $\hat{S}_n(x)$ (véase la expresión (4.11)) es una variable aleatoria con distribución $\text{Ber}(p)$ con $p = P(X_i > x) = S(x)$. Por lo tanto,

$$\sum_{i=1}^{n} 1_{(x,\infty)}(X_i) \sim \text{bin}(n, p).$$

De aquí se obtienen las dos afirmaciones del enunciado. ∎

Uno de los problemas centrales en el análisis de supervivencia es encontrar formas de estimar la función de supervivencia de un tiempo de vida, a partir de datos censurados. En este caso, cuando las observaciones presentan censura, no es evidente la forma en la que un dato censurado debe incorporarse al cálculo de la estimación de la función de supervivencia. El método actuarial, el estimador de Kaplan-Meier, o el estimador de Nelson-Aalen, son maneras de resolver este problema en el caso de censura por la derecha. Estudiaremos estos procedimientos en las siguientes secciones.

4.5. Estimador de Kaplan-Meier[1]

Este es posiblemente el estimador no paramétrico más importante para una función de supervivencia. Fue propuesto en 1958 por Edward L. Kaplan y Paul Meier [31]. Este trabajo de investigación es uno de los más citados en el área de la estadística. Definiremos primero a este estimador en un caso particular (llamado caso simple) y después veremos el caso general.

Sean nuevamente $x_1, \ldots, x_n$ observaciones de un tiempo de vida X, cuya distribución deseamos estimar. Supondremos que los datos fueron generados de manera independiente y que algunos de los registros pueden presentar censura por la derecha. El estimador de Kaplan-Meier $\hat{S}(x)$ proporciona una manera de estimar la función de supervivencia desconocida $S(x)$.

Caso simple

El primer paso es ordenar las observaciones en orden ascendente:

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}.$$

Consideraremos primero el caso simple cuando todas las observaciones son distintas, es decir, no hay dos tiempos observados iguales, de modo que las desigualdades anteriores son estrictas. Se define $x_{(0)} = 0$. De manera similar al método actuarial, consideraremos ciertos intervalos de tiempo pero ahora

[1] Edward Lynn Kaplan (1920–2006) matemático estadounidense.
Paul Meier (1924–2011) estadístico estadounidense.

definidos a partir de los datos observados de la siguiente forma:

$$(0, x_{(1)}],\ (x_{(1)}, x_{(2)}],\ \ldots\ , (x_{(n-1)}, x_{(n)}].$$

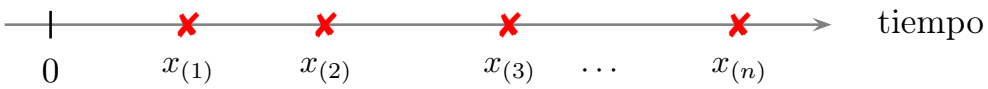

Figura 4.8: En el caso simple ocurre 1 censura ó 1 fallecimiento en cada tiempo registrado.

A diferencia del método actuarial, las constantes que se utilizan para definir la partición del tiempo en subintervalos no son las edades de los individuos sino los tiempos en los que ocurren los fallecimientos y las censuras. Nos referiremos a estos conjuntos como los intervalos $1, 2, \ldots, n$, respectivamente. Cada extremo derecho $x_{(j)}$ es un tiempo en el que ocurre un fallecimiento o una censura, pero no ambos eventos a la vez. Por lo tanto, en cada uno de los intervalos $(x_{j-1}, x_j]$ ocurre uno, y sólo uno, de estos dos eventos, y tal evento se presenta justo al final del intervalo. Véase la Figura 4.8.

Definición 4.9 *Para cada $j = 1, \ldots, n$, se definen las cantidades:*

n_j = *"Núm. de individuos en observación al inicio del intervalo j."*
d_j = *"Núm. de fallecimientos en el intervalo j".*
c_j = *"Núm. de censuras en el intervalo j".*

Puede comprobarse que se cumple la igualdad $n_j = n - j + 1$, pues en cada subintervalo se pierde a un individuo, ya sea por muerte o por censura. Claramente $n_1 = n$ (población completa en el primer intervalo) y $n_n = 1$ (último sobreviviente en el último intervalo). Por otro lado, la variable d_j toma los valores 0 ó 1, dependiendo de si al tiempo $x_{(j)}$ ocurre una censura o una muerte, es decir, se trata de una función indicadora. La probabilidad de

que un individuo que ingresa con vida al intervalo j sobreviva dicho intervalo es

$$p_j = P(X > x_{(j)} \,|\, X > x_{(j-1)}).$$

Bajo estas condiciones, resulta natural definir un estimador para esta probabilidad de la siguiente forma.

Definición 4.10 $\hat{p}_j := 1 - \dfrac{d_j}{n_j}$, *para* $j = 1, \ldots, n$.

Este es un estimador para la probabilidad condicional de sobrevivir al j-ésimo intervalo cuando se ha llegado con vida al inicio de dicho intervalo. Como d_j sólo toma los valores 0 ó 1, se puede también escribir

$$\hat{p}_j = (1 - 1/n_j)^{d_j} = \begin{cases} 1 - 1/n_j & \text{si} \quad d_j = 1, \\ 1 & \text{si} \quad d_j = 0. \end{cases}$$

Usando la fórmula del producto (2.1), se puede dar ahora una primera versión del estimador de Kaplan-Meier, véase [47].

Definición 4.11 (Estimador de Kaplan-Meier, caso simple) *Suponga el caso cuando los datos son todos distintos:* $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, *es decir, en cada tiempo observado ocurre, o bien un fallecimiento o bien una censura por la derecha. El estimador de Kaplan-Meier es*

$$\hat{S}(x) := \prod_{j\,:\,x_{(j)} \leqslant x} \Big(1 - \frac{d_j}{n_j}\Big), \quad \textit{para} \;\; 0 \leqslant x \leqslant x_{(n)}. \tag{4.12}$$

Como d_j sólo toma los valores 0 ó 1, el estimador (4.12) también puede ex-

presarse de las siguientes formas equivalentes:

$$\begin{aligned}
\hat{S}(x) &= \prod_{j\,:\,x_{(j)}\leqslant x} \hat{p}_j \\
&= \prod_{j\,:\,x_{(j)}\leqslant x} (1-1/n_j)^{d_j} \\
&= \prod_{j\,:\,x_{(j)}\leqslant x} (1-1/(n-j+1))^{d_j} \\
&= \prod_{j\,:\,x_{(j)}\leqslant x} \Big(\frac{n-j}{n-j+1}\Big)^{d_j}, \quad \text{para } \ 0\leqslant x\leqslant x_{(n)}. \qquad (4.13)
\end{aligned}$$

El estimador (4.12) también se puede escribir de manera recursiva en los tiempos $x_{(1)} < \cdots < x_{(n)}$ de la siguiente forma:

$$\hat{S}(x_{(j)}) = \Big(\frac{n-j}{n-j+1}\Big)^{d_j} \cdot \hat{S}(x_{(j-1)}), \quad \text{para } \ j = 1, \ldots, n. \qquad (4.14)$$

Es evidente que esta expresión es conveniente desde el punto de vista computacional pues permite calcular los valores de $\hat{S}(x)$ de una manera sucesiva. Como función del tiempo x, el estimador (4.12) posee las siguientes propiedades:

- $\hat{S}(0) = 1$. Esto es así bajo el supuesto de que el producto vacío se define como el valor 1.
- $\hat{S}(x)$ es monótona decreciente. Esto es evidente a partir de observar que al incrementarse el valor de x en (4.13), posiblemente se incorporan nuevos factores, pero éstos son menores o iguales a 1.
- $\hat{S}(x)$ es constante por pedazos pues la función permanece constante en el interior de los subintervalos de análisis.
- $\hat{S}(x)$ no tiene saltos en los tiempos en donde ocurre una observación censurada. Esto es así pues si x_j es una observación censurada, entonces $d_j = 0$, y el factor correspondiente $\hat{p}_j$ es igual a 1. Esto es más claro en la fórmula recursiva (4.14).
- $\hat{S}(x)$ tiene saltos hacia abajo únicamente en los tiempos en donde ocurre un fallecimiento, es decir, en tiempos $x_{(j)}$ en donde $d_j = 1$.

Esto es claro a partir de la fórmula (4.14). Si $x_{(j)}$ es uno de tales tiempos, entonces la magnitud del salto en $x_{(j)}$ es

$$\begin{aligned}\hat{S}(x_{(j-1)}) - \hat{S}(x_{(j)}) &= \hat{S}(x_{(j-1)}) - (\frac{n-j}{n-j+1})^{d_j}\,\hat{S}(x_{(j-1)}) \\ &= (1 - \frac{n-j}{n-j+1})\,\hat{S}(x_{(j-1)}) \\ &= \frac{1}{n-j+1}\,\hat{S}(x_{(j-1)}) \\ &= \frac{1}{n-j+1}\prod_{i=1}^{j-1}(\frac{n-i}{n-i+1})^{d_i}.\end{aligned}$$

- $\hat{S}(x)$ es una función continua por la derecha. Esto es consecuencia de la expresión (4.12).
- Si el último tiempo observado $x_{(n)}$ es un fallecimiento, entonces el último factor es $\hat{p}_n = 1 - d_n/n_n = 1 - 1/1 = 0$ y, en consecuencia, $\hat{S}(x_{(n)}) = 0$. Así, el estimador de Kaplan-Meier produce una función de supervivencia discreta genuina. Sin embargo, si $x_{(n)}$ es una censura, entonces $\hat{S}(x_{(n)}) > 0$ y la fórmula (4.12) no proporciona una función de supervivencia discreta verdadera.

Veremos ahora un ejemplo con muy pocos datos en donde el objetivo es ilustrar la forma de obtener $\hat{S}(x)$ en el caso simple, es decir, cuando todos los tiempos de eventos son distintos.

Ejemplo 4.2 *Consideremos el siguiente conjunto de $n = 5$ observaciones:*

$$x_{(1)} = 4, \quad x_{(2)} = 5+, \quad x_{(3)} = 7+, \quad x_{(4)} = 8, \quad x_{(5)} = 10.$$

Todos los datos son diferentes, algunos datos son completos y otros están censurados por la derecha. De manera gráfica los datos se muestran en la Figura 4.9, en donde los tiempos de fallecimientos se indican con el símbolo ✘ *y para los tiempos censurados se usa la letra* **c**.
El objetivo de este ejemplo es mostrar la forma de calcular el estimador de Kaplan-Meier usando la fórmula (4.13) *en el contexto de su validez. Los intervalos de análisis son:*

$$(0, 4], \quad (4, 5], \quad (5, 7], \quad (7, 8], \quad (8, 10].$$

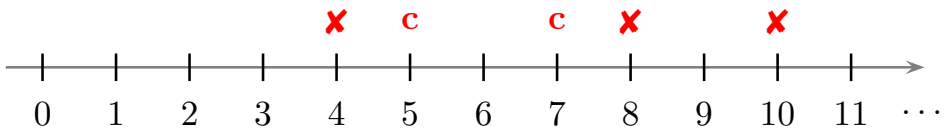

Figura 4.9: Datos del Ejemplo 4.2.

Los elementos para el cálculo del estimador de Kaplan-Meier se muestran en la Tabla 4.5. La columna indicada como $\hat{p}_j$ *se refiere al término* $1 - d_j/n_j$ *y es el factor que se aplica al valor previo de* $\hat{S}(x_{(j-1)})$ *para calcular el siguiente valor de* $\hat{S}(x_{(j)})$.

j	$x_{(j)}$	Intervalo	n_j	d_j	c_j	$\hat{p}_j$	$\hat{S}(x_{(j)})$
0	0	–	–	–	–	–	1
1	4	$(0, 4]$	5	1	0	0.8	0.8
2	5	$(4, 5]$	4	0	1	1	0.8
3	7	$(5, 7]$	3	0	1	1	0.8
4	8	$(7, 8]$	2	1	0	0.5	0.4
5	10	$(8, 10]$	1	1	0	0	0

Tabla 4.5: Cálculo de $\hat{S}(x)$ para el Ejemplo 4.2.

La gráfica de $\hat{S}(x)$ *aparece en la Figura* (4.10). *Observe que no hay saltos en los tiempos en donde se presentan censuras, en cambio, la función decrece en los tiempos de fallecimientos. Es evidente en la gráfica que las censuras sólo tienen efecto en la función de supervivencia estimada hasta el momento en el que ocurre el siguiente fallecimiento.*

Como el último tiempo observado es un fallecimiento, el estimador $\hat{S}(x)$ *obtenido es una función de supervivencia discreta genuina. Es importante señalar que, para llevar a cabo el análisis de la Tabla 4.5, se consideran los intervalos de la forma* $(x_{(j-1)}, x_{(j)}]$, *sin embargo, los distintos valores que aparecen en esta tabla para* $\hat{S}(x)$ *corresponden a los intervalos de la forma*

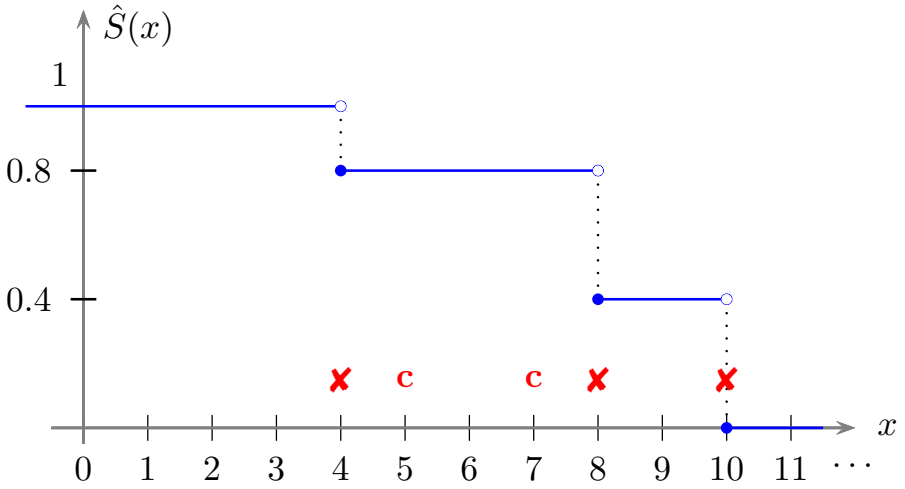

Figura 4.10: Función de supervivencia estimada del Ejemplo 4.2.

$[x_{(j-1)}, x_{(j)})$, $j = 1, \ldots, n$. *De esta manera, la función* $\hat{S}(x)$ *es continua por la derecha. Esto se hace explícito en la gráfica de la Figura 4.10. En el siguiente recuadro se muestra un código en* R *que usa la función* `survfit` *para obtener el estimador de Kaplan-Meier del presente ejemplo. Se usa el comando* `plot` *para graficar esta función. La gráfica se muestra en la Figura 4.11.*

R

```
# Graficación del estimador de Kaplan-Meier
# para los datos del Ejemplo 4.2.
library(survival)
time <- c(4,5,7,8,10)
status <- c(1,0,0,1,1)
df <-data.frame(time,status)
plot(survfit(Surv(time, status)~1, data=df),
conf.int=FALSE)
```

La expresión "~ 1*" que aparece en el código anterior indica que se deben usar todos los datos. Observe que la gráfica de la Figura 4.11 que se obtuvo con* R *es idéntica a la que se mostró antes en la Figura 4.10. Se puede obtener un*

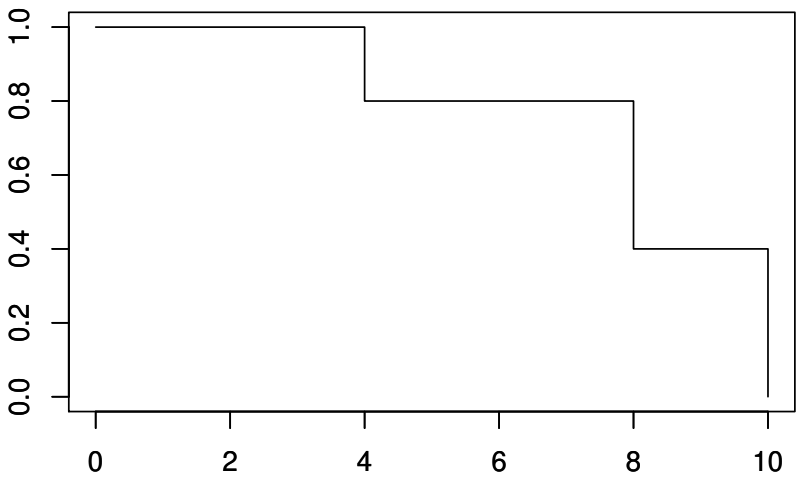

Figura 4.11: Función de supervivencia estimada del Ejemplo 4.2 generada en `R`.

resumen con la información básica del estimador de Kaplan-Meier encontrado en `R` *mediante el comando* `summary` *como se muestra en la Figura 4.12.*

```
> summary(survfit(Surv(time,status)~1,data=df))
Call: survfit(formula = Surv(time, status) ~ 1, data = df)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    4      5       1      0.8   0.179       0.5161            1
    8      2       1      0.4   0.297       0.0935            1
   10      1       1      0.0     NaN           NA           NA
>
```

Figura 4.12: Resumen en `R` del estimador de Kaplan-Meier construido para los datos del Ejemplo 4.2.

Caso general

Ahora consideraremos la situación general cuando a un mismo tiempo pueden ocurrir múltiples fallecimientos y, posiblemente, múltiples censuras por la derecha. Como antes, tenemos una población inicial de n individuos. Sean $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ las observaciones de sus tiempos de vida, ordenadas de menor a mayor, ahora no necesariamente distintos pues varios de estos tiempos pueden coincidir. Sean $x'_{(1)} < x'_{(2)} < \cdots < x'_{(k)}$ aquellos tiempos que son distintos y en los que ocurren algún fallecimiento, alguna censura, o ambos. El entero k es tal que $1 \leqslant k \leqslant n$.

Tenemos ahora los siguientes k subintervalos en los cuales se llevará a cabo el análisis del número de individuos que ingresan con vida y el número de ellos que fallecen:

$$(0, x'_{(1)}],\ (x'_{(1)}, x'_{(2)}],\ \ldots\ , (x'_{(k-1)}, x'_{(k)}].$$

El intervalo j es $(x'_{(j-1)}, x'_{(j)}]$, para $j = 1, \ldots, k$, en donde se define $x'_{(0)} := 0$. Como antes, consideraremos las siguientes cantidades:

$$\begin{aligned} n_j &= \text{“Núm. de individuos en observación al inicio del intervalo } j\text{.”} \\ d_j &= \text{“Núm. de fallecimientos en el intervalo } j\text{.”} \\ c_j &= \text{“Núm. de censuras en el intervalo } j\text{.”} \end{aligned}$$

Consideraremos varios casos en los cuales estimaremos la probabilidad condicional $p_j = P(X > x'_{(j)} \,|\, X > x'_{(j-1)})$.

Caso 1. Sólo fallecimientos. Suponga que en el tiempo $x'_{(j)}$ ocurren d_j fallecimientos simultáneos y ninguna censura, $1 \leqslant d_j \leqslant n_j$. De manera artificial se colocan los d_j fallecimientos a una distancia infinitesimal uno después otro justo antes del tiempo $x'_{(j)}$. Entonces el factor que debe aparecer en el estimador de Kaplan-Meier es

$$(1 - \frac{1}{n_j})(1 - \frac{1}{n_j - 1}) \cdots (1 - \frac{1}{n_j - (d_j - 1)}).$$

Esta cantidad contiene d_j factores y se puede escribir de la siguiente forma

$$\begin{aligned}(\frac{n_j-1}{n_j})(\frac{n_j-2}{n_j-1})(\frac{n_j-3}{n_j-2})\cdots(\frac{n_j-d_j}{n_j-d_j+1}) &= \frac{n_j-d_j}{n_j}\\ &= 1-\frac{d_j}{n_j}.\end{aligned}$$

Así, cuando ingresan n_j individuos al intervalo j y ocurren d_j fallecimientos y ninguna censura, el estimador para la probabilidad de supervivencia en el intervalo considerado es

$$\hat{p}_j := 1-\frac{d_j}{n_j}.$$

Esta es la misma expresión que para el caso simple $d_j \in \{0,1\}$.

Caso 2. Sólo censuras. Suponga que en el tiempo $x'_{(j)}$ ocurren $c_j \geqslant 1$ censuras y ningún fallecimiento. En este caso, el estimador para la probabilidad de supervivencia p_j se define como

$$\hat{p}_j := 1,$$

pues no se presentaron fallecimientos en el intervalo de análisis, aunque ciertamente c_j individuos fueron censurados justo en el último momento $x'_{(j)}$. La reducción del número de individuos tiene efecto en el análisis del siguiente intervalo al ingresar con vida un número menor de individuos. De este modo, el estimador de Kaplan-Meier se mantiene constante en el tiempo $x'_{(j)}$ pues el factor correspondiente es 1. Esta situación es idéntica al caso de una única observación censurada en el caso simple estudiado antes.

Caso 3. Fallecimientos y censuras múltiples simultáneas. Suponga que en el tiempo $x'_{(j)}$ ocurren $d_j \geqslant 1$ fallecimientos y al mismo tiempo $c_j \geqslant 1$ censuras. Como antes, las censuras no contribuyen al factor asociado al intervalo de análisis, únicamente los fallecimientos son relevantes. La estimación para la probabilidad de supervivencia es, nuevamente,

$$\hat{p}_j := 1-\frac{d_j}{n_j}.$$

En conclusión, la definición general del estimador de Kaplan-Meier es la siguiente.

Definición 4.12 (Estimador de Kaplan-Meier) *Sean $x_1, \ldots, x_n$ observaciones del tiempo de vida de n individuos, algunas de ellas censuradas por la derecha. Sean $x'_{(1)} < \cdots < x'_{(k)}$ las observaciones ordenadas de menor a mayor, omitiendo repeticiones, $1 \leqslant k \leqslant n$. El estimador de Kaplan-Meier es*

$$\hat{S}(x) := \prod_{j\,:\,x'_{(j)} \leqslant x} \Big(1 - \frac{d_j}{n_j}\Big), \quad \textit{para} \;\; 0 \leqslant x \leqslant x'_{(k)}. \tag{4.15}$$

Como función del tiempo x, el estimador $\hat{S}(x)$ definido por (4.15) posee propiedades similares a las que se enunciaron antes para el estimador simple (4.12), aunque ahora hay que considerar que pueden ocurrir múltiples fallecimientos y múltiples censuras a un mismo tiempo. Estas propiedades son: $\hat{S}(0) = 1$, $\hat{S}(x)$ es monótona decreciente, es constante por pedazos, no tiene saltos en donde sólo ocurren censuras, tiene saltos hacia abajo en los tiempos donde ocurren fallecimientos, es continua por la derecha y, finalmente, si en el último tiempo observado $x'_{(k)}$ sólo ocurren fallecimientos, entonces $\hat{S}(x'_{(k)}) = 0$ y el estimador es una función de supervivencia genuina. De lo contrario, si en $x'_{(k)}$ ocurre alguna censura, entonces $\hat{S}(x'_{(k)}) > 0$ y el estimador nunca alcanza el valor cero.

La fórmula (4.15) se puede escribir de manera recursiva en los tiempos $x'_{(1)}, \ldots, x'_{(k)}$ como aparece en la expresión (4.16). Como se ha mencionado, esta expresión es muy conveniente para calcular de manera sucesiva los valores de $\hat{S}(x)$.

$$\hat{S}(x'_{(j)}) = (1 - \frac{d_j}{n_j}) \cdot \hat{S}(x'_{(j-1)}), \quad \text{para} \;\; j = 1, \ldots, k. \tag{4.16}$$

Cuando los fallecimientos y las censuras ocurren en tiempos distintos, estamos en el caso al que le hemos llamado simple, y la definición general que aparece en (4.15) se reduce a la definición particular dada por la iden-

tidad (4.13) pues los factores coinciden, esto es,

$$(\frac{n-j}{n-j+1})^{d_j} = (1 - \frac{d_j}{n_j}), \quad \text{para} \;\; j = 1, \ldots, k = n.$$

Para verificar esta identidad se deben considerar los dos casos: $d_j = 0$ y $d_j = 1$. Cuando $d_j = 0$, ambas cantidades toman el valor 1. Este es el caso de una censura y el estimador de Kaplan-Meier no cambia. Por otro lado, bajo las condiciones mencionadas, $n_j = n - j + 1$, de modo que cuando $d_j = 1$, ambas cantidades toman el valor $(n-j)/(n-j+1)$.

Ejemplo 4.3 *Considere las siguientes observaciones de $n = 9$ tiempos de vida:*

$$x_{(1)} = 1, \quad x_{(2)} = 2+, \quad x_{(3)} = 2+, \quad x_{(4)} = 2+, \quad x_{(5)} = 4, \quad x_{(6)} = 4,$$
$$x_{(7)} = 6, \quad x_{(8)} = 6+, \quad x_{(9)} = 7+ \; .$$

Observe que los datos están ordenados de menor a mayor, hay repeticiones, algunos de ellos son datos completos y otros están censurados por la derecha. Gráficamente estos datos se muestran en la Figura 4.13 . Como antes, el símbolo ✘ denota un fallecimiento y la letra c *denota una censura por la derecha.*

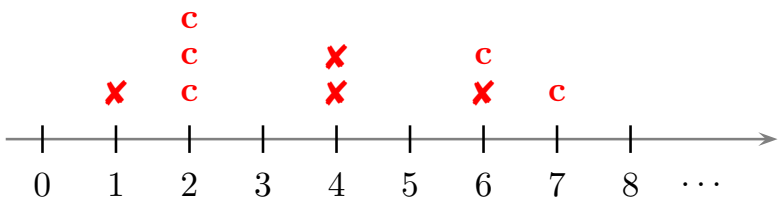

Figura 4.13: Datos del Ejemplo 4.3.

Omitiendo repeticiones, los tiempos observados son:

$$x'_{(1)} = 1, \quad x'_{(2)} = 2, \quad x'_{(3)} = 4, \quad x'_{(4)} = 6, \quad x'_{(5)} = 7.$$

Estos tiempos definen la siguiente partición:

$$(0, 1], \quad (1, 2], \quad (2, 4], \quad (4, 6], \quad (6, 7].$$

Los elementos para el cálculo del estimador de Kaplan-Meier se muestran en la Tabla 4.6. Para cada intervalo se registra: el número de individuos n_j que ingresan con vida, el número de fallecidos d_j y el número de censurados c_j. El factor $\hat{p}_j = 1 - d_j/n_j$ se multiplica por el valor de $\hat{S}(x'_{(j-1)})$ para obtener $\hat{S}(x'_{(j)})$.

j	$x'_{(j)}$	Intervalo	n_j	d_j	c_j	$\hat{p}_j$	$\hat{S}(x'_{(j)})$
0	0	–	–	–	–	–	1
1	1	$(0,1]$	9	1	0	$8/9$	$0.\bar{8}$
2	2	$(1,2]$	8	0	3	1	$0.\bar{8}$
3	4	$(2,4]$	5	2	0	$3/5$	$0.5\bar{3}$
4	6	$(4,6]$	3	1	1	$2/3$	$0.3\bar{5}$
5	7	$(6,7]$	1	0	1	1	$0.3\bar{5}$

Tabla 4.6: Datos del Ejemplo 4.3.

La gráfica de $\hat{S}(x)$ aparece en la Figura 4.14. Observe que no hay saltos en donde sólo se presentan censuras, en cambio, la función decrece en los tiempos en donde hay fallecimientos. Como el último tiempo registrado es una censura, el estimador no alcanza el nivel 0.

En el siguiente recuadro se muestra un código simple en `R` *que usa la función* `survfit` *para obtener el estimador de Kaplan-Meier del presente ejemplo. Se usa el comando* `plot` *para graficar esta función. La gráfica se muestra en la Figura 4.15.*

R

```
# Graficación del estimador de Kaplan-Meier
# para los datos del Ejemplo 4.3 .
library(survival)
time <- c(1,2,2,2,4,4,6,6,7)
status <- c(1,0,0,0,1,1,1,0,0)
df <-data.frame(time,status)
plot(survfit(Surv(time, status)~1, data=df),
conf.int=FALSE)
```

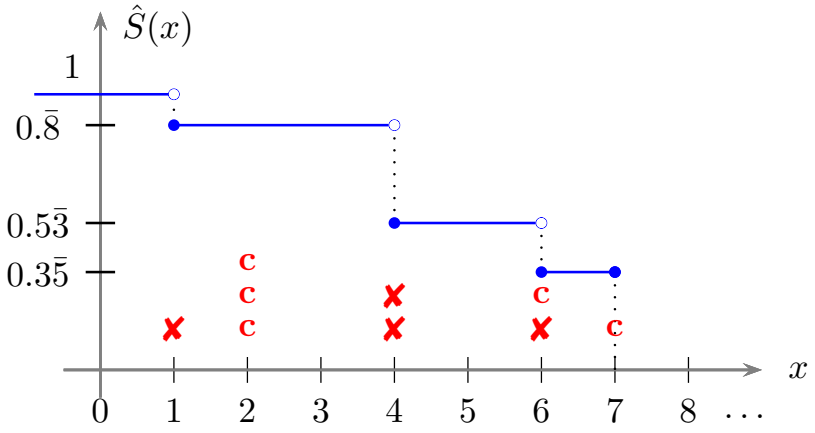

Figura 4.14: Función de supervivencia estimada del Ejemplo 4.3.

La expresión "∼1" que aparece en el código indica que se deben usar todos los datos. Observe que la gráfica de la Figura 4.15 que se obtuvo con `R` *es idéntica a la que se mostró antes en la Figura 4.14.*

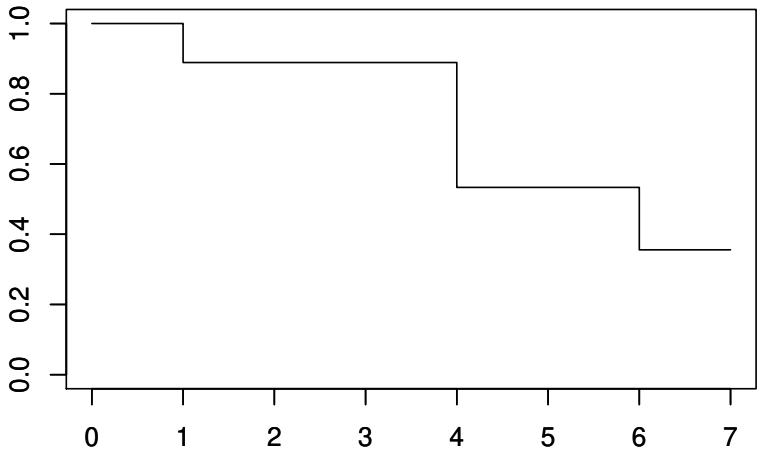

Figura 4.15: Función de supervivencia estimada del Ejemplo 4.3 generada en `R`.

```
>  summary(survfit(Surv(time,status)~1,data=df))
Call: survfit(formula = Surv(time, status) ~ 1, data = df)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1      9       1    0.889   0.105        0.706            1
    4      5       2    0.533   0.205        0.251            1
    6      3       1    0.356   0.199        0.119            1
```

Figura 4.16: Resumen en R del estimador de Kaplan-Meier para los datos del Ejemplo 4.3 .

En la Figura 4.16 se muestra un resumen con la información básica del estimador de Kaplan-Meier generado en R *para los datos del Ejemplo 4.3.*

■

En la sección de ejercicios se pide demostrar el siguiente resultado.

Proposición 4.9 *Cuando un conjunto de datos no presenta censura, el estimador de Kaplan-Meier* (4.15) *coincide con la función de supervivencia empírica definida en* (4.9).

A pesar de que el estimador de Kaplan-Meier $\hat{S}(x)$ puede no alcanzar el valor 0 y no estar definida para cualquier $x \geqslant 0$, nos referiremos a ella como una función de supervivencia.

El estimador de Kaplan-Meier como estimador de máxima verosimilitud

La función de verosimilitud de un conjunto de datos de supervivencia censurados por la derecha $(x_1, \delta_1), \ldots, (x_n, \delta_n)$, en términos de la función de supervivencia, está dada por

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} [S(x_i-) - S(x_i)]^{\delta_i} [S(x_i)]^{1-\delta_i}. \tag{4.17}$$

Cuando el dato observado x_i es un fallecimiento ($\delta_i = 1$), el factor es $S(x_i-) - S(x_i)$, lo cual corresponde a la probabilidad de que el individuo i fallezca al tiempo x_i. Cuando el dato observado x_i es una censura ($\delta_i = 0$), el factor es $S(x_i)$, lo cual es la probabilidad de que el individuo i sea censurado al tiempo x_i.

En el siguiente resultado se demuestra que el estimador de Kaplan-Meier maximiza (4.17).

Proposición 4.10 *El estimador de Kaplan-Meier maximiza la función de verosimilitud* (4.17).

Demostración. Se busca una función de supervivencia discreta $\hat{S}(x)$ que maximice la función de verosimilitud (4.17). Este problema se puede restringir un poco a partir de las siguientes dos observaciones cruciales:

- La función óptima $\hat{S}(x)$ necesariamente debe tener saltos en los tiempos de fallecimientos. Si esto no fuera así, el factor $\hat{S}(x_i-) - \hat{S}(x_i)$ sería cero para algún tiempo de fallecimiento x_i y la función de verosimilitud tomaría el valor cero. En tal caso, $\hat{S}(x)$ no sería el punto en donde se alcance el máximo.

- La función óptima $\hat{S}(x)$ no debe asignar probabilidades positivas a datos x_i que no correspondan a fallecimientos. Si ello ocurriera, redistribuir esa probabilidad al siguiente tiempo de fallecimiento incrementa la verosimilitud.

Así, el problema de encontrar $\hat{S}(x)$ se reduce a determinar la probabilidad que se le debe asignar a cada tiempo de fallecimiento. Denotaremos estos tiempo por $x'_1 < \cdots < x'_k$, en donde el entero k es tal que $1 \leqslant k \leqslant n$. Hemos omitido los paréntesis en los subíndices para aligerar la notación pero debe tenerse presente que los datos están ordenados de menor a mayor. Como antes, se consideran los intervalos de análisis $(0, x'_1], \ldots, (x'_{k-1}, x'_k]$ y,

en cada uno de ellos, las cantidades:

$$\begin{aligned} n_j &= \text{“Núm. en riesgo al inicio del intervalo } (x'_{j-1}, x'_j]\text{”}, \\ d_j &= \text{“Núm. de fallecimientos en } (x'_{j-1}, x'_j]\text{”}, \\ c_j &= \text{“Núm. de censuras en } (x'_{j-1}, x'_j]\text{”}, \end{aligned}$$

para $j = 1, \ldots, k$. La función de verosimilitud se puede ahora escribir como

$$L = \prod_{j=1}^{k} [S(x'_j-) - S(x'_j)]^{d_j} [S(x'_j)]^{c_j}. \tag{4.18}$$

Sean $\lambda_1, \ldots, \lambda_k$ las tasas de riesgo en los tiempos de fallecimientos $x'_1 < \cdots < x'_k$, es decir, $\lambda_j = 1 - p_j$. Estas son las probabilidades de fallecer en los intervalos de análisis cuando se ha ingresado con vida a ellos. En términos de estas tasas de riesgo, la función de verosimilitud es

$$L = \prod_{j=1}^{k} [\ \lambda_j \prod_{k=1}^{j-1} (1 - \lambda_k)\ ]^{d_j} [\ \prod_{k=1}^{j} (1 - \lambda_k)\ ]^{c_j}. \tag{4.19}$$

Ahora se trata de encontrar los valores $\lambda_1, \ldots, \lambda_k$ que maximizan esta función. Tomando logaritmo

$$\begin{aligned} \log L &= \sum_{j=1}^{k} [d_j \log \lambda_j + d_j \sum_{k=1}^{j-1} \log(1 - \lambda_k) + c_j \sum_{k=1}^{j} \log(1 - \lambda_k)] \\ &= \sum_{j=1}^{k} [d_j \log \lambda_j + (d_j + c_j) \sum_{k=1}^{j-1} \log(1 - \lambda_k) + c_j \log(1 - \lambda_j)] \\ &= \sum_{j=1}^{k} d_j \log \lambda_j + \sum_{j=1}^{k} (d_j + c_j) \sum_{k=1}^{j-1} \log(1 - \lambda_k) + \sum_{j=1}^{k} c_j \log(1 - \lambda_j). \end{aligned}$$

Cambiando el orden de las sumas en la doble suma y usando la identidad $n_{j+1} - n_j = d_j + c_j$ se obtiene que ese término es

$$\begin{aligned} \sum_{j=1}^{k} (d_j + c_j) \sum_{k=1}^{j-1} \log(1 - \lambda_k) &= \sum_{j=1}^{k-1} n_{j+1} \log(1 - \lambda_j) \\ &= \sum_{j=1}^{k} n_{j+1} \log(1 - \lambda_j). \end{aligned}$$

Se ha añadido el último sumando pero éste es cero pues $n_{k+1} = 0$. Por lo tanto,

$$\begin{aligned}
\log L &= \sum_{j=1}^{k} \log \lambda_j^{d_j} + \sum_{j=1}^{k} \log(1-\lambda_j)^{n_{j+1}} + \sum_{j=1}^{k} \log(1-\lambda_j)^{c_j} \\
&= \sum_{j=1}^{k} \log \lambda_j^{d_j} + \sum_{j=1}^{k} \log(1-\lambda_j)^{n_{j+1}+c_j} \\
&= \log \prod_{j=1}^{k} \lambda_j^{d_j} + \log \prod_{j=1}^{k} (1-\lambda_j)^{n_j - d_j} \\
&= \log \prod_{j=1}^{k} \lambda_j^{d_j} (1-\lambda_j)^{n_j - d_j}.
\end{aligned}$$

Derivando respecto de la variable λ_j para $j = 1, \ldots, k$, e igualando a cero, se obtiene que

$$\lambda_j = \frac{d_j}{n_j}, \quad j = 1, \ldots, k.$$

Esto significa que $p_j = 1 - \lambda_j = 1 - d_j/n_j$ y que la función $\hat{S}(x)$ que maximiza la verosimilitud (4.17) es el estimador de Kaplan-Meier. ∎

El estimador de Kaplan-Meier como variable aleatoria

Definiremos ahora al estimador de Kaplan-Meier como una variable aleatoria y no como una función determinista estimada a partir de un conjunto de datos. El procedimiento es similar al caso de observaciones particulares.

Sean $X_1, \ldots, X_n$ los tiempos de vida aleatorios de n individuos, algunos de los cuales pueden estar censurados por la derecha. Sean $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ los tiempos ordenados de menor a mayor, es decir, las estadísticas de orden. Los intervalos, ahora aleatorios, en los que se lleva a cabo el análisis son:

$$(0, X_{(1)}], (X_{(1)}, X_{(2)}], \ldots, (X_{(n-1)}, X_{(n)}].$$

El j-ésimo intervalo es $(X_{(j-1)}, X_{(j)}]$, para $j = 1, \ldots, n$, en donde se define $X'_{(0)} := 0$. Como antes, sea N_j el número de individuos que ingresan con vida al intervalo j (número en riesgo), y sea D_j el número de fallecimientos que ocurren al final del intervalo j, es decir, al tiempo $X_{(j)}$. Observe que estas cantidades ahora son variables aleatorias con valores enteros.

Definición 4.13 (Estimador de Kaplan-Meier como variable aleatoria) *Sean $X_1, \ldots, X_n$ tiempos de vida independientes con posible censura por la derecha. Sean $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ los tiempos ordenados en orden creciente. Sea N_j el número de individuos que ingresan con vida al intervalo $(X_{(j-1)}, X'_{(j)}]$ y sea D_j el número de fallecimientos que ocurren al tiempo $X_{(j)}$, $j = 1, \ldots, n$. El estimador de Kaplan-Meier es la variable aleatoria*

$$\hat{\mathbb{S}}(x) := \prod_{j\,:\,X_{(j)} \leqslant x} (1 - \frac{D_j}{N_j}), \quad \textit{para} \;\; 0 \leqslant x \leqslant X_{(n)}. \tag{4.20}$$

Puede considerarse que la definición del estimador $\hat{\mathbb{S}}(x)$ corresponde al de un proceso estocástico a tiempo continuo, pues la función de supervivencia es una función del tiempo x. Inicia en el valor 1 y tiene saltos hacia abajo en los tiempos en donde ocurren fallecimientos. El término "estimador" adquiere ahora un sentido más formal desde el punto de vista de la estadística matemática. El estimador de Kaplan-Meier $\hat{S}(x)$ estudiado antes es un posible valor o trayectoria de $\hat{\mathbb{S}}(x)$, véase el texto de P. J. Smith [47].

Es posible comprobar que $\hat{\mathbb{S}}(x)$ tiene esperanza y varianza aproximadas como se indica en las siguientes fórmulas:

1. $E(\hat{\mathbb{S}}(x)) \approx \hat{S}(x)$.

2. $\text{Var}(\,\hat{\mathbb{S}}(x)\,) \approx (\,\hat{S}(x)\,)^2 \sum_{j=1}^{n} \dfrac{d_j}{n_j(n_j - d_j)}$.

Observe que estas aproximaciones están expresadas en términos de la función de supervivencia de Kaplan-Meier $\hat{S}(x)$, la cual está calculada a partir de

datos particulares $(x_1, \delta_1), \ldots, (x_n, \delta_n)$. A la aproximación para la varianza se le llama fórmula de Greenwood y la demostraremos en la siguiente sección.

Fórmula de Greenwood

Se verá ahora una forma de aproximar la varianza del estimador de Kaplan-Meier, considerado éste como la variable aleatoria definida en (4.20). Esto permite encontrar intervalos de confianza aproximados para la función de supervivencia. Utilizaremos las siguientes aproximaciones para la media y la varianza de la transformación φ de una variable aleatoria X:

$$E(\varphi(X)) \approx \varphi(E(X)), \tag{4.21}$$

$$\text{Var}(\varphi(X)) \approx (\varphi'(E(X)))^2 \,\text{Var}(X). \tag{4.22}$$

Estas dos aproximaciones se obtienen al aplicar el método delta (véase la página 279 en el Apéndice A.2) y se requiere que la transformación sea diferenciable. La aproximación (4.21) para la esperanza es una simplificación de la que parece en el apéndice. El resultado es el siguiente.

Proposición 4.11 (Fórmula de Greenwood) *Si $n_1, \ldots, n_k$ y $d_1, \ldots, d_k$ es una muestra de las variables $N_1, \ldots, N_k$ y $D_1, \ldots, D_k$, $1 \leqslant k \leqslant n$, en el estimador de Kaplan-Meier definido en (4.20), entonces $\hat{\mathbb{S}}(x)$ tiene varianza aproximada*

$$Var(\,\hat{\mathbb{S}}(x)\,) \approx (\,\hat{S}(x)\,)^2 \sum_{j=1}^{n} \frac{d_j}{n_j(n_j - d_j)}. \tag{4.23}$$

Demostración. Primero se aplica la aproximación (4.21) para la esperanza a la variable aleatoria $X = \hat{\mathbb{S}}(x)$ y la transformación $\varphi(x) = \log x$. Así, tenemos que

$$E(\log \hat{\mathbb{S}}(x)) \approx \log E(\hat{\mathbb{S}}(x)) \approx \log \hat{S}(x).$$

Tomando exponencial,

$$\exp\{E(\log \hat{\mathbb{S}}(x))\} \approx \hat{S}(x).$$

Ahora se aplica la aproximación (4.22) para la varianza a la variable $X = \log \hat{\mathbb{S}}(x)$ y la transformación $\varphi(x) = \exp(x)$. Iniciamos con una igualdad evidente,

$$\begin{aligned} \mathrm{Var}(\,\hat{\mathbb{S}}(x)\,) &= \mathrm{Var}(\,\exp\{\,\log \hat{\mathbb{S}}(x)\,\}\,) \\ &\approx [\,\exp\{\,\log \hat{S}(x)\,\}\,]^2 \cdot \mathrm{Var}(\,\log \hat{\mathbb{S}}(x)\,) \\ &\approx [\,\hat{S}(x)\,]^2 \cdot \mathrm{Var}(\,\log \hat{\mathbb{S}}(x)\,). \end{aligned}$$

Entonces

$$\begin{aligned} \mathrm{Var}(\,\hat{\mathbb{S}}(x)\,) &= [\,\hat{S}(x)\,]^2 \cdot \mathrm{Var}(\,\sum_{j=1}^{n} \log(1 - D_j/N_j)\,) \\ &\approx [\,\hat{S}(x)\,]^2 \cdot \sum_{j=1}^{n} \mathrm{Var}(\,\log(1 - D_j/N_j)\,). \end{aligned}$$

Para la última aproximación se ha supuesto que la varianza de la suma es, aproximadamente, la suma de las varianzas. Se usa nuevamente la aproximación para la varianza (4.22) para la variable $X = \log(1 - D_j/N_j)$ y la transformación $\varphi(x) = \log x$, cuya derivada es $\varphi'(x) = 1/x$. Entonces

$$\mathrm{Var}(\,\hat{\mathbb{S}}(x)\,) \approx [\,\hat{S}(x)\,]^2 \cdot \sum_{j=1}^{n} \left(\frac{1}{E(1 - D_j/N_j)}\right)^2 \mathrm{Var}\,(1 - D_j/N_j).$$

La variable aleatoria $(N_j - D_j)/N_j$ es la proporción de sobrevivientes en el intervalo j. Condicionado al evento $(N_j = n_j)$, el numerador tiene distribución $\mathrm{bin}(n_j, \hat{p}_j)$, en donde $\hat{p}_j = 1 - d_j/n_j$. Esta es una estimación para la probabilidad de sobrevivir al intervalo en estudio. Entonces, suponiendo la ocurrencia del evento $(N_j = n_j)$ y sin escribirlo explícitamente, se tiene que

$$\begin{aligned} E(\frac{N_j - D_j}{N_j}) &= \frac{n_j\,\hat{p}_j}{n_j} = \hat{p}_j, \\ \mathrm{Var}(\frac{N_j - D_j}{N_j}) &= \frac{n_j\,\hat{p}_j\,(1 - \hat{p}_j)}{n_j^2} = \frac{\hat{p}_j\,(1 - \hat{p}_j)}{n_j}. \end{aligned}$$

Por lo tanto,

$$\begin{aligned}
\mathrm{Var}(\,\hat{\mathbb{S}}(x)\,) &\approx [\,\hat{S}(x)\,]^2 \cdot \sum_{j=1}^{n} \left(\frac{1}{\hat{p}_j}\right)^2 \frac{\hat{p}_j\,(1-\hat{p}_j)}{n_j} \\
&= [\,\hat{S}(x)\,]^2 \cdot \sum_{j=1}^{n} \frac{1-\hat{p}_j}{n_j\,\hat{p}_j} \\
&= [\,\hat{S}(x)\,]^2 \cdot \sum_{j=1}^{n} \frac{d_j/n_j}{n_j\,(1-d_j/n_j)} \\
&= [\,\hat{S}(x)\,]^2 \cdot \sum_{j=1}^{n} \frac{d_j}{n_j\,(n_j-d_j)}.
\end{aligned}$$

■

La fórmula de Greenwood puede consultarse tambien en R. C. Elandt-johnson y N. L. Johnson [19], E. T. Lee y J. W. Wang [37] ó P. J. Smith [47].

Intervalo de confianza aproximado

Como se indicó antes, la fórmula de Greenwood (4.23) se puede usar para construir un intervalo de confianza aproximado para la función de supervivencia desconocida. A partir de (4.20),

$$\log \hat{\mathbb{S}}(x) = \sum_{j\,:\,X_{(j)} \leqslant x} \log\left(\frac{N_j - D_j}{N_j}\right).$$

Esta expresión de la variable aleatoria $\log \hat{\mathbb{S}}(x)$ como una suma sugiere usar el teorema central del límite. Usando nuevamente las aproximaciones (4.21) y (4.22), la media y varianza de $\log \hat{\mathbb{S}}(x)$ son:

$$E(\log \hat{\mathbb{S}}(x)) \approx \log E(\hat{\mathbb{S}}(x)) \approx \log \hat{S}(x),$$

$$\text{y} \qquad \begin{aligned} \text{Var}(\log \hat{\mathbb{S}}(x)) &\approx \left(\frac{1}{E(\hat{\mathbb{S}}(x))}\right)^2 \text{Var}(\hat{\mathbb{S}}(x)) \\ &\approx \left(\frac{1}{\hat{S}(x)}\right)^2 (\hat{S}(x))^2 \sum_{j=1}^{n} \frac{d_j}{n_j(n_j - d_j)} \\ &= \sum_{j=1}^{n} \frac{d_j}{n_j(n - d_j)}. \end{aligned}$$

Para cualquier probabilidad α, se puede encontrar un cuantil $z_{\alpha/2}$ de la distribución normal estándar tal que

$$P(-z_{\alpha/2} \leqslant \frac{\log \hat{\mathbb{S}}(x) - \log \hat{S}(x)}{\sqrt{\sum_{j=1}^{n} \frac{d_j}{n_j\,(n_j - d_j)}}} \leqslant z_{\alpha/2}) \approx 1 - \alpha.$$

De donde se obtiene que

$$P(\hat{S}(x)\, \exp\{-z_{\alpha/2}\sqrt{\Sigma}\} \leqslant \mathbb{S}(x) \leqslant \hat{S}(x)\, \exp\{+z_{\alpha/2}\sqrt{\Sigma}\}) \approx 1 - \alpha.$$

Esta expresión proporciona una intervalo de confianza aproximado al $(1 - \alpha)100\,\%$ para la probabilidad desconocida $S(x)$, para cada $x > 0$. Observe que los extremos del intervalo están expresados en términos de la función de supervivencia de Kaplan- Meier $\hat{S}(x)$, la cual está calculada a partir de datos particulares $(x_1, \delta_1), \ldots, (x_n, \delta_n)$, y no se tiene garantía de que el extremo derecho del intervalo no rebase el valor 1.

Una aplicación simple del estimador de Kaplan-Meier

Sea $\hat{S}(x)$ el estimador de Kaplan-Meier para un conjunto de datos de supervivencia con posible censura por la derecha. A continuación se muestra un procedimiento en el que se utiliza $\hat{S}(x)$ para justificar el uso de los modelos paramétricos exponencial y Weibull para la representación teórica de los datos.

Modelo exponencial

Para este modelo sabemos que $S(x) = \exp(-\lambda x)$, para $x > 0$, en donde el parámetro $\lambda > 0$ es desconocido. Tomando logaritmo,

$$\log S(x) = -\lambda x, \quad \text{para} \quad x > 0.$$

Si la gráfica de la función $x \mapsto \log \hat{S}(x)$ se asemeja a la gráfica de una línea recta de la forma $x \mapsto -\lambda x$, entonces se tendría evidencia empírica para buscar ajustar el modelo exponencial. El parámetro desconocido λ puede ser estimado, por ejemplo, por mínimos cuadrados.

Modelo Weibull

Para la distribución Weibull(α, λ) sabemos que $S(x) = \exp(-\lambda x^\alpha)$, para $x > 0$, en donde $\alpha > 0$ y $\lambda > 0$ son dos parámetros desconocidos. Tomando logaritmo,

$$\log S(x) = -\lambda\, x^\alpha, \quad \text{para} \quad x > 0.$$

Tomando logaritmo por segunda vez,

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

Definiendo la variable $u = \log x$, se tiene la función lineal $u \mapsto \log \lambda + \alpha\, u$. Por lo tanto, si la gráfica del conjunto de puntos $(\log x, \log(-\log \hat{S}(x)))$, para x tal que $\hat{S}(x) > 0$, tiene el aspecto de una línea recta con pendiente positiva, entonces se tendría evidencia empírica para buscar ajustar el modelo Weibull al conjunto de datos dado. Nuevamente, la estimación por mínimos cuadrados de la recta $u \mapsto \log \lambda + \alpha\, u$ produce una estimación para los parámetros desconocidos α y λ de este modelo.

4.6. Estimador de Nelson-Aalen[2]

Este estimador es una simplificación del estimador de Kaplan-Meier. Para su definición se utiliza la aproximación

$$\log(1 - x) \approx -x, \quad \text{cuando} \quad 0 < x < 1, \tag{4.24}$$

[2] Wayne Nelson (1936–) estadístico estadounidense.
Odd Olai Aalen (1947–) estadístico noruego.

en donde, recordemos, log indica logaritmo natural. Esta aproximación corresponde a la serie de Taylor de la función $\log(1-x)$ alrededor de $x = 0$ hasta el término lineal, y es más precisa para valores pequeños de x, pues cuando $x = 0$ ambas expresiones toman el valor cero y conforme x crece la diferencia se incrementa. Este comportamiento se muestra en la Figura 4.17.

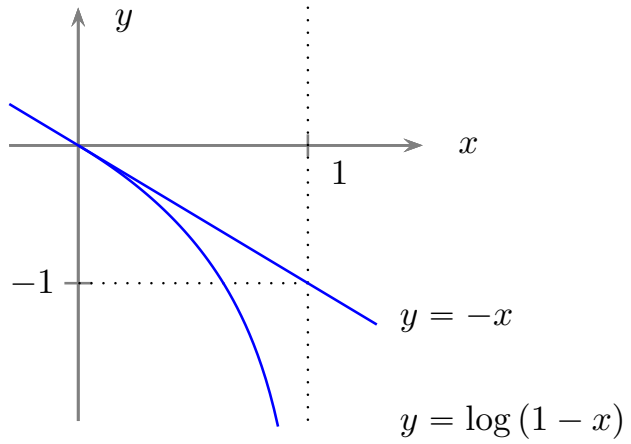

Figura 4.17: Aproximación $\log(1-x) \approx -x$, para $x > 0$ pequeño.

Recordemos que, para tiempos de vida continuos, se cumple la relación

$$S(x) = \exp\{-\Lambda(x)\}, \quad \text{para } x > 0,$$

en donde $\Lambda(x) = \int_0^x \lambda(u)\, du$ es la función de riesgo acumulado. Por lo tanto,

$$\Lambda(x) = -\log S(x), \quad \text{para } x \text{ tal que } S(x) > 0.$$

Sea $\hat{S}_{KM}(x)$ el estimador de Kaplan-Meier construido a partir de los datos de supervivencia $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$ con posible censura por la derecha. Recordemos que n_j denota el número de individuos que ingresan con vida al intervalo $(x_{(j-1)}, x_{(j)}]$ y d_j es el número de fallecimientos al tiempo $x_{(j)}$, para $j = 1, \ldots, n$ y en donde se define $x_{(0)} = 0$. Usando el estimador de Kaplan-Meier y la aproximación (4.24) se propone el siguiente estimador para $\Lambda(x)$: para $0 \leqslant x \leqslant x_{(n)}$,

$$\begin{aligned}
\hat{\Lambda}(x) &= -\log \hat{S}_{KM}(x) \\
&= -\log \prod_{j\,:\,x_{(j)}\leqslant x} (1 - \frac{d_j}{n_j}) \\
&= -\sum_{j\,:\,x_{(j)}\leqslant x} \log (1 - \frac{d_j}{n_j}) \\
&\approx \sum_{j\,:\,x_{(j)}\leqslant x} \frac{d_j}{n_j}.
\end{aligned}$$

Este es el estimador de Nelson-Aalen para $\Lambda(x)$ y de aquí se construye el estimador para $S(x)$. El resultado es el siguiente.

Definición 4.14 (Estimador de Nelson-Aalen) *Sean $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$ datos de supervivencia ordenados en orden creciente y con posible censura por la derecha. Sea n_j el número de individuos que ingresan con vida al intervalo $(x_{(j-1)}, x_{(j)}]$ y sea d_j el número de fallecimientos al tiempo $x_{(j)}$, para $j = 1, \ldots, n$. El estimador de Nelson-Aalen es*

$$\hat{S}(x) := \exp\{-\sum_{j\,:\,x_{(j)}\leqslant x} \frac{d_j}{n_j}\}, \quad \textit{para} \;\; 0 \leqslant x \leqslant x_{(n)}. \tag{4.25}$$

Para enfatizar que el estimador (4.25) es el de Nelson-Aalen, se le denota por $\hat{S}_{NA}(x)$. Observe que en los tiempos en donde no hay fallecimientos, es decir, cuando $d_j = 0$, la función $\hat{S}_{NA}(x)$ no cambia. Esta función decrece sólo cuando $d_j \geqslant 1$. Para fines de comparación, observe que los factores $\hat{p}_j$ son

$$\hat{p}_j = \begin{cases} 1 - d_j/n_j & \text{para } \hat{S}_{KM}(x), \\ \exp\{-d_j/n_j\} & \text{para } \hat{S}_{NA}(x). \end{cases}$$

Como $e^{-x} \geqslant 1 - x$, para cualquier valor de x, puede comprobarse que $\hat{S}_{KM}(x) \leqslant \hat{S}_{NA}(x)$, para $0 \leqslant x \leqslant x_{(n)}$.

Como antes, una vez obtenida la función de supervivencia estimada $\hat{S}_{NA}(x)$, otras características del tiempo de vida en estudio pueden calcularse. Veamos un ejemplo.

Ejemplo 4.4 *Consideremos nuevamente el siguiente conjunto de $n = 9$ observaciones que habíamos estudiado antes en el Ejemplo 4.3 de la página 178. Observe que los datos están ordenados de menor a mayor, algunos de ellos son datos completos y otros están censurados por la derecha.*

$$x_{(1)} = 1, \quad x_{(2)} = 2+, \quad x_{(3)} = 2+, \quad x_{(4)} = 2+, \quad x_{(5)} = 4, \quad x_{(6)} = 4,$$
$$x_{(7)} = 6, \quad x_{(8)} = 6+, \quad x_{(9)} = 7+ \ .$$

La representación gráfica de estos datos se muestra en la Figura 4.13. Así como ocurre para el estimador de Kaplan-Meier, los tiempos registrados en donde suceden fallecimientos son también los tiempos relevantes en la construcción del estimador de Nelson-Aalen. En la Tabla 4.7 se muestra el cálculo del estimador de Nelson-Aalen siguiendo la fórmula (4.25) *y denotado por $\hat{S}_{NA}(x)$. La columna $\hat{p}_j$ se refiere al factor $\hat{p}_j = \exp\{-d_j/n_j\}$. Observe que el último dato $x_{(9)} = 7+$ es una censura y ello provoca que el estimador no alcance el valor* 0.

j	$x_{(j)}$	Intervalo	n_j	d_j	$\hat{p}_j$	$\hat{S}_{NA}(x_{(j)})$	$\hat{S}_{KM}(x_{(j)})$
0	0	–	–	–	–	1	1
1	1	$(0, 1]$	9	1	0.8948	0.8948	$0.\bar{8}$
2	2	$(1, 2]$	8	0	1	0.8948	$0.\bar{8}$
3	4	$(2, 4]$	5	2	0.6703	0.5998	$0.5\bar{3}$
4	6	$(4, 6]$	3	1	0.7165	0.4297	$0.3\bar{5}$
5	7	$(6, 7]$	1	0	1	0.4297	$0.3\bar{5}$

Tabla 4.7: Construcción de $\hat{S}_{NA}(x)$ para los datos del Ejemplo 4.4.

Para fines de comparación se muestran también los valores del estimador de Kaplan-Meier encontrados antes, ahora denotado por $\hat{S}_{KM}(x)$. Observe que se verifica numéricamente la relación general $\hat{S}_{KM}(x) \leqslant \hat{S}_{NA}(x)$, para $0 \leqslant x \leqslant x_{(n)}$. Las gráficas de estos estimadores se muestran en la Figura 4.18.

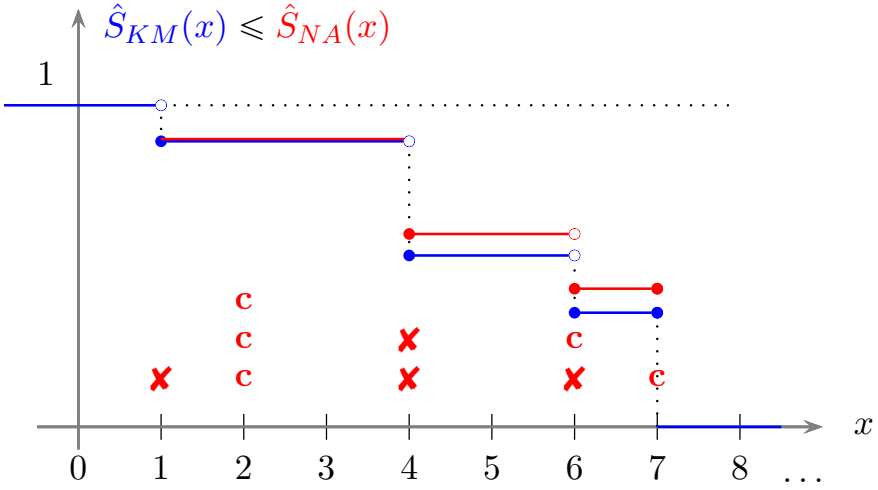

Figura 4.18: Comparación de los estimadores de Kaplan-Meier y Nelson-Aalen para los datos del Ejemplo 4.4.

■

El estimador de Nelson-Aalen se puede escribir también como variable aleatoria y se pueden estudiar sus propiedades estadísticas. Puede consultarse mayor información en los libros de J. P. Klein y M. L. Moeschberger [33] ó T. M. Therneau y P. M. Grambsch [49].

4.7. Prueba log-rank para comparar poblaciones

En esta sección se estudia una prueba de hipótesis no paramétrica llamada log-rank, que se utiliza para determinar si existe alguna diferencia significativa entre dos conjuntos de datos de supervivencia.

Supongamos que se cuenta con observaciones de los tiempos de vida de dos grupos o poblaciones de individuos, los cuales están sujetos a posible censura por la derecha. Como un ejemplo tenemos que el primer grupo puede estar constituido por hombres y el segundo grupo por mujeres. Como un segundo ejemplo, y en el contexto del estudio estadístico de nuevos medicamentos, el primer grupo puede estar integrado por aquellos pacientes a quienes se les suministra un nuevo medicamento, mientras que en el segundo grupo se

colocan a los pacientes a los que se les suministra un placebo o medicamento inocuo. Se busca determinar si ambos grupos presentan la misma experiencia de mortalidad o, bien, si existe diferencia entre ellos.

Consideremos el caso de dos poblaciones a las que llamaremos grupo 0 y grupo 1. A las correspondientes funciones de supervivencia (desconocidas) se les denotará por $S_0(x)$ y $S_1(x)$, respectivamente. Nos interesa llevar a cabo el contraste de hipótesis:

$$H_0: \quad S_0(x) = S_1(x) \text{ para todo } x > 0. \qquad \text{vs} \qquad H_1: \quad S_0(x) \neq S_1(x) \text{ para algún } x > 0.$$

Sea $x_1, \ldots, x_n$ el conjunto de observaciones de los dos grupos conjuntamente, algunos de los cuales pueden presentar censura por la derecha. Supondremos que el grupo 0 tiene n_0 individuos y el grupo 1 tiene n_1 individuos. Se cumple que $n_0 + n_1 = n$.

Figura 4.19: El intervalo j de análisis es $(x'_{(j-1)}, x'_{(j)}]$, $j = 1, \ldots, k$, determinado por dos tiempos de fallecimientos sucesivos.

Procediendo de la misma forma como en el estimador de Kaplan-Meier, sean $x'_{(1)} < \cdots < x'_{(k)}$ los tiempos de fallecimientos efectivos de ambos grupos conjuntamente y ordenados de menor a mayor. Hemos omitido de esta lista los tiempos de censura y sólo incluimos los tiempos en los que se presenta uno o más fallecimientos. Observe que $1 \leqslant k \leqslant n$. Consideraremos nuevamente los intervalos de tiempo $(x'_{(j-1)}, x'_{(j)}]$, para $j = 1, \ldots, k$, nos referiremos a uno cualquiera de ellos como el intervalo j. Véase la Figura 4.19. Nuevamente, es conveniente definir $x'_{(0)} := 0$.

Se hace un análisis de lo que ocurre en cada uno de estos intervalos de la siguiente manera:

- Se toma como información conocida al inicio de cada intervalo las siguientes cantidades: Para cada intervalo $j = 1, \ldots, k$ se define el número

$$n_j = \#\text{ Individuos en observación al inicio del intervalo } j.$$

  Claramente $n_1 = n$. Se define también la misma cantidad para cada grupo por separado como

$$n_{0j} = \#\text{ Individuos del grupo 0 en observación al inicio del intervalo } j,$$
$$n_{1j} = \#\text{ Individuos del grupo 1 en observación al inicio del intervalo } j.$$

  Claramente se debe cumplir la igualdad $n_j = n_{0j} + n_{1j}$. Se define también el número

$$d_j = \#\text{ Individuos que fallecen en el intervalo } j.$$

  Observemos que, por construcción de los intervalos, el número de fallecimientos d_j es mayor o igual a 1, y que estos fallecimientos ocurren en el extremo derecho del intervalo j, es decir, en el tiempo $x'_{(j)}$.

- Se les considera como variables aleatorias a las cantidades:

$$D_{0j} = \#\text{ Individuos del grupo 0 que fallecen en el intervalo } j,$$
$$D_{1j} = \#\text{ Individuos del grupo 1 que fallecen en el intervalo } j.$$

  Es evidente que se cumple la igualdad $D_{0j} + D_{1j} = d_j$ y que el uso de la letra D proviene del término en inglés *Death*. Por otro lado, se debe observar que las variables D_{0j}, para $j = 1, \ldots, k$, no son independientes pues el número de fallecidos en un intervalo cualquiera j depende del número de individuos en riesgo al inicio del intervalo y esta cantidad depende del número de fallecidos en los intervalos anteriores. Lo mismo puede afirmarse de la variables D_{1j}, $j = 1, \ldots, k$.

Así, para cada intervalo j, las definiciones y notación anteriores se pueden representar mediante un arreglo tabular como el que se muestra en la Tabla 4.8.

	Fallecidos	Sobrevivientes	Suma
Grupo 0	D_{0j}	$n_{0j} - D_{0j}$	n_{0j}
Grupo 1	D_{1j}	$n_{1j} - D_{1j}$	n_{1j}
Suma	d_j	$n_j - d_j$	n_j

Tabla 4.8: Variables aleatorias: número de individuos fallecidos y número de sobrevivientes, por grupo, en el intervalo j.

De este modo, para cada intervalo j se especifica el número de individuos en riesgo de morir para cada uno de los grupos, es decir, n_{0j} y n_{1j}, en donde debe cumplirse que $n_{0j} + n_{1j} = n_j$, y se especifica también el total de fallecimientos d_j que ocurren en el intervalo en cuestión. Lo que queda sin determinar es el número de fallecimientos en cada grupo, es decir, las variables aleatorias D_{0j} y D_{1j}.

Cuando H_0 es cierta, es decir, cuando ambos grupos tienen la misma experiencia de mortalidad, los n_j individuos en riesgo de morir tienen todos la misma probabilidad de fallecer en el intervalo en estudio, sin importar si pertenecen al grupo 0 ó al grupo 1. Tenemos aquí la situación de contar con un conjunto de n_j objetos (individuos). Cada objeto pertenece a una de dos categorías: grupo 0 con n_{0j} elementos, y grupo 1 con $n_{1j} = n_j - n_{0j}$ elementos. Del grupo completo se desea que en una muestra de tamaño d_j (número de fallecidos), se obtenga un cierto número de objetos de cada categoría o grupo. Este es el contexto en el que puede aplicarse la distribución hipergeométrica. Los parámetros de esta distribución son $(N, K, n) = (n_j, n_{0j}, d_j)$.

Entonces, bajo la hipótesis nula H_0, el número de fallecimientos del grupo 0, es decir, la variable D_{0j}, tiene una distribución hipergeo(n_j, n_{0j}, d_j), en donde los parámetros son las siguientes cantidades conocidas:

$$
\begin{aligned}
n_j &= \text{“Total de elementos”} \\
&\quad \text{(Número de individuos vivos al inicio del intervalo } j\text{)}. \\
n_{0j} &= \text{“Número de elementos de la categoría 0”} \\
&\quad \text{(Número de individuos del grupo 0)}. \\
d_j &= \text{“Tamaño de la muestra”} \\
&\quad \text{(Número de individuos fallecidos en el intervalo } j\text{)}.
\end{aligned}
$$

Es conveniente recordar aquí que la función de probabilidad de la distribución hipergeo(n_j, n_{0j}, d_j) está dada por

$$
P(D_{0j} = d_{0j}) = \frac{\binom{n_{0j}}{d_{0j}}\binom{n_j - n_{0j}}{d_j - d_{0j}}}{\binom{n_j}{d_j}}, \quad \text{para} \;\; d_{0j} = 0, 1, \ldots, n_{0j},
$$

en donde el valor d_{0j} representa el número observado de fallecimientos del grupo 0. Es necesario reiterar que la distribución indicada para D_{0j} se obtiene cuando se asume que la hipótesis H_0 se cumple. Por otro lado, conociendo la distribución de D_{0j}, se puede encontrar su esperanza y varianza, las cuales son:

$$
e_{0j} := E(D_{0j}) = d_j \, \frac{n_{0j}}{n_j}, \tag{4.26}
$$

$$
v_{0j} := \text{Var}(D_{0j}) = d_j \, \frac{n_{0j}}{n_j} \, \frac{n_j - n_{0j}}{n_j} \, \frac{n_j - d_j}{n_j - 1}. \tag{4.27}
$$

Observe que todos los términos de estas fórmulas son conocidos para cada intervalo j. Para el último intervalo (es decir, para el máximo valor de j), se tiene que $n_j = d_j$ y se define la varianza como $v_{0j} = 0$.

De manera análoga, tenemos también que la variable D_{1j} tiene distribución hipergeo(n_j, d_j, n_{1j}). Sin embargo, para nuestro análisis, es equivalente considerar una variable aleatoria o la otra. Regresando a la variable D_{0j}, se

puede definir el número total de observaciones como la variable aleatoria

$$O := \sum_{j=1}^{k} D_{0j}. \tag{4.28}$$

La letra O indica observaciones de fallecimientos. Bajo la hipótesis nula H_0, la esperanza y varianza de la variable O son:

$$E := E(O) = \sum_{j=1}^{k} e_{0j}, \tag{4.29}$$

$$V := \mathrm{Var}(O) \approx \sum_{j=1}^{k} v_{0j}. \tag{4.30}$$

La varianza sólo es aproximada pues las variables D_{0j}, $j = 1, \ldots, k$, no son independientes. Se define entonces el estadístico de la prueba como la variable aleatoria

$$Z = \frac{O - E}{\sqrt{V}}. \tag{4.31}$$

Bajo la hipótesis nula H_0 y usando el teorema central del límite, la variable Z tiene una distribución aproximada $\mathrm{N}(0, 1)$. Esto implica que Z^2 tiene distribución $\chi^2(1)$.

- *Procedimiento clásico.* Cuando la hipótesis nula H_0 es cierta, es decir, cuando no existe diferencia entre las dos poblaciones, la variable Z toma valores pequeños pues la suma $\sum_{j=1}^{k} D_{0j}$ tiende a estar cercana a su media. Se rechaza H_0 cuando Z^2 es grande, más específicamente, cuando $Z^2 > \chi^2_{(1),1-\alpha}$, en donde el término $\chi^2_{(1),1-\alpha}$ denota el cuantil al nivel $100(1-\alpha)\,\%$ de la distribución $\chi^2(1)$. Por ejemplo, para $1 - \alpha = 0.9$, tenemos que $\chi^2_{(1),1-\alpha} = 2.70554$. También es común tomar $1 - \alpha = 0.95$ y, en ese caso, $\chi^2_{(1),1-\alpha} = 3.84146$.

- *Procedimiento usando el p-value.*[3] Recordemos que el p-value es la probabilidad de que la estadística de la prueba tome un valor como el

[3]Debido a su amplio uso en la literatura estadística, se usa el término en inglés *p-value*.

observado u otros más alejados de la hipótesis nula, suponiendo que la hipótesis nula es cierta. En nuestro caso, la estadística de la prueba es $T = Z^2$ y se trata de una prueba de cola derecha: rechazar H_0 cuando T es grande. El p-value se calcula entonces de la siguiente forma:

$$p\text{-value} = P(T \geqslant t \mid H_0 \text{ es cierta}) = 1 - F_T(t),$$

en donde t es el valor tomado por la estadística T. El p-value puede servir como una alternativa para rechazar la hipótesis nula pues provee el nivel de significancia más pequeño para el cual la hipótesis nula puede rechazarse: si el *p-value* es pequeño y resulta ser menor que un cierto nivel de significancia, por ejemplo, $\alpha = 0.05$, se puede rechazar la hipótesis nula. En tal caso, se concluye que los datos presentan evidencia para afirmar que existe una diferencia en las experiencias de mortalidad de ambos grupos.

Ejemplo 4.5 *Este es un ejemplo general que resume el procedimiento del cálculo para llevar a cabo la prueba log-rank a partir de un conjunto de tiempos de vida $x_1, \ldots, x_n$, con posible censura por la derecha, para dos poblaciones conjuntamente. Los pasos son los siguientes:*

- *Se omiten los datos censurados conservando sólo los tiempos de fallecimientos observados y se ordenan estos tiempos de menor a mayor: $x'_{(1)} < x'_{(2)} < \cdots < x'_{(k)}$, en donde $1 \leqslant k \leqslant n$.*
- *Se define el intervalo j como $(x'_{(j-1)}, x'_{(j)}]$, cuyos extremos son los tiempos de fallecimientos, $j = 1, \ldots, k$, en donde $x'_{(0)} := 0$.*
- *Para cada intervalo j se calculan las cantidades $n_j, n_{0j}, n_{1j}, d_j, d_{0j}, d_{1j}$, en donde $n_j = n_{0j} + n_{1j}$ y $d_j = d_{0j} + d_{1j}$, para $j = 1, \ldots, k$. Observe que aquí se toman en cuenta los datos censurados pues $n_{j+1} = n_j - d_j - c_j$, en donde c_j es el número de tiempos censurados en el intervalo j. También se calculan las medias o valores esperados e_{0j} y las varianzas v_{0j} como fue indicado en las fórmulas* (4.26) *y* (4.27)*.*
- *Finalmente, se calcula el valor z de la estadística de la prueba (o bien*

z^2) mediante la fórmula

$$z = \frac{\sum_{j=1}^{k} d_{0j} - \sum_{j=1}^{k} e_{0j}}{\sqrt{\sum_{j=1}^{k} v_{0j}}}. \tag{4.32}$$

■

Ejemplo 4.6 *Consideremos que el símbolo x^0 denota el siguiente conjunto de 6 datos de supervivencia del grupo 0,*

$$x^0 = \{\ 0.3,\ 0.5+,\ 1.4,\ 3.8,\ 6.8,\ 7.7\ \}.$$

Los tiempos se encuentran ordenados de menor a mayor. Por otro lado, suponga que el término x^1 denota el conjunto de los siguientes 8 datos también ya ordenados del grupo 1,

$$x^1 = \{\ 0.3,\ 1.4,\ 2.4+,\ 2.9+,\ 3.5,\ 5.5,\ 6.2,\ 23\ \}.$$

El total de individuos es $n = 6 + 8 = 14$. Considerando únicamente los tiempos de fallecimientos de ambos grupos, se conforman los 9 intervalos que aparecen abajo, al final de cada uno de los cuales ocurre uno o varios fallecimientos. Estos son los intervalos en los que se lleva a cabo el análisis:

$$(0, 0.3], (0.3, 1.4], (1.4, 3.5], (3.5, 3.8], (3.8, 5.5],$$
$$(5.5, 6.2], (6.2, 6.8], (6.8, 7.7], (7.7, 23].$$

La información de ambos grupos de datos se muestra en la Tabla 4.9. La variable c_j denota el número de datos censurados en el intervalo j. El número de censuras por grupo se denota por las variables c_{0j} y c_{1j}. En las últimas dos columnas aparecen la esperanza e_{0j} y la varianza v_{0j}, las cuales se calculan como indican las fórmulas (4.26) y (4.27), respectivamente. Con estos cálculos puede comprobarse que la estadística $Z^2 = (O - E)^2/V$ toma el valor 0.076 y el p-value es 0.8 .

j	Intervalo	n_j	n_{0j}	n_{1j}	d_j	d_{0j}	d_{1j}	c_j	c_{0j}	c_{1j}	e_{0j}	v_{0j}
1	$(0, 0.3]$	14	6	8	2	1	1	0	0	0	0.857	0.452
2	$(0.3, 1.4]$	12	5	7	2	1	1	1	1	0	0.833	0.441
3	$(1.4, 3.5]$	9	3	6	1	0	1	2	0	2	0.333	0.222
4	$(3.5, 3.8]$	6	3	3	1	1	0	0	0	0	0.500	0.250
5	$(3.8, 5.5]$	5	2	3	1	0	1	0	0	0	0.400	0.240
6	$(5.5, 6.2]$	4	2	2	1	0	1	0	0	0	0.500	0.250
7	$(6.2, 6.8]$	3	2	1	1	1	0	0	0	0	0.666	0.222
8	$(6.8, 7.7]$	2	1	1	1	1	0	0	0	0	0.500	0.250
9	$(7.7, 23]$	1	0	1	1	0	1	0	0	0	0.000	0.000

Tabla 4.9: Análisis de datos del Ejemplo 4.6

A continuación llevaremos a cabo el contraste de hipótesis con el paquete estadístico `R` *usando la biblioteca* `survival`*. Se declaran primero los datos de cada grupo por separado y después se concatenan. Usamos las variables* `time`*,* `status` *y* `group` *para denotar: los tiempos de vida registrados, la censura o no censura (función delta), y el grupo (0 ó 1), respectivamente.*

```
# Organización de los datos del Ejemplo 4.6
library(survival)
# Datos del grupo 0:
time0 <- c(0.3,0.5,1.4,3.8,6.8,7.7)
status0 <- c(1,0,1,1,1,1)
group0 <- c(0,0,0,0,0,0)
# Datos del grupo 1:
time1 <- c(0.3,1.4,2.4,2.9,3.5,5.5,6.2,23)
status1 <- c(1,1,0,0,1,1,1,1)
group1 <- c(1,1,1,1,1,1,1,1)
# Concatenación:
time <- c(time0,time1)
status <- c(status0,status1)
group <- c(group0,group1)
df <-data.frame(time,status,group)
```

De esta manera la información completa se encuentra en el arreglo `df` *(dataframe). Se puede dar la instrucción* `print(df)` *para visualizar el listado completo del arreglo* `df`*. La prueba log-rank en* `R` *se lleva a cabo llamando a*

la función `survdiff`*. En el siguiente recuadro se muestra la sintaxis de ese comando junto con los resultados que se obtienen.*

R

```
# Prueba log-rank para los datos del Ejemplo 4.6
survdiff(Surv(time,status)~group, data=df)

        N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 6 5 4.58 0.0386 0.076
group=1 8 6 6.42 0.0275 0.076

Chisq= 0.1 on 1 degrees of freedom, p= 0.8
```

Como puede observarse, la función `survdiff` *proporciona un resumen de algunas estadísticas de los dos grupos en estudio. Para cada grupo, la función muestra el número inicial de individuos al inicio del experimento:* 6 *y* 8*, el número de muertes observadas:* 5 *y* 6*, el número de muertes esperadas:* 4.58 *y* 6.42*, y el valor de las estadísticas* $(O-E)^2/E$ *y* $(O-E)^2/V$*, en donde* O*,* E *y* V *es la suma de observaciones, valor esperado y la varianza de cada grupo. En particular, se corrobora que el valor que toma el estadístico de la prueba* $Z^2 = (O-E)^2/V$ *es* 0.076 *.*

En la parte final, la función `survdiff` *muestra el valor* `Chisq=0.1`*. Este es el valor de la estadística de la prueba, la cual tiene una distribución ji-cuadrada con* 1 *grado de libertad. Finalmente se proporciona el p-value de la prueba, que es* 0.8*.*

Tomando un nivel de significancia de $\alpha = 0.05$*, se rechaza la hipótesis nula cuando p-value* $< \alpha$*. Como esto no ocurre, no se rechaza* H_0*, es decir, los datos sugieren que ambos grupos tienen la misma experiencia de mortalidad.*

Se pueden graficar las curvas de supervivencia de los dos grupos por separado usando `R`*. La instrucción se encuentra en el siguiente recuadro y la gráfica que se obtiene se muestra en la Figura 4.20.*

R

```
# Graficación de las curvas de supervivencia
# de los dos grupos de datos del Ejemplo 4.6
plot(survfit(Surv(time,status)~group,data=df),
axes=FALSE))
```

Puede observarse que las dos curvas de supervivencia mantienen una cierta cercanía una de la otra en la mayor parte del rango de valores. El dato con valor 23 *dentro del grupo* x^1 *parece ser un dato aislado (outlier) que hace que la curva de supervivencia del grupo* 1 *muestre una llegada al valor* 0 *más tardíamente.*

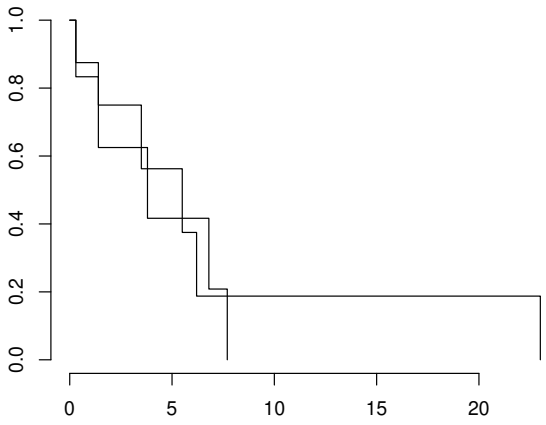

Figura 4.20: Funciones de supervivencia de los datos del Ejemplo 4.6.

■

Ejemplo 4.7 *Consideremos el siguiente conjunto de* 10 *datos de supervivencia de un primer grupo* 0*:*

$$x^0 = \{\ 0.1,\ 1.4+,\ 1.6,\ 2.5+,\ 2.8,\ 3.2,\ 5.6,\ 6.2,\ 6.4,\ 16.2\ \}.$$

Por comodidad, los tiempos aparecen ya ordenados de menor a mayor; dos de ellos están censurados por la derecha. Supongamos, además, que tenemos

los siguientes datos para un segundo grupo 1*:*

$$x^1 = \{\ 0.6,\ 0.7,\ 2.8+,\ 3.6,\ 4.2,\ 4.7+,\ 6.1+,\ 6.2,\ 13.3,\ 16.2\ \}.$$

Tenemos nuevamente 10 *registros para este segundo grupo y en él aparecen tres datos censurados. Considerando únicamente los tiempos de fallecimientos de ambos grupos, se conforman los siguientes* 13 *intervalos, al final de cada uno de ellos ocurre uno o varios fallecimientos:*

$$(0, 0.1], (0.1, 0.6], (0.6, 0.7], (0.7, 1.6], (1.6, 2.8], (2.8, 3.2], (3.2, 3.6],$$
$$(3.6, 4.2], (4.2, 5.6], (5.6, 6.2], (6.2, 6.4], (6.4, 13.3], (13.3, 16.2].$$

La información de los datos de supervivencia se muestra en la Tabla 4.10.

j	Intervalo	n_j	n_{0j}	n_{1j}	d_j	d_{0j}	d_{1j}	c_j	c_{0j}	c_{1j}	e_{0j}	v_{0j}
1	$(0, 0.1]$	20	10	10	1	1	0	0	–	–	–	–
2	$(0.1, 0.6]$	19	9	10	1	0	1	0	–	–	–	–
3	$(0.6, 0.7]$	18	9	9	1	0	1	0	–	–	–	–
4	$(0.7, 1.6]$	17	9	8	1	1	0	1	–	–	–	–
5	$(1.6, 2.8]$	15	7	8	1	1	0	2	–	–	–	–
6	$(2.8, 3.2]$	12	5	7	1	1	0	0	–	–	–	–
7	$(3.2, 3.6]$	11	4	7	1	0	1	0	–	–	–	–
8	$(3.6, 4.2]$	10	4	6	1	0	1	0	–	–	–	–
9	$(4.2, 5.6]$	9	4	5	1	1	0	1	–	–	–	–
10	$(5.6, 6.2]$	7	3	4	2	1	1	1	–	–	–	–
11	$(6.2, 6.4]$	4	2	2	1	1	0	0	–	–	–	–
12	$(6.4, 13.3]$	3	1	2	1	0	1	0	–	–	–	–
13	$(13.3, 16.2]$	2	1	1	2	1	1	0	–	–	–	–

Tabla 4.10: Análisis de datos del Ejemplo 4.7.

La variable c_j *denota el número de datos censurados en el intervalo* j *y las variables* c_{0j} *y* c_{1j} *el desglose del número de censuras por grupo. Se calcula el valor de la estadística de prueba* Z *y se encuentra que* $z = 0.69$*. Tomando* $1 - \alpha = 0.9$*, tenemos que* $\chi^2_{(1),1-\alpha} = 2.70554$*, de modo que, como el valor de* $z^2 = (0.69)^2 = 0.4761$ *no rebasa el valor* 2.70554*, no se rechaza* H_0*. Así, la muestra aleatoria analizada permite suponer con un* 90 % *de confianza que*

los dos grupos en estudio tienen aproximadamente la misma experiencia de mortalidad, es decir, sus funciones de supervivencia son parecidas. ∎

A continuación se mencionan algunas variantes de la prueba log-rank. La intención es mostrar que se pueden considerar algunas extensiones de la prueba presentada. Un estudio más detallado debe considerar las propiedades estadísticas de estas pruebas.

Prueba log-rank con ponderaciones

En esta variante de la prueba log-rank se tienen pesos $\omega_1 \geqslant 0, \ldots, \omega_k \geqslant 0$ que se aplican a las variables aleatorias D_{0j}, dando mayor o menor relevancia a las observaciones de estas variables. La letra ω proviene del término en inglés *weights*. Así, la variable D_{0j} con ponderación ω_j se define como $\omega_j D_{0j}$, modificando su media y varianza. De este modo, la estadística de la prueba log-rank con ponderaciones es

$$Z_\omega := \frac{\sum_{j=1}^{k} \omega_j (D_{0j} - e_{0j})}{\sqrt{\sum_{j=1}^{k} \omega_j^2 v_{0j}}}.$$

Nuevamente, esta variable tiene distribución aproximada normal estándar y la prueba se lleva a cabo como en el caso usual. Cuando los pesos son una misma constante $\omega_j = \omega > 0$, la estadística Z_ω se reduce a la estadística Z usual.

Prueba log-rank estratificada

Supongamos que se desean comparar las funciones de supervivencia de dos grupos conformados, por ejemplo, por pacientes con dos tratamientos distintos. Supongamos también que la población completa está estratificada de acuerdo a una variable categórica, por ejemplo, hombres y mujeres. La prueba log-rank se puede llevar a cabo considerando esta clasificación.

Supongamos que $S_0^{(\ell)}(x)$ y $S_1^{(\ell)}(x)$ denotan las funciones de supervivencia de los grupos 0 y 1, en el estrato ℓ. Se puede llevar a cabo una prueba log-rank ordinaria para comparar $S_0^{(\ell)}$ y $S_1^{(\ell)}$, para cada estrato ℓ.

También se puede buscar una comparación global considerando la hipótesis nula

$$H_0 : \ S_0^{(\ell)}(x) = S_1^{(\ell)}(x), \quad \ell = 1, \ldots, L, \quad \text{para todo } x > 0,$$

en donde L es el total de estratos. Es decir, la hipótesis establece que $S_0(x)$ y $S_1(x)$ son iguales en cada uno de los estratos. Por ejemplo, en el caso de la variable sexo se tienen $L = 2$ estratos y la hipótesis afirma que no hay diferencia en la experiencia de mortalidad entre hombres y mujeres bajo los dos tratamientos.

Para esta prueba se calculan las cantidades $O^{(\ell)}$, $E^{(\ell)}$, y $V^{(\ell)}$ para cada estrato ℓ como en la prueba ordinaria, véase (4.28), (4.29) y (4.30), y después se suman estas cantidades para todos los estratos. La estadística de la prueba es la variable aleatoria Z que aparece abajo, la cual tiene una distribución aproximada normal estándar suponiendo H_0 cierta.

$$Z = \frac{O - E}{\sqrt{V}} = \frac{\sum_{\ell=1}^{L} O^{(\ell)} - \sum_{\ell=1}^{L} E^{(\ell)}}{\sqrt{\sum_{\ell=1}^{L} V^{(\ell)}}}.$$

Prueba log-rank para varios grupos

Se pueden llevar a cabo pruebas log-rank para 3 o más poblaciones. Una primera forma de proceder consiste en tomar las poblaciones por pares y determinar si existe diferencia entre cada par de poblaciones. También existe una extensión del procedimiento aquí expuesto en el que la hipótesis nula se define como la afirmación que establece que todas las funciones de supervivencia son iguales contra la hipótesis alternativa que afirma que alguna de ellas es diferente. Más específicamente, si se cuenta con $p + 1$ grupos,

denotados por los números $0, 1, \ldots, p$ y las funciones de supervivencia correspondientes se les escribe con subíndice esos valores, entonces la hipótesis nula que se plantea es

$$H_0 : S_0(x) = S_1(x) = \cdots = S_p(x) \quad \text{para todo } x > 0.$$

Esta hipótesis indica que no hay diferencia en los $p + 1$ grupos en estudio. En contraparte, la hipótesis alternativa sería que al menos uno de los grupos presenta experiencia de mortalidad distinta de los otros grupos.

El procedimiento es análogo al caso de dos grupos. Considerando únicamente los tiempos de fallecimientos $0 = x'_{(0)} < x'_{(1)} < \cdots < x'_{(k)}$ de la población completa, se analiza lo que ocurre entre dos tiempos sucesivos $x'_{(j-1)}$ y $x'_{(j)}$. A este intervalo de tiempo $(x'_{(j-1)}, x'_{(j)}]$ le llamamos intervalo j. Como antes, se crea un arreglo tabular como el que aparece en la Tabla 4.11.

	Fallecidos	Sobrevivientes	Suma
Grupo 0	D_{0j}	$n_{0j} - D_{0j}$	n_{0j}
Grupo 1	D_{1j}	$n_{1j} - D_{1j}$	n_{1j}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Grupo p	D_{pj}	$n_{pj} - D_{pj}$	n_{pj}
Suma	d_j	$n_j - d_j$	n_j

Tabla 4.11: Variables aleatorias: número de individuos fallecidos y número de sobrevivientes, por grupo, en el intervalo j.

En la Tabla 4.11, n_j denota el número global en riesgo y n_{ij} es el número en riesgo de individuos en el grupo i. Claramente, $n_j = n_{0j} + \cdots + n_{pj}$. Además, D_{ij} es el número de fallecidos del grupo i.

Cuando H_0 es cierta, es decir, cuando todos los grupos tienen la misma experiencia de mortalidad, los n_j individuos en riesgo de morir en el intervalo j tienen todos la misma probabilidad de fallecer en el intervalo en estudio,

sin importar si pertenecen a un grupo en particular. Nuevamente, se tiene la situación de contar con un conjunto de n_j objetos (individuos), en donde cada objeto pertenece a una de las $p+1$ categorías: grupo 0 con n_{0j} elementos, grupo 1 con n_{1j} elementos, etc. Del grupo completo se desea que en una muestra de tamaño d_j (número de fallecidos), se obtenga un cierto un número de objetos de cada categoría o grupo. Así, se tiene que el vector aleatorio

$$D_j = (D_{0j}, D_{1j}, \ldots, D_{pj})$$

tiene distribución hipergeométrica multivariada con función de probabilidad

$$f(x_0, \ldots, x_p) = \frac{\binom{n_{0j}}{x_0}\binom{n_{1j}}{x_1}\cdots\binom{n_{pj}}{x_p}}{\binom{n_j}{d_j}},$$

para valores enteros $x_0, \ldots, x_p$ que satisfacen $0 \leqslant x_i \leqslant n_{ij}$ y $x_0 + x_1 + \cdots + x_p = d_j$. Se puede comprobar que cada coordenada D_{ij} del vector D_j tiene distribución hipergeométrica univariada de parámetros $(N, K, n) = (n_j, n_{ij}, d_j)$. Por lo tanto, su media y varianza están dados por

$$\begin{aligned} e_{ij} &:= E(D_{ij}) = d_j \frac{n_{ij}}{n_j}, \\ v_{ij} &:= \mathrm{Var}(D_{ij}) = d_j \frac{n_{ij}}{n_j} \frac{n_j - n_{ij}}{n_j} \frac{n_j - d_j}{n_j - 1}. \end{aligned}$$

La media y varianza del vector D_j son

$$E(D_j) = e_j = (e_{0j}, e_{1,j}, \ldots, e_{pj}),$$

$$(\mathrm{Var}(D_j))_{kl} = v_{kl}^{(j)} = \begin{cases} d_j \dfrac{n_{kj}}{n_j} \dfrac{n_j - n_{kj}}{n_j} \dfrac{n_j - d_j}{n_j - 1} & \text{si } \ k = l, \\ d_j \dfrac{n_{kj}}{n_j} \dfrac{n_{lj}}{n_j} \dfrac{d_j - n_j}{n_j - 1} & \text{si } \ k \neq l. \end{cases}$$

Finalmente se definen las cantidades globales

$$O := \sum_{j=1}^{k} D_j, \quad E := E(O) = \sum_{j=1}^{k} e_j, \quad V := (\mathrm{Var}(O))_{kl} \approx \sum_{j=1}^{k} v_{kl}^{(j)}.$$

y la estadística de la prueba es la variable aleatoria Z^2 que aparece abajo, cuya distribución es aproximada χ^2_p, cuando la hipótesis nula H_0 es cierta.

$$Z^2 = (O - E)(V)^{-1}(O - E)^t.$$

En el artículo de Villanueva *et al* [53] se puede consultar una implementación del procedimiento anterior en `R`.

§

Debe observarse que se pueden considerar combinaciones de las variantes de la prueba log-rank mencionadas u otras extensiones. Por ejemplo, se puede llevar a cabo una prueba estratificada con ponderaciones.

Algunos de los libros disponibles sobre análisis de supervivencia dedican por lo menos una sección sobre pruebas de hipótesis para comparar funciones de supervivencia de dos o más poblaciones, ya sea su aplicación en algún software estadístico o su derivación matemática. Entre estos últimos están los trabajos de D. G. Altman [3], D. Collet [15], X. Liu [42] y P. J. Smith [47]. Particularmente el capítulo 3 del libro de D. Machin, Y. B. Cheung y M. K. B. Parmar [43] está dedicado a la comparación de curvas de supervivencia. Véase también el capítulo 4 de K. Bogaerts, A. Komárek y E. Lesaffre [6].

4.8. Ejercicios

Tablas de mortalidad

132. Complete las entradas faltantes de la tabla de mortalidad que aparece abajo. Una hoja de cálculo puede ayudar a agilizar las operaciones.

x	ℓ_x	d_x	q_x	p_x
0	100			
1	82			
2	65			
3	45			
4	30			
5	10			
6	0			

133. Suponga que se cuenta con la tabla de mortalidad que aparece abajo. Encuentre las funciones básicas $f(x)$, $S(x)$, $\lambda(x)$ y $R(x)$ y escriba sus valores en las columnas indicadas. Elabore una gráfica de estas funciones.

x	ℓ_x	$f(x)$	$S(x)$	$\lambda(x)$	$R(x)$
0	100				
1	87				
2	62				
3	30				
4	16				
5	0				

134. Suponga que un modelo tabular de supervivencia está dado por la siguiente tabla para las probabilidades p_x.

x	p_x
0	0.9
1	0.8
2	0.5
3	0.2
4	0

a) Encuentre y grafique las funciones básicas $f(x)$, $S(x)$, $\lambda(x)$ y $R(x)$ para $x = 0, 1, 2, 3$.

b) Usando un radix ℓ_0 igual a $10,000$, elabore una tabla de mortalidad que muestre los valores ℓ_x y d_x.

c) ¿Cuál es el valor de ω en esta tabla de mortalidad?

d) Compruebe que $d_0 + d_1 + \cdots + d_{\omega-1} = \ell_0$.

e) Encuentre ${}_3d_0$, ${}_2q_1$, ${}_3p_1$, ${}_3q_2$.

135. Para edades $x = 0, 1, \ldots$ y tiempos $n, m = 1, 2, \ldots$, demuestre las siguientes identidades:

a) ${}_nq_x = 1 - {}_np_x$.

b) $\ell_x = \ell_0 \cdot p_0 \cdot p_1 \cdots p_{x-1}, \quad x \geqslant 1$.

c) ${}_np_x = (1 - q_x)(1 - q_{x+1}) \cdots (1 - q_{x+n-1})$.

d) ${}_{n|m}q_x = {}_np_x \cdot {}_mq_{x+n}$.

e) ${}_{n|m}q_x = {}_np_x - {}_{n+m}p_x$.

136. La probabilidad de muerte entre las edades x y $x + 1$ se denota por el símbolo ${}_x|q_0$.

a) Encuentre una expresión para ${}_x|q_0$ en términos de $S(x)$ y ℓ_x.

b) Demuestre que $\displaystyle\sum_{x=0}^{\omega-1} {}_x|q_0 = 1$.

137. Un tiempo de vida tiene función de supervivencia $S(x) = (c-x)/(c+x)$, para $0 \leqslant x \leqslant c$, en donde $c > 0$ es una constante. Suponga que se construye una tabla de mortalidad para las edades $x = 0, 1, 2, \ldots$ con $\ell_0 = 100,000$, en donde $\ell_{35} = 44,000$.

a) Encuentre el valor de c.

b) Grafique $S(x)$.

c) Encuentre el valor de ω en la tabla.

d) Encuentre la probabilidad de que una persona alcance la edad 60.

e) Encuentre la probabilidad de que un niño de 10 años fallezca entre las edades 30 y 45.

138. Suponga que se conocen las funciones básicas $F(x)$, $S(x)$ y $\lambda(x)$, para $x = 0, 1, \ldots, \omega$, de un modelo tabular de mortalidad. Demuestre las fórmulas que aparecen abajo, las cuales expresan ℓ_x en términos de las funciones indicadas y del valor inicial ℓ_0.

   a) $\ell_x = \ell_0\,(1 - F(x))$.

   b) $\ell_x = \ell_0\, S(x)$.

   c) $\ell_x = \ell_0 \prod_{u=1}^{x-1}(1 - \lambda(u))$.

   Nota: La expresión de ℓ_x en términos de $R(x)$ es más complicada y se ha omitido de esta lista.

139. Demuestre que la función $R(x)$ en el modelo tabular satisface:

   a) $R(x) = \dfrac{1}{\ell_x}[(x+1)\ell_x + \sum_{u=x+1}^{\omega} \ell_u] - x, \quad x = 0, 1, \ldots, \omega.$

   b) $R(0) = \sum_{u=0}^{\omega} S(u) = E(X)$.

   c) $R(\omega) = 1$.

El método actuarial

140. Elabore una hoja de cálculo en donde se reproduzcan los resultados de la Tabla 4.4, correspondientes a los datos de supervivencia del Ejemplo 4.1.

141. Un conjunto de datos de supervivencia con censura por la derecha se encuentran agrupados como se muestra en la tabla de abajo. Encuentre las cantidades faltantes para estimar la función de supervivencia por el método actuarial. Elabore una gráfica de $\hat{S}(x)$.

j	a_j	Intervalo	n_j	d_j	w_j	$1-d_j/n'_j$	$\hat{S}(a_j)$
1	2	$(0,2]$	50	10	2	_____	1
2	4	$(2,4]$	38	12	1	_____	_____
3	5	$(4,5]$	25	8	0	_____	_____
4	7	$(5,7]$	17	5	4	_____	_____
5	8	$(7,8]$	8	6	2	_____	_____
–	–	$(8,\infty)$	0	–	–	–	0

Tabla 4.12

142. Sea $(a_{j-1}, a_j]$ un intervalo de análisis para el estimador de la función de supervivencia por el método actuarial, en donde el número de individuos que ingresan con vida a dicho intervalo es $n_j > 0$. El estimador por el método actuarial $\hat{S}(x)$ se ha definido usando las probabilidades de supervivencia $\hat{p}_j = 1-d_j/n'_j$. Denote por $\hat{S}_0(x)$ al estimador usando $\hat{p}_j = 1 - d_j/n_j$. Compruebe que:

a) $n'_j \leqslant n_j$.

b) $1 - d_j/n'_j \leqslant 1 - d_j/n_j$.

c) $\hat{S}(x) \leqslant \hat{S}_0(x), \quad x \geqslant 0$.

Esto significa que las probabilidades de supervivencia son menores cuando se considera el número efectivo en riesgo n'_j.

143. Complete las entradas faltantes de la siguiente tabla para estimar la función de supervivencia por el método actuarial. La unidad de tiempo es 1 año. El tiempo de vida en estudio es el tiempo que transcurre entre el final del tratamiento de una persona enferma y su eventual fallecimiento por la enfermedad. Elabore una gráfica de $\hat{S}(x)$ y responda los incisos que aparecen abajo.

j	a_j	$(a_{j-1}, a_j]$	n_j	d_j	w_j	n_j	n'_j	$\hat{p}_j$	$\hat{S}(a_j)$
1	1	$(0,1]$	913	312	96	___	___	___	1
2	2	$(1,2]$	505	96	74	___	___	___	___
3	3	$(2,3]$	335	45	62	___	___	___	___
4	4	$(3,4]$	228	29	30	___	___	___	___
5	5	$(4,5]$	169	25	20	___	___	___	___
6	6	$(5,6]$	124	18	15	___	___	___	___
7	7	$(6,7]$	91	20	18	___	___	___	___
8	8	$(7,8]$	53	18	10	___	___	___	___
9	9	$(8,9]$	25	20	5	___	___	___	___
–	–	$(9,\infty]$	0	–	–	–	–	–	0

Tabla 4.13

a) Estime la probabilidad de que un paciente sobreviva más de 5 años después de su tratamiento.

b) Para un paciente que ha sobrevivido 3 años después de su tratamiento, ¿cuál es la probabilidad de que sobreviva 1 año adicional? Es decir, estime $P(X > 4 \,|\, X > 3)$.

c) Un paciente acaba de terminar su tratamiento y pregunta por el número de años que le quedan de vida. ¿Qué le respondería?

Interpolación lineal en tablas de mortalidad

144. Demuestre las fórmulas sobre la interpolación lineal que aparecen en la Proposición 4.4 en la página 156.

145. Demuestre las fórmulas de las funciones básicas para el tiempo de vida continuo X inducido por la interpolación lineal que aparecen en la Proposición 4.5 en la página 157.

Interpolación exponencial en tablas de mortalidad

146. Demuestre los resultados sobre la interpolación exponencial que aparecen en la Proposición 4.6 en la página 158.

147. *Convexidad.*
Considere la función $\varphi(s) = \ell_x(\ell_{x+1}/\ell_x)^s$, para $0 \leqslant s \leqslant 1$, que define la interpolación exponencial del valor ℓ_x al valor ℓ_{x+1}. Use el criterio de la segunda derivada para demostrar que $\varphi(s)$ es una función convexa.

Interpolación hiperbólica en tablas de mortalidad

148. Demuestre los resultados sobre la interpolación hiperbólica que aparecen en la Proposición 4.7 en la página 159.

149. *Convexidad.*
Considere la función $\varphi(s) = \left(\frac{1}{\ell_x} + s \cdot \left[\frac{1}{\ell_{x+1}} - \frac{1}{\ell_x}\right]\right)^{-1}$, para $0 \leqslant s \leqslant 1$, que define la interpolación hiperbólica del valor ℓ_x al valor ℓ_{x+1}. Use el criterio de la segunda derivada para demostrar que $\varphi(s)$ es una función convexa.

Función de supervivencia empírica

150. Encuentre y grafique la función de supervivencia empírica para cada uno de los siguientes conjuntos de observaciones de un tiempo de vida. Recuerde que no se considera aquí el fenómeno de la censura, todas las observaciones son completas. En cada caso calcule, además, el tiempo promedio de vida.

    a) $x_1 = 7$, $x_2 = 7$, $x_3 = 8$, $x_4 = 8$.

    b) $x_1 = 10$, $x_2 = 11$, $x_3 = 13$, $x_4 = 9$.

    c) $x_1 = 2$, $x_2 = 4$, $x_3 = 2$, $x_4 = 3$, $x_5 = 3$.

    d) $x_1 = 10$, $x_2 = 17$, $x_3 = 12$, $x_4 = 15$, $x_5 = 13$, $x_6 = 10$, $x_7 = 15$.

151. Sea $x_1, \ldots, x_n$ una colección de observaciones sin censura de una variable aleatoria X.

    a) Demuestre que la función $F_n(x)$ definida en (4.8) es una función de distribución.

    b) Demuestre que la función $S_n(x)$ definida en (4.9) es una función de supervivencia.

152. *Propiedades de la función de distribución empírica.*
Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable aleatoria X con función de distribución desconocida $F(x)$. Sea $\hat{F}_n(x)$ la función de distribución empírica definida en (4.10). Demuestre que:

a) $E(\hat{F}_n(x)) = F(x)$. (Insesgamiento)

b) $\text{Var}(\hat{F}_n(x)) = \dfrac{1}{n} F(x)[1 - F(x)]$.

c) $\lim\limits_{n\to\infty} \hat{F}_n(x) = F(x)$.

153. *Propiedades de la función de supervivencia empírica.*
Sea $X_1, \ldots, X_n$ una muestra aleatoria de un tiempo de vida X con función de supervivencia desconocida $S(x)$. Sea $\hat{S}_n(x)$ la función de supervivencia empírica definida en (4.11). Demuestre que, para cada $x \geqslant 0$,

a) $E(\hat{S}_n(x)) = S(x)$. (Insesgamiento)

b) $\text{Var}(\hat{S}_n(x)) = \dfrac{1}{n} S(x)[1 - S(x)]$.

c) $\lim\limits_{n\to\infty} \hat{S}_n(x) = S(x)$.

Las primeras dos propiedades fueron demostradas en el texto, se pide aquí hacer la demostración con detalle.

154. Utilice `R` para obtener la gráfica de la función de distribución empírica y la función de supervivencia empírica de los siguientes conjuntos de datos completos:

a) $-2,\ -1,\ 1,\ 2$.

b) $-4,\ -3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3,\ 4$.

c) $5,\ 2,\ 4,\ 2,\ 3,\ 3,\ 3$.

El estimador de Kaplan-Meier

155. Considere el siguiente conjunto de $n = 10$ observaciones distintas de un tiempo de vida:

$$2, 3+, 5+, 6, 7, 8, 9+, 10, 11+, 15.$$

a) Elabore una tabla similar a la Tabla 4.5 y encuentre el estimador de Kaplan-Meier $\hat{S}(x)$ para la función de supervivencia desconocida $S(x)$.

b) Grafique con precisión $\hat{S}(x)$ para $0 \leqslant x \leqslant x_{(n)}$.

c) ¿Cambia el estimador de Kaplan-Meier si en lugar del dato 15 se tiene el dato 15+?

156. Considere el siguiente conjunto de $n = 6$ observaciones distintas de un tiempo de vida:

$$3, 4+, 6, 7+, 8, 9+.$$

a) Elabore una tabla similar a la Tabla 4.5 y encuentre el estimador de Kaplan-Meier $\hat{S}(x)$ para la función de supervivencia desconocida $S(x)$.

b) Grafique con precisión $\hat{S}(x)$ para $x \geqslant 0$.

157. Se registró el nacimiento de 7 ratones con una rara enfermedad, los cuales murieron en los días $1, 3, 5, 6, 7$ y 8. Observe que no hubo fallecimientos múltiples.

a) Elabore una tabla similar a la Tabla 4.5 y encuentre el estimador de Kaplan-Meier $\hat{S}(x)$ para la función de supervivencia desconocida $S(x)$.

b) Grafique con precisión $\hat{S}(x)$ para $x \geqslant 0$.

c) Estime $S(3)$ y $S(4)$ e interprete estas cantidades.

158. Se mantuvo bajo observación a un conjunto de $n = 10$ individuos bajo un esquema de censura por la derecha tipo II. Los tiempos de vida observados fueron los siguientes:

$$10, 11, 11, 11.5, 12, 12, 15,$$

en donde al tiempo 15 se censuraron al resto de los individuos.

a) Elabore una tabla similar a la Tabla 4.6 y encuentre el estimador de Kaplan-Meier $\hat{S}(x)$ para la función de supervivencia desconocida $S(x)$.

b) Grafique con precisión $\hat{S}(x)$ para $x \geqslant 0$.

159. Elabore una tabla similar a la Tabla 4.6 y encuentre y grafique el estimador de Kaplan-Meier $\hat{S}(x)$ para la función de supervivencia desconocida $S(x)$ de $n = 12$ individuos cuyos tiempos de vida son:

$$2, 3, 3, 4, 5, 7, 4+, 6+, 8+, 9+, 10+, 10+.$$

160. Suponga que en un estudio clínico se observaron fallecimientos de ciertos pacientes en los años: $1.1, 2.5, 3.2, 4.0, 5.8$ y 7.9 . Se tuvieron, además, observaciones censuradas en los años: $2.5, 3.2, 6.0$ y 7.9 .

    a) Elabore una tabla similar a la Tabla 4.6 y encuentre el estimador de Kaplan-Meier $\hat{S}(x)$ para la función de supervivencia desconocida $S(x)$.

    b) Grafique con precisión $\hat{S}(x)$ para $x \geqslant 0$.

161. Debido a la sospecha de que un cierto componente mecánico instalado en automóviles sufría de una determinada falla, se mantuvo en observación a un conjunto de 50 automóviles hasta que el componente presentaba la falla. Los datos registrados (medido en años) fueron los siguientes:

$$0.5, 0.8+, 1.2, 1.5, 1.6, 1.6+, 2, 2+, 2.1, 2.3, 2.5+, 2.7, 2.8, 3, 3.$$

    Al término del tercer año se decidió concluir con el estudio y la información del resto de los automóviles fue censurada. Algunos automóviles fueron censurados dentro de los primeros tres años y ello se debió a que tuvieron un accidente mayor u otra descompostura seria que imposibilitó darle seguimiento a la vida del componente mecánico.

    a) Elabore una tabla similar a la Tabla 4.6 y encuentre el estimador de Kaplan-Meier $\hat{S}(x)$ para la función de supervivencia desconocida $S(x)$.

    b) Grafique con precisión $\hat{S}(x)$ para $x \geqslant 0$.

162. *Estimador de Kaplan-Meier y la función de supervivencia empírica.* Demuestre que el estimador de Kaplan-Meier (4.15) se reduce la función de supervivencia empírica (4.9) en el caso cuando los datos de supervivencia no presentan censura.

163. Utilice la función `survfit()` de `R` para obtener el estimador de Kaplan-Meier para los siguientes conjuntos de datos de supervivencia:

   a) 24, 17, 18, 10+, 25, 10, 14+, 13 .

   b) 4.5, 6.4, 1.5, 13.1, 0.3+, 3.5, 12.1, 3, 19.2+, 13.2, 4.8+ .

   c) 12.7, 14.6+, 11.8+, 9.4, 20.6, 19.2+, 19.5, 44.7, 40.5, 8.4+, 3.2 .

El estimador de Nelson-Aalen

164. Se registró el nacimiento de 5 ratones pero debido al mal estado de salud de la madre, éstos murieron en los días $2, 3, 6, 9$ y 12. Encuentre el estimador de Nelson-Aalen y estime el tiempo de vida promedio.

165. Sean $\hat{S}_{KM}(x)$ y $\hat{S}_{NA}(x)$ los estimadores de Kaplan-Meier y Nelson-Aalen, respectivamente, para una función de supervivencia desconocida $S(x)$. Demuestre la desigualdad que aparece abajo. Esta desigualdad es estricta cuando ambas cantidades se encuentran en el intervalo abierto $(0, 1)$.

$$\hat{S}_{KM}(x) \leqslant \hat{S}_{NA}(x), \quad \text{para} \;\; 0 \leqslant x \leqslant x'_{(n)}.$$

166. Las cantidades que aparecen abajo son los tiempos de vida, en semanas, que fueron registrados de $n = 16$ individuos. Como antes, el signo "+" indica censura por la derecha.

$$4, 5, 5, 7, 10, 10, 12, 13, 14, 15, 20, 7+, 8+, 14+, 14+, 15+$$

   a) Encuentre el estimador de Kaplan-Meier para la función de supervivencia.

   b) Encuentre el estimador de Nelson-Aalen para la función de supervivencia.

   c) En una misma gráfica dibuje ambas funciones de supervivencia estimadas.

   d) Calcule el tiempo promedio de vida usando cada una de estas funciones de supervivencia estimadas.

   e) ¿Cuántas semanas en promedio le restan por vivir a una persona de edad $x = 5, 10, 15$ semanas? Use ambas funciones de supervivencia estimadas para responder a esta pregunta.

Prueba log-rank para comparar poblaciones

167. Considere los datos de supervivencia de dos grupos:

$$\begin{aligned} \text{Grupo } 0 &= \{\ 2.5,\ 7.7+,\ 10,\ 12,\ 13.2+,\ 16.5\ \}, \\ \text{Grupo } 1 &= \{\ 4,\ 8.7,\ 10,\ 11+,\ 19+,\ 24\}. \end{aligned}$$

Determine los intervalos entre fallecimientos sucesivos $(x'_{(j-1)}, x'_{(j)}]$ y, para cada intervalo j, elabore una tabla como la que aparece abajo siguiendo la notación usada en este trabajo.

Tiempo $x'_{(j)}$

	Fallecidos	Sobrevivientes	
Grupo 0	d_{0j}	$n_{0j} - d_{0j}$	n_{0j}
Grupo 1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
	d_j	$n_j - d_j$	n_j

Para cada tabla calcule también la esperanza e_j y la varianza v_j. Con la información de todas las tablas calcule las cantidades globales O, E y V. Finalmente calcule la estadística de la prueba y determine si existe diferencia en la experiencia de mortalidad de los dos grupos mediante la prueba log-rank. Use el *p-value* y considere un nivel de significancia de $\alpha = 0.05$.

168. Considere los siguientes dos conjuntos de datos de supervivencia:

$$\begin{aligned} x^0 &= \{\ 0.5,\ 0.7+,\ 1.5,\ 2,\ 3.2,\ 6.5,\ 7\ \}, \\ x^1 &= \{\ 1,\ 1.7,\ 1.8,\ 1.8+,\ 1.9+,\ 2,\ 2.4,\ 5,\ 6\ \}. \end{aligned}$$

Observe que los tiempos se encuentran ordenados de menor a mayor y que en ellos aparecen varios datos censurados. Nos interesa llevar a cabo la prueba log-rank para determinar si existen diferencia en las funciones de supervivencia de ambos grupos.

a) Considerando únicamente los tiempos de fallecimientos de ambos grupos determine los subintervalos de análisis.

b) Elabore un arreglo como el que aparece en la Tabla 4.14 completando todas las entradas indicadas.

c) Encuentre el valor z del estadístico de la prueba.

d) Encuentre el *p-value* y con un nivel de significancia de $\alpha = 0.05$ determine si se rechaza, o no se rechaza, la hipótesis nula.

j	Intervalo	n_j	n_{0j}	n_{1j}	d_j	d_{0j}	d_{1j}	c_j	c_{0j}	c_{1j}	e_{0j}	v_{0j}
1	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	–	–	–	–	–	–	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	–	–	–	–	–	–	–	–	–	–	–	–

Tabla 4.14

169. Lleve a cabo una prueba log-rank para los datos del ejercicio anterior pero ahora usando `R` como en los Ejemplos 4.6 y 4.7.

170. Lleve a cabo una prueba de hipótesis log-rank usando `R` para determinar si existe diferencia en las experiencias de mortalidad de los siguientes pares de conjuntos de datos de supervivencia. Use el *p-value* y un nivel de significancia $\alpha = 0.05$ para rechazar, o no rechazar, la hipótesis nula. Utilice, además, la función `survfit` para obtener las curvas de supervivencia de ambos grupos en un mismo plano cartesiano.

a) Grupo 0 $= \{6, 1, 2+, 6, 9, 1+, 2+, 2, 4, 10\}$,
Grupo 1 $= \{7, 3, 18, 2+, 10, 21, 18, 25, 6+, 26\}$.

b) Grupo 0 $= \{3, 1, 4, 4, 5, 4+, 7\}$,
Grupo 1 $= \{2+, 5, 1, 6, 5, 6, 8\}$.

c) Grupo 0 $= \{2.6, 5.1, 2.2, 1.2+, 1.2, 5.6, 2.9, 9.9, 3.3, 4.3, 0.8+, 0.9+\}$,
Grupo 1 $= \{4.8, 0.9+, 2.5, 2.7, 5.2, 9.8, 0.4+, 3.0, 2.2, 4.1, 3.8\}$.

Capítulo 5

Modelos con covariables

En este último capítulo se exponen brevemente algunos modelos de supervivencia que incorporan covariables. En particular, se provee de una introducción al modelo de riesgos proporcionales de Cox y de la estimación de sus parámetros. Veremos primero el concepto de covariable.

5.1. Covariables

Hasta ahora hemos considerado que las funciones de supervivencia $S(x)$ son funciones del tiempo x, y hemos también supuesto, implícitamente, que los conjuntos de individuos son homogéneos en el sentido de que sus tiempos de vida siguen una misma distribución de probabilidad dada por la función desconocida $S(x)$.

Con frecuencia los elementos de una población presentan características o condiciones que hacen que sus tiempos de vida sean diferentes unos de otros. Estas características pueden ser propias de los individuos como: la edad, el peso, el consumo de cigarro o alcohol, la falta de ejercicio físico, el nivel económico, los hábitos alimenticios, etc. Otras variables pueden ser exógenas o del medio ambiente como: el lugar de residencia, la clínica u hospital de atención médica, el tratamiento particular al que se somete un paciente, etc. Esto lleva a la siguiente definición.

Definición 5.1 *A las variables que presumiblemente pudieran afectar el tiempo de vida de un individuo se les llama covariables.*

Según el área de estudio o aplicación, a las covariables también se les llama variables concomitantes o variables prognósticas. El objetivo es incorporar covariables a la función de supervivencia o, equivalentemente, a alguna de sus funciones básicas asociadas.

Separación vs inclusión

Se pueden considerar funciones de supervivencia para cada uno de los valores de las covariables y, en este caso, se dice que las covariables han sido incorporadas por separación. Se tendría así, un función de supervivencia distinta para cada individuo, o grupo de individuos, con los mismos valores de las covariables. Esta es la situación que, implícitamente, hemos supuesto antes cuando todos los individuos tienen, en cierta medida, características homogéneas. En este caso, cada función de supervivencia depende sólo del tiempo x.

En contraparte, se pueden considerar modelos en donde las funciones de supervivencia, o cualquier otra de las funciones básicas equivalentes, dependen directamente del tiempo x y también de algunas covariables. En este caso, se dice que las covariables han sido incorporadas por inclusión. Así, se tendrá una sóla función de supervivencia, la cual es función del tiempo x y de las covariables. Adoptaremos este tipo de modelos en lo que resta del capítulo.

Vector de covariables

Agruparemos las covariables de interés, o de las que se cuente con información, en un vector de dimensión $s \geqslant 1$, de la siguiente forma

$$\underline{z} = (z_1, z_2, \ldots, z_s),$$

en donde cada entrada es una covariable que representa una posible característica o condición que tiene efectos sobre los tiempos de vida de los

individuos. De esta manera, en lugar de escribir $f(x)$, $S(x)$ ó $\lambda(x)$, por ejemplo, escribiremos $f(x, \underline{z})$, $S(x, \underline{z})$ y $\lambda(x, \underline{z})$, respectivamente. En ocasiones será conveniente escribir explícitamente cada una de las covariables como en $\lambda(x, z_1, \ldots, z_s)$.

Ejemplo 5.1 *Para una población humana consideremos el vector de covariables* $\underline{z} = (z_1, z_2, z_3)$*, en donde se define*

$$z_1 := \begin{cases} 1 & \textit{Nivel económico bajo,} \\ 2 & \textit{Nivel económico medio,} \\ 3 & \textit{Nivel económico alto.} \end{cases}$$

$$z_2 := \begin{cases} 1 & \textit{Femenino,} \\ 2 & \textit{Masculino.} \end{cases}$$

$$z_3 := \begin{cases} 1 & \textit{No fumador,} \\ 2 & \textit{Fumador.} \end{cases}$$

Si un individuo o grupo i es de nivel económico bajo, es mujer y fuma, su función de riesgo se escribe $\lambda(x, \underline{z}_i)$ *con* $\underline{z}_i = (1, 1, 2)$*. De este modo, cada individuo o grupo homogéneo de individuos tiene su propia función de supervivencia, o función de riesgo, según su clasificación a través del vector de covariables* $\underline{z}$*.*

Nota: Las variables z_1, z_2 y z_3 definidas en el ejemplo anterior son categóricas, de modo que, al utilizar algún programa de cómputo, se debe tener cuidado en su declaración pues la forma de su especificación determina la manera en la que estas variables son tratadas por el software utilizado. ∎

Se pueden considerar también modelos dinámicos en donde el vector de covariables $\underline{z}$ es función del tiempo, es decir, las covariables pueden cambiar de valor al paso del tiempo para un individuo o grupo de individuos. Estos modelos dinámicos pueden ser más cercanos a algunas realidades, sin embargo, no los consideraremos en este trabajo.

A continuación veremos algunas maneras explícitas en las que las covariables pueden ser incorporadas a la distribución de un tiempo de vida.

5.2. Modelos para incluir covariables

Sea $\lambda(x)$ una función de riesgo. Deseamos incorporar a esta función los efectos de un vector de covariables $\underline{z} = (z_1, \ldots, z_s)$. Para enfatizar que es la función de riesgo antes de la modificación, le llamaremos función de riesgo base o subyacente, y en ocasiones la escribiremos como $\lambda_0(x)$. Esta será la función de riesgo de todos los individuos en estudio antes de considerar las covariables. A la función de riesgo con el vector de covariables $\underline{z}$ incorporado la escribiremos como $\lambda(x, \underline{z})$.

Modelo aditivo

En este modelo se propone la siguiente forma de la función de riesgo modificada

$$\lambda(x, \underline{z}) = \lambda_0(x) + \sum_{j=1}^{s} h_j(x)\, g(z_j), \tag{5.1}$$

en donde cada sumando $h_j(x)\, g(z_j) \geqslant 0$ es una función que representa el riesgo adicional al tiempo x debido a la covariable z_j. Puede comprobarse que $\lambda(x, \underline{z})$ dada en (5.1) es una función de riesgo. Se puede encontrar mayor información de este modelo en Cox D. R. y Oakes D. [16], N. E. Breslow y N. E. Day [7] ó en D. C. Thomas [50].

Ejemplo 5.2 *Para el modelo* (5.1) *se puede tomar la función constante* $h_j(x) = a_j \geqslant 0$ *y y la función identidad* $g(z_j) = z_j \geqslant 0$. *Esto produce la función de riesgo* $\lambda(x, \underline{z})$ *que aparece abajo. Como se ha supuesto que* $a_j z_j \geqslant 0$, *las covariables provocan una propensión a fallecer pues incrementan la función de riesgo base.*

$$\lambda(x, \underline{z}) = \lambda_0(x) + \sum_{j=1}^{s} a_j\, z_j.$$

■

Ejemplo 5.3 (Modelo lineal-exponencial) *Continuando con el ejemplo anterior, si además se toma a la función de riesgo base* $\lambda_0(x)$ *como una*

constante $a_0 > 0$, lo cual corresponde a tiempos de vida exponenciales de parámetro a_0, y si se define a la covariable auxiliar constante $z_0 = 1$, entonces se puede escribir

$$\lambda(x, \underline{z}) = \sum_{j=0}^{s} a_j \, z_j. \tag{5.2}$$

Debido a la linealidad de esta función respecto de las covariables y a la hipótesis de distribución exponencial de la función de riesgo base, al modelo definido por (5.2) *se le llama modelo lineal-exponencial. Observe que* (5.2) *representa una función de riesgo constante (no dependiente del tiempo x) y, por lo tanto, es un tiempo de vida exponencial.* ▪

Modelo multiplicativo

En este modelo se propone la siguiente forma de la función de riesgo modificada

$$\lambda(x, \underline{z}) = \lambda_0(x) \prod_{j=1}^{s} h(x, z_j), \tag{5.3}$$

en donde $h(x, z_j) \geqslant 0$ es una función que modifica de manera multiplicativa la función de riesgo base al tiempo x y debido a la covariable z_j. Puede comprobarse que (5.3) es una función de riesgo. El modelo de Cox es un modelo multiplicativo importante en donde la función h es la función exponencial $\exp(a_1 z_1 + \cdots + a_s z_s)$ para ciertas constantes $a_1, \ldots, a_s$ y en donde, en su versión simple, no aparece el tiempo x. Estudiaremos este modelo en las siguientes secciones. Puede consultarse mayor información de (5.3) en el libro de E. T. Lee y J. W. Wang [37].

Modelo de vida acelerada

En este caso, un tiempo de vida X se modifica directamente a través del cociente X/a, en donde $a > 0$ es una constante. El cociente X/a representa también un tiempo de vida pero la velocidad con la que transcurre el tiempo se modifica de la siguiente manera: si $0 < a < 1$, los tiempos de vida se alargan, y si $a > 1$, los tiempos de vida se reducen o, en otras palabras, el

tiempo de vida se acelera. En ambos casos se dice que el cociente representa un tiempo de vida acelerado. No es difícil encontrar las funciones básicas asociadas al tiempo de vida X/a en términos de las funciones de X. En efecto, si $Y = X/a$ entonces se puede comprobar que:

a) $F_Y(y) = F_X(ay), \quad y \geqslant 0.$

b) $f_Y(y) = a\, f_X(ay). \quad y \geqslant 0.$

c) $S_Y(y) = S_X(ay), \quad y \geqslant 0.$

d) $\lambda_Y(y) = a\, \lambda_X(ay) \quad$ para $\; y \geqslant 0$ tal que $S_X(ay) > 0.$

e) $R_Y(y) = \dfrac{1}{a\, S_X(ay)} \displaystyle\int_{ay}^{\infty} S_X(u)\, du \quad$ para $\; y \geqslant 0$ tal que $S_X(ay) > 0.$

Por claridad, se han indicado como subíndices de estas funciones el tiempo de vida de referencia, es decir, X ó Y. El modelo de vida acelerada puede expresarse también por la expresión equivalente $Y = aX$, en donde el tiempo de vida se reduce cuando $0 < a < 1$.

Más generalmente, si $\underline{z} = (z_1, \ldots, z_s)$ es un vector de covariables, una función positiva $g(\underline{z})$ puede acelerar un tiempo de vida X considerando el cociente $X/g(\underline{z})$. Las fórmulas anteriores permanecen válidas substituyendo a la constante a for $g(\underline{z})$. En particular, cuando a un tiempo de vida Weibull se le acelera por una función positiva $g(\underline{z})$, se obtiene el modelo de riesgos proporcionales de Cox (definido en la siguiente sección). Se puede comprobar que la distribución Weibull es la única distribución continua que cumple esta propiedad. Véase el Ejercicio 181.

5.3. El modelo de Cox[1]

En este modelo multiplicativo se propone que la función de riesgo base $\lambda_0(x)$ dependa de manera arbitraria del tiempo x y, por otro lado, la función de riesgo modificada dependa de forma paramétrica y lineal de un vector de covariables de la siguiente manera.

[1] David Roxbee Cox (1924–2022) estadístico inglés.

Definición 5.2 *Se le llama modelo de Cox al modelo multiplicativo* (5.3) *cuando se toma* $h(x, z_j) := \exp(a_j\, z_j)$, *en donde* $a_1, \ldots, a_s$ *son constantes y* $\underline{z} = (z_1, \ldots, z_s)$ *es un vector de covariables. Es decir,*

$$\lambda(x, \underline{z}) := \lambda_0(x)\, \exp\{ \sum_{j=1}^{s} a_j\, z_j \}. \tag{5.4}$$

Al modelo de Cox se le llama también modelo de riesgos proporcionales. Se le puede encontrar en la literatura también con los términos en inglés *PH model ó proportional hazzard model*, pues para cada valor del vector de covariables $\underline{z}$, el producto (5.4) representa una proporción de la función de riesgo base $\lambda_0(x)$.

Para la función de riesgo base desconocida $\lambda_0(x)$ puede adoptarse un modelo de una familia paramétrica y, en este caso, al modelo de Cox se le llama paramétrico. En cambio, si se adopta una perspectiva no paramétrica para $\lambda_0(x)$, al modelo de Cox se le llama semiparamétrico, pues permanece la parte paramétrica dada por la regresión lineal de las covariables. Puede consultarse una exposición del modelo de Cox directamente de su creador en el libro de D. R Cox y D. Oakes [16], o en el libro de E. T. Lee y J. W. Wang [37], por ejemplo.

El coeficiente a_j representa una medida del impacto de la covariable z_j en los tiempos de vida. Observemos que la función $\lambda_0(x)$ y los coeficientes $a_1, \ldots, a_s$ son desconocidos. Uno de los problemas es estimar estos parámetros a partir de un conjunto de observaciones.

Consideraremos que el vector de ceros $\underline{0} = (0, \ldots, 0)$ es un posible valor del vector de covariables $\underline{z} = (z_1, \ldots, z_s)$ en el modelo de Cox. Esto es conveniente pues, cuando $\underline{z} = \underline{0}$ en (5.4), se obtiene la función de riesgo base, la cual escribiremos también como $\lambda(x, \underline{0})$. De esta manera, el modelo de Cox (5.4) se puede escribir como

$$\lambda(x, \underline{z}) := \lambda_0(x)\, \rho(\underline{z}), \tag{5.5}$$

en donde la función $\rho(\underline{z}) = \exp\{\, a_1 z_1 + \cdots + a_s z_s \,\}$ satisface las siguientes dos propiedades:

a) $\rho(\underline{z}) \geqslant 0$.

b) $\rho(\underline{0}) = 1$.

Como la función $\rho(\underline{z})$ no depende del tiempo x y actúa como un factor positivo sobre la función de riesgo base, es claro que $\lambda(x, \underline{z})$ es también una función de riesgo. En la siguiente sección encontraremos sus funciones básicas asociadas.

Cuando no hay covariables en el modelo (5.4), $\rho(\underline{z}) = 1$ y la función de riesgo $\lambda(x, \underline{z})$ se reduce a $\lambda_0(x)$. Por otro lado, observe que el modelo de Cox puede escribirse como una regresión lineal múltiple sobre las covariables al considerar el logaritmo del cociente $\lambda(x, \underline{z})/\lambda_0(x)$, es decir,

$$\log \frac{\lambda(x, \underline{z})}{\lambda_0(x)} = a_1 z_1 + \cdots + a_s z_s.$$

Ejemplo 5.4 *Sunpoga que el vector de covariabes es* $\underline{z} = (z_1, z_2)$*, en donde* z_1 *distingue dos tipos de tratamiento médico y* z_2 *representa el sexo del paciente, es decir,*

$$z_1 = \begin{cases} 1 & \textit{Tratamiento } 1, \\ 2 & \textit{Tratamiento } 2. \end{cases}$$

$$z_2 = \begin{cases} 1 & \textit{Masculino}, \\ 2 & \textit{Femenino}. \end{cases}$$

Entonces $\rho(\underline{z}) = \exp\{\, a_1 z_1 + a_2 z_2 \,\}$ *y el modelo de Cox contempla las siguientes cuatro distintas funciones de riesgo*

$$\lambda(x, \underline{z}) = \begin{cases} \lambda_0(x)\, e^{a_1 + a_2} & \textit{Tratamiento } 1 \textit{ a hombre}, \\ \lambda_0(x)\, e^{a_1 + 2a_2} & \textit{Tratamiento } 1 \textit{ a mujer}, \\ \lambda_0(x)\, e^{2a_1 + a_2} & \textit{Tratamiento } 2 \textit{ a hombre}, \\ \lambda_0(x)\, e^{2a_1 + 2a_2} & \textit{Tratamiento } 2 \textit{ a mujer}. \end{cases}$$

■

Ejemplo 5.5 (Modelo log-lineal) *Si la función de riesgo base $\lambda_0(x)$ es constante igual a $\exp\{a_0\}$, para alguna constante a_0, es decir, el tiempo de vida base es exponencial, y definiendo la covariable artificial $z_0 = 1$, entonces la función de riesgo modificada es*

$$\lambda(x, \underline{z}) = \exp\{\sum_{j=0}^{s} a_j z_j\}.$$

Como esta función es constante respecto de x, el modelo sigue siendo el de un tiempo de vida exponencial, cuyo parámetro depende de los valores de las covariables z_j y los coeficientes a_j. A este caso particular se le llama modelo log-lineal, pues el logaritmo de $\lambda(x, \underline{z})$ es una función lineal de las covariables.

■

Ejemplo 5.6 *Como se ha mencionado, las covariables pueden depender del tiempo y no es difícil dar un ejemplo de esa situación. La covariable $z_1(x)$ puede representar la evolución en el tiempo de un índice de contaminantes presentes en el aire en una ciudad densamente poblada. Si un paciente está enfermo de asma, nos podría interesar el tiempo que transcurre hasta el siguiente ataque de asma. Considerando sólo a esta covariable, el vector de covariables $\underline{z}(x)$ es $(z_1(x))$ y la función de riesgo según el modelo de Cox se puede escribir como*

$$\lambda(x, \underline{z}(x)) = \lambda_0(x)\, e^{a_1 z_1(x)}, \quad x \geqslant 0,$$

para algún coeficiente a_1. De este modo, las covariables dependientes del tiempo son características del individuo, o de su medio ambiente, que cambian de valor con el tiempo y pueden afectar su tiempo de vida. En este trabajo consideraremos principalmente el caso simple cuando las covariables son constantes en el tiempo. ■

Funciones básicas asociadas en el modelo de Cox

A partir de la forma de la función de riesgo $\lambda(x, \underline{z}) = \lambda_0(x)\, \rho(\underline{z})$ en el modelo de Cox, se pueden encontrar las otras funciones básicas asociadas como se muestra a continuación.

Proposición 5.1 *Sean $S_0(x)$ y $\Lambda_0(x)$ las funciones de supervivencia y de riesgo acumulado asociadas a la función de riesgo base $\lambda_0(x)$ del modelo de Cox (5.5). Entonces las funciones asociadas a la función de riesgo $\lambda(x, \underline{z}) = \lambda_0(x)\,\rho(\underline{z})$ son:*

1. $S(x, \underline{z}) = [S_0(x)]^{\rho(\underline{z})}, \quad x \geqslant 0.$
2. $F(x, \underline{z}) = 1 - [S_0(x)]^{\rho(\underline{z})}, \quad x \geqslant 0.$
3. $f(x, \underline{z}) = \lambda_0(x)\,\rho(\underline{z})\,[S_0(x)]^{\rho(\underline{z})}, \quad x \geqslant 0.$
4. $\Lambda(x, \underline{z}) = \Lambda_0(x)\,\rho(\underline{z}), \quad x \geqslant 0.$

Demostración.

1. Suponiendo el caso continuo, para $x \geqslant 0$,

$$\begin{aligned} S(x, \underline{\mathrm{z}}) &= \exp\{-\int_0^x \lambda_0(u)\,\rho(\underline{\mathrm{z}})\,du\} \\ &= \exp\{-\rho(\underline{\mathrm{z}}) \int_0^x \lambda_0(u)\,du\} \\ &= [\exp\{-\int_0^x \lambda_0(u)\,du\}]^{\rho(\underline{\mathrm{z}})} \\ &= [S_0(x)]^{\rho(\underline{\mathrm{z}})}. \end{aligned}$$

2. Esto es consecuencia inmediata del inciso anterior.

3. Por definición y por el resultado del primer inciso, para $x \geqslant 0$,

$$\begin{aligned} f(x, \underline{\mathrm{z}}) &= \lambda(x, \underline{\mathrm{z}})\,S(x, \underline{\mathrm{z}}) \\ &= \lambda_0(x)\,\rho(\underline{\mathrm{z}})\,[S_0(x)]^{\rho(\underline{\mathrm{z}})}. \end{aligned}$$

   Alternativamente, se puede calcular $f(x, \underline{\mathrm{z}}) = -(d/dx)S(x, \underline{\mathrm{z}})$, obteniendo el mismo resultado.

4. Por definición, para $x \geqslant 0$,

$$\begin{aligned}\Lambda(x, \underline{z}) &= \int_0^x \lambda(u, \underline{z})\, du \\ &= \int_0^x \lambda_0(u)\, \rho(\underline{z})\, du \\ &= \rho(\underline{z}) \int_0^x \lambda_0(u)\, du \\ &= \rho(\underline{z})\, \Lambda_0(x).\end{aligned}$$

■

En la expresión $S(x, \underline{z}) = [S_0(x)]^{\rho(\underline{z})}$ se vuelve transparente la forma en la que las covariables afectan a la función de supervivencia base, suponiendo válido el modelo de Cox. Cuando $0 < S_0(x) < 1$, un valor $\rho(\underline{z}) > 1$ produce un valor de la función de supervivencia más pequeño (mayor mortalidad). Por el contrario, el valor de la función de supervivencia es más grande (mayor supervivencia) para $\rho(\underline{z}) < 1$. Cuando $\rho(\underline{z}) = 1$, la función de supervivencia base no tiene cambio.

Por otro lado, observando los cálculos en la demostración anterior, no es difícil darse cuenta que si las covariables son funciones del tiempo x, las expresiones para las funciones básicas asociadas a $\lambda(x, \underline{z}(x))$ en el modelo de Cox quedan expresadas en términos de integrales.

Ejemplo 5.7 *Cuando el tiempo de vida base en el modelo de Cox es exp*(λ)*, es decir, cuando* $S_0(x) = \exp\{-\lambda x\}$*, se tiene que* $\lambda_0(x) = \lambda$ *y* $\Lambda_0(x) = \lambda x$*. Por lo tanto, las funciones asociadas a la función de riesgo* $\lambda(x, \underline{z}) = \lambda_0(x)\, \rho(\underline{z})$ *son:*

a) $S(x, \underline{z}) = \exp\{-\lambda \rho(\underline{z}) x\}, \quad x \geqslant 0.$

b) $F(x, \underline{z}) = 1 - \exp\{-\lambda \rho(\underline{z}) x\}, \quad x \geqslant 0.$

c) $f(x, \underline{z}) = \lambda\, \rho(\underline{z})\, \exp\{-\lambda \rho(\underline{z}) x\}, \quad x \geqslant 0.$

d) $\Lambda(x, \underline{z}) = \lambda \rho(\underline{z}) x, \quad x \geqslant 0.$

Cualquiera de estas expresiones indica que la distribución del tiempo de vida con covariables $\underline{z}$ es exponencial de parámetro $\lambda\rho(\underline{z})$, como se había señalado antes en el Ejemplo 5.5. ∎

Interpretación de parámetros

A continuación haremos algunas observaciones sobre el modelo de riesgos proporcionales de Cox definido por (5.4), e indicaremos algunas interpretaciones de los parámetros $a_1, \ldots, a_s$.

- Los coeficientes a_j pueden ser positivos, negativos o cero. Cuando $a_j > 0$, la función $z_j \mapsto \exp(a_1\, z_j)$ es creciente. Esto significa que cuando la covariable z_j crece, la proporción $\exp(a_1 z_j)$ también crece y ello incrementa la función de riesgo $\lambda(x, \underline{z})$. Esto se traduce en una mayor propensión a fallecer acortando así el tiempo de vida. Cuando $a_j < 0$, el efecto contrario ocurre, cada vez que z_j crece, se presenta una menor propensión a fallecer, alargando el tiempo de vida. Cuando $a_j = 0$, la covariable z_j no tiene ningún efecto en la función de riesgo.

- Cocientes de riesgo (*Hazard Ratios* HR). A los exponentes de los coeficientes a_j, es decir, a $\exp(a_j)$ se les llama cocientes de riesgo. De manera análoga a lo explicado en el párrafo anterior, estas cantidades también representan una medida del efecto de las covariables sobre los tiempos de vida, pues son la base de exponenciación

$$(\exp\, a_j)^{z_j} = \exp\,(a_j\, z_j).$$

  Puede comprobarse que, cuando $\exp a_j > 1$, la función de riesgo es creciente en la covariable z_j. Cuando $\exp a_j < 1$, el efecto es contrario, la función de riesgo es decreciente en la covariable z_j. Finalmente, cuando $\exp a_j = 1$, esto significa que la covariable z_j no tiene efecto sobre la función de riesgo y ésta permanece constante ante las variaciones de z_j.

- Es evidente que el siguiente cociente no depende del tiempo x,

$$\frac{\lambda(x, \underline{z})}{\lambda_0(x)} = \rho(\underline{z}) = \exp\{\sum_{j=1}^{s} a_j z_j\}.$$

Tomando logaritmo,

$$\log\, \lambda(x, \underline{z}) - \log\, \lambda_0(x) = \sum_{j=1}^{s} a_j z_j.$$

Esto quiere decir que las funciones $x \mapsto \log\, \lambda(x, \underline{z})$ y $x \mapsto \log\, \lambda_0(x)$ son paralelas, es decir, a cada tiempo x, sus gráficas difieren una de la otra en la misma cantidad. Véase la Figura 5.1. En otras palabras, el logaritmo de la función de riesgo con covariables se encuentra siempre a una misma distancia del logaritmo de la función de riesgo base. La distancia de separación está dada por $a_1 z_1 + \cdots + a_s z_s$.

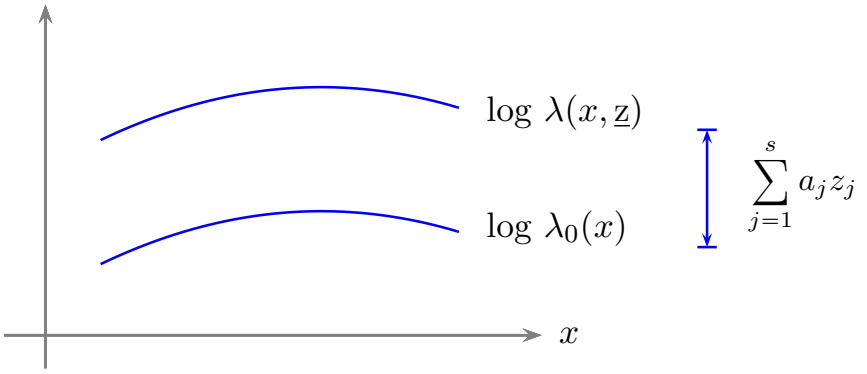

Figura 5.1: La separación entre $\log\, \lambda_0(x)$ y $\log\, \lambda(x, \underline{z})$ es $\log \rho(\underline{z})$.

- Si $\underline{z}_1 = (z_{11}, \ldots, z_{1s})$ y $\underline{z}_2 = (z_{21}, \ldots, z_{2s})$ son dos valores del vector de covariables, entonces el siguiente cociente no depende del tiempo x ni de la función de riesgo base $\lambda_0(x)$,

$$\frac{\lambda(x, \underline{z}_1)}{\lambda(x, \underline{z}_2)} = \exp\{ \sum_{j=1}^{s} a_j (z_{1j} - z_{2j}) \}.$$

- En el caso de una sola covariable, $\underline{z} = (z_1)$, se tiene que

$$\lambda(x, \underline{z}) = \lambda_0(x)\, \exp\{ a_1 z_1 \}.$$

Supongamos ahora que z y $z+1$ son dos posibles valores de la covariable z_1. Estos dos valores distan uno del otro en una unidad. Entonces se cumplen las siguientes identidades:

a) $\log \dfrac{\lambda(x, z+1)}{\lambda_0(x)} = a_1(z+1) = a_1 + a_1 z.$

b) $\log \dfrac{\lambda(x, z)}{\lambda_0(x)} = a_1 z.$

De estas dos igualdades se obtiene que

$$\frac{\lambda(x, z+1)}{\lambda(x, z)} = \exp\{a_1\}.$$

Esto quiere decir que la constante a_1 se puede interpretar como el cambio en el logaritmo del cociente de riesgos por unidad de cambio en la covariable z_1. Más generalmente, para cualquier $r \geqslant 0$ tal que z y $z + r$ son dos posibles valores de z_1, un cálculo similar al anterior muestra que

$$\log \frac{\lambda(x, z+r)}{\lambda(x, z)} = a_1 r.$$

Es decir,

$$\frac{\lambda(x, z+r)}{\lambda(x, z)} = \exp\{a_1 r\}.$$

En palabras, cuando la covariable z_1 se incrementa en r unidades, esto produce un cambio en la función de riesgo $\lambda(x, \underline{z})$ consistente en la multiplicación por el factor $e^{a_1 r}$.

- La interpretación para el caso de una covariable se puede extender sin dificultad al caso cuando $\underline{z}$ cuenta con múltiples covariables. Cuando se incrementa una covariable z_j en $r \geqslant 0$ unidades y se deja a las otras covariables sin cambio, se encuentra que el factor que adquiere la función de riesgo es $e^{a_j r}$, $1 \leqslant j \leqslant s$. En efecto, si $\bar{z}^{j,r} = (z_1, \ldots, z_j + r, \ldots, z_s)$, entonces

$$\begin{aligned} \frac{\lambda(x, \underline{z}^{j,r})}{\lambda(x, \underline{z})} &= \frac{\lambda_0(x)\, \exp\left(a_1 z_1 + \cdots + a_j(z_j + r) + \cdots + a_s z_s\right)}{\lambda_0(x)\, \exp\left(a_1 z_1 + \cdots + a_j z_j + \cdots + a_s z_s\right)} \\ &= \exp\left(a_j r\right). \end{aligned}$$

Al aplicar logaritmo se obtiene

$$\log \lambda(x, \underline{z}^{j,r}) - \log \lambda(x, \underline{z}) = a_j r.$$

Gráficamente se tiene la misma situación que se muestra en la Figura 5.1, excepto que ahora se ha modificado sólo la covariable z_j (añadiéndole la cantidad r) y la distancia entre las curvas es $a_j r$.

5.4. Estimación de coeficientes

En esta sección veremos la manera en la que los coeficientes de regresión $a_1, \ldots, a_s$ que aparecen en el modelo de Cox (5.4) pueden ser estimados por máxima verosimilitud.

Supondremos que los tiempos de vida de n individuos con posible censura por la derecha ya se encuentran ordenados de menor a mayor. Estos tiempos de fallecimiento o censura son: $(x_{(1)}, \delta_1), \ldots, (x_{(n)}, \delta_n)$. Llamaremos individuo i a la persona que fallece o se censura al tiempo $x_{(i)}$, $i = 1, \ldots, n$. Así, el individuo 1 fallece ($\delta_1 = 1$) o se censura ($\delta_1 = 0$) al tiempo $x_{(1)}$, el individuo 2 fallece ($\delta_2 = 1$) o se censura ($\delta_2 = 0$) al tiempo $x_{(2)}$, etc. Por otro lado, cada individuo tiene asociado ciertos valores de las covariables $\underline{z} = (z_1, \ldots, z_s)$ que inciden en su función de riesgo. Definimos esto a continuación.

Notación. Los valores de las covariables para el individuo i se denotan por $\underline{z}_i = (z_{i1}, \ldots, z_{is})$. El primer subíndice identifica al individuo $i \in \{1, \ldots, n\}$ y el segundo subíndice identifica a la covariable.

Observe que varios individuos pueden compartir los mismos valores de las covariables cuando presenten las mismas características y condiciones del riesgo de morir. Como se hizo en el caso del estimador de Kaplan-Meier, también aquí se crea la siguiente partición usando los tiempos de fallecimiento o censura:

$$(0, x_{(1)}],\ (x_{(1)}, x_{(2)}],\ \ldots\ , (x_{(n-1)}, x_{(n)}],$$

en donde, como antes, se define $x_{(0)} := 0$. Esta partición del tiempo en n intervalos tiene sentido cuando no hay fallecimientos o censuras múltiples, es decir, cuando todas las observaciones son distintas.: $0 < x_{(1)} < x_{(2)} <$

$\cdots < x_{(n)}$. Este es el caso que estudiaremos con más detalle a continuación. Antes de ello definamos los siguientes conjuntos.

Definición 5.3 *Para cada $i = 1, \ldots, n$, se define:*

$$\begin{aligned} N_i &= \text{“Conjunto de individuos en riesgo de morir al inicio del intervalo } (x_{(i-1)}, x_{(i)}]\text{.”} \\ D_i &= \text{“Conjunto de individuos que fallecen al tiempo } x_{(i)}\text{.”} \end{aligned}$$

Es claro que $D_i \subseteq N_i$. Observe que un individuo en estudio adquiere la etiqueta o numeración i cuando es el i-ésimo individuo en retirarse, ya sea por fallecimiento o por censura: el individuo 1 se retira al tiempo $x_{(1)}$ y sus covariables son $\underline{z}_1$, el individuo 2 se retira al tiempo $x_{(2)}$ y sus covariables son $\underline{z}_2$, etc.

Caso simple

Supondremos que no hay fallecimientos o censuras múltiples. Así, el conjunto D_i es vacío o sólo contiene al individuo i. Además, como en el tiempo $x_{(i)}$ se descarta al individuo i, ya sea por fallecimiento o por censura, se tiene que $N_i = \{i, i+1, \ldots, n\}$, en donde $\#N_i = n - i + 1$. El resultado de estimación es el siguiente.

Proposición 5.2 (Estimación de coeficientes $a_1, \ldots, a_s$) *Sean $(x_{(1)}, \delta_1), \ldots, (x_{(n)}, \delta_n)$ datos de supervivencia con censura por la derecha, ordenados de menor a mayor, y en donde no se presentan fallecimientos o censuras múltiples, es decir, $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Los estimadores por máxima verosimilitud (parcial) para los coeficientes $a_1, \ldots, a_s$ en el modelo de Cox (5.4) están dados por la solución al sistema de ecuaciones*

$$\sum_{i=1}^{n} \delta_i \left[z_{ik} - \frac{\sum_{j \in N_i} z_{jk}\, \rho(\underline{z}_j)}{\sum_{j \in N_i} \rho(\underline{z}_j)} \right] = 0, \quad k = 1, \ldots, s. \tag{5.6}$$

Demostración. La función de verosimilitud de los datos es

$$\begin{aligned}
L &= \prod_{i=1}^{n} [f(x_{(i)}, \underline{z}_i)]^{\delta_i} [S(x_{(i)}, \underline{z}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} [\lambda(x_{(i)}, \underline{z}_i)\, S(x_{(i)}, \underline{z}_i)]^{\delta_i} [S(x_{(i)}, \underline{z}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} [\lambda(x_{(i)}, \underline{z}_i)]^{\delta_i} S(x_{(i)}, \underline{z}_i) \\
&= \prod_{i=1}^{n} \Big[\frac{\lambda(x_{(i)}, \underline{z}_i)}{\sum_{j \in N_i} \lambda(x_{(i)}, \underline{z}_j)} \Big]^{\delta_i} \Big[\sum_{j \in N_i} \lambda(x_{(i)}, \underline{z}_j) \Big]^{\delta_i} S(x_{(i)}, \underline{z}_i) \\
&= \prod_{i=1}^{n} \Big[\frac{\lambda_0(x_{(i)})\, \rho(\underline{z}_i)}{\sum_{j \in N_i} \lambda_0(x_{(i)})\, \rho(\underline{z}_j)} \Big]^{\delta_i} \Big[\sum_{j \in N_i} \lambda(x_{(i)}, \underline{z}_j) \Big]^{\delta_i} S(x_{(i)}, \underline{z}_i) \\
&= \prod_{i=1}^{n} \Big[\frac{\rho(\underline{z}_i)}{\sum_{j \in N_i} \rho(\underline{z}_j)} \Big]^{\delta_i} \Big[\sum_{j \in N_i} \lambda(x_{(i)}, \underline{z}_j) \Big]^{\delta_i} S(x_{(i)}, \underline{z}_i).
\end{aligned}$$

El método sugiere descartar los dos factores de la derecha en la última expresión y considerar sólo el primer factor, al que se le llama función de verosimilitud parcial de Cox. Es decir,

$$L_{parcial} = \prod_{i=1}^{n} \Big[\frac{\rho(\underline{z}_i)}{\sum_{j \in N_i} \rho(\underline{z}_j)} \Big]^{\delta_i}. \tag{5.7}$$

Tomando logaritmo,

$$\begin{aligned}
\log L_{parcial} &= \sum_{i=1}^{n} \delta_i \Big[\log \rho(\underline{z}_i) - \log \sum_{j \in N_i} \rho(\underline{z}_j) \Big] \\
&= \sum_{i=1}^{n} \delta_i \Big[\sum_{j=1}^{s} a_j z_{ij} - \log \sum_{j \in N_i} \rho(\underline{z}_j) \Big].
\end{aligned}$$

Derivando respecto a a_k, para $k = 1, \ldots, s$, se obtiene

$$\frac{\partial}{\partial a_k} \log L_{parcial} = \sum_{i=1}^{n} \delta_i \Big[z_{ik} - \frac{\sum\limits_{j \in N_i} z_{jk}\, \rho(\underline{z}_j)}{\sum\limits_{j \in N_i} \rho(\underline{z}_j)} \Big].$$

Igualando a cero cada una de estas ecuaciones se encuentra el sistema de s ecuaciones que aparecen en el enunciado. Las incógnitas son los coeficientes $a_1, \ldots, a_s$. ∎

Debe observarse que la utilización de la verosimilitud parcial producirá solamente una estimación aproximada para los valores de los coeficientes. Por otro lado, en la función de verosimilitud parcial (5.7) y en el sistema de ecuaciones (5.6) no aparecen explícitamente los tiempos de fallecimiento o censura $x_{(1)} < \cdots < x_{(n)}$. La utilización de los datos se limita a las funciones indicadoras δ_i, a los conjuntos N_i y las características $\rho(\underline{z}_i)$ de sus elementos. También es necesario observar que los sumandos de (5.6) son distintos de cero sólo cuando $\delta_i = 1$, es decir, para individuos que fallecen. También es interesante observar que, en la estimación de los coeficientes $a_1, \ldots, a_s$, no aparece involucrada la función de riesgo base desconocida $\lambda_0(x)$.

En general, el sistema de ecuaciones (5.6) no es sencillo de resolver y es necesario el uso de programas de cómputo. Para clarificar la situación, presentaremos a continuación un ejemplo con pocos datos.

Ejemplo 5.8 *Supongamos que se tienen los siguientes tiempos de vida ordenados de $n = 5$ individuos:*

$$(x_{(1)}, 1), (x_{(2)}, 0), (x_{(3)}, 1), (x_{(4)}, 0), (x_{(5)}, 1).$$

Supondremos que todos los tiempos son distintos, es decir, $x_{(1)} < \cdots < x_{(5)}$, de esta forma no hay fallecimientos o censuras múltiples. La representación gráfica de los datos se muestra en la Figura 5.2.

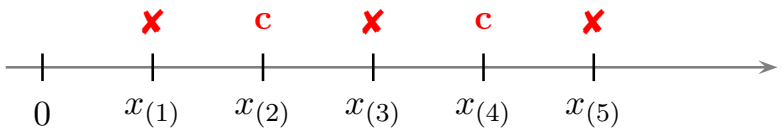

Figura 5.2: Datos del Ejemplo 5.8.

Se le designa como individuo i al elemento que es descartado al tiempo $x(i)$, $i = 1, \ldots, 5$. En los tiempos de fallecimientos ($\delta_i = 1$), tenemos que $N_1 = \{1, 2, 3, 4, 5\}$, $N_3 = \{3, 4, 5\}$ y $N_5 = \{5\}$, en donde $D_1 = \{1\}$, $D_3 = \{3\}$, $D_5 = \{5\}$.

Sea $\underline{z}_i = (z_{i1}, \ldots, z_{is})$ el vector de valores de s covariables asociado el individuo i. Por brevedad en la escritura, llamaremos ρ_i al factor de proporcionalidad en la función de riesgo del individuo i, es decir,

$$\rho_i := \rho(\underline{z}_i) = \exp\left(a_1 z_{i1} + \cdots + a_s z_{is}\right).$$

Entonces el sistema de ecuaciones (5.6) *es*

$$\begin{aligned}
&\Big[\, z_{1k} + \frac{z_{1k}\,\rho_1 + z_{2k}\,\rho_2 + z_{3k}\,\rho_3 + z_{4k}\,\rho_4 + z_{5k}\,\rho_5}{\rho_1 + \rho_2 + \rho_3 + \rho_4 + \rho_5} \,\Big] \\
&+ \Big[\, z_{3k} + \frac{z_{3k}\,\rho_3 + z_{4k}\,\rho_4 + z_{5k}\,\rho_5}{\rho_3 + \rho_4 + \rho_5} \,\Big] \\
&+ \Big[\, z_{5k} + \frac{z_{5k}\,\rho_5}{\rho_5} \,\Big] = 0, \quad k = 1, \ldots, s.
\end{aligned}$$

Las incógnitas de este sistema de ecuaciones son $a_1, \ldots, a_s$ y se encuentran dentro de las expresiones $\rho_i = \exp\left(a_1 z_{i1} + \cdots + a_s z_{is}\right)$. Los términos $z_{i1}, \ldots, z_{is}$ son los valores de las covariables para cada individuo i y son conocidos. Este ejemplo muestra que la solución al sistema de ecuaciones (5.6) *no es fácil de encontrar, aún cuando el número de datos sea pequeño.* ∎

Caso general

En el caso de múltiples fallecimientos a un mismo tiempo, se conoce la fórmula de aproximación de Breslow para la función de verosimilitud. Esta

fórmula establece que

$$L \approx \prod_{i=1}^{n} \Big[\frac{\prod_{j\in D_i} \rho(\underline{z}_j)}{(\sum_{j\in N_i} \rho(\underline{z}_j))^{d_i}} \Big]^{\delta_i}, \tag{5.8}$$

en donde d_i es el número de fallecimientos al tiempo $x_{(i)}$. La expresión (5.8) es una generalización de la función de verosimilitud parcial (5.7). Sólo los fallecimientos ($\delta_i = 1$) contribuyen en el producto (5.8) y en cada uno de ellos hay $d_i \geqslant 1$ fallecidos. Observe que si cada vez que $\delta_i = 1$ se tiene que $d_i = 1$, entonces (5.8) se reduce a (5.6).

Considerando igualdad en la aproximación de Breslow y tomando logaritmo,

$$\begin{aligned} \log L &= \sum_{i=1}^{n} \delta_i \Big[\sum_{j\in D_i} \log \rho(\underline{z}_j) - d_i \log \sum_{j\in N_i} \rho(\underline{z}_j) \Big] \\ &= \sum_{i=1}^{n} \delta_i \Big[\sum_{j\in D_i} \sum_{k=1}^{s} a_k z_{jk} - d_i \log \sum_{j\in N_i} \rho(\underline{z}_j) \Big]. \end{aligned}$$

Derivando respecto a a_k, para $k = 1, \ldots, s$, e igualando a cero se obtiene

$$\frac{\partial}{\partial a_k} \log L = \sum_{i=1}^{n} \delta_i \Big[\sum_{j\in D_i} z_{jk} - d_i \frac{\sum_{j\in N_i} z_{jk}\, \rho(\underline{z}_j)}{\sum_{j\in N_i} \rho(\underline{z}_j)} \Big] = 0. \tag{5.9}$$

Este es un sistema de s ecuaciones para las incógnitas $a_1, \ldots, a_s$. Se puede encontrar mayor información sobre la aproximación de Breslow en el artículo de D. Y. Lin [41].

Intervalo de confianza

Sean $\hat{a}_1, \ldots, \hat{a}_s$ los estimadores por máxima verosimilitud parcial de los coeficientes en el modelo de Cox. Entonces el estimador para el cociente de

riesgo (HR) es

$$\begin{aligned}\hat{\rho}(\underline{z}) &= \exp\{\,\hat{a}_1 z_1 + \cdots + \hat{a}_s z_s\,\} \\ &= \exp\{\hat{a}_1 z_1\} \cdots \exp\{\hat{a}_s z_s\} \\ &= (\exp\{\hat{a}_1\})^{z_1} \cdots (\exp\{\hat{a}_s\})^{z_s}.\end{aligned}$$

Usando el método delta puede comprobarse que

$$\operatorname{Var}(\exp\{\hat{a}_j\}) \approx (\exp\{\hat{a}_j\})^2 \operatorname{Var}(\hat{a}_j).$$

De donde se obtiene que la desviación estándar es

$$\operatorname{SE}(\exp\{\hat{a}_j\}) \approx \exp\{\hat{a}_j\} \operatorname{SE}(\hat{a}_j),$$

en donde SE significa *standard error*. Suponiendo una distribución aproximada normal para el estimador $\exp\{\hat{a}_j\}$, se puede encontrar el siguiente intervalo de confianza al $(1-\alpha)100\,\%$ para el cociente de riesgo,

$$\exp\{\hat{a}_j\} \pm z_{\alpha/2} \exp\{\hat{a}_j\} \operatorname{SE}(\hat{a}_j),$$

en donde $z_{\alpha/2}$ es un valor tal que $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$, con $Z \sim \mathrm{N}(0,1)$.

5.5. Estimación de la función de riesgo base

Nos interesa ahora estimar la función de riesgo base $\lambda_0(x)$ a la luz de una serie de observaciones que están sujetos a una posible censura por la derecha y que siguen el modelo de riesgos proporcionales de Cox con vector de covariables $\underline{z} = (z_1, \ldots, z_s)$, esto es,

$$\lambda(x, \underline{z}) = \lambda_0(x)\, \rho(\underline{z}),$$

en donde $\rho(\underline{z}) = \exp\{\,\sum_{j=1}^{s} a_j z_j\,\}$. La función desconocida $\lambda_0(x)$ puede tomar una forma paramétrica o bien puede corresponder a una distribución arbitraria. Supondremos este último caso. Supondremos también que los coeficientes $a_1, \ldots, a_s$ que aparecen en la expresión de $\rho(\underline{z})$ son conocidos o bien fueron estimados previamente, véase la sección anterior. Como antes, consideraremos que los datos se encuentran ordenados en forma ascendente $(x_{(1)}, \delta_1), \ldots, (x_{(n)}, \delta_n)$, y esta vez se puede llevar a cabo el análisis considerando que puede haber observaciones repetidas. A los tiempos

distintos en donde ocurren fallecimientos o censuras los denotaremos por $x'_{(1)} < \cdots < x'_{(k)}$, en donde k es un entero tal que $1 \leqslant k \leqslant n$. Los intervalos de análisis serán, nuevamente,

$$(0, x'_{(1)}],\ (x'_{(1)}, x'_{(2)}],\ \ldots\ , (x'_{(k-1)}, x'_{(k)}].$$

Recordemos también la definición de los siguientes conjuntos: para cada $i = 1, \ldots, k$,

$$\begin{aligned} N_i &= \text{“Conjunto de individuos en riesgo de morir} \\ &\quad \text{al inicio del intervalo } (x'_{(i-1)}, x'_{(i)}]\text{.”} \\ D_i &= \text{“Conjunto de individuos que fallecen al tiempo } x'_{(i)}\text{.”} \end{aligned}$$

Debe observarse que la fórmula (5.10) que aparece abajo es la estimación de una función de riesgo discreta como se indicó en el Ejercicio 95.

Proposición 5.3 *Suponga que $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ son datos de supervivencia con posible censura por la derecha. Sean $x'_{(1)} < \cdots < x'_{(k)}$ los tiempos distintos de fallecimientos, $1 \leqslant k \leqslant n$. El estimador por máxima verosimilitud para la función de riesgo base $\lambda_0(x)$ en el modelo de Cox (5.4) está dada por*

$$\hat{\lambda}_0(x'_{(i)}) = \frac{\hat{S}_0(x'_{(i-1)}) - \hat{S}_0(x'_{(i)})}{\hat{S}_0(x'_{(i-1)})}, \quad i = 1, \ldots, k, \tag{5.10}$$

en donde $\hat{S}_0(x)$ se calcula mediante la fórmula producto

$$\hat{S}_0(x) = \prod_{i: x'_{(i)} \leqslant x} \hat{\pi}_i, \quad 0 \leqslant x \leqslant x'_{(k)}, \tag{5.11}$$

y las probabilidades $\hat{\pi}_1, \ldots, \hat{\pi}_k$ se determinan a través del siguiente sistema de ecuaciones llamadas ecuaciones normales,

$$\sum_{j \in D_i} \frac{\rho(z_j)}{1 - \hat{\pi}_i^{\rho(z_j)}} = \sum_{j \in N_i} \rho(z_j), \quad i = 1, \ldots, k. \tag{5.12}$$

Demostración. Denotaremos por $S_0(x)$ a la función de supervivencia asociada a la función de riesgo base desconocida $\lambda_0(x)$, y por X_0 a una variable aleatoria con dicha distribución. Para cada $i = 1, \ldots, k$, se define π_i como la probabilidad de que un individuo con tiempo de vida X_0 y que está vivo al tiempo $x'_{(i-1)}$ sobreviva al intervalo $(x'_{(i-1)}, x'_{(i)}]$, es decir,

$$\pi_i := P(X_0 > x'_{(i)} \,|\, X_0 > x'_{(i-1)}) = \frac{S_0(x'_{(i)})}{S_0(x'_{(i-1)})}.$$

Consideremos ahora un individuo cualquiera en la población en estudio y con tiempo de vida X que sigue el modelo de riesgos proporcionales de Cox con vector de covariables $\underline{z}_j$. Si este individuo se encuentra con vida y bajo observación al tiempo $x'_{(i-1)}$, la probabilidad de que sobreviva al intervalo $(x'_{(i-1)}, x'_{(i)}]$ es

$$\begin{aligned} P(X > x'_{(i)} \,|\, X > x'_{(i-1)}) &= \frac{S(x'_{(i)}, \underline{z}_j)}{S(x'_{(i-1)}, \underline{z}_j)} \\ &= \frac{[\, S_0(x'_{(i)})\,]^{\rho(\underline{z}_j)}}{[\, S_0(x'_{(i-1)})\,]^{\rho(\underline{z}_j)}} \\ &= [\, \frac{S_0(x'_{(i)})}{S_0(x'_{(i-1)})}\,]^{\rho(\underline{z}_j)} \\ &= \pi_i^{\rho(\underline{z}_j)}. \end{aligned}$$

De esta forma, estas probabilidades de supervivencia de cualquier individuo quedan expresadas en términos de la misma probabilidad de supervivencia π_i asociadas al riesgo base y los valores de las covariables del individuo.

A continuación encontraremos la función de verosimilitud de los datos en términos de las probabilidades π_i.

- Sea $j \in D_i$. Entonces j es un individuo que inicia con vida el intervalo $(x'_{(i-1)}, x'_{(i)}]$ y fallece al tiempo $x'_{(i)}$. La probabilidad de que eso ocurra es
$$1 - \pi_i^{\rho(\underline{z}_j)}.$$

- Sea $j \in N_i \setminus D_i$. Entonces j es un individuo que está en riesgo de morir en el intervalo $(x'_{(i-1)}, x'_{(i)}]$ pero sobrevive a dicho intervalo. La probabilidad de que eso suceda es

$$\pi_i^{\rho(\underline{z}_j)}.$$

Por lo tanto, la función de verosimilitud completa es

$$L(\pi_1, \ldots, \pi_k) = \prod_{i=1}^{k} \Big[\prod_{j \in D_i} (1 - \pi_i^{\rho(\underline{z}_j)}) \Big] \cdot \Big[\prod_{j \in N_i \setminus D_i} \pi_i^{\rho(\underline{z}_j)} \Big].$$

Aquí se ha usado la hipótesis de independencia de los tiempos de vida de los individuos. Tomando logaritmo,

$$\log L(\pi_1, \ldots, \pi_k) = \sum_{i=1}^{k} \Big[\sum_{j \in D_i} \log (1 - \pi_i^{\rho(\underline{z}_j)}) \Big] + \Big[\sum_{j \in N_i \setminus D_i} \rho(\underline{z}_j) \log \pi_i \Big].$$

Derivando respecto de $\hat{\pi}_i$ e igualando a cero se obtiene

$$\sum_{j \in D_i} \frac{\rho(\underline{z}_j)\, \hat{\pi}_i^{\rho(\underline{z}_j)-1}}{1 - \hat{\pi}_i^{\rho(\underline{z}_j)}} = \sum_{j \in N_i \setminus D_i} \rho(\underline{z}_j)\, \frac{1}{\hat{\pi}_i}.$$

Separando la segunda suma,

$$\sum_{j \in D_i} \frac{\rho(\underline{z}_j)}{\hat{\pi}_i} \cdot \frac{\hat{\pi}_i^{\rho(\underline{z}_j)}}{1 - \hat{\pi}_i^{\rho(\underline{z}_j)}} = \sum_{j \in N_i} \frac{\rho(\underline{z}_j)}{\hat{\pi}_i} - \sum_{j \in D_i} \frac{\rho(\underline{z}_j)}{\hat{\pi}_i}.$$

Es decir,

$$\sum_{j \in D_i} \frac{\rho(\underline{z}_j)}{\hat{\pi}_i} \cdot \Big(\frac{\hat{\pi}_i^{\rho(\underline{z}_j)}}{1 - \hat{\pi}_i^{\rho(\underline{z}_j)}} + 1 \Big) = \sum_{j \in N_i} \frac{\rho(\underline{z}_j)}{\hat{\pi}_i}.$$

Simplificando la expresión que aparece entre paréntesis y eliminando el denominador $\hat{\pi}_i$ en ambos lados de la igualdad, se encuentran las así llamadas ecuaciones normales (5.12) que aparecen en el enunciado.

A partir de las probabilidades de supervivencia $\hat{\pi}_1, \ldots, \hat{\pi}_k$ se construye la estimación de la función de supervivencia discreta $\hat{S}_0(x)$ como aparece en (5.11).

Después, se usa $\hat{S}_0(x)$ para definir la función de riesgo discreta asociada $\hat{\lambda}_0(x)$ como se muestra en (5.10).

■

Observemos que $\hat{S}_0(x)$ será una función de supervivencia discreta genuina (es decir, tomará eventualmente el valor 0) cuando en el útimo dato $x_{(n)}$ sólo se observen fallecimientos, de modo que el último factor es $\hat{\pi}_k = 0$.

Corolario 5.1 *En ausencia de covariables, las ecuaciones normales tienen como solución $\hat{\pi}_i = 1 - d_i/n_i$, los cuales corresponden a los factores del estimador de Kaplan-Meier. Es decir,*

$$\hat{S}_0(x) = \hat{S}_{KM}(x), \quad 0 \leqslant x \leqslant x'_{(k)}.$$

Demostración. Cuando no hay covariables, es decir, cuando $\rho(\underline{z}) = 1$, las ecuaciones normales (5.12) se reducen a

$$\sum_{j \in D_i} \frac{1}{1 - \hat{\pi}_i} = \sum_{j \in N_i} 1, \quad \text{para} \;\; i = 1, \ldots, k.$$

Es decir, $d_i/(1 - \hat{\pi}_i) = n_i$, en donde $d_i = \#D_i$ y $n_i = \#N_i$. De aquí se obtiene que las probabilidades de supervivencia son $\hat{\pi}_i = 1 - d_i/n_i$. Estos son lo mismos factores que aparecen en el estimador de Kaplan-Meier. Véanse las expresiones (4.15) ó (4.16) en las páginas 177–177. ■

Ejemplo 5.9 *Consideremos nuevamente los datos del Ejemplo 5.8 dados por*

$$(x_{(1)}, 1), (x_{(2)}, 0), (x_{(3)}, 1), (x_{(4)}, 0), (x_{(5)}, 1),$$

en donde todos los tiempos son distintos, es decir, $x_{(1)} < \cdots < x_{(5)}$. Se le designa como individuo i al elemento que es descartado al tiempo $x(i)$, $i = 1, \ldots, 5$, ya sea por fallecimiento o por censura. En los tiempos de falle-

cimiento $x'_{(i)}$*, es decir, cuando* $\delta_i = 1$*, tenemos que*

$$\begin{aligned} N_1 &= \{1,2,3,4,5\}, & D_1 &= \{1\}, \\ N_3 &= \{3,4,5\}, & D_3 &= \{3\}, \\ N_5 &= \{5\}, & D_5 &= \{5\}. \end{aligned}$$

Escribiendo ρ_j *en lugar de* $\rho(\underline{z}_j)$*, las ecuaciones normales* (5.12) *son*

$$\frac{\rho_1}{1-\hat{\pi}_1^{\rho_1}} = \sum_{j=1}^{5} \rho_j, \qquad \frac{\rho_3}{1-\hat{\pi}_3^{\rho_3}} = \sum_{j=3}^{5} \rho_j, \qquad \frac{\rho_5}{1-\hat{\pi}_5^{\rho_5}} = \sum_{j=5}^{5} \rho_j.$$

En casos como este, en donde no se tienen observaciones repetidas, las ecuaciones normales pueden resolverse con facilidad. Para este ejemplo se tiene que la solución es

$$\hat{\pi}_i = \left(1 - \frac{\rho(\underline{z}_i)}{\sum_{j=i}^{5} \rho(\underline{z}_j)}\right)^{1/\rho(\underline{z}_i)}, \quad i = 1,3,5.$$

Si suponemos que no se tienen covariables, es decir, $\rho(\cdot) = 1$*, las estimaciones son* $\hat{\pi}_1 = 4/5$*,* $\hat{\pi}_3 = 2/3$ *y* $\hat{\pi}_5 = 0$*. Estas son los factores del estimador de Kaplan-Meier* (4.15)–(4.16)*, vea las páginas 177–177.* ∎

Los procesos de estimación anteriores y sus varias extensiones se encuentran implementados en distintos programas de cómputo. A continuación se verán algunos ejemplos usando el paquete estadístico `R`.

Ejemplo 5.10
(Base de datos "lung" dentro de la librería "survival" de `R`)
Este es un ejemplo bastante utilizado que hace uso de la base de datos "lung", la cual es parte de la librería "survival" del paquete estadístico `R`*. La base de datos contiene información de* 228 *pacientes con cáncer avanzado de pulmón. Cada registro contiene mediciones de las* 10 *variables que aparecen abajo. Las valoraciones (scores) miden la habilidad del paciente para realizar actividades diarias rutinarias.*

inst	*Código de la institución.*
time	*Tiempo de supervivencia en días.*
status	*1=Censurado, 2=Fallecido.*
age	*Edad del paciente en años cumplidos.*
sex	*1=Masculino, 2=Femenino.*
ph.ecog	*Valoración (score) ECOG otorgada por el médico.*
	0=Asintomático.
	1=Sintomático pero completamente ambulatorio.
	2=En cama menos del 50 % del día.
	3=En cama más del 50 % del día pero no postrado.
	4=Postrado en cama (bedbound).
ph.karno	*Valoración (score) Karnofsky otorgada por el médico.*
	0=Mal, . . ., 100=Bien.
pat.karno	*Valoración (score) Karnofsky del propio paciente.*
	0=Mal, . . ., 100=Bien.
meal.cal	*Calorías consumidas en las comidas.*
wt.loss	*Libras de peso perdidas en los últimos* 6 *meses.*
	(1 libra de peso es aproximadamente 0.45 kilos.)

Los comandos y códigos en R *que se presentan en este ejemplo fueron tomados del sitio web http://www.sthda.com, véase [48].*

Visualización de los datos

La variable de interés es el tiempo de supervivencia "time" expresado en días. Se pueden consultar los primeros 10 *registros con el código que aparece en el siguiente recuadro.*

R

```
# Acceso y visualización de la base de datos ‘‘lung’’
library(survival)
data(lung)
# primeros 10 registros completos
head(lung,10)
```

El resultado se muestra en la Figura 5.3.

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	NA
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	NA	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	NA	0
6	12	1022	1	74	1	1	50	80	513	0
7	7	310	2	68	2	2	70	60	384	10
8	11	361	2	71	2	2	60	80	538	1
9	1	218	2	53	1	1	70	80	825	16
10	7	166	2	61	1	2	70	70	271	34

Figura 5.3: Primeros registros de la base de datos “lung”.

Los tiempos de vida “time” y el “status” (censurado o fallecido) usando la notación x ó $x+$, de los primeros 10 *pacientes, se obtienen con el siguiente comando.*

R

```
# Base de datos ‘‘lung’’
# primeros 10 tiempos registrados
with(lung, Surv(time,status))[1:10]
```

El resultado se muestra en la Figura 5.4.

```
[1]  306   455   1010+  210   883   1022+  310   361   218   166
```

Figura 5.4: Primeros tiempos de la base de datos “lung”.

Ajuste del modelo

Se usa la función `coxph(·)` *para ajustar el modelo de Cox a los datos. Por ejemplo, suponga que se desea ajustar el modelo con únicamente la cova-*

riable z_1 = "sex". Esta covariable particular es categórica y, como lo hemos indicado, tiene como valores las etiquetas: 1 (Hombre) y 2 (mujer). Antes de hacer el ajuste del modelo, esta covariable debe ser tratada como tipo factor haciendo la siguiente transformación.

R

```
# Base de datos ''lung''
# transformación de la covariable ''sex''
lung$sex<-as.factor(lung$sex)
```

La estimación del correspondiente coeficiente a_1 se obtiene con el código

R

```
# Base de datos ''lung''
# ajuste del modelo de Cox con covariable ''sex''
coxph(Surv(time,status)~sex, data=lung)
```

El resultado se muestra en la Figura 5.5.

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

       coef exp(coef) se(coef)      z       p
sex -0.5310    0.5880   0.1672 -3.176 0.00149

Likelihood ratio test=10.63  on 1 df, p=0.001111
n= 228, number of events= 165
```

Figura 5.5: Resultado del comando `coxph(·)`.

Se obtiene que la estimación del coeficiente (coef) es $\hat{a}_1 = -0.5310$. Por lo tanto, el modelo se puede escribir como

$$\lambda(x, \underline{z}) = \lambda_0(x) \exp\{-0.5310 z_1\}, \quad x \geqslant 0.$$

El signo negativo de $\hat{a}_1$ significa que, cuando z_1 crece, el factor $\exp(\hat{a}_1 z_1)$ decrece. Tomando como referencia la categoría 1 (Masculino), el factor $\exp(\hat{a}_1 z_1)$ para mujeres tiene el siguiente efecto e interpretación,

$$0.5880 = \underbrace{\exp(-0.5310 * 1)}_{Mujeres} < \underbrace{\exp(-0.5310 * 0)}_{Hombres} = 1.$$

Esto significa que ser mujer (categoría 2) reduce el riesgo de muerte, comparativamente con el ser hombre (categoría 1), en un factor de 0.58 . Eso representa una reducción del 42 % en el riesgo. También se muestra el valor $\exp(\hat{a}_1) = 0.5880$*, el cual es el cociente de riesgo (Hazard ratio, HR). Esta cantidad proporciona una medida del efecto de la covariable. El valor se(coef)=0.1672 corresponde al error estándar de la estimación.*

El número $z = -3.176$ *es el valor de la estadística de la prueba de Wald sobre la significancia del coeficiente* a_1*. Véase la última sección de este capítulo en donde se discute brevemente la prueba de Wald. En este caso, el valor estimado es altamente significativo, es decir, se concluye que* $a_1 \neq 0$ *y, por lo tanto, la covariable "sex" tiene efectos reelevantes sobre la función de riesgo base. Por supuesto, se puede hacer un análisis similar al anterior para cada una de las covariables, por separado, que aparecen en la base de datos "lung", suponiendo válido el modelo de Cox y cuando la covariable es significativa.*

Gráficas del modelo

Usando el código que aparece abajo se obtiene la función de supervivencia estimada del modelo de Cox. Además de la librería "survival", también se necesita tener instalada la librería "survminer" para producir las gráficas.

R

```
# Base de datos ‘‘lung’’
library(survival)
library(survminer)
# ajuste del modelo con covariable ‘‘sex’’
res.cox<-coxph(Surv(time,status)~sex, data=lung)
# graficación
ggsurvplot(survfit(res.cox), data=lung,
palette="#2E9FDF", ggtheme=theme_light())
```

El resultado se muestra en la Figura 5.6, en donde, por defecto, se utilizan los valores medios de las covariables. En este caso, el valor medio sólo es

para la covariable "sex". En el eje horizontal se representa el tiempo medido en días.

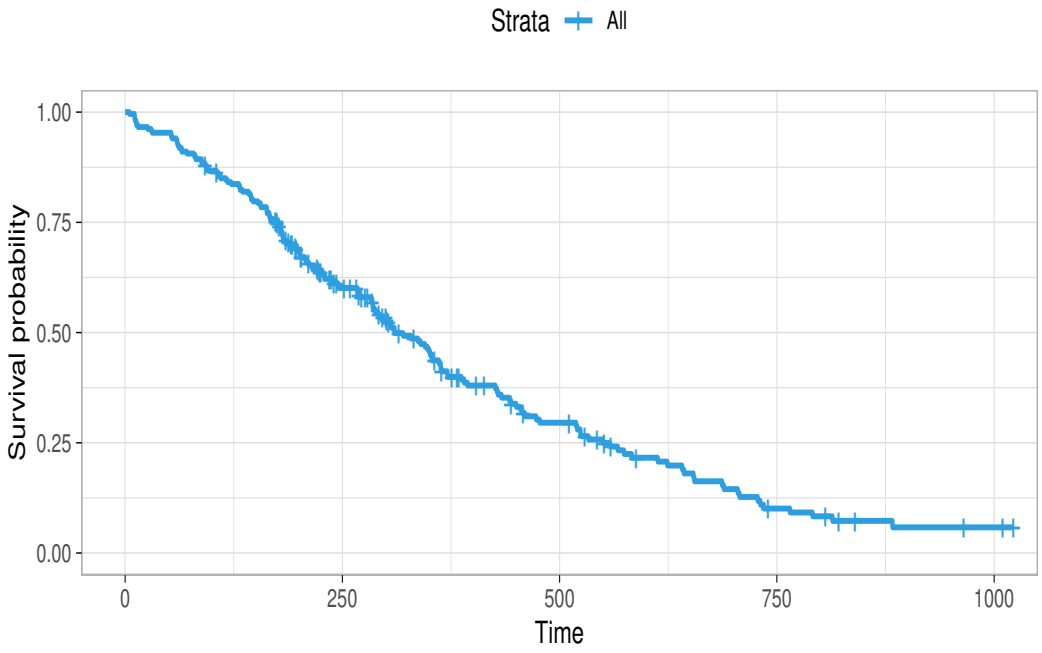

Figura 5.6: Función de supervivencia estimada de la base de datos "lung" para un modelo de Cox con covariable "sex".

Se pueden también graficar las funciones de supervivencia por sexo. Esto puede lograrse con el código que aparece abajo.

R

```
# Base de datos ‘‘lung’’
library(survival)
library(survminer)
res.cox<-coxph(Surv(time,status)∼sex, data=lung)
# creación de los nuevos datos
sex_df <- with(lung, data.frame(sex = c(1, 2),
age = rep(mean(age,na.rm = TRUE),2), ph.ecog = c(1,1)))
# curvas de supervivencia por sexo
fit <- survfit(res.cox, newdata = sex_df)
ggsurvplot(fit, data=sex_df, conf.int = TRUE,
legend.labs=c("Sex=1", "Sex=2"),
ggtheme = theme_minimal())
```

El resultado se muestra en la Figura 5.7. Las curvas mostradas hacen evidente que los hombres ($sex = 1$) tienen tiempos de vida más cortos que las mujeres ($sex = 2$).

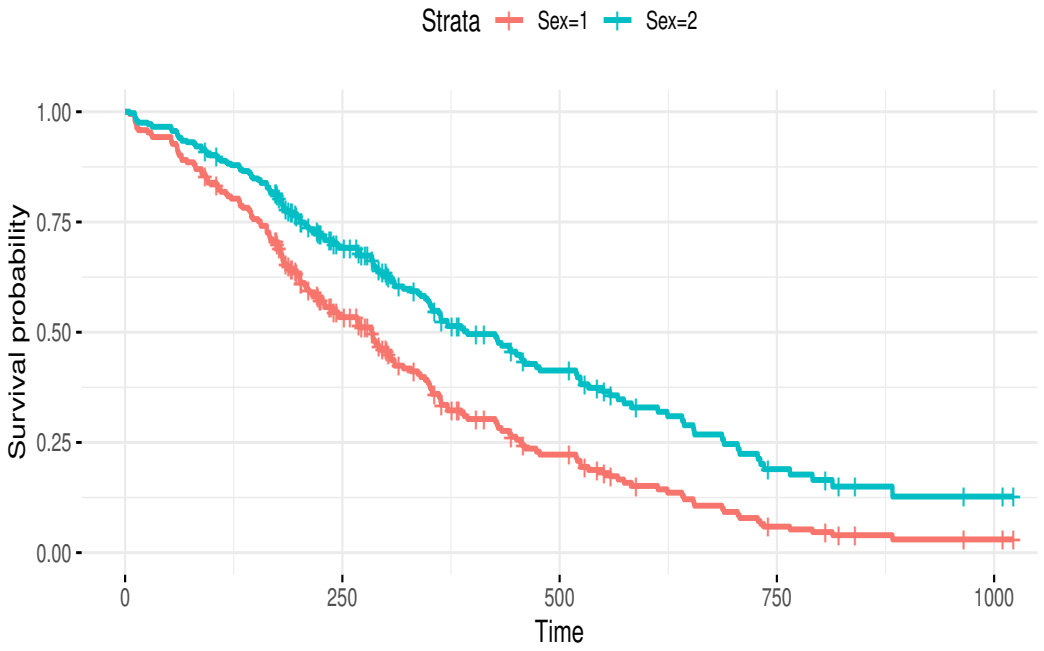

Figura 5.7: Función de supervivencia estimada por sexo para la base de datos "lung".

Mayor información del ajuste

Se puede obtener mayor información del ajuste mediante el comando "summary" que aparece en el siguiente recuadro.

R

```
# Base de datos ''lung''
res.cox<-coxph(Surv(time,status)~sex, data=lung)
# mayor información del ajuste
summary(res.cox)
```

El resultado se muestra en la Figura 5.8.

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

  n= 228, number of events= 165

       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816

Concordance= 0.579  (se = 0.021 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001
```

Figura 5.8: Resultado del comando `summary(res.cox)`.

Entre otra información, se proporciona un intervalo de confianza al 95 % *para el cociente de riesgo (HR)* $\exp(\hat{a}_1)$*, es decir,*

$$0.4237 < \exp(\hat{a}_1) < 0.816\,.$$

En la última parte de los resultados, se muestran los p−values para tres pruebas alternativas para la significancia general del modelo: la prueba del cociente de verosimilitud, la prueba de Wald y la prueba log-rank.

Ajuste del modelo para 2 ó más covariables

Un análisis similar se puede hacer para cada una de las 10 *covariables por separado de la base de datos "lung". Un siguiente paso consiste en ajustar modelos con múltiples covariables, suponiendo la validez de dichos modelos y considerando covariables significativas. Por ejemplo, si se desea ajustar el modelo de Cox para las covariables "sex" y "age" conjuntamente, se puede usar el código que se muestra en el recuadro siguiente. La llamada a la variable "res.cox" proporciona la información básica de la estimación, mientras que el comando "summary" provee mayor información.*

R

```
# Base de datos ‘‘lung’’
# ajuste del modelo con covariables ‘‘sex’’ y ‘‘age’’
res.cox<-coxph(Surv(time,status)~sex+age, data=lung)
# información básica del ajuste
res.cox
# mayor información del ajuste
summary(res.cox)
```

El resultado del comando “summary” del modelo ajustado a las covariables z_1 = “sex” y z_2 = “age” se muestra en la Figura 5.9. La estimación $\hat{a}_1$, correspondiente a la covariable “sex”, ha tomado ahora un valor ligeramente mayor en este nuevo modelo.

```
Call:
coxph(formula = Surv(time, status) ~ sex + age, data = lung)

  n= 228, number of events= 165

        coef exp(coef)  se(coef)      z Pr(>|z|)
sex -0.513219  0.598566  0.167458 -3.065  0.00218 **
age  0.017045  1.017191  0.009223  1.848  0.06459 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
sex    0.5986     1.6707    0.4311    0.8311
age    1.0172     0.9831    0.9990    1.0357

Concordance= 0.603  (se = 0.025 )
Likelihood ratio test= 14.12  on 2 df,   p=9e-04
Wald test            = 13.47  on 2 df,   p=0.001
Score (logrank) test = 13.72  on 2 df,   p=0.001
```

Figura 5.9: Resultado del comando `summary(res.cox)` del ajuste del modelo de Cox con covariables “age” y “sex”, conjuntamente.

El modelo se puede escribir como

$$\lambda(x, \underline{z}) = \lambda_0(x) \exp\{-0.513219 z_1 + 0.017045 z_2\}.$$

Los intervalos de confianza al 95 % *para los cocientes de riesgo son*

$$0.4311 < \exp\{a_1\} < 0.8311,$$
$$0.9990 < \exp\{a_2\} < 1.0357\,.$$

Se puede generar una gráfica de la función de supervivencia para este modelo con las dos covariables indicadas. El código en R *es similar al caso de una covariable y se muestra a continuación.*

R

```
# Base de datos ‘‘lung’’
library(survival)
library(survminer)
# ajuste del modelo con covariables ‘‘sex’’ y ‘‘age’’
res.cox<-coxph(Surv(time,status)~sex+age, data=lung)
# graficación
ggsurvplot(survfit(res.cox), data=lung,
palette="#2E9FDF", ggtheme=theme_light())
```

El resultado se muestra en la Figura 5.10, en donde, por defecto, se utilizan los valores medios de las dos covariables. ■

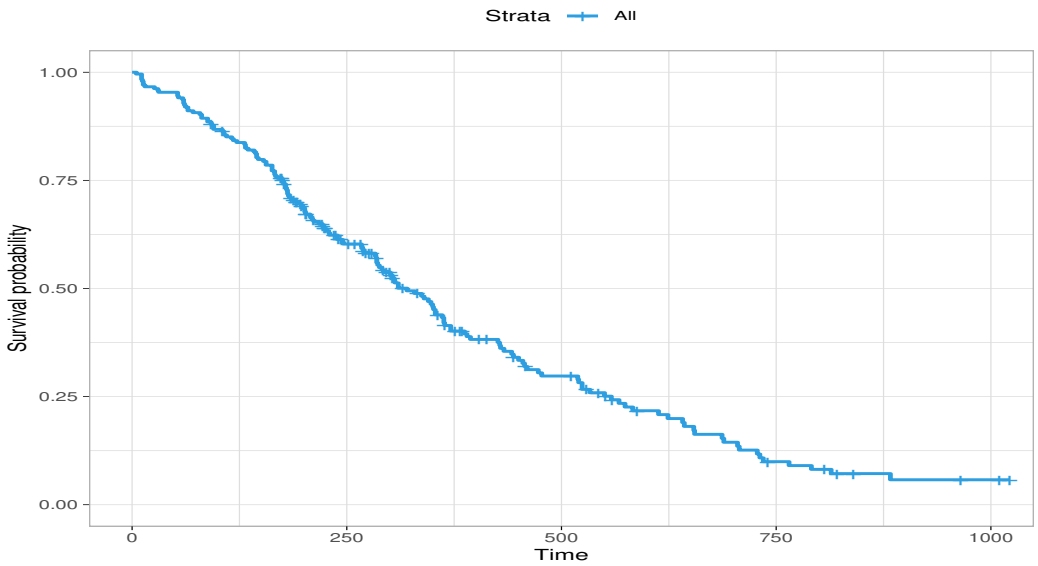

Figura 5.10: Función de supervivencia estimada de la base de datos “lung” para el modelo de Cox con covariables “sex” y “age”.

A continuación se presentan ejemplos de bases de datos de supervivencia que se encuentran disponibles en algunas librerías del paquete estadístico R. La presentación sólo es a nivel informativo sin hacer ningún análisis estadístico de los tiempos de vida. En la sección de ejercicios se hace referencia a estos conjuntos de datos. Estas bases de datos también pueden consultarse en el Apéndice A del libro de J. D. Kalbfleisch y R. L. Prentice [30].

Ejemplo 5.11
(Base de datos "veteran" dentro de la librería "survival" de R)
Esta base de datos contiene información de 125 *pacientes con cáncer de pulmón. A cada paciente se le sometió (al azar) a uno de dos tratamientos disponibles. Cada registro contiene mediciones de las* 8 *variables que se muestran abajo.*

trt	*Tipo de tratamiento.*
	1=Estándar.
	2=Prueba.
celltype	*Tipo de célula.*
	1=Squamous.
	2=Smallcell.
	3=Adeno.
	4=Large.
time	*Tiempo de supervivencia.*
status	*1=Censurado, 2=Fallecido.*
karno	*Valoración (score) Karnofsky.*
	0=Mal, . . ., 100=Bien.
diagtime	*Tiempo en meses desde el diagnóstico hasta el tratamiento.*
age	*Edad del paciente en años cumplidos.*
prior	*Terapia previa.*
	0=No.
	10=Si.

En particular, la variable "celltype" describe 4 *tipos de cáncer de acuerdo a la forma que toman las células enfermas ("squamous", "smallcell", "adeno" y "large"). La variable de interés es el tiempo de supervivencia "time" expresado en días. Se pueden consultar los primeros* 5 *registros con el código que aparece en el siguiente recuadro.*

R

```
# Base de datos ''veteran''
# dentro de la librería ''survival'' de R
library(survival)
# primeros 5 registros completos
head(veteran,5)
```

El resultado se muestra en la Figura 5.11.

	trt	celltype	time	status	karno	diagtime	age	prior
1	1	squamous	72	1	60	7	69	0
2	1	squamous	411	1	70	5	64	10
3	1	squamous	228	1	60	3	38	0
4	1	squamous	126	1	60	9	63	10
5	1	squamous	118	1	70	11	65	10

Figura 5.11: Primeros registros de la base de datos "veteran".

Ejemplo 5.12
(Base de datos "larynx" dentro de la librería "KMsurv" de R)
La librería "KMsurv" del R *contiene un conjunto de bases de datos que aparecen también en el libro de J. P. Klein y M. L. Moeschberger [33]. Las primeras letras de "KMsurv" hacen referencia a los apellidos de estos dos autores. La base de datos "larynx" contiene información de* 90 *pacientes con cáncer de laringe, quienes presentaban diferentes estados de su enfermedad. Cada registro contiene las siguientes* 5 *variables:*

stage *Estado de la enfermedad.*

- *1 = Estado 1.*
- *2 = Estado 2.*
- *3 = Estado 3.*
- *4 = Estado 4.*

time *Tiempo de vida en meses.*

age *Edad a la que se detectó el cáncer.*

diagyr *Año calendario en el que se detectó el cáncer.*

delta *0=Censurado, 1=Fallecido.*

La variable de interés es el tiempo de supervivencia "time" expresado en meses. Se pueden consultar los primeros 10 *registros con el código que aparece en el siguiente recuadro. El resultado se muestra en la Figura 5.12.*

R

```
# Base de datos "larynx"
library(KMsurv)
data(larynx)
# primeros 10 registros completos
head(larynx,10)
```

	stage	time	age	diagyr	delta
1	1	0.6	77	76	1
2	1	1.3	53	71	1
3	1	2.4	45	71	1
4	1	2.5	57	78	0
5	1	3.2	58	74	1
6	1	3.2	51	77	0
7	1	3.3	76	74	1
8	1	3.3	63	77	0
9	1	3.5	43	71	1
10	1	3.5	60	73	1

Figura 5.12: Primeros registros de la base de datos "larynx".

Los tiempos de vida "time" y el valor de "delta" (censurado o fallecido) usando la notación x ó $x+$, de los primeros 10 *pacientes, se obtienen con los siguientes comandos. El resultado se muestra en la Figura 5.13.*

R

```
# Base de datos ''larynx''
library(KMsurv)
# primeros 10 tiempos registrados
with(larynx, Surv(time,delta))[1:10]
```

```
[1] 0.6  1.3  2.4  2.5+ 3.2  3.2+ 3.3  3.3+ 3.5  3.5
```

Figura 5.13: Primeros tiempos de la base de datos "larynx".

■

Ejemplo 5.13

(Base de datos "hoel" dentro de la librería "survival" de R)

Esta base de datos contiene información del tiempo en días que tardaron en morir de cáncer un grupo de 181 *ratones machos a los que se les sometió a* 300 *rads de radiación. Un rad es una medida estadounidense antigua utilizada para medir dosis de radiación absorbida. Un grupo de ratones se mantuvo en condiciones higiénicas ("Germ-free") y un segundo grupo fue el de control ("Control"). Cada registro de la base de datos contiene las siguientes* 4 *variables:*

trt — *Tipo de tratamiento: "Control" ó "Germ-free".*

days — *Tiempo de vida en días.*

outcome — *Tipo de cáncer.*

- *Codecensor.*
- *Thymic lymphoma.*
- *Reticulum cell sarcoma.*
- *Other causes.*

id — *Número de identificación del ratón.*

La variable de interés es el tiempo de supervivencia "days" expresado en días. Se pueden consultar los primeros 5 *registros con el código que aparece en el siguiente recuadro.*

R

```
# Base de datos ‘‘hoel’’
library(survival)
# primeros 5 registros completos
head(hoel,5)
```

El resultado se muestra en la Figura 5.14.

	trt	days	outcome	id
1	Control	159	thymic lymphoma	1
2	Control	189	thymic lymphoma	2
3	Control	191	thymic lymphoma	3
4	Control	198	thymic lymphoma	4
5	Control	200	thymic lymphoma	5

Figura 5.14: Primeros registros de la base de datos "hoel".

El nombre de la base de datos proviene de D. G. Hoel, quien utilizó esta información en su trabajo: "A representation of mortality data by competing risks". Biometrics 33, pp. 1-30, 1972. El estudio de D. G. Hoel afirma que el ambiente de condiciones higiénicas ("Germ-free") tiene poco efecto en la ocurrencia del cáncer "thymic lymphoma", pero retrasa las otras causas de muerte de los ratones. En la sección de ejercicios se pide corroborar estadísticamente estas afirmaciones.

5.6. Algunas pruebas de hipótesis

Concluimos este capítulo con algunas explicaciones breves sobre ciertas pruebas de hipótesis relativas al ajuste del modelo de Cox.

Prueba de significancia de Wald

Esta es una de las pruebas más usadas para determinar la significancia o relevancia de un parámetro dentro de un modelo a partir de una serie de observaciones, en este caso, la significancia de las covariables en el modelo de Cox. Es decir, determina si una covariable tiene, o no tiene, una contribución significativa en la función de riesgo base. Para la covariable z_j con coeficiente a_j, la prueba constrasta las hipótesis

$$H_0 : a_j = 0 \quad \text{v.s.} \quad H_1 : a_j \neq 0.$$

La prueba está basada en la estadística $Z = (\hat{a}_j - 0)/SE(\hat{a}_j)$, en donde $\hat{a}_j$ es el estimador máximo verosímil para a_j y $SE(\hat{a}_j)$ es el error estándar (*standard error*) de la estimación. La variable Z^2 tiene distribución aproximada χ^2 con 1 grado de libertad. Se puede entonces calcular el *p-value* como la probabilidad de que la estadística de la prueba tome un valor como el observado u otros más alejados de lo que establece la hipótesis nula, suponiendo ésta cierta, es decir,

$$p\text{-value} = P(Z^2 \geqslant z \mid H_0 \text{ es cierta}),$$

en donde $z > 0$ es el valor tomado por la estadística Z^2.

Se rechaza H_0 cuando $\hat{a}_j$ se aleja del valor 0 y, por lo tanto, Z^2 toma un valor grande. O bien, en términos del *p-value*, cuando éste es menor a un cierto nivel del significancia (típicamente 0.05) se rechaza H_0. Véase la Figura 5.15.

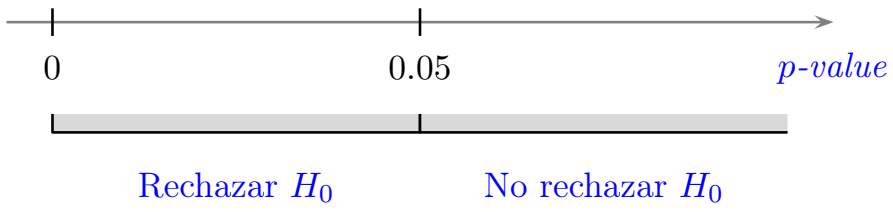

Figura 5.15: Decisión usando el *p-value* en la prueba de significancia de Wald.

La prueba de significancia de Wald se lleva a cabo de manera automática en el paquete `R` al ajustar el modelo de Cox (usando el comando `coxph(·)`), para cada una de las covariables especificadas. Véase el Ejemplo 5.10. Se puede encontrar mayor información sobre la prueba en el artículo de A. Wald [54].

Prueba para verificar proporcionalidad

En la mayoría de los ejemplos mencionados en este trabajo se ha supuesto que el modelo de Cox puede ajustarse a un conjunto de datos. Sin embargo, en algunas áreas de aplicación del modelo de Cox, como en estudios clínicos, no se espera que los riesgos sean verdaderamente proporcionales. Los riesgos no son proporcionales cuando, por ejemplo, los efectos de un tratamiento médico cambian a lo largo del tiempo. Dichos efectos pueden manifestarse después de periodos largos, por ejemplo, 6 o más meses. Otra situación en donde los riesgos pueden no cumplir la hipótesis de proporcionalidad se presenta cuando, por cuestiones genéticas, algunos individuos pueden tener diferentes susceptibilidades para desarrollar con mayor o menor rapidez una enfermedad dada. De esta manera, el cociente de riesgo $\exp\{a_j z_j\}$ (HR, *hazard ratio*) asociado a una covariable (tratamiento) z_j no se mantiene constante, sino que cambia al paso del tiempo. Un caso extremo ocurre cuando la covariable no tiene ningún efecto en la función de riesgo. En tales casos, el cociente de riesgo (HR) es constante igual a 1 a lo largo del tiempo pues el coeficiente es cero.

Existen pruebas de hipótesis para verificar que las covariables inciden en la función de riesgo base como proporciones, tal y como lo supone el modelo de Cox. La hipótesis nula que se propone se puede expresar de la sigueinte forma:

H_0 : "Los riesgos derivados de las covariables son proporcionales."

Una prueba para esta hipótesis y que se encuentra implementada en el paquete `R` es la propuesta por P. M. Grambsch y T. M. Therneau en el artículo [22]. Esta prueba está basada en una estadística que tiene distribución asintótica χ^2 con ciertos grados de libertad.

Por ejemplo, considerando nuevamente la base de datos "lung" del paquete `R` (ver Ejemplo 5.10), para llevar a cabo una prueba para determinar si la covariable "sexo" incide en la función de riesgo de forma proporcional, se puede emplear el siguiente código que hace uso de la función `cox.zph`.

R

```
# Base de datos ''lung''
fit <- coxph(Surv(time,status)~sex, data=lung)
temp <- cox.zph(fit)
print(temp)
plot(temp)
```

El resultado del comando `print` se muestra en la Figura 5.16.

```
        chisq df     p
sex      2.86  1 0.091
GLOBAL   2.86  1 0.091
```

Figura 5.16: Resultados numéricos de la prueba `cox.zph` para la covariable "sex" de la base de datos "lung".

Se muestra el valor de la estadística χ^2 (2.86), los grados de libertad ($df = 1$) y el *p-value* de la prueba ($p = 0.091$). Como el *p-value* no es menor al nivel de significancia $\alpha = 0.05$, no se rechaza la hipótesis nula, es decir, puede suponerse que la covariable "sex" incide en la función de riesgo de manera proporcional.

Por otro lado, el comando `plot` muestra la gráfica de la Figura 5.17. En el eje horizontal aparece una escala no lineal del tiempo y en el eje vertical se muestra el residual de Schoenfeld. La línea continua corresponde a un *spline* suavizado junto con ± 2 desviaciones estándar. Véase el capítulo 13 del libro de W. N. Venables y B. D. Ripley [52]. La línea continua cambia poco a lo largo del tiempo, lo que indica que puede considerarse que el cociente de riesgo es constante.

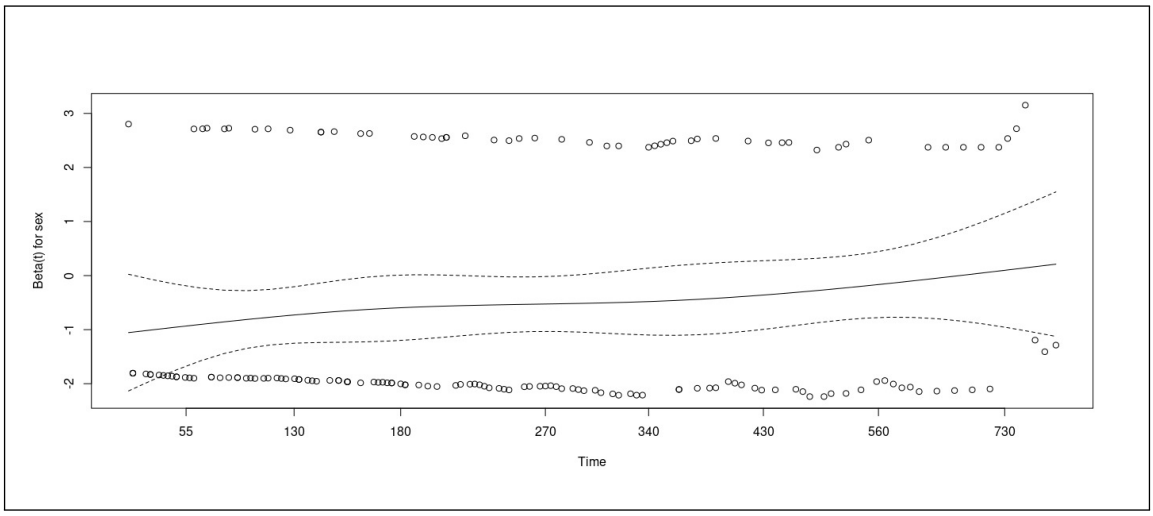

Figura 5.17: Resultado gráfico de la prueba `cox.zph` para la covariable "sex" de la base de datos "lung".

Se pueden llevar a cabo pruebas de proporcionalidad para cada una de las covariables por separado, y también considerando dos o más covariables al mismo tiempo añadiendo al código en `R` la cadena "`~sex+age`", por ejemplo.

Se puede encontrar mayor información sobre la prueba acerca de la hipótesis de proporcionalidad de los riesgos en el modelo de Cox en los libros de T. M. Therneau y P. M. Grambsch [49] ó J. P. Klein y M. L. Moeschberger [33].

5.7. Ejercicios

Covariables

171. Para un conjunto de pacientes a quienes se les ha detectado una cierta enfermedad terminal, indique 5 posibles covariables que podrían incidir en el tiempo de vida de estas personas.

172. *Interpretación de los cocientes de riesgo (HR).*
Considere el modelo de Cox con vector de covariables $\underline{z} = (z_1, \ldots, z_s)$. Demuestre que:

   a) Si $\exp a_i > 1$, la función $z_i \mapsto \lambda(x, \underline{z})$ es creciente.

   b) Si $\exp a_i < 1$, la función $z_i \mapsto \lambda(x, \underline{z})$ es decreciente.

c) Si $\exp a_i = 1$, la función $z_i \mapsto \lambda(x, \underline{z})$ es constante.

173. *Función de riesgo base Weibull.*
Suponga que un tiempo de vida X_z depende de una covariable z de la siguiente forma

$$X_z \sim \text{Weibull}(\alpha, \lambda_z), \quad \text{con} \quad \lambda_z = e^{\beta z}.$$

a) Encuentre la función de riesgo $\lambda(x, z)$.

b) Encuentre una expresión para la función de riesgo base $\lambda_0(x)$.

c) Compruebe que X_z presenta riesgos proporcionales para diferentes valores de z.

174. *Función de riesgo base Weibull.*
Considere el modelo de Cox con vector de covariables $\underline{z} = (z_1, \ldots, z_s)$ y con función de riesgo base Weibull(α, λ), es decir,

$$\lambda_0(x) = \alpha\lambda x^{\alpha-1}, \quad x > 0.$$

a) Encuentre expresiones para $\lambda(x, \underline{z})$, $S(x, \underline{z})$, $F(x, \underline{z})$, $f(x, \underline{z})$ y $\Lambda(x, \underline{z})$.

b) Asigne valores a α y λ. Grafique las funciones encontradas en el inciso anterior como función del tiempo x y analice el efecto que tiene en cada una de ellas el factor $\rho(\underline{z})$.

175. *Logaritmo de un tiempo de vida con riesgos proporcionales.*
Sea X un tiempo de vida que depende de un vector de covariables $\underline{z}$ tal que su función de riesgo es

$$\lambda_X(x, \underline{z}) = \lambda_X^0(x) g(\underline{z}),$$

en donde $\lambda_X^0(x)$ denota la función de riesgo base ($\underline{z} = \underline{0}$) de X, y g es una función positiva tal que $g(\underline{0}) = 1$. Sea $Y = \log X$.

a) Demuestre que $S_Y(y, \underline{z}) = S_X(e^y, \underline{z})$.

b) Demuestre que $\lambda_Y(y, \underline{z}) = \lambda_Y^0(y) g(\underline{z})$, en donde $\lambda_Y^0(y) = e^y \lambda_X^0(e^y)$. Esto significa que también Y sigue el modelo de riesgos proporcionales. La proporción $g(\underline{z})$ que se aplica al riesgo es la misma pero la función de riesgo base cambia.

176. *Covariables dependientes del tiempo.*
Considere el modelo de Cox con vector de covariables dependiente del tiempo x, es decir, $\lambda(x, \underline{z}(x)) = \lambda_0(x)\, \rho(\underline{z}(x))$. Encuentre expresiones lo más reducidas posible para las funciones básicas asociadas $S(x, \underline{z}(x))$, $F(x, \underline{z}(x))$, $f(x, \underline{z}(x))$ y $\Lambda(x, \underline{z}(x))$.

Modelos para incorporar covariables

177. *Modelo aditivo.*
Considere el modelo aditivo $\lambda(x) = \lambda_0(x) + a$, en donde $a \geqslant 0$ es una constante. Encuentre expresiones para $S(x)$, $F(x)$, $f(x)$, $\Lambda(x)$ y $R(x)$. Observe que la constante a puede ser la combinación lineal $\sum_{j=1}^{s} a_j z_j$.

178. *Modelo multiplicativo.*
Considere el modelo multiplicativo $\lambda(x) = a\, \lambda_0(x)$, en donde $a > 0$ es una constante. Encuentre expresiones para $S(x)$, $F(x)$, $f(x)$, $\Lambda(x)$ y $R(x)$. Observe que la constante a puede ser la función de covariables $\prod_{j=1}^{s} h(z_j)$.

179. *Tiempo de vida acelerada.*
Sea X un tiempo de vida y sea $a > 0$ una constante. El tiempo de vida acelerada de X se define como $Y := X/a$. Demuestre que:

   a) $F_Y(y) = F_X(ay)$, para $y \geqslant 0$.

   b) $f_Y(y) = a\, f_X(ay)$, para $y \geqslant 0$.

   c) $S_Y(y) = S_X(ay)$, para $y \geqslant 0$.

   d) $\lambda_Y(y) = a\, \lambda_X(ay)$, para $y \geqslant 0$ tal que $S_X(ay) > 0$.

   e) $R_Y(y) = \dfrac{1}{a\, S_X(ay)} \displaystyle\int_{ay}^{\infty} S_X(u)\, du$, para $y \geqslant 0$ tal que $S_X(ay) > 0$.

180. *Tiempo de vida acelerada.*
Considere el modelo de vida acelerada $Y = X/a$. Demuestre que la función $S_Y(y) = S_X(ay)$ es de supervivencia. Dibuje en un mismo plano $S_X(x)$ y $S_Y(x)$ para alguna constante $a > 0$. Distinga los tres casos: $0 < a < 1$, $a = 1$ y $a > 1$.

181. *Tiempo de vida Weibull acelerado coincide con el modelo de Cox.*
Sea $g(\underline{z})$ una función positiva de un vector de covariables $\underline{z} = (z_1, \ldots, z_s)$.

   a) Sea X un tiempo de vida con distribución Weibull(α, λ), es decir, su función de riesgo es $\lambda_0(x) = \alpha\lambda x^{\alpha-1}$, para $x > 0$. Demuestre que la función de riesgo de $X/g(\underline{z})$ es

$$\lambda(x) = \lambda_0(x)\,[g(\underline{z})]^{\alpha}.$$

   b) Recíprocamente, si X es un tiempo de vida con función de riesgo $\lambda_0(x)$ tal que $X/g(\underline{z})$ tiene función de riesgo $\lambda(x) = \lambda_0(x)\,[g(\underline{z})]^{\alpha}$ con $\alpha > 0$, entonces X tiene distribución Weibull.

El modelo de Cox

182. Suponga que X_1 representa el tiempo de vida de un individuo con covariable $Z = z_1$ y X_2 representa el tiempo de vida de un individuo con covariable $Z = z_2$. Suponiendo un modelo de riesgos proporcionales, demuestre que

$$P(X_1 > x) > P(X_2 > x) \quad \text{ó} \quad P(X_1 > x) \leqslant P(X_2 > x),$$

para todo x tal que $0 < P(X_i > x) < 1$, para $i = 1, 2$.

183. Demuestre que la función de densidad asociada a la función de riesgo $\lambda(x, \underline{z})$ en el modelo de Cox se puede escribir de la forma siguiente

$$f(x, \underline{z}) = \lambda_0(x)\rho(\underline{z}) \exp\{\, -\rho(\underline{z})\Lambda_0(x)\,\}.$$

184. *Modelo exponencial.*
Sea $\lambda(x)$ la función de riesgo del modelo exp(λ) y sea $\lambda(x, \underline{z})$ la función de riesgo del modelo (5.4) de Cox con covariables $\underline{z} = (z_1, \ldots, z_s)$.

   a) Demuestre que $\lambda(x, \underline{z})$ también es una función de riesgo del modelo exponencial y encuentre el parámetro correspondiente.

   b) Grafique en el mismo plano las funciones $\lambda(x)$ y $\lambda(x, \underline{z})$.

185. Sea $\lambda(x)$ una función de riesgo continua con soporte $(0, \infty)$. Usando la Definición 2.4 de la página 51, demuestre que la función riesgo $\lambda(x, \underline{z})$ en el modelo (5.4) de Cox también es una función de riesgo.

186. *Distribución de riesgo base Weibull.*
Suponga que el tiempo de vida base en el modelo de Cox tiene distribución Weibull(α, λ). Demuestre que las funciones asociadas a la función de riesgo $\lambda(x, \underline{z}) = \lambda_0(x)\, \rho(\underline{z})$ son:

a) $S(x, \underline{z}) = \exp\{-\lambda\rho(\underline{z})x\alpha\}$.

b) $F(x, \underline{z}) = 1 - \exp\{-\lambda\rho(\underline{z})x\alpha\}$.

c) $f(x, \underline{z}) = \alpha\lambda\, x^{\alpha-1}\, \rho(\underline{z})\, \exp\{-\lambda\rho(\underline{z})x^{\alpha}\}$.

d) $\Lambda(x, \underline{z}) = \lambda\rho(\underline{z})x^{\alpha}$.

Cualquiera de estas expresiones indica que la distribución del tiempo de vida con covariables $\underline{z}$ es Weibull$(\alpha, \lambda\rho(\underline{z}))$.

187. *Interpretación de coeficientes en el modelo de Cox.*
Sea $r \geqslant 0$ una constante y sean z_j y $z_j + r$ dos posibles valores de la covariable z_j en el modelo de Cox con vector de covariables $\underline{z} = (z_1, \ldots, z_s)$, $1 \leqslant j \leqslant s$. Demuestre e interprete la identidad

$$\frac{\lambda(x, z_1, \ldots, z_j + r, \ldots, z_s)}{\lambda(x, z_1, \ldots, z_j, \ldots, z_s)} = \exp\{a_j r\}.$$

Nota: Por simplicidad en la notación, los valores z_j y $z_j + r$ usan la misma letra que el nombre de la covariable z_j.

188. *Base de datos "veteran".*
Considere la base de datos "veteran" de la librería "survival" en `R`, la cual se muestra en el Ejemplo 5.11, página 258.

a) Grafique la estimación de Kaplan-Meier para la función de supervivencia. Aquí solamente se usan los tiempos de vida registrados.

b) Considerando cada uno de los 4 tipos de cáncer y suponiendo válido el modelo de Cox, determine visualmente si el tipo de cáncer afecta a la función de supervivencia.

c) Determine si existe diferencia significativa entre los dos tratamientos aplicados.

d) Para cada tipo de cáncer, determine si un tratamiento es mejor que el otro.

189. *Base de datos "veteran".*
Considere nuevamente la base de datos "veteran" de la librería "survival" en R, que se expone en el Ejemplo 5.11, página 258. Ajuste un modelo de Cox que incorpore las covariables tratamiento "trt" y tipo de cáncer "celltype".

   a) Determine si alguno de los tratamientos tiene efectos para una supervivencia mayor.

   b) Determine si alguno de los tipos de cáncer tiene efectos para una supervivencia menor.

190. *Base de datos "hoel".*
Considere la base de datos "hoel" de la librería "survival" en R, la cual se expone en el Ejemplo 5.13, en la página 261. Compruebe que el ambiente de condiciones higiénicas ("Germ-free") tiene poco efecto en la ocurrencia del cáncer "thymic lymphoma", pero retrasa las otras causas de muerte de los ratones.

Estimación de coeficientes

191. Considere las edades de fallecimiento de 8 personas como aparecen abajo. Suponga que se puede adoptar el modelo de Cox con vector de covariables $\underline{z} = (z_1, \ldots, z_s)$. Escriba el sistema de ecuaciones (5.9) para estimar los coeficientes $a_1, \ldots, a_s$. Observe que existen observaciones repetidas.

$$87, 92, 70+, 55, 70, 92, 85+, 70.$$

192. *Base de datos "lung": covariables significativas.*
Utilizando el paquete estadístico R y para la base de datos "lung" (vea el Ejemplo 5.10), aplique la prueba de Wald para determinar si cada una de las 10 covariables son significativas, es decir, si contribuyen a la modificación de la función de riesgo base, suponiendo válido el modelo de Cox. Para aquellas covariables significativas estime su coeficiente de manera individual y grafique la correspondiente función de riesgo.

193. *Base de datos "lung": dos covariables más significativas.*
Utilizando el paquete estadístico R y para la base de datos "lung" (vea el Ejemplo 5.10), ajuste un modelo de Cox con las 2 covariables más

significativas. Aplique la prueba de Wald para determinar si cada una de las 10 covariables son significativas, es decir, si contribuyen a la modificación de la función de riesgo base, suponiendo válido el modelo de Cox. Para aquellas covariables significativas estime su coeficiente de manera individual y grafique la correspondiente función de riesgo.

Estimación de la función de riesgo base

194. Considere el conjunto de edades de fallecimiento de 10 personas que aparece abajo. Suponiendo aplicable el modelo de Cox con vector de covariables $\underline{z} = (z_1, \ldots, z_s)$, escriba el sistema de ecuaciones normales (5.12) para estimar los factores de $\hat{S}_0(x)$.

$$75, 89+, 72, 65, 72, 75+, 81, 59, 63+, 48.$$

195. Considere el conjunto de ecuaciones normales (5.12) y suponga que no hay observaciones repetidas.

a) Demuestre que la solución está dada por

$$\hat{\pi}_i = \Big(1 - \frac{\rho(\underline{\mathrm{z}}_i)}{\sum_{j \in N_i} \rho(\underline{\mathrm{z}}_j)}\Big)^{1/\rho(\underline{\mathrm{z}}_i)}, \quad i = 1, \ldots, k.$$

b) Compruebe que el estimador $\hat{S}_0(x)$ coincide con el estimador de Kaplan-Meier en ausencia de covariables.

§

Apéndice A

A.1. Transformación de un tiempo de vida

Sea X un tiempo de vida continuo y sea φ una función definida en el rango de valores de X y tal que $Y := \varphi(X)$ sigue siendo un tiempo de vida, es decir, una variable aleatoria positiva. En esta sección encontraremos fórmulas para las funciones básicas asociadas a la variable Y cuando la transformación φ cumple ciertas condiciones.

Por simplicidad en la escritura, las funciones básicas de X no aparecen con subíndice X. En cambio, las funciones básicas asociadas a Y aparecen con el subíndice Y.

Proposición A.1 *Sea X un tiempo de vida continuo con función de densidad $f(x)$, función de supervivencia $S(x)$, función de riesgo $\lambda(x)$ y función tiempo promedio de vida restante $R(x)$. Sea φ una función continua, estrictamente creciente o decreciente, con inversa diferenciable y tal que $Y := \varphi(X)$ sigue siendo un tiempo de vida. Entonces, para valores de y en el rango de valores de $\varphi(X)$,*

1. $F_Y(y) = \begin{cases} F(\varphi^{-1}(y)) & \text{si } \varphi \text{ es creciente,} \\ 1 - F(\varphi^{-1}(y)) & \text{si } \varphi \text{ es decreciente.} \end{cases}$

2. $f_Y(y) = f(\varphi^{-1}(y)) \left| \dfrac{d}{dy} \varphi^{-1}(y) \right|.$

3. $S_Y(y) = \begin{cases} S(\varphi^{-1}(y)) & \text{si } \varphi \text{ es creciente,} \\ 1 - S(\varphi^{-1}(y)) & \text{si } \varphi \text{ es decreciente.} \end{cases}$

4. $\lambda_Y(y) = \begin{cases} \lambda(\varphi^{-1}(y)) \cdot \left| \dfrac{d}{dy}\varphi^{-1}(y) \right| & \text{si } \varphi \text{ es creciente,} \\ \lambda(\varphi^{-1}(y)) \ \dfrac{S(\varphi^{-1}(y))}{1 - S(\varphi^{-1}(y))} \cdot \left| \dfrac{d}{dy}\varphi^{-1}(y) \right| & \text{si } \varphi \text{ es decreciente.} \end{cases}$

5. $R_Y(y) = \begin{cases} R(y) \dfrac{S(y)}{S(\varphi^{-1}(y))} \dfrac{\int_{\varphi^{-1}(y)}^{\varphi^{-1}(\infty)} S(v)\,\varphi'(v)\,dv}{\int_y^\infty S(u)\,du} & \text{si } \varphi \text{ es creciente,} \\ R(y) \dfrac{S(y)}{1 - S(\varphi^{-1}(y))} \dfrac{\int_{\varphi^{-1}(y)}^{\varphi^{-1}(\infty)} [1 - S(v)]\,\varphi'(v)\,dv}{\int_y^\infty S(u)\,du} & \text{si } \varphi \text{ es decreciente.} \end{cases}$

Demostración.

Recordemos que $\varphi(x)$ es creciente si, y sólo si, $\varphi^{-1}(x)$ es creciente. La situación análoga ocurre en el caso decreciente.

1. Cuando $\varphi(x)$ es creciente,

$$F_Y(y) = P(\varphi(X) \leqslant y) = P(X \leqslant \varphi^{-1}(y)) = F(\varphi^{-1}(y)).$$

   En el caso $\varphi(x)$ decreciente,

$$F_Y(y) = P(\varphi(X) \leqslant y) = P(X \geqslant \varphi^{-1}(y)) = 1 - F(\varphi^{-1}(y)).$$

2. Derivando la expresión del inciso anterior se encuentra la fórmula enunciada.

3. Cuando $\varphi(x)$ es creciente,

$$S_Y(y) = P(\varphi(X) > y) = P(X > \varphi^{-1}(y)) = S(\varphi^{-1}(y)).$$

   Cuando $\varphi(x)$ es decreciente,

$$S_Y(y) = P(\varphi(X) > y) = P(X < \varphi^{-1}(y)) = 1 - S(\varphi^{-1}(y)).$$

   Observe que, por la hipótesis de la continuidad de la distribución, se cumple la identidad $P(X < x) = P(X \leqslant x)$. Esto se utiliza en la última igualdad.

4. Supongamos el caso $\varphi(x)$ creciente. Por los resultados anteriores se tiene que

$$\begin{aligned}\lambda_Y(y) &= \frac{f_Y(y)}{S_Y(y)} \\ &= \frac{f(\varphi^{-1}(y))}{S(\varphi^{-1}(y))} \left| \frac{d}{dy}\varphi^{-1}(y) \right| \\ &= \lambda(\varphi^{-1}(y)) \left| \frac{d}{dy}\varphi^{-1}(y) \right|.\end{aligned}$$

El caso $\varphi(x)$ decreciente es similar,

$$\begin{aligned}\lambda_Y(y) &= \frac{f_Y(y)}{S_Y(y)} \\ &= \frac{f(\varphi^{-1}(y))}{1 - S(\varphi^{-1}(y))} \left| \frac{d}{dy}\varphi^{-1}(y) \right| \\ &= \lambda(\varphi^{-1}(y)) \frac{S(\varphi^{-1}(y))}{1 - S(\varphi^{-1}(y))} \left| \frac{d}{dy}\varphi^{-1}(y) \right|.\end{aligned}$$

5. Cuando $\varphi(x)$ es creciente y diferenciable,

$$\begin{aligned}R_Y(y) &= \frac{1}{S_Y(y)} \int_y^\infty S_Y(u)\, du \\ &= \frac{1}{S(\varphi^{-1}(y))} \int_y^\infty S(\varphi^{-1}(u))\, du \\ &= \frac{1}{S(\varphi^{-1}(y))} \int_{\varphi^{-1}(y)}^{\varphi^{-1}(\infty)} S(v)\, \varphi'(v)\, dv, \quad v := \varphi^{-1}(u) \\ &= R(y) \frac{S(y)}{S(\varphi^{-1}(y))} \frac{\int_{\varphi^{-1}(y)}^{\varphi^{-1}(\infty)} S(v)\, \varphi'(v)\, dv}{\int_y^\infty S(u)\, du}.\end{aligned}$$

En el caso $\varphi(x)$ decreciente y diferenciable,

$$\begin{aligned} R_Y(y) &= \frac{1}{S_Y(y)} \int_y^\infty S_Y(u)\, du \\ &= \frac{1}{1 - S(\varphi^{-1}(y))} \int_y^\infty [1 - S(\varphi^{-1}(u))]\, du \\ &= \frac{1}{1 - S(\varphi^{-1}(y))} \int_{\varphi^{-1}(y)}^{\varphi^{-1}(\infty)} [1 - S(v)]\, \varphi'(v)\, dv, \quad v := \varphi^{-1}(u) \\ &= R(y)\, \frac{S(y)}{1 - S(\varphi^{-1}(y))}\, \frac{\int_{\varphi^{-1}(y)}^{\varphi^{-1}(\infty)} [1 - S(v)]\, \varphi'(v)\, dv}{\int_y^\infty S(u)\, du}. \end{aligned}$$

■

Ejemplo A.14 (Tranformación cuadrática)
Sea X un tiempo de vida continuo con función de distribución $F(x)$ y función de densidad $f(x)$. Considere la transformación creciente $\varphi(x) = x^2$, para $x > 0$. Usando probabilidad elemental puede comprobarse que la distribución de la variable aleatoria $Y := \varphi(X) = X^2$ es

$$\begin{aligned} F_Y(y) &= F(\sqrt{y}), \quad y > 0, \\ f_Y(y) &= \frac{1}{2\sqrt{y}}\, f(\sqrt{y}), \quad y > 0. \end{aligned}$$

Suponga que X tiene función de supervivencia $S(x)$, función de riesgo $\lambda(x)$ y función tiempo promedio de vida restante $R(x)$. Usando las fórmulas de la proposición anterior puede comprobarse que el tiempo de vida Y es tal que

$$\begin{aligned} S_Y(y) &= S(\sqrt{y}), \quad y > 0, \\ \lambda_Y(y) &= \frac{1}{2\sqrt{y}} \cdot \lambda(\sqrt{y}), \quad y > 0, \\ R_Y(y) &= \frac{1}{S(\sqrt{y})} \int_{\sqrt{y}}^\infty 2v\, S(v)\, dv, \quad \textit{para } y \textit{ tal que } S(\sqrt{y}) > 0. \end{aligned}$$

■

El tema de transformaciones de variables aleatorias puede encontrarse en los varios textos de la teoría de la probabilidad como el de A. Gut [25]. En la siguiente sección veremos que es posible encontrar fórmulas aproximadas para la esperanza y la varianza de la transformación de una variable aleatoria.

Ejercicios

196. *Transformación lineal.*
Sea X un tiempo de vida continuo y defina $Y := aX + b$, en donde $a > 0$ y $b \geqslant 0$ son dos constantes. Demuestre que:

a) $F_Y(y) = F((y-b)/a), \quad y > b.$

b) $f_Y(y) = \dfrac{1}{a}\, f((y-b)/a), \quad y > b.$

c) $S_Y(y) = S((y-b)/a), \quad y > b.$

d) $\lambda_Y(y) = \dfrac{1}{a}\, \lambda((y-b)/a), \quad y > b.$

e) $R_Y(y) = \dfrac{a}{S((y-b)/a)} \displaystyle\int_{(y-b)/a}^{\infty} S(u)\, du,$ para $y > b$ tal que $S((y-b)/a) > 0.$

Encuentre explícitamente estas expresiones cuando el tiempo de vida X tiene distribución $\exp(\lambda)$.

197. *Transformación cuadrática.*
Sea X un tiempo de vida continuo y defina $Y := X^2$. Demuestre que:

a) $F_Y(y) = F(\sqrt{y}), \quad y > 0.$

b) $f_Y(y) = \dfrac{1}{2\sqrt{y}}\, f(\sqrt{y}), \quad y > 0.$

c) $S_Y(y) = S(\sqrt{y}), \quad y > 0.$

d) $\lambda_Y(y) = \dfrac{1}{2\sqrt{y}}\, \lambda(\sqrt{y}), \quad y > 0.$

e) $R_Y(y) = \dfrac{1}{S(\sqrt{y})} \displaystyle\int_{\sqrt{y}}^{\infty} 2v\, S(v)\, dv,$ para y tal que $S(\sqrt{y}) > 0.$

Estas fórmulas aparecen en el Ejemplo A.14. Encuentre explícitamente estas expresiones cuando el tiempo de vida X tiene distribución $\exp(\lambda)$.

198. *Transformación raíz cuadrada.*
Sea X un tiempo de vida continuo y defina $Y := \sqrt{X}$. Demuestre que:

a) $F_Y(y) = F(y^2), \quad y > 0.$

b) $f_Y(y) = 2y\, f(y^2), \quad y > 0.$

c) $S_Y(y) = S(y^2), \quad y > 0.$

d) $\lambda_Y(y) = 2y\, \lambda(y^2), \quad y > 0.$

e) $R_Y(y) = \dfrac{1}{S(y^2)} \displaystyle\int_{y^2}^{\infty} S(u)\, \frac{1}{2\sqrt{u}}\, du,$ para y tal que $S(y^2) > 0.$

Encuentre explícitamente las expresiones anteriores cuando X tiene distribución $\exp(\lambda)$.

199. *Transformación inverso multiplicativo.*
Sea X un tiempo de vida continuo y defina $Y := 1/X$. Demuestre que:

a) $F_Y(y) = 1 - F(1/y), \quad y > 0.$

b) $f_Y(y) = \dfrac{1}{y^2}\, f(1/y), \quad y > 0.$

c) $S_Y(y) = 1 - S(1/y), \quad y > 0.$

d) $\lambda_Y(y) = \dfrac{1}{y^2}\, \dfrac{S(1/y)}{1 - S(1/y)}\, \lambda(1/y), \quad y > 0.$

e) $R_Y(y) = \dfrac{1}{1 - S(1/y)} \displaystyle\int_0^{1/y} \frac{1 - S(u)}{u^2}\, du,$ para y tal que $S(1/y) < 1.$

Encuentre explícitamente las expresiones anteriores cuando X tiene distribución $\exp(\lambda)$.

A.2. Método delta

Se le denomina con este nombre al procedimiento que explicaremos abajo y que consiste en aproximar la media y varianza de una transformación de una variable aleatoria. Estas aproximaciones se utilizan en el presente trabajo en

la derivación de la Fórmula de Greenwood que aparece en la página 186. Sea $Y = \varphi(X)$ una transformación de una variable aleatoria X. En general, no existe una forma sencilla de calcular la esperanza y varianza de Y en términos de estas cantidades para X. El método delta, también llamado método de diferenciales estadísticas, propone un procedimiento para aproximar $E(Y)$ y $\text{Var}(Y)$. El resultado es el siguiente.

Proposición A.2 (Método delta) *Sea X una variable aleatoria con esperanza μ y varianza σ^2, ambas finitas. Sea Y la variable aleatoria definida por la transformación $Y = \varphi(X)$, en donde φ es una función infinitamente diferenciable. Entonces*

1. $E(Y) \approx \varphi(\mu) + \frac{1}{2}\,\varphi''(\mu)\,\sigma^2$.
2. $Var(Y) \approx (\varphi'(\mu))^2\,\sigma^2$.

Demostración.

La serie de Taylor de la función $\varphi(x)$ alrededor del punto $x = a$ es

$$\varphi(x) = \varphi(a) + (x-a)\varphi'(a) + \frac{1}{2}(x-a)^2\varphi''(a) + \cdots$$

Tomando $a = \mu$ y evaluando en la variable aleatoria X se obtiene la siguiente igualdad de variables aleatorias en el sentido puntual

$$\varphi(X) = \varphi(\mu) + (X-\mu)\varphi'(\mu) + \frac{1}{2}(X-\mu)^2\varphi''(\mu) + \cdots \qquad (13)$$

Recortando la serie (13) a los primeros tres sumandos y tomando esperanza se obtiene la aproximación para $E(Y)$. Si ahora el recorte incluye sólo a los dos primeros sumandos y se toma varianza se tiene que

$$\begin{aligned} \text{Var}(Y) &\approx \text{Var}[\varphi(\mu) + (X-\mu)\varphi'(\mu)] \\ &= \text{Var}[(X-\mu)\varphi'(\mu)] \\ &= (\varphi'(\mu))^2\sigma^2. \end{aligned}$$

■

Es importante señalar que las aproximaciones encontradas pueden no ser buenas en términos generales. Se trata, simplemente, de aproximaciones muy

gruesas de cantidades para las cuales no se conoce una fórmula exacta. Sin embargo, pueden existir condiciones bajo las cuales la calidad de la aproximación mejora, por ejemplo, cuando σ^2 es pequeña y la derivada $\varphi''(x)$ es acotada, la aproximación para $E(Y)$ es buena.

Observemos además que, para la validez de las fórmulas anteriores, no se ha supuesto que X ó Y sean tiempos de vida. Estas variables aleatorias pueden ser cualesquiera que satisfagan las condiciones de la proposición. Veamos un ejemplo.

Ejemplo A.15 *Supongamos que X es una variable aleatoria con distribución exp(λ). El n-ésimo momento de X es $E(X^n) = n!/\lambda^n$. Consideremos la transformación $\varphi(x) = x^2$. Entonces la esperanza y varianza de $\varphi(X) = X^2$ se pueden calcular con exactitud y son:*

$$\begin{aligned} E(X^2) &= 2/\lambda^2, \\ Var(X^2) &= E(X^4) - E^2(X^2) = 24/\lambda^4 - (2/\lambda^2)^2 = 20/\lambda^4. \end{aligned}$$

Por otro lado, las fórmulas de la proposición anterior arrojan las siguientes aproximaciones:

$$\begin{aligned} E(X^2) &\approx 2/\lambda^2, \\ Var(X^2) &\approx 4/\lambda^4. \end{aligned}$$

■

Las aproximaciones demostradas en el método delta pueden extenderse al caso de funciones de dos o más variables aleatorias. Estudiaremos a continuación el caso bidimensional.

Proposición A.3 (Método delta, dos dimensiones) *Sean X_1 y X_2 dos variables aleatorias con esperanzas μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2 finitas, respectivamente. Sea $Y = \varphi(X_1, X_2)$, en donde φ es una función de dos variables infinitamente diferenciable. Entonces*

1. $$E(Y) \approx \varphi(\mu_1, \mu_2) + \frac{1}{2}\frac{\partial^2 \varphi}{\partial x_1^2}(\mu_1, \mu_2)\,\sigma_1^2 + \frac{1}{2}\frac{\partial^2 \varphi}{\partial x_2^2}(\mu_1, \mu_2)\,\sigma_2^2 + \frac{\partial^2 \varphi}{\partial x_2 \partial x_1}(\mu_1, \mu_2)\, Cov(X_1, X_2).$$

2. $Var(Y) \approx [\frac{\partial \varphi}{\partial x_1}(\mu_1, \mu_2)]^2 \sigma_1^2 + [\frac{\partial \varphi}{\partial x_2}(\mu_1, \mu_2)]^2 \sigma_2^2$
$+ 2\,[\frac{\partial \varphi}{\partial x_1}(\mu_1, \mu_2)]\,[\frac{\partial \varphi}{\partial x_2}(\mu_1, \mu_2)]\, Cov(X_1, X_2).$

Demostración.

La comprobación se basa nuevamente en la serie de Taylor de la función $\varphi(x, y)$ en el punto (μ_1, μ_2) y evaluada en el vector aleatorio (X_1, X_2). Los primeros términos de esta serie son

$$\begin{aligned}
\varphi(X_1, X_2) &= \varphi(\mu_1, \mu_2) \\
&\quad + (X_1 - \mu_1)\, \frac{\partial \varphi}{\partial x_1}(\mu_1, \mu_2) + (X_2 - \mu_2)\, \frac{\partial \varphi}{\partial x_2}(\mu_1, \mu_2) \\
&\quad + \frac{1}{2}(X_1 - \mu_1)^2\, \frac{\partial^2 \varphi}{\partial x_1^2}(\mu_1, \mu_2) + \frac{1}{2}(X_2 - \mu_2)^2\, \frac{\partial^2 \varphi}{\partial x_2^2}(\mu_1, \mu_2) \\
&\quad + (X_1 - \mu_1)(X_2 - \mu_2)\, \frac{\partial^2 \varphi}{\partial x_2 \partial x_1}(\mu_1, \mu_2) + \cdots \qquad (14)
\end{aligned}$$

1. Tomando esperanza de cada uno de los términos que aparecen de manera explícita en la expansión (14) se obtiene la aproximación enunciada para $E(Y)$.

2. Recortando la serie (14) a los primeros tres sumandos, éstos corresponden a la primera línea de la identidad (14), y tomando varianza, se tiene que

$$\begin{aligned}
\text{Var}(Y) &\approx \text{Var}[(X_1 - \mu_1)\, \frac{\partial \varphi}{\partial x_1}(\mu_1, \mu_2) + (X_2 - \mu_2)\, \frac{\partial \varphi}{\partial x_2}(\mu_1, \mu_2)] \\
&= [\frac{\partial \varphi}{\partial x_1}(\mu_1, \mu_2)]^2 \sigma_1^2 + [\frac{\partial \varphi}{\partial x_2}(\mu_1, \mu_2)]^2 \sigma_2^2 \\
&\quad + 2\,[\frac{\partial \varphi}{\partial x_1}(\mu_1, \mu_2)]\,[\frac{\partial \varphi}{\partial x_2}(\mu_1, \mu_2)]\, \text{Cov}(X_1, X_2).
\end{aligned}$$

■

Se puede encontrar mayor información matemática e histórica acerca del método delta en el artículo de Jay M. Ver Hoef [27].

Ejercicios

200. *Transformación lineal.*
Sea X una variable aleatoria con esperanza y varianza finitas. Considere la transformación $Y = aX + b$, en donde a y b son dos constantes. Compruebe que las expresiones para calcular de manera aproximada la media y varianza de Y por el método delta producen los valores exactos de estas cantidades.

201. Sea X un tiempo de vida con distribución unif$(0, 1)$ y defina $Y = X^2$.

    a) Encuentre $E(Y)$ y $\text{Var}(Y)$ de manera exacta.

    b) Aplique el método delta para encontrar $E(Y)$ y $\text{Var}(Y)$ de manera aproximada.

202. Sea X un tiempo de vida con distribución unif$(0, 1)$. Sea $\lambda > 0$ una constante y defina la función $\varphi(x) = (-1/\lambda) \ln x$, para $x \in (0, 1)$. Sea $Y := \varphi(X)$.

    a) Use el método delta para encontrar valores aproximados para $E(Y)$ y $\text{Var}(Y)$.

    b) Demuestre que $Y \sim \exp(\lambda)$ y encuentre los valores exactos de la media y varianza de Y.

203. Sea X una variable aleatoria con distribución $\text{N}(\mu, \sigma^2)$ y considere la transformación $Y = e^X$.

    a) Encuentre $E(Y)$ y $\text{Var}(Y)$.

    b) Aplique el método delta para encontrar $E(Y)$ y $\text{Var}(Y)$ de manera aproximada.

    Si $E_{ap}(Y)$ y $\text{Var}_{ap}(Y)$ denotan la esperanza y varianza aproximadas de Y por el método delta, demuestre que:

    c) $E(Y) > E_{ap}(Y)$.

    d) $\text{Var}(Y) > \text{Var}_{ap}(Y)$.

204. Sean X_1 y X_2 dos variables aleatorias independientes con idéntica distribución unif$(0, 1)$. Defina $Y := X_1 \cdot X_2$.

a) Use el método delta para encontrar valores aproximados para $E(Y)$ y $\text{Var}(Y)$.

b) Encuentre los valores exactos de $E(Y)$ y $\text{Var}(Y)$.

Bibliografía

[1] Abramowitz M., Stegun I. A. (Editors) *Handbook of mathematical functions with formulas, graphs and mathematical tables.* National Bureau of Standards, Department of Commerce, USA, 1972.

[2] Aalen O. O., Borgan Ø., Gjessing H. K. *Survival and event history analysis: a process point of view.* Springer, 2008.

[3] Altman D. G. *Practical statistics for medical research.* Chapman & Hall/CRC, 1999.

[4] Andersen P. K., Borgan Ø., Gill R. D., Keiding N. *Statistical models based on counting processes.* Springer-Verlag, 1993.

[5] Balakrishnan N., Nevzorov V. B. *A primer on statistical distributions.* Wiley, 2003.

[6] Bogaerts K., Komárek A., Lesaffre E. *Survival analysis with interval-censored data: a practical approach with examples in* `R`, `SAS`, *and* `BUGS`. Chapman & Hall/CRC, 2018.

[7] Breslow N. E., Day N. E. *Statistical Methods in Cancer Research*, 2, The design and Analysis of Cohort Studies, IARC, 1987.

[8] Box-Steffensmeier J., Jones B. S. *Event history modeling: a guide to social scientists.* Cambridge University Press, 2004.

[9] Bowers N. L. Jr., Gerber H. U., Hickman J. C., Jones D. A., Nesbitt C. J. *Actuarial mathematics.* The Society of Actuaries, 1997. Second edition.

[10] Broström G. *Event history analysis with* R. CRC Press, 2012.

[11] Cantor A. B. SAS *survival analysis techniques for medical research.* SAS Publishing, 2003. Second edition.

[12] Cleves M., Gutierrez R. G., Gould W., Marchenko Y. V. *An introduction to survival analysis using* STATA. Stata Press, 2010. Third edition.

[13] Clifford Cohen A. *Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples.* Technometrics Vol. 7, No. 4 November, 1965.

[14] Coghlan A. *A little book of* R *for biomedical statistics.* Release 0.2, December 17, 2016.

[15] Collet D. *Modelling survival data in medical research.* Chapman & Hall/CRC, 2003. Second edition.

[16] Cox D. R., Oakes D. *Analysis of survival data.* Chapman & Hall/CRC, 1998.

[17] Cunningham R. J., Herzog T. N., London R. L. *Models for quantifying risk.* ACTEX Publications, 2006. Second edition.

[18] Dickson D. C. M., Hardy M. R., Waters H. R. *Actuarial mathematics for life contingent risks.* Cambridge University Press, 2009

[19] Elandt-Johnson R. C., Johnson N. L. *Survival models and data analysis.* John Wiley & Sons, 1980.

[20] Fleming T. R., Harrington D. P. *Counting processes and survival analysis.* Wiley-interscience, 1991.

[21] Forbes C., Evans M., Hastings N., Peacock B. *Statistical distributions.* Wiley, 2011, Fourth edition.

[22] Grambsch P. M., Therneau T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 3, pp. 515-526, 1994.

[23] Guess F., Proschan F. *Mean residual life: theory and applications.* In *Handbook of Statistics, Vol. 7, Quality control and reliability.* Krishnaiah P. R., Rao C. R. (Editors) Elsevier Science Publishers B. V., 1988.

[24] Guo S. *Survival analysis.* Oxford University Press, 2010.

[25] Gut A. *Probability: a graduate course.* Springer, 2013. Second edition.

[26] Helsel D. R. *Statistics for censored environmental data using MINITAB and R.* Wiley, 2012. Second edition.

[27] Hoef J. M. Who invented the delta method? *The American Statistician*, 66(2), pp. 124–127, 2012.

[28] Hosmer D. W., Lemeshow S., May S. *Applied survival analysis: regression modeling of time-to-event data.* Wiley-Interscience, 2008. Second edition.

[29] Johnson N. L., Kotz S., Balakrishnan N. *Continuous univariate distributions, Vol. 2,* John Wiley & Sons, 1995. Second edition.

[30] Kalbfleisch J. D., Prentice R. L. *The statistical analysis of failure time data.* Wiley-Interscience, 2002. Second edition.

[31] Kaplan E. L., Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association (JASA)*, 53, pp. 457–481, 1958.

[32] Keyfitz N, Caswell H. *Applied mathematical demography.* Springer, 2005. Third edition.

[33] Klein J. P., Moeschberger M. L. *Survival analysis: techniques for censored and truncated data.* Springer, 2003. Second edition.

[34] Kleinbaum D. G., Klein M. *Survival analysis: a self learning text.* Springer, 2012. Third edition.

[35] Korosteleva O. *Clinical statistics: introducing clinical trials, survival analysis, and longitudinal data analysis.* Jones & Bartlett Publishers, 2008.

[36] Lawless J. F. *Statistical models and methods for lifetime data.* Wiley-Interscience, 2003. Second edition.

[37] Lee E. T., Wang J. W. *Statistical methods for survival data analysis.* Wiley-Interscience, 2003.

[38] Leguina J. *Fundamentos de demografía.* Siglo Veintiuno Editores, 1981. Tercera edición.

[39] London, D. *Survival models and their estimation.* ACTEX Publications, 1997.

[40] Li J., Ma S. *Survival analysis in medicine and genetics.* Chapman & hall/CRC, 2013.

[41] Lin D. Y. On the Breslow estimator. *Lifetime Data Analysis*, **13**, pp. 471–480, 2007.

[42] Liu X. *Survival analysis: models and applications.* Wiley, 2012.

[43] Machin D., Cheung Y. B., Parmar M. K. B. *Survival analysis: a practical approach.* John Wiley & Sons, 2006. Second edition.

[44] Moore D. F. *Applied survival analysis using* `R`. Springer, 2016.

[45] Nag A. *Survival analysis with Python.* CRC Press, 2022.

[46] Promislow S. D. *Fundamentals of actuarial mathematics.* Wiley, 2011.

[47] Smith P. J. *Analysis of failure and survival data.* CRC Press, 2002.

[48] STHDA, Statistical tools for high-throughtput data analysis, accedido el 23 de noviembre de 2023, <http://www.sthda.com/english/wiki/cox-proportional-hazards-model>.

[49] Therneau T. M., Grambsch P. M. *Modeling survival data: extending the Cox model.* Springer, 2000.

[50] Thomas D. C. Use of auxiliary information in fitting nonproportional hazards models. in *Modern Statistical Methods in Chronic Disease Epidemiology.* Ed. S. H. Moolgavkar and R. L. Prentice. Wiley, 1986.

[51] Tutz G., Schmid M. *Modeling discrete time-to-event data.* Springer, 2016.

[52] Venables W. N., Ripley B. D. *Modern applied statistics with* `S`. Springer, 2002. Fourth edition.

[53] Villanueva N. M., Sestelo M., Meira-Machado L. An R package for determining groups in multiple survival curves. *Proceedings of the 33rd International Workshop on Statistical Modelling (IWSM)*, Vol. 2, pp. 198-203, University of Bristol, UK, 16-20 July 2018.

[54] Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3), pp. 426-482, 1943.

Índice alfabético

Censura, 9
  aleatoria, 20
  por intervalo, 19
  por la derecha, 10
  por la izquierda, 17
  tipo I, 13
  tipo I progresiva, 14
  tipo II, 15
  tipo II progresiva, 16
Cocientes de riesgos (HR), 234
Cohorte, 139
Covariables, 223
Cox
  modelo de, 228
Cuantil
  de un tiempo de vida, 85
  del tiempo de vida restante, 91
Curva
  de supervivencia, 38

Datos
  hoel, 261
  larynx, 259
  lung, 248
  veteran, 258
  censurados, 9
  de supervivencia, 9
  faltantes, 10
  incompletos, 10
Distribución
  χ^2, 107
  bilateralmente truncada, 24
  de valor extremo, 129
  doblemente truncada, 24
  Erlang, 107
  exponencial, 101
  gama, 106
  Gompertz, 103, 127
  lineal-exponencial, 86
  lognormal, 108
  Makeham, 104
  Pareto, 109
  Rayleigh, 130
  truncada, 24
    por abajo, 24
    por arriba, 24
    por la derecha, 24
    por la izquierda, 24
  uniforme continua, 100
  Weibull, 85, 105

Estimador
actuarial, 160
de Kaplan-Meier, 167
de Nelson-Aalen, 190

Force of mortality, 47
Fuerza de mortalidad, 47
constante, 84, 102
Función
de densidad truncada, 25
de distribución, 145
de distribución empírica, 161
de distribución truncada, 25
de falla, 47
de riesgo, 46, 47, 146
de riesgo acumulado, 52, 57
de riesgo base, 226
de riesgo constante, 48, 54, 84
de riesgo continua, 47
de riesgo discreta, 53
de riesgo exponencial, 50, 83
de riesgo Rayleigh, 49
de riesgo subyacente, 226
de riesgo truncada, 58
de supervivencia, 38, 145
de supervivencia empírica, 162, 165
de tasa de intensidad, 47
de tasa de riesgo, 47
de verosimilitud parcial, 239
gama, 106
gama incompleta, 107
número de fallecidos, 141
número de personas vivas, 141
probabilidad de fallecer, 142
probabilidad de sobrevivir, 143
tiempo medio de vida residual o restante, 60, 147
tiempo promedio de vida restante, 64
Fórmula
de Greenwood, 154, 186
del producto, 44, 144

Gompertz
distribución de, 103
Greenwood
fórmula de, 154, 186

Hazard rate, 47
i.e. función de riesgo, 47
HR *Hazard Ratios*, 234

Intensity rate, 47
i.e. función de riesgo, 47
Interpolación, 155
exponencial, 157
hiperbólica, 158
lineal, 155

Kaplan-Meier
estimador de, 167

Límite
por la derecha, 37
por la izquierda, 37

Makeham
distribución de, 104
Mean Residual Lifetime, 65
Media

de la transf. de una v.a., 278
Modelo
–s no paramétricos, 139
aditivo, 226
de Cox, 228
funciones básicas asociadas, 231
interpretación de parámetros, 234
de riesgos proporcionales, 229
de vida acelerada, 227
lineal-exponencial, 227
multiplicativo, 227
MRL
i.e. Mean Residual Lifetime, 65
Método
actuarial, 147
de diferenciales estadísticas, 279
delta, 278

Nelson-Aalen
estimador de , 190
Número en riesgo, 148

Pareto
distribución de, 109
Producto
fórmula del, 44
Prueba
de significancia de Wald, 263
para determinar proporcionalidad, 264
Prueba log-rank
con ponderaciones, 206
estratificada, 206
para varios grupos, 207

Radix, 140
Rayleigh
distribución, 130
Función de riesgo de, 49
Riesgos proporcionales
modelo de, 229

Supervivencia
curva de, 38
función de, 38, 145

Tabla
de mortalidad, 139
Tasa
de intensidad, 47
de riesgo, 47
Tiempo
–s de falla, 8
–s de vida, 7
de vida residual, 60
de vida restante, 60
Transformación
de una v.a., 273
Truncamiento, 21
por la derecha, 23
por la izquierda, 22

Variables
concomitantes, 224
prognósticas, 224
Varianza
de la transf. de una v.a., 278

Weibull
distribución de, 85, 105

Análisis de supervivencia: conceptos y modelos básicos
fue editado por la Facultad de Ciencias
de la Universidad Nacional Autónoma de México.
Se terminó de editar el 20 de marzo de 2025.

Publicación electrónica.

El cuidado de la edición estuvo a cargo de
Leticia Pacheco Gasca.