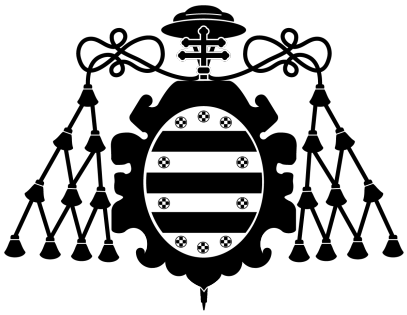


Universidad de Oviedo

Facultad de Ciencias

Grado en Matemáticas

Trabajo Fin de Grado

Técnicas básicas del Análisis de Supervivencia y sus aplicaciones

Autor: Fernando de la Puente Alonso de la Torre

Supervisado por:

Beatriz Sinova Fernández y Sonia Pérez Fernández

Codirigido por:

Emilio Esteban González

Curso 2023-2024

Índice general

Prólogo

Una cuestión de tiempo...

> *"Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution. It is, strictly speaking, a real factor in scientific research."*
>
> – Albert Einstein, *Cosmic Religion and Other Opinions and Aphorisms*

Es frecuente en los estudios médicos, entre otros, que el tiempo hasta que ocurre cierto evento sea una variable de gran interés. Lo más habitual es encontrar ejemplos en los que dicho evento es desfavorable, como puede ser la muerte, la ocurrencia de un infarto o la recidiva de un cáncer. Sin embargo, si el evento es la curación o la reincorporación al mundo laboral tras haber padecido una enfermedad, este tiene una connotación positiva. Centrándonos en el ámbito clínico, el diseño de este tipo de estudios consiste en hacer un seguimiento a lo largo del tiempo a determinados pacientes, con el fin de medir la variable *tiempo hasta la ocurrencia del evento en estudio*. Este seguimiento se realiza desde un instante inicial que puede venir marcado, por ejemplo, por el comienzo de un tratamiento o una intervención quirúrgica, hasta un momento final en el que se da por concluida la recogida de datos. Sin embargo, en el abordaje del análisis estadístico de la variable continua *tiempo hasta evento* es habitual encontrarse con dos dificultades principales que impiden hacer uso de métodos habituales para este tipo de variables. Por un lado, esta variable suele presentar una distribución asimétrica, lejos de la distribución normal que se asume en muchos métodos comúnmente empleados en Estadística. Por otro lado y marcando el rasgo distintivo de este tipo de análisis, en general ocurre que al final del período de seguimiento, existen pacientes para los que no se ha podido observar el evento, lo cual conlleva un desconocimiento del valor de la variable *tiempo hasta evento* de esos individuos. Pero se trata de un desconocimiento parcial, ya que sí se sabe que dicho valor es mayor o igual que el tiempo observado en su último registro. Este fenómeno se denomina censura por la derecha y es la más frecuente, aunque existen otros tipos de censura como veremos en este trabajo. Estas particularidades hacen necesario el desarrollo de análisis propios para abordar estos estudios y de ahí surge el denominado análisis de supervivencia, que trata de aprovechar toda la información recogida en el seguimiento de los individuos.

Capítulo 1

Preliminares

El análisis de supervivencia constituye un área particular de la estadística que analiza la variable 'tiempo hasta la ocurrencia de un evento de interés" o también denominada 'tiempo de supervivencia'. El hecho de que esta variable aleatoria no se mida instantáneamente, sino que sea preciso hacer un seguimiento de los individuos a lo largo de un período de tiempo, causa que, en ocasiones, aparezcan los denominados datos censurados. La existencia de datos censurados es, justamente, el aspecto diferencial más importante del análisis de supervivencia, que hace necesario el desarrollo de métodos particulares para abordar este tipo de estudios. Debido a la naturaleza de las variables se presentan los siguientes inconvenientes.

- Generalmente, la distribución de los tiempos de supervivencia tiene sesgo positivo (asimetría por la derecha o cola derecha 'pesada'), por lo que técnicas estadísticas que requieran normalidad (ANOVA, test t, etc.) estarían descartadas *a priori*. Sin embargo, podríamos eliminar este sesgo haciendo una transformación que normalice la variable tiempo de vida, T, por ejemplo, $\log T$. El problema que puede surgir entonces es la falta de interpretación de los resultados, sobre todo en términos relativos.
- El segundo motivo se debe a la presencia de observaciones censuradas (información parcial), una de las características más importantes de los datos de supervivencia, tanto es así que incluso cálculos simples en el análisis descriptivo requieren procesos mas complejos que los habituales. Por ejemplo, no se deben dibujar histogramas, pues no sabríamos cómo poner los valores censurados.

1.1. Reseña histórica

Los primeros pasos relevantes del análisis de supervivencia pertenecen al campo de la demografía y fueron dados por John Graunt y William Petty (Figura 1.1), nacidos en Hampshire (Inglaterra).

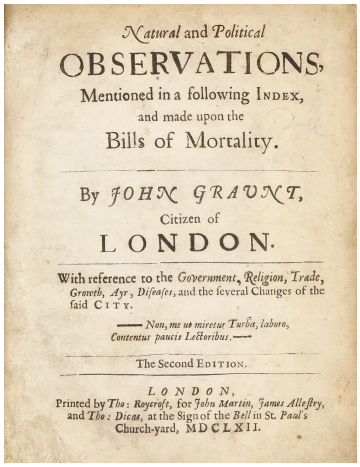
Les describen como 'los pioneros no solo de la estadística médica y de la estadística demográfica, sino de los métodos numéricos aplicados a fenómenos de la sociedad humana' ([12], pág. 2), por sus grandes contribuciones en la epidemiología, la demografía, la econometría y la economía [36].

(a) John Graunt (1620)

(b) William Petty (1623)

Figura 1.1: Estadísticos del siglo XVII que fueron pioneros en el análisis de supervivencia. **(a)** Autor: Terence O'Donnell. Licencia: Wikimedia Commons. **(b)** Autor desconocido. Licencia: © National Portrait Gallery, London.

Debido a las epidemias de peste que tuvieron lugar en Londres durante el siglo XVI, se ideó un sistema que contabilizara los entierros semanales que se producían en cada parroquia: los boletines de mortalidad (*Bills of Mortality*) (Figura 1.2). A partir de 1592 este sistema se hizo regular. Así, a principios del siglo XVII se empiezan a imprimir los primeros boletines y se comienzan a incluir fallecimientos por causas de muerte diferentes a la de la peste y el número de bautizos de cada parroquia [13].



Natural and *Political*

OBSERVATIONS,

Mentioned in a following INDEX,
and made upon the
Bills of Mortality.

By JOHN GRAUNT,
Citizen of
LONDON.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Ayr*, *Diseases*, and the several Changes of the said CITY.

—— *Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus. ——

The Second EDITION.

LONDON,
Printed by *Tho: Roycroft*, for *John Martin*, *James Allestry*, and *Tho: Dicas*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXII.

Figura 1.2: Portada del libro [11]. Fuente: Milestone Books

La actividad social de Graunt en la ciudad de Londres le permitió acceder a estos boletines, que fueron la base documental sobre la que estableció sus investigaciones estadísticas, actuariales y demográficas.

Graunt analizó todos los boletines anuales de 1604 a 1660, utilizando ocasionalmente boletines semanales para cuestiones o etapas concretas. Los datos que usa son, fundamentalmente, recuentos de entierros y bautizos por parroquias, en los que basa la mayor parte de sus deducciones.

Se le atribuye a Graunt la creación de la primera tabla de vida (o tabla de mortalidad), que expresaba las probabilidades de supervivencia para cada edad. Aunque hay sospechas sobre si la autoría le corresponde a Petty, pues se tiene registro de su participación en la construcción de tablas de mortalidad por edad de ocurrencia. Tanta fue la implicación de Petty que hasta bautizó la forma de manejar la información poblacional como 'aritmética política' [33].

1.2. Conceptos básicos

A continuación, se introducen una serie de conceptos básicos imprescindibles para seguir la formalización matemática de desarrollos propios del análisis de supervivencia incluidos en esta memoria. Muchos de estos conceptos forman parte del temario de asignaturas del Grado en Matemáticas, particularmente Estadística Descriptiva y Probabilidad de primer curso y Probabilidades y Estadística de segundo curso [9].

Definición 1.1 (σ-álgebra)**.** Sea Ω un conjunto no vacío. Una **σ-álgebra** $\mathcal{A}$ sobre Ω es una clase de subconjuntos de Ω que verifica

(1) $\Omega \in \mathcal{A}$.

(2) Si $A \in \mathcal{A}$, entonces $A^c \in \mathcal{A}$.

(3) Si $\{A_n\}_{n\in\mathbb{N}} \in \mathcal{A}$, entonces $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

Definición 1.2 (Espacio medible)**.** Sea $\mathcal{A}$ una σ-álgebra sobre Ω. El par $(\Omega, \mathcal{A})$ se dice **espacio medible**.

Definición 1.3 (σ-álgebra generada por una clase de conjuntos)**.** Dada una clase de conjuntos arbitraria, $\mathcal{C} \subseteq \mathcal{P}(\Omega)$. Se define la **$\sigma$-álgebra generada por $\mathcal{C}$** como la intersección de todas las σ-álgebras que contienen a $\mathcal{C}$. Es decir, es la mínima clase, $\sigma(\mathcal{C}) \subseteq \mathcal{P}(\Omega)$, con estructura de σ-álgebra, que contiene a $\mathcal{C}$.

Definición 1.4 (σ-álgebra de Borel en $\mathbb{R}$)**.** Se define la **σ-álgebra de Borel en $\mathbb{R}$** como la generada por la clase de todos los intervalos; esto es, si denotamos $\mathcal{I} \subset \mathcal{P}(\mathbb{R})$ a la clase de todos los intervalos de $\mathbb{R}$, la σ-álgebra de Borel de $\mathbb{R}$ se define como

$$\mathcal{B}_{\mathbb{R}} = \sigma(\mathcal{I}).$$

Definición 1.5 (Función medible)**.** Sean $(\Omega_1, \mathcal{A}), (\Omega_2, \mathcal{B})$ espacios medibles (probabilizables), y $f : \Omega_1 \to \Omega_2$ una aplicación. Se dice que f es una **función medible** con respecto a $\mathcal{A}$ y $\mathcal{B}$ (o simplemente se dice que es medible) si verifica

$$f^{-1}(B) \in \mathcal{A}, \quad \forall B \in \mathcal{B},$$

o más generalmente, $f^{-1}(\mathcal{B}) \subseteq \mathcal{A}$. La función f también se denomina **elemento aleatorio de Ω_2** si $(\Omega_1, \mathcal{A}, P)$ es un espacio de probabilidad.

Observación 1.1. En Teoría de Probabilidad se dice que f es una **variable aleatoria**, aunque algunos textos reservan este término sólo cuando $\Omega_2 = \mathbb{R}$, usando en su lugar 'elemento aleatorio de Ω_2'.

Definición 1.6 (Función Borel-medible)**.** Una función $f : \Omega \to \mathbb{R}^n$ sobre un espacio medible $(\Omega, \mathcal{A})$ se dice que es **Borel-medible** si f es medible cuando tomamos la σ-álgebra de Borel $\mathcal{B}_{\mathbb{R}^n}$ como σ-álgebra en $\mathbb{R}^n$. Si $\Omega \subseteq \mathbb{R}^k$ y decimos que f es Borel-medible, asumiendo que $\mathcal{A} = \mathcal{B}_{\mathbb{R}^k}$.

Definición 1.7 (Espacio de probabilidad o probabilístico)**.** Dado un espacio probabilizable $(\Omega, \mathcal{A})$, donde Ω es el conjunto de posibles resultados asociados a un experimento aleatorio y $\mathcal{A}$ es la σ-álgebra generada por una clase de subconjuntos de Ω (clase de sucesos de interés asociados al experimento aleatorio), y una probabilidad P asociada a ese espacio, la terna $(\Omega, \mathcal{A}, P)$ se denomina **espacio de probabilidad** o **espacio probabilístico**.

Definición 1.8 (Variable aleatoria)**.** Dado un espacio de probabilidad $(\Omega, \mathcal{A}, P)$, una **variable aleatoria** (v.a.) unidimensional es una aplicación $X : \Omega \to \mathbb{R}$ Borel-medible.

Definición 1.9 (σ-álgebra generada por una variable aleatoria)**.** La menor σ-álgebra respecto a la cual una variable aleatoria X es medible se conoce como la **σ-álgebra generada por X**, y se denota por $\sigma(X)$. Esta σ-álgebra se corresponde con la intersección de todas las σ-álgebras respecto a las cuales X es medible.

Definición 1.10 (Independencia)**.** Sean X e Y variables aleatorias definidas en el espacio probabilístico $(\Omega, \mathcal{A}, P)$. Se dice que X e Y son **independientes** si las σ-álgebras que generan, $\sigma(X)$ y $\sigma(Y)$, son independientes.

Teorema 1.1. *Sean X e Y variables aleatorias independientes, y sean f y g funciones Borel-medibles en $\mathbb{R}$. Entonces, $f(X)$ y $g(Y)$ son variables aleatorias independientes.* □

Definición 1.11. Una variable aleatoria X se dice **discreta** si existe un conjunto $S_X \subset \mathbb{R}$ numerable (finito o infinito numerable) tal que $P(X \in S) = 1$ (o, equivalentemente, $P(X \in \Omega \setminus S_X) = 0$). El conjunto S_X se llama **soporte** de X.

Observación 1.2. El soporte de una variable aleatoria (discreta) es un conjunto formado por los puntos de probabilidad positiva. Más aún, para un conjunto boreliano $B \in \mathcal{B}_{\mathbb{R}}$,

$$P(X \in B) = \sum_{s \in S \cap B} P(X = s).$$

Definición 1.12. Una variable aleatoria se dice **continua** si toma valores de un conjunto infinito no numerable, como puede ser el conjunto de los números reales o un intervalo (de $\mathbb{R}$).

Definición 1.13 (Función de distribución)**.** [9] Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y X una variable aleatoria asociada a él. Se llama **función de distribución** de X a la función de distribución de la medida de probabilidad P_X. Es decir, es una función $F_X : \mathbb{R} \to [0, 1]$ que viene dada por

$$F_X(x) = P_X\big((-\infty, x]\big) = P(X \leq x),$$

donde $P(X \leq x)$ es la notación abreviada que se suele usar para referirse a

$$P\big(\{\omega \in \Omega \mid X(\omega) \leq x\}\big).$$

Proposición 1.2. *Sea F la función de distribución de una variable aleatoria X, entonces se verifican las siguientes propiedades [9]:*

1. $\lim_{x\to\infty} F(x) = 1$.
2. $\lim_{x\to-\infty} F(x) = 0$.
3. *F es no decreciente.*
4. *F es continua por la derecha.*

Definición 1.14 (Función de masa de probabilidad)**.** Sea X una variable aleatoria discreta. Se llama **función de masa de probabilidad** de X a una función $p_X : \mathbb{R} \to [0, 1]$ tal que para cada $k \in \mathbb{R}$,

$$p_X(k) = P(X = k).$$

Proposición 1.3. *Sea X una variable aleatoria discreta con soporte S_X, la función de masa de probabilidad de X verifica las siguientes propiedades:*

- $p_X(x) \geq 0$, *para todo* $x \in \mathbb{R}$. *(no-negatividad)*
- $\sum_{x\in S_X} p_X(x) = 1$. *(normalización)*
- $P(X \in A) = \sum_{x\in A} p_X(x)$, *para todo* $A \subset \mathbb{R}$.

Definición 1.15 (Función de densidad)**.** Sea X una variable aleatoria continua. Se llama **función de densidad** de X a la derivada de la función de distribución de X,

$$f(x) = \frac{d}{dx} F_X(x).$$

Por convenio, suele considerarse que en los puntos en los que F no es derivable, la función de densidad se anula [9].

El siguiente resultado, aunque es consecuencia de la Definición 1.15, se presenta con bastante frecuencia en la literatura como defindión de función de densidad.

Proposición 1.4. *Sea X una variable aleatoria continua, la función de densidad de X es una función integrable $f_X : \mathbb{R} \to \mathbb{R}$ que verifica las siguientes propiedades:*

- $f_X(x) \geq 0$, *para todo* $x \in \mathbb{R}$. *(no-negatividad)*
- $\int_{\mathbb{R}} f_X(x)\, dx = 1$. *(normalización)*
- $P(X \in A) = \int_A f_X(x)\, dx$, *para todo* $A \subset \mathbb{R}$.

Observación 1.3. Cuando X es una variable aleatoria continua, para cada $x \in \mathbb{R}$ se tiene que $P(X = x) = 0$ y, por tanto,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b),$$

para cualesquiera $a, b \in \mathbb{R}$ con $a < b$.

Se definen a continuación dos operadores que usaremos con bastante frecuencia: la esperanza y la varianza de variables aleatorias.

Definición 1.16 (Esperanza)**.** Sea X una variable aleatoria discreta con función de masa de probabilidad p_X e Y una variable aleatoria continua con función de densidad f_Y. Se define la **esperanza** de X y de Y como

$$E[X] = \sum_{k \in S_X} k\, p_X(k) \quad \text{y} \quad E[Y] = \int_{\mathbb{R}} y\, f_Y(y)\, dy.$$

Definición 1.17 (Varianza)**.** Sea X una variable aleatoria, se define la **varianza** de X como

$$\operatorname{Var}[X] = E\Big[\big(X - E[X]\big)^2\Big] = E\left[X^2\right] - E[X]^2,$$

siendo $E\left[X^2\right] = \displaystyle\sum_{k \in S_X} k^2\, p_X(k)$ en el caso discreto y $E\left[X^2\right] = \displaystyle\int_{\mathbb{R}} x^2\, f_X(x)\, dx$ en el caso continuo.

Definición 1.18 (Probabilidad condicionada)**.** Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y A un suceso ($A \in \mathcal{A}$) tal que $P(A) > 0$. Para cualquier otro suceso $B \in \mathcal{A}$, se define la **probabilidad condicionada** de B conocida la ocurrencia de A como

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}.$$

Definición 1.19 (Función cuantil)**.** Sea X una variable aleatoria con función de distribución F y sea $p \in (0, 1)$. Se define la **función cuantil** (o cuasinversa de la función de distribución) como

$$Q(p) = F^{(-1)}(p) = \text{ínf}\,\{x \in \mathbb{R} \mid F(x) \geq p\}.$$

Si F es estrictamente creciente, entonces tiene una única cuasinversa que coincide con la inversa habitual, que denotaremos por F^{-1}.

1.3. Función de supervivencia

Definición 1.20. Llamaremos **tiempo de supervivencia** (o indistintamente, tiempo de vida/fallo) y denotaremos por T a una variable aleatoria continua no negativa que representará el

tiempo hasta la ocurrencia de cierto suceso de interés.

Definición 1.21. Sea T la variable aleatoria tiempo de supervivencia, se define la **función de supervivencia**[1] como

$$S(t) = P\left(T > t\right).$$

Intuitivamente, representa la probabilidad de que un individuo seleccionado al azar 'sobreviva' transcurrido un tiempo t, o de forma más general, la probabilidad de que el suceso de interés no haya ocurrido antes de dicho instante.

Proposición 1.5. *Sea T una variable aleatoria denotando el tiempo de vida, F su función de distribución y S su función de supervivencia. Entonces se cumple: (a) S es monótona no creciente, (b) S es continua por la derecha y (c) $S(t)$ siempre toma valores en $[0, 1]$.*

Demostración. La función de supervivencia se puede expresar en términos de la función de distribución (cuyas propiedades son bien conocidas; véase Proposición 1.2) sin más que considerar el suceso complementario al de su definición, $\{T \leq t\}$. Así,

$$S(t) = 1 - P\big(\{T > t\}^c\big) = 1 - P\left(T \leq t\right) = 1 - F(t). \tag{1.1}$$

Por ser F función de distribución, esta es no decreciente, así, para todo $t_1, t_2 \in [0, \infty)$ con $t_1 < t_2$ se tiene:

$$F(t_1) \leq F(t_2) \Rightarrow -F(t_2) \leq -F(t_1) \Rightarrow 1 - F(t_2) \leq 1 - F(t_1) \Rightarrow S(t_2) \leq S(t_1),$$

por lo que S será no creciente (a).

Por otra parte, F es continua por la derecha, luego para cualquier $c \in \mathbb{R}$ es

$$F(c) = \lim_{t \to c^+} F(t) = \lim_{t \to c^+} (1 - S(t)) = 1 - \lim_{t \to c^+} S(t) \Rightarrow \lim_{t \to c^+} S(t) = S(c),$$

continua por la derecha (b).

Por último, por ser $[0, \infty)$ el soporte de T, se tiene que $F(0) = 0$ y $F(\infty) = 1$, luego $S(t)$ siempre toma valores en $[0, 1]$ (c). □

Además de la condición (c) del resultado anterior, es habitual que se cumpla que (d) $S(0) = 1$ y (e) $S(\infty) = \lim_{t \to \infty} S(t) = 0$. A lo largo de esta memoria, se considerará su cumplimiento.

Definición 1.22. La función de supervivencia se dice **impropia** (*long-term survival function*) si $S(\infty) > 0$.

[1] La función $S(t)$ también es conocida como la tasa de supervivencia acumulada (*cumulative survival rate*).

Proposición 1.6 (Tiempo medio de vida)**.** *Si T es una variable aleatoria en las condiciones anteriores y $S(t)$ verifica (d) y (e), entonces*

$$E(T) = \int_0^\infty S(t)\,dt.$$

Demostración. Usando la Definición 1.16 para variables aleatorias continuas se tiene que

$$E(T) = \int_0^\infty t \cdot f(t)\,dt.$$

Integrando por partes y teniendo en cuenta que $f(t)\,dt = -dS(t)$ se sigue que

$$\int_0^\infty t \cdot f(t)\,dt = \underbrace{-t \cdot S(t)\Big|_0^\infty}_{=0} + \int_0^\infty S(t)\,dt = \int_0^\infty S(t)\,dt,$$

en donde se ha usado $\lim_{t\to\infty} t \cdot S(t) = 0$ por ser S una función monótona no creciente tal que $\int_0^\infty S(x)\,dx$ converge. □

Existe una demostración alternativa de este resultado mediante el uso del método de integración por partes para la integral de Lebesgue-Stieltjes [7] que se puede encontrar en [21] (pág. 8).

Observación 1.4. La Proposición 1.6 es útil a la hora de estimar la esperanza de observaciones censuradas. No es muy evidente cómo generalizar la media para datos censurados; sin embargo, si existe un estimador $\widehat{S}(t)$ de la función de supervivencia, consistente sobre el intervalo $(0, \infty)$, entonces $\int_0^\infty \widehat{S}(t)\,dt$ sería un estimador consistente de la esperanza de vida.

Es importante mencionar que, debido a la asimetría de la distribución y la existencia de *outliers*, habitualmente se trabaja con la mediana, por ser una medida de tendencia central más robusta que la media.

Por otra parte, usando que $f(t)\,dt = -dS(t)$ y el método de integración por partes, se obtiene que la varianza del tiempo de supervivencia se puede calcular como sigue

$$\mathrm{Var}(T) = 2\int_0^\infty t\,S(t)\,dt - \left[\int_0^\infty S(t)\,dt\right]^2.$$

1.4. Función de riesgo

Definición 1.23. La **función de riesgo**[2] (*hazard function*) se define, para $0 < t < \infty$ como sigue:

[2]También conocida como *force of mortality* en demografía o como *intensity function* en procesos estocásticos. Es similar a la densidad en el sentido de que es una función positiva, sin embargo, no está normalizada.

- En notación de incrementos,

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P\left(t < T \leq t + \Delta t \mid T > t\right)}{\Delta t}. \tag{1.2}$$

- En notación diferencial,

$$h(t)\, dt = P\left(T \in N_t \mid T > t\right), \text{donde } N_t = (t, t + dt].$$

Intuitivamente, la definición anterior cuantifica el **riesgo instantáneo** por unidad de tiempo de que para un individuo se produzca el evento de interés en t, sabiendo que antes de t no se había producido. Esto se debe al numerador de (1.2), que es la verosimilitud de que cierto evento de interés ocurra en un tiempo t, habida cuenta de no ha ocurrido antes que t.

Debido a que el tiempo es continuo, la probabilidad de que un evento ocurra exactamente en t es 0. Es por esta razón por la que se escribe en su lugar la probabilidad de que el suceso ocurra en un pequeño intervalo entre t y $t + \Delta t$. El motivo de imponer que el individuo haya sobrevivido al menos hasta t se debe a que, si este ya ha experimentado el evento, es claro que no va a estar más en riesgo. Por tanto, se consideran solamente aquellos individuos presentes al inicio del intervalo $[t, t + \Delta t)$. Finalmente, para obtener la probabilidad por unidad de tiempo (tasa o ratio), esta se relativiza al tamaño del intervalo dividiendo por Δt, dando lugar a la expresión (1.2).

Usando la definición de probabilidad condicionada (Definición 1.18), se obtiene una relación entre las funciones de densidad, riesgo y supervivencia:

$$\begin{aligned} h(t) &= \frac{1}{P\left(T > t\right)} \cdot \lim_{\Delta t \to 0} \frac{P\left(t < T \leq t + \Delta t\right)}{\Delta t} \\ &= \frac{1}{P\left(T > t\right)} \cdot \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} = \frac{f(t)}{S(t)}, \end{aligned} \tag{1.3}$$

Las diferentes formas que puede adoptar la función de riesgo hace que sea una herramienta muy útil y versátil para representar datos de supervivencia e identificar (o proponer) posibles modelos paramétricos (Capítulo 2).

De la relación (1.1), se deduce que

$$f(t) = F'(t) = -S'(t), \tag{1.4}$$

y usando (1.3), se sigue que

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t). \tag{1.5}$$

Usando ahora (1.5), se obtiene otra expresión para la función de supervivencia:

$$S(t) = \exp\big[-H(t)\big], \tag{1.6}$$

donde $H(t) = \int_0^t h(u)\,du$ representa la **función de riesgo acumulada**.

1.5. Vida media residual

Definición 1.24. Sea T la v.a. tiempo de supervivencia, se define la **vida media residual** como

$$m(t) = \begin{cases} E\left(T - t \mid T > t\right), & \text{si} \quad S(t) > 0, \\ 0, & \text{si} \quad S(t) = 0, \end{cases}$$

para $t \geq 0$.

Aunque algunos textos emplean el término[3] *expected remaining life(time)*, es más conocida por el nombre *mean residual life*, o abreviadamente, MRL. Esta función representa cuánto tiempo de vida resta, de media, para un individuo de edad t.

Proposición 1.7. *Sea T la v.a. tiempo de supervivencia, entonces se cumple[4]:*

$$m(t) = \frac{\int_t^\infty S(x)\,dx}{S(t)}. \tag{1.7}$$

Demostración. Tomemos T una variable aleatoria continua, no negativa y con esperanza finita y definamos $Y := T - t$. Por ser Y una transformación medible (traslación) de una variable aleatoria, Y también es variable aleatoria. Así,

$$E\left(T - t \,|\, T > t\right) = E\left(Y \,|\, Y > 0\right) = \frac{1}{P(Y > 0)} \int_0^\infty y \cdot f_Y(y)\,dy,$$

en donde se ha usado la definición de esperanza condicionada a un suceso de interés (ver, por ejemplo, [3]). Usando el teorema de cambio de variable con la función inyectiva $g(x) = x - t$, se tiene que

$$f_Y(y) = f_T\left(g^{-1}(y)\right) \cdot \left|\frac{d}{dy} g^{-1}(y)\right| = f_T(y+t) \cdot \underbrace{\left|\frac{d}{dy}(y+t)\right|}_{=1} = f_T(y+t) = f_T(x).$$

[3] Otros nombres habituales son *complete expectation of life* y *mean excess-loss function* (contexto actuarial).

[4] Suponiendo que $E(T)$ existe, por la Proposición 1.6 está bien definida. $S \geq 0$ y S Riemann integrable por ser S continua y acotada, luego la integral converge.

Por tanto,

$$E\left(T-t \mid T>t\right)=\frac{1}{P(T>t)}\int_t^{\infty}(x-t)\,f_T(x)\,dx=\frac{1}{S(t)}\int_t^{\infty}(x-t)\,f_T(x)\,dx.$$

Ahora, integrando por partes y teniendo en cuenta que $f_T(x)\,dx=-dS(x)$, se sigue que

$$\int_t^{\infty}(x-t)\,f_T(x)\,dx=\underbrace{-S(x)(x-t)\Big|_{x=t}^{x\to\infty}}_{=0}+\int_t^{\infty}S(x)\,dx=\int_t^{\infty}S(x)\,dx.$$

□

El tiempo medio de vida, esto es, la esperanza de la v.a. T, se puede expresar en términos de la función MRL sin más que tener en cuenta que $T-0\,|\,T>0$ es la propia variable T: $E(T)=m(0)$.

Proposición 1.8. *Sea T la v.a. tiempo de supervivencia, h su función de riesgo y m su vida media residual, entonces se cumple:*

$$m'(t)=h(t)\,m(t)-1. \tag{1.8}$$

Demostración. De la expresión (1.7) se tiene que $m(t)\,S(t)=\int_t^{\infty}S(x)\,dx$. Calculemos la derivada de la función integral $\int_t^{\infty}S(x)\,dx$.

Por definición de integral impropia,

$$\int_t^{\infty}S(x)\,dx=\lim_{b\to\infty}\int_t^{b}S(x)\,dx.$$

Por ser S continua en $[t,b]$, existe su primitiva en dicho intervalo, digamos $\mathcal{S}$. Así, por la Regla de Barrow se tiene que

$$\lim_{b\to\infty}\int_t^{b}S(x)\,dx=\lim_{b\to\infty}\left[\mathcal{S}(b)-\mathcal{S}(t)\right]=\lim_{b\to\infty}\mathcal{S}(b)-\mathcal{S}(t).$$

Así,

$$\frac{d}{dt}\int_t^{\infty}S(x)\,dx=\frac{d}{dt}\left(-\mathcal{S}(t)\right)=-S(t).$$

Luego, derivando ambos miembros de la ecuación $m(t)\,S(t)=\int_t^{\infty}S(x)\,dx$ se sigue que

$$\frac{d}{dt}\big(m(t)\,S(t)\big)=\frac{d}{dt}\int_t^{\infty}S(x)\,dx=-S(t),$$

es decir,

$$m'(t)\,S(t)+m(t)\,S'(t)=-S(t). \tag{1.9}$$

Dividiendo por $S(t)$ y usando (1.5) se llega al resultado. □

Por ser las funciones $m(t)$ y $h(t)$ no negativas, de la igualdad (1.8) se deduce directamente que $m'(t) \geq -1$. En otras palabras, la pendiente de la curva MRL nunca será menor que -1.

Proposición 1.9. *Sea T una variable aleatoria continua denotando el tiempo de supervivencia, entonces se cumple:*

$$S(t) = \frac{m(0)}{m(t)} \exp\left(-\int_0^t \frac{du}{m(u)}\right).$$

Demostración. Se trata de resolver la ecuación diferencial (1.8), ordinaria, lineal y de primer orden.

Supongamos $S(t) > 0$. Entonces, necesariamente: $m(t) \neq 0$. Así, dividiendo por $m(t)$ se tiene que

$$\frac{S'(t)}{S(t)} = -h(t) = -\frac{m'(t)}{m(t)} - \frac{1}{m(t)},$$

o, equivalentemente,

$$\frac{d}{dt}\log(S(t)) = -\frac{d}{dt}\log(m(t)) - \frac{1}{m(t)}.$$

Integrando en $[0, t]$ se obtiene:

$$\log(S(t)) - \log(S(0)) = -\big[\log(m(t)) - \log(m(0))\big] - \int_0^t \frac{du}{m(u)},$$

de donde se sigue el resultado. □

1.6. Mediana de supervivencia y mediana de vida residual

Además del tiempo medio de vida, existen otras medidas de tendencia central utilizadas en análisis de supervivencia. Debido a la naturaleza asimétrica de los tiempos de supervivencia, la mediana suele ser la medida resumen más adecuada.

Definición 1.25 (Cuantiles del tiempo de supervivencia)**.** Dado $p \in (0, 1)$, el **p-cuantil** (o percentil $100\,p\,\%$ si $100\,p \in \mathbb{N}$) de la distribución de T es el valor t_p tal que

$$t_p = \text{ínf}\,\{t \in \mathbb{R} \mid S(t) \geq 1 - p\}.$$

Cuando T es una variable aleatoria continua, el p-cuantil se obtiene a partir de la expresión: $S(t_p) = 1 - p$ (ver Definición 1.19).

Observación 1.5. Si la distribución de los tiempos de supervivencia es continua, la **mediana** de T es el valor $t_{0.5}$ tal que $S(t_{0.5}) = 0.5$. En otras palabras, es el tiempo (en la unidad que corresponda) en el que el $50\,\%$ de la población ha experimentado el suceso de interés.

En oncología, es frecuente que haya pacientes que permanezcan libres de recurrencia después de varios años de tratamiento (denominados 'largos supervivientes'). Se presenta a continuación una medida que extiende a la mediana y que es útil cuando hay individuos que han sobrevivido cierto tiempo de interés (véase [39]).

Definición 1.26. Se define la **mediana de vida residual** en el tiempo t como la mediana de la distribución de los tiempos de supervivencia entre aquellos individuos que permanecen vivos en t, es decir, es el valor θ que verifica:

$$P(T - t > \theta \mid T > t) = 0.5,$$

o, equivalentemente, es el valor θ que resuelve la ecuación:

$$\frac{S(t+\theta)}{S(t)} = 0.5.$$

1.7. Generalidades del modelo a tiempo discreto

A pesar de que a nivel teórico el tiempo es una magnitud continua, a nivel práctico los datos de supervivencia son discretos. Esto hace que el caso discreto surja frecuentemente en análisis de supervivencia:

- Las técnicas no paramétricas (Capítulo 5) normalmente están directamente asociadas a la muestra de tiempos de supervivencia, dando lugar a distribuciones discretas.
- Los datos no pueden ser almacenados continuamente, se pueden recoger por días, meses, años, o en general, a razón de cualquier periodo de tiempo.

Sea T una variable aleatoria discreta que toma valores $0 \leq t_1 < t_2 < \cdots < t_j < \cdots$ con probabilidades

$$p_j \coloneqq p_T(t_j) = P(T = t_j), \quad j = 1, 2, \ldots,$$

entonces, la función de supervivencia (Definición 1.21) es simplemente:

$$S(t) = P(T > t) = \sum_{j:t_j > t} p_j.$$

Observación 1.6. Algunos autores definen la función de supervivencia como $P(T \geq t)$, dando lugar a la expresión $S(t) = \sum_{t_j \geq t} p_j$. En el caso continuo esta diferencia puede ser ignorada.

La función de riesgo en el caso discreto requiere introducir una nueva definición.

Definición 1.27. Sea T una variable aleatoria discreta que toma valores $0 \leq t_1 < t_2 < \cdots$, se define la **función de riesgo** en t_j como

$$h_j \coloneqq P(T = t_j \mid T \geq t_j), \quad j = 1, 2, \ldots$$

Observación 1.7. Conviene advertir que la función de riesgo en el caso continuo (Definición 1.23) no es una probabilidad, pero sí lo es en el caso discreto.

Proposición 1.10. *Sea T una variable aleatoria discreta que toma valores $0 \leq t_1 < t_2 < \cdots$, entonces se cumple:*

$$S(t) = \prod_{j:t_j \leq t} (1 - h_j). \tag{1.10}$$

Demostración. A partir de la definición anterior, es claro que

$$h_j = \frac{p_j}{S(t_j^-)} = \frac{p_j}{S(t_{j-1})},$$

donde $S(t_j^-) = \lim_{t \to t_j^-} S(t)$. Notando ahora que

$$p_j = P(T = t_j) = P(T \geq t_j) - P(T > t_j) = S(t_{j-1}) - S(t_j),$$

se tiene:

$$h_j = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}. \tag{1.11}$$

Por último, haciendo

$$S(t) = \frac{S(t_1)}{S(0)} \frac{S(t_2)}{S(t_1)} \cdots \frac{S(t_j)}{S(t_{j-1})} \frac{S(t)}{S(t_j)},$$

y usando (1.11) se sigue el resultado. □

Definición 1.28. Sea T una variable aleatoria discreta que toma valores $0 \leq t_1 < t_2 < \cdots$, se define la **función de riesgo acumulado** como

$$H(t) = \sum_{j:t_j \leq t} h_j.$$

Proposición 1.11. *Sea T una variable aleatoria discreta que toma valores $0 \leq t_1 < t_2 < \cdots$, entonces se cumple:*

$$H(t) \approx -\sum_{j:t_j \leq t} \log(1 - h_j).$$

Demostración. Tomando logaritmos en (1.10) se tiene

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

Usando el desarrollo en serie de Taylor se puede probar que $\log(1-h_j) = -h_j + \mathcal{O}(h_j^2)$, por lo que para h_i pequeño[5] tendremos $\log(1-h_j) \approx -h_j$. Así,

$$\log S(t) \approx -\sum_{j:t_j \leq t} h_j = -H(t).$$

□

Por último, se define la función de vida media residual para el caso discreto.

Definición 1.29. Sea T una variable aleatoria discreta que toma valores $0 \leq t_1 < t_2 < \cdots$, se define la **función de vida media residual** en t, con $t_i < t \leq t_{i+1}$, como

$$m(t) = \frac{1}{S(t)}\,(t_{i+1} - t)\,S(t_i) + \frac{1}{S(t)} \sum_{j \geq i+1} (t_{j+1} - t_j)\,S(t_j). \tag{1.12}$$

Al igual que en el caso continuo, la función MRL en t es igual a $1/S(t)$ por el área comprendida entre la función de supervivencia, la recta vertical $x = t$ y el eje de abscisas. En la Figura 1.3 se puede comprobar gráficamente la expresión (1.12).

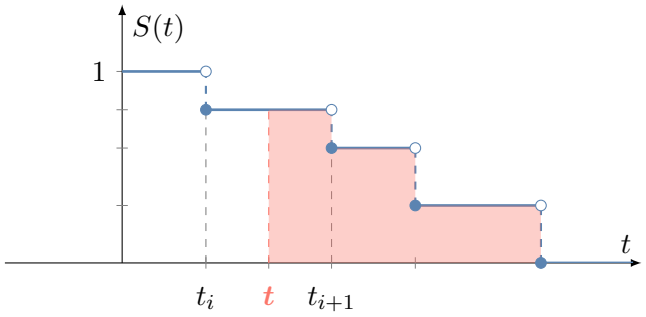


Figura 1.3: Función de supervivencia de una variable aleatoria T discretizada

[5]Nótese que, por ser la función de riesgo una probabilidad (en el caso discreto), siempre se tiene $h_j \in [0, 1]$.

Capítulo 2

Modelos paramétricos

Existen varios modelos paramétricos que se emplean con regularidad en análisis de supervivencia. En este capítulo se describen las distribuciones de probabilidad más frecuentes que presentan las variables de tiempo de vida.

2.1. Modelo Exponencial

Definición 2.1. Una variable aleatoria continua X se dice que sigue **distribución exponencial** de parámetro λ, con $\lambda > 0$, denotado $X \sim \text{Exp}(\lambda)$, si su función de densidad viene dada por

$$f_X(x) = \lambda\, e^{-\lambda x} \cdot I(x \in [0, \infty)).$$

La distribución exponencial describe el tiempo (o la distancia, por ejemplo, en otros contextos) entre dos sucesos consecutivos que tienen lugar a una tasa constante. Es decir, si el número de sucesos por unidad de tiempo se modela mediante una Poisson(λ), entonces el tiempo que pasa entre dos de esos sucesos consecutivos sigue una Exp(λ).

Observación 2.1. La distribución exponencial también puede venir caracterizada a través de su parámetro de escala θ, que es el inverso del parámetro λ en la Definición 2.1, $\theta = 1/\lambda$. De hecho, el parámetro θ se puede considerar un 'parámetro de supervivencia', pues indica el tiempo medio entre dos sucesos. Este parámetro es llamado a veces llamado parámetro o tasa de decaimiento, o bien, parámetro de intensidad del proceso. En inglés: *rate parameter* o *hazard parameter*.

Para la variable aleatoria tiempo de vida T con $T \sim \text{Exp}(1/\theta)$, la función de distribución es

$$F(t) = \int_{-\infty}^{t} \frac{1}{\theta} \exp\left(-t/\theta\right) \cdot I(t \in [0, \infty))\, dt = \int_{0}^{t} \frac{1}{\theta} \exp\left(-t/\theta\right) dt = 1 - \exp\left(-t/\theta\right), \text{ si } t \geq 0.$$

A partir de esta[1] se obtiene la función de supervivencia, la función de riesgo y la función de riesgo acumulado. Para $t \geq 0$,

$$\begin{aligned} S(t) &= 1 - F(t) = \exp(-t/\theta), \\ h(t) &= \frac{f(t)}{S(t)} = \frac{1}{\theta}, \\ H(t) &= \int_0^t h(x)\,dx = t/\theta. \end{aligned}$$

Se observa que la distribución exponencial no es muy flexible, pues la función de riesgo es contante con el tiempo. En términos médicos, supone que el riesgo instantáneo de muerte ('justo después de t') no depende de la edad del individuo o, en otras palabras, la distribución exponencial **no puede modelar el envejecimiento**. Esto es debido a la bien conocida propiedad de pérdida de memoria (*lack-of-memory*, *memoryless*, *forgetfulness property*) de la función exponencial, que se enuncia a continuación.

Proposición 2.1 (Pérdida de memoria de la función exponencial)**.** *Sea* $X \sim \text{Exp}(\lambda)$ *con* $\lambda > 0$. *Entonces, para* $x, a \geq 0$, *se cumple*

$$P(X > x + a \mid X > a) = P(X > x).$$

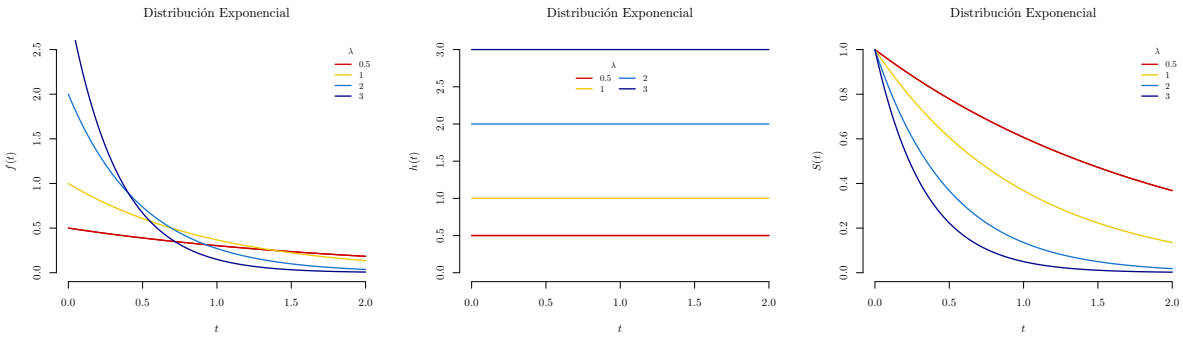


Figura 2.1: Distribución Exponencial: funciones de densidad, riesgo y supervivencia (de izquierda a derecha) para varias elecciones del parámetro

2.2. Modelo Weibull

El modelo Weibull es una generalización del modelo exponencial, se incorpora un parámetro de forma[2] α y se mantiene el parámetro de escala $\theta := 1/\lambda$. Este modelo es uno de los más

[1] También, a partir de la relación $S'(t) = -f(t)$, se puede hallar directamente la función de supervivencia como $S(t) - S(0) = \int_0^t S'(t)\,dt = -\int_0^t \frac{1}{\theta} \exp(-t/\theta)\,dt$.

[2] En las distribuciones de probabilidad existen tres tipos de parámetros: de **localización**, de **escala** y de **forma**. El primero define el origen de tiempos (traslada la distribución); el segundo afecta a la variabilidad ('estira' o

utilizados para tiempos de supervivencia: tiene utilidad en la durabilidad de algunos artículos manufacturados, así como en los tiempos de aparición de tumores en medicina.

Definición 2.2. Una variable aleatoria continua X se dice que sigue **distribución Weibull** con parámetros de forma y escala, respectivamente, $\alpha, \theta > 0$, denotado $X \sim \text{Weibull}(\alpha, \theta)$, si su función de densidad viene dada por

$$f_X(x) = \frac{\alpha}{\theta}\left(\frac{x}{\theta}\right)^{\alpha-1} e^{-(x/\theta)^{\alpha}} \cdot I(x \in [0, \infty)). \tag{2.1}$$

Observación 2.2. Para $\alpha = 1$, la expresión (2.1) se reduce a la distribución exponencial.

Para la variable aleatoria tiempo de vida T con $T \sim \text{Weibull}(\alpha, \theta)$, la función de supervivencia y de riesgo con, para $t \geq 0$:

$$S(t) = \exp\big(-(t/\theta)^{\alpha}\big),$$
$$h(t) = \frac{f(t)}{S(t)} = (\alpha/\theta)\,(t/\theta)^{\alpha-1}.$$

El tiempo de vida esperado y su varianza vienen dados por

$$E[T] = \frac{1}{\lambda}\Gamma(1+\theta) \quad \text{y} \quad \text{Var}(T) = \theta^2\left[\Gamma(1+2\theta) - (\Gamma(1+\theta))^2\right],$$

donde Γ es la función Gamma definida por

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

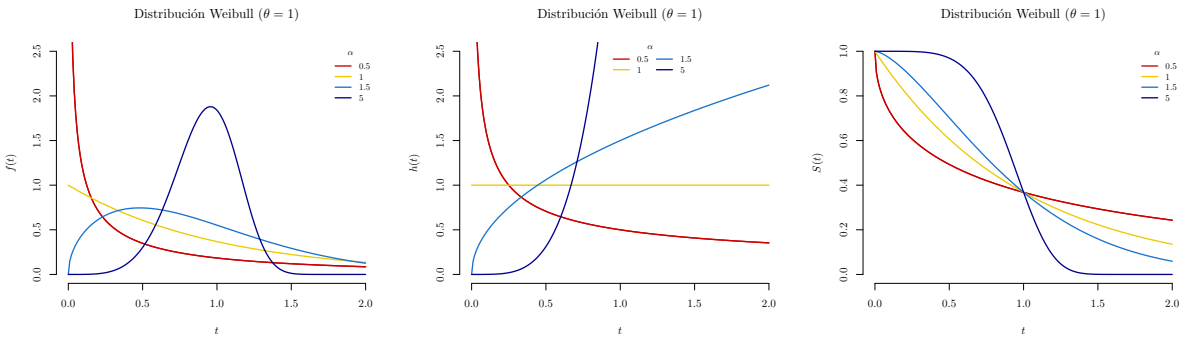


Figura 2.2: Distribución Weibull: funciones de densidad, de riesgo y de supervivencia (de izquierda a derecha) para varias elecciones del parámetro de forma

'comprime' la distribución) y el último no es ni de localización ni de escala, ni función de ambos, sólo afecta a la forma. A partir de los parámetros de escala y de forma se derivan los coeficientes de asimetría y curtosis.

2.3. Modelo Gamma

Para el modelo Gamma, las funciones de densidad, riesgo y supervivencia cuando intervienen los parámetros de forma, $\alpha > 0$, e inverso de escala, $\lambda > 0$, vienen dadas por:

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t},$$

$$h(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha) - \gamma(\alpha, \lambda t)},$$

$$S(t) = 1 - \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \lambda t),$$

para $t \geq 0$, y donde Γ y γ son, respectivamente, las funciones gamma y gamma incompleta:

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x}\, dx \quad \text{y} \quad \gamma(a,t) = \int_0^t x^{a-1} e^{-x} dx,$$

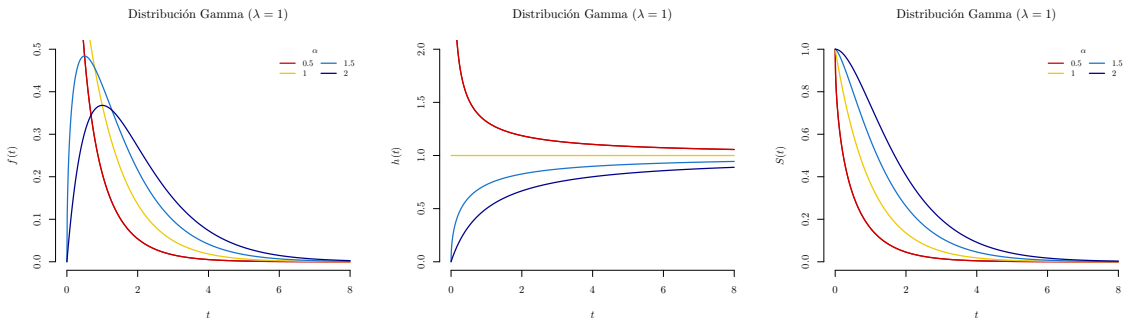


Figura 2.3: Distribución Gamma: funciones de densidad, riesgo y supervivencia (de izquierda a derecha) para varias elecciones del parámetro de forma

2.4. Modelo Log-Logístico

Definición 2.3. Se dice que T tiene una **distribución Log-Logística** si la variable $\log T$ tiene una distribución Logística. La distribución Logística, como la Normal, es una distribución con parámetros de localización μ y escala σ, es decir, $Y = \mu + \sigma W$, donde W es la distribución Logística estándar. W tiene una función de densidad muy similar a la de $\mathcal{N}(0,1)$, pero tiene colas más pesadas (mayor curtosis). Su función de densidad es

$$f(y) = \frac{e^{-(y-\mu)/\sigma}}{\sigma \left(1 + e^{-(y-\mu)/\sigma}\right)^2}, \quad -\infty < y < \infty.$$

Las funciones que caracterizan la distribución Log-Logística con parámetros de forma, $\alpha = 1/\sigma > 0$, y de escala, $\theta = \exp(\mu) \in \mathbb{R}$, vienen dadas, para $t > 0$, por:

$$f(t) = \frac{(\alpha/\theta)(x/\theta)^{\alpha-1}}{(1+(x/\theta)^{\alpha})^2},$$

$$h(t) = \frac{(\alpha/\theta)(t/\theta)^{\alpha-1}}{1+(t/\theta)^{\alpha}},$$

$$S(t) = (1+(t/\theta)^{\alpha})^{-1}.$$

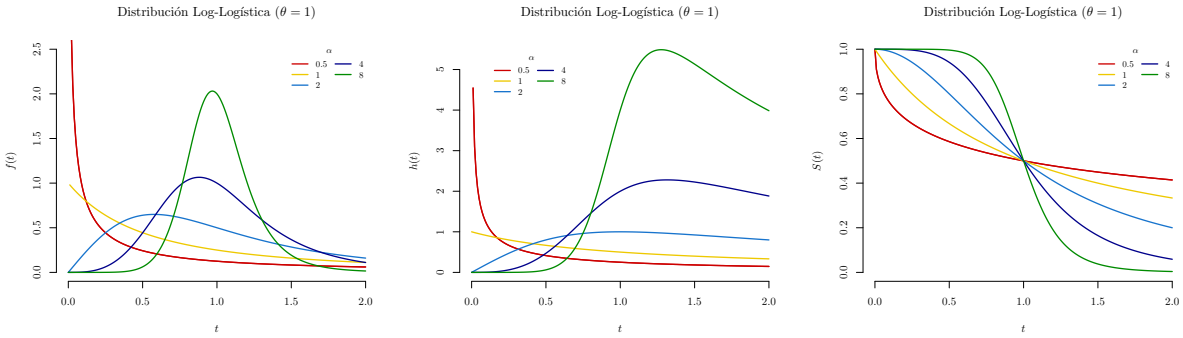


Figura 2.4: Distribución Log-Logística: funciones de densidad, riesgo y supervivencia (de izquierda a derecha)

2.5. Modelo Log-Normal

Para el modelo Log-Normal, las funciones de densidad, riesgo y supervivencia vienen dadas por

$$f(t) = \frac{1/t}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\,\sigma^2}\right),$$

$$h(t) = \frac{\frac{1/t}{\sigma\sqrt{2\pi}}}{1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)} \exp\left(-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right),$$

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right).$$

La distribución normal tiene dos parámetros que hereda el modelo Log-Normal: la media, $\mu \in \mathbb{R}$, y la desviación estándar, $\sigma > 0$. Estas dos distribuciones se corresponden unívocamente a través de las transformaciones exponencial y logaritmo (ver Figura 2.6).

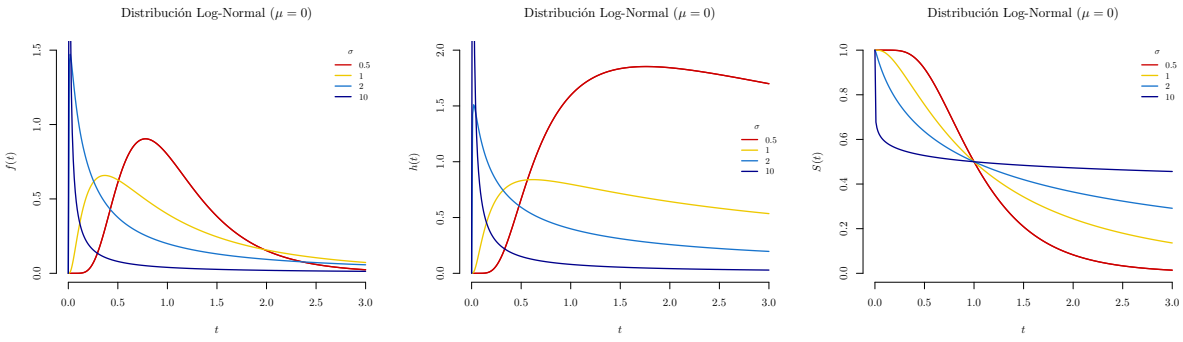


Figura 2.5: Distribución Log-Normal: funciones de densidad, riesgo y supervivencia (de izquierda a derecha)

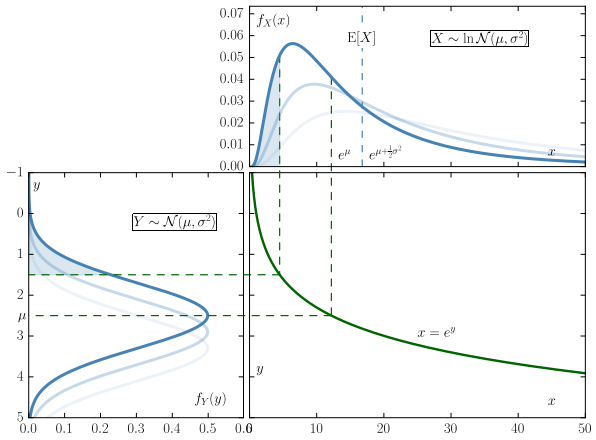


Figura 2.6: Distribución Log-Normal. Fuente: Wikipedia. Autor: Stijn De Vuyst. Licencia: CC BY-SA 4.0.

En la Tabla 2.1 se muestran algunas posibles aplicaciones de los modelos paramétricos, indicando la forma de la distribución correspondiente caracterizada por la función de riesgo.

Forma función de riesgo	Aplicación	Modelo paramétrico
constante	producto manufacturado	Weibull ($\alpha = 1$)
monótona creciente	paciente con leucemia	Weibull ($\alpha > 2$)
monótona decreciente	paciente después de la cirugía	Log-Logístico ($\alpha < 1$)
forma de joroba	infección por tuberculosis	Log-Normal
forma de U	trasplante de corazón (tolerado)	Weibull ($\alpha < 1$)
	trasplante de corazón (rechazado)	Gompertz-Makeham[3]

Tabla 2.1: Distintas formas de la función de riesgo y aplicaciones.

[3]Es el resultado de combinar los modelos de Gompertz (1825) y Makeham (1860).

Capítulo 3

Censura y Truncamiento

La censura se presenta cuando se tiene información incompleta sobre la supervivencia de algunos individuos. La mayoría de los métodos usados en análisis de supervivencia presuponen que los individuos censurados conforman una **submuestra aleatoria**, es decir, que están sujetos a la misma probabilidad de experimentar el evento que los que permanecen en el estudio; esta propiedad es la llamada condición de censura no informativa.

Los modelos de pérdida de datos (*missing data*), clasificados por Rubin[1] en 1976 [5], requieren un tratamiento específico [6], pues hay muchas razones por las cuales pueden ocurrir. Es preciso conocer el patrón de la pérdida de los datos faltantes para determinar el método de imputación de los mismos. Los patrones de los datos faltantes mas frecuentes son: MCAR (datos perdidos completamente al azar), MAR (perdidos aleatoriamente) y NMAR (no perdidos aleatoriamente).

Existen varios modelos de censura, y cada uno dará lugar a una verosimilitud diferente (Tabla 4.1), que será la base para hacer inferencia. Aunque la clasificación no es completamente estándar, los datos de supervivencia (*time-to-event data*) se pueden distinguir según sean:

- observados, truncados o censurados,
- censurados por la derecha, por la izquierda o por intervalo,
- censurados de tipo I, de tipo II, o censurados aleatoriamente,
- simple o múltiplemente censurados.

[1] Donald B. Rubin (22 de diciembre de 1943) es un profesor emérito de estadística de la Universidad de Harvard conocido por el 'modelo causal de Rubin', en donde se incluye la causalidad en los casos de estudios observacionales y cuasiexperimentales.

3.1. Censura por la derecha

La censura por la derecha surge cuando el evento de interés solo puede ser observado si el tiempo de supervivencia T no supera el valor del tiempo de censura, C. En el planteamiento del problema, las variables estarán relacionadas mediante la expresión $Y = \text{mín}(T, C)$, siendo Y el tiempo de seguimiento del individuo (hasta el evento de interés o hasta la censura). Más aún, veremos que no sólo observamos Y, sino que también observamos si el individuo sufre una censura o un evento, es decir, observamos una nueva variable aleatoria $\Delta = I(T \leq C)$, siendo I la función indicador del suceso "se observa el evento de interés". Por ser la variable observada Y función de otras dos variables aleatorias, las funciones definidas en el Capítulo 1 se pueden expresar en términos de estas dos variables[2]. A continuación, se formaliza matemáticamente con más detalle el modelo de censura por la derecha.

Para cada individuo dentro del estudio, se supone que existe un tiempo de supervivencia T y un tiempo de censura C (este último puede ser fijo o aleatorio). Además, se asume que los tiempos de vida son independientes e idénticamente distribuidos. El tiempo de vida de un individuo será conocido si, y sólo si, se cumple $T \leq C$. Si ocurre lo contrario, el tiempo de vida para este individuo estará censurado en C. Es decir, se observa la variable aleatoria

$$Y \coloneqq \text{mín}(T, C).$$

De esta forma, los datos recogidos del experimento podrán ser representados como realizaciones muestrales de pares de variables aleatorias (Y, Δ), donde Δ indica si el tiempo de vida T corresponde con el del suceso o si está censurado, es decir,

$$\Delta = I(T \leq C) = \begin{cases} 1, & \text{si } T \leq C \quad \text{(observado)} \\ 0, & \text{si } T > C \quad \text{(censurado).} \end{cases}$$

Algunos autores se refieren a esta función como función indicadora de censura (en la literatura inglesa se puede encontrar como *status indicator*, *event indicator* o *last-known status* [29]).

3.1.1. Clasificación de las censuras por la derecha

Censura de tipo I

En este esquema el experimento se programa con una duración establecida previamente, digamos c, y todos los individuos entran al estudio simultáneamente. El tiempo de vida de un sujeto se observará si es menor o igual que dicho valor prefijado. En otro caso, la observación correspon-

[2]Lo explicaremos en detalle para el mecanismo de censura aleatoria por ser este, como veremos, el modelo de censura más frecuente (véase Funciones asociadas a la supervivencia en presencia de censura aleatoria)

diente tendrá un valor censurado c. En este contexto, el número de observaciones censuradas de la muestra es aleatorio[3] y el valor de censura es fijo.

En general, para una muestra de tamaño n, a saber, $(Y_1, \Delta_1), (Y_2, \Delta_2), \ldots, (Y_n, \Delta_n)$, observaremos para cada individuo:

$$Y_i = \text{mín}(T_i, c), \qquad \Delta_i = \begin{cases} 1, & \text{si } T_i \leq c \quad \text{(individuo } i \text{ observado)}, \\ 0, & \text{si } T_i > c \quad \text{(individuo } i \text{ censurado)}. \end{cases}$$

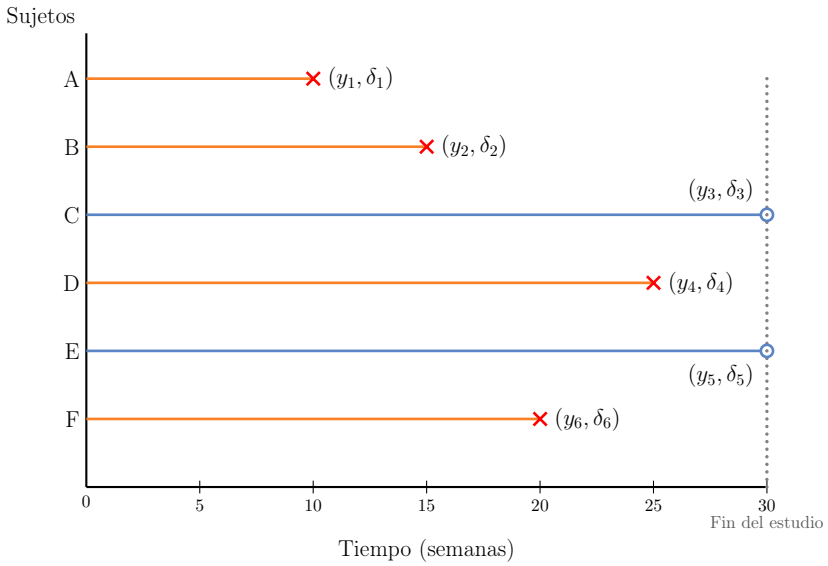


Figura 3.1: Esquema censura de tipo I

En la Figura 3.1, si suponemos que los datos provienen de un estudio médico (ficticio), los pacientes C y E seguían vivos al finalizar el estudio, mientras que el resto experimentaron el evento. Así, se tendría que $y_i = t_i$, $\delta_i = 1$, para $i = 1, 2, 4, 6$ (observados), e $y_j = c$, $\delta_j = 0$, para $j = 3, 5$ (censurados). En resumen,

$$Y = (10, 15, \mathbf{30}, 25, \mathbf{30}, 20),$$
$$\delta = (1, 1, \mathbf{0}, 1, \mathbf{0}, 1).$$

Observación 3.1. Es común recoger los datos de supervivencia seguidos del símbolo '+' cuando la observación está censurada. Por ejemplo, en el esquema de la Figura 3.1 los datos se resumirían como sigue: 10, 15, 30+, 25, 30+, y 20 semanas. El símbolo '+' advierte que el tiempo de supervivencia real es superior al número que le precede.

[3]De hecho, es una variable aleatoria discreta que sigue una distribución binomial $B(n, p)$, donde n es el número de individuos de la muestra y p se corresponde con la probabilidad de que una observación sea censurada, es decir, $p = P(T > c)$.

Observación 3.2. En general, en ningún modelo de censura se van a poder recuperar los pares (t_i, c_i) si la información disponible es la de los pares (y_i, δ_i), donde $\{(y_i, \delta_i)\}_{i=1}^n$ es una realización muestral (muestra de observaciones) de $(Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)$; únicamente se recuperaría una componente (t_i o c_i, la menor de las dos).

Cuando no todos los individuos entran al estudio al mismo tiempo, se habla de **censura de tipo I generalizada** [20]. En estos estudios, los sujetos tienen sus propios tiempos censurados, que son conocidos (y fijos) en el momento de su entrada (véanse sujetos C y F en Figura 3.6). Este tipo de datos pueden ser representados a través de un Diagrama de Lexis [19], el cual utiliza 'líneas de vida' para representar la duración temporal de un determinado fenómeno en cada uno de los individuos que componen la población objeto de estudio.

En la Figura 3.2 se representan las observaciones de cuatro individuos y su momento de entrada en un estudio realizado durante del mes de junio. Considérese que se está estudiando la recidiva de cierto tumor.

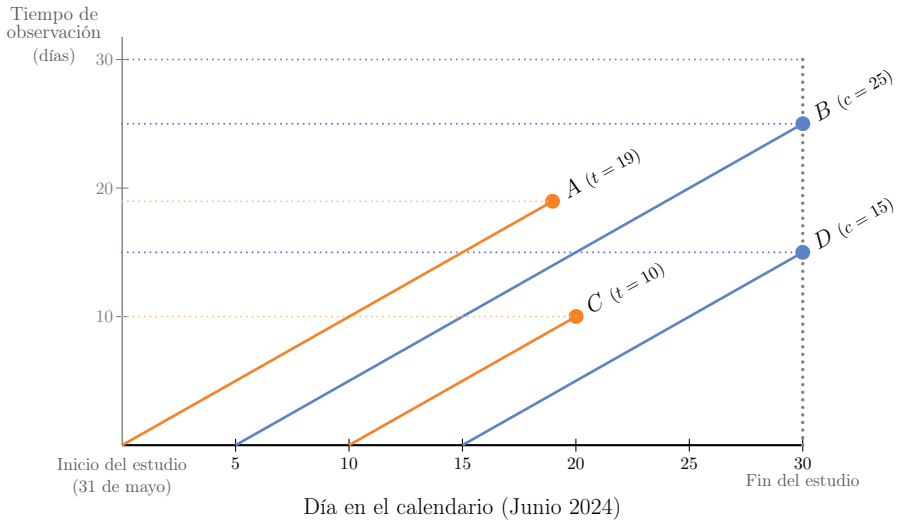


Figura 3.2: Esquema censura de tipo I generalizada mediante Diagrama de Lexis

Censura de tipo II

Puede ocurrir que ningún evento sea observado durante la duración establecida del experimento, lo que produciría un análisis estadístico muy pobre. Surge así la necesidad de establecer otro modelo que ponga el enfoque en el número de eventos observados.

En este modelo todos los individuos entran a la vez, y se diseña de tal forma que el estudio finaliza en el momento en el que una fracción prefijada r/n de individuos experimenta el evento, siendo r un número entero tal que $1 \leq r \leq n$. En otras palabras, el experimento se interrumpe cuando se produce el r-ésimo fallo. Ese instante, que depende de r, será el valor censurado de

todos los sujetos que en ese momento no hayan fallado todavía. De esta forma, se observarán los r tiempos de fallo más pequeños de la muestra y aparecerán $n - r$ tiempos censurados (véase Figura 3.3). Una razón común para determinar el número de eventos que deben observarse es la potencia requerida en el estudio. En este esquema, el número de observaciones censuradas de la muestra es fijo y el valor de censura aleatorio, pero toma un mismo valor para todos los individuos [27].

Cuando la fracción r/n es próxima a uno, la muestra se dice débilmente censurada (*lightly censored*); mientras que si r/n es próxima a cero, se dice fuertemente censurada (*heavily censored*).

Sean $T_1, \ldots, T_n$ tiempos de fallo independientes e idénticamente distribuidos. Si se considera la muestra ordenada $T_{(1)} \leq \cdots \leq T_{(n)}$, entonces, las n realizaciones de la variable observada Y serán de la forma:

$$Y_i \coloneqq Y_{(i)} = \begin{cases} T_{(i)}, & \text{si } i \leq r \quad \text{(observado)}, \\ T_{(r)}, & \text{si } i > r \quad \text{(censurado)}. \end{cases}$$

Observación 3.3. Al igual que en el caso de Censura de tipo I, las observaciones para el individuo i-ésimo de la muestra se pueden escribir a través del par (Y_i, Δ_i), con la salvedad de que ahora Y_i y Δ_i dependen del resto de tiempos de supervivencia (T_j con $j \neq i$). Por lo tanto, los n pares de la muestra $(Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)$ son, en este modelo de censura, **dependientes**.

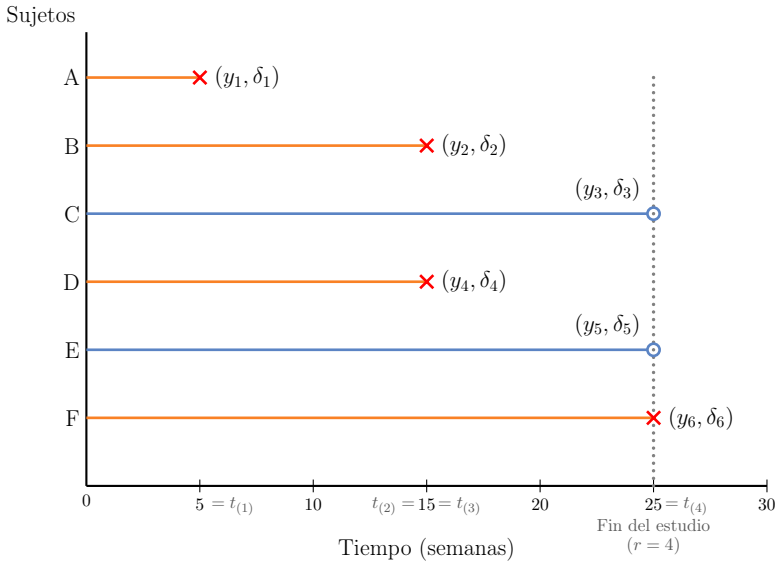


Figura 3.3: Esquema censura de tipo II

Se ha empleado el verbo 'fallar' para poner en valor la importancia que la censura de tipo II tiene en las pruebas de fiabilidad (*reliability*), ensayos de vida útil (*life-testing*) y otras técnicas de control de calidad. No solo la motivación económica hace que este experimento sea atractivo en la industria, también interviene el tiempo, ya que esperar hasta que todos los sujetos fallen puede llegar a ser inviable.

El tratamiento estadístico de los datos censurados de tipo II es más simple que los de tipo I, pues como los datos consisten en los r tiempos de supervivencia más pequeños de entre los n de una muestra aleatoria, la teoría de los estadísticos de orden es directamente aplicable.

Para garantizar la independencia entre el mecanismo de censura y la observación del fenómeno, se debe fijar de antemano el valor c (duración del estudio) en el esquema de tipo I y la fracción r/n (porcentaje de individuos para los que se observa el evento) en el esquema de tipo II.

El planteamiento general de la censura de tipo II, a veces apellidada unietápica (*one-stage*), tiene la desventaja práctica de que el periodo de estudio, $T_{(r)}$, que define el tiempo de censura, es aleatorio y desconocido al inicio del experimento. Para acelerar la finalización del estudio, existen modelos más complejos que planifican la retirada de individuos, ahorrando tiempo y costes [2].

La **censura progresiva de tipo II** permite retirar sujetos en diferentes etapas del estudio. En este modelo, un total de n unidades con tiempos de vida $T_1, T_2, \ldots, T_n$ son introducidas simultáneamente en el ensayo. Una vez que tiene lugar el primer fallo, un mecanismo de selección aleatoria retira R_1 unidades de las $n-1$ restantes que permanecen a riesgo. Este proceso continúa reiteradamente de manera que en el j-ésimo fallo serán elegidas aleatoriamente R_j unidades para su retirada, de entre las $n - j - \sum_{i=1}^{j-1} R_i$, con $1 \leq j \leq m$, $m \leq n$. Esta censura progresiva proporciona m tiempos ordenados de fallo, denotados[4] por $T_{1:m:n}, T_{2:m:n}, \ldots, T_{m:m:n}$, y una colección $R := (R_1, R_2, \ldots, R_m)$ que caracteriza[5] el esquema del experimento (Figura 3.4).

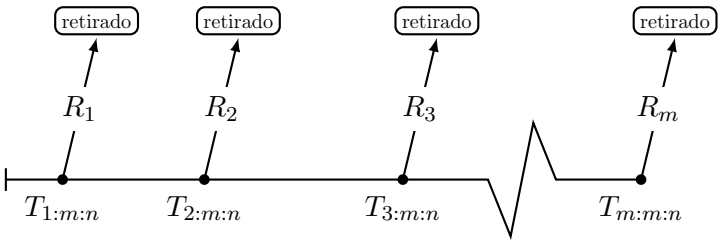


Figura 3.4: Esquema censura progresiva de tipo II

Este mecanismo puede ser aplicado en ensayos clínicos en los que se pierdan individuos debido a causas ajenas al evento de interés, como migración, desmotivación o decisiones éticas [34].

Censura aleatoria

En los modelos anteriores la censura estaba controlada por el investigador: en Censura de tipo I, de manera directa al elegir el periodo del estudio, y en Censura de tipo II, de manera indirecta al elegir el número de eventos deseados. Se introduce ahora un nuevo mecanismo de censura

[4] Notar que $T_{1:m:n} = T_{1:n} = T_{(1)}$, pues no hay censuras antes del primer fallo.

[5] La censura de tipo II unietápica puede ser vista como un caso particular de censura progresiva de tipo II, con esquema $R = (0, 0, \ldots, n-r)$.

no controlada por el investigador: la **censura aleatoria** (a veces[6] llamada de tipo III), que en resumen se da cuando los individuos pueden abandonar el estudio, o fallar por alguna causa que no es la de interés.

La censura aleatoria se presenta cuando el número de observaciones censuradas y el valor de la censura son ambos aleatorios (y la censura no tiene por qué tomar el mismo valor para todos los individuos). Este modelo supone la existencia de dos variables aleatorias **independientes** para cada individuo i de la muestra; a saber, el tiempo de supervivencia T_i y el tiempo de censura C_i. La muestra quedará determinada por dos variables:

- El mínimo de cada par (T_i, C_i):
$$Y_i := \text{mín}(T_i, C_i),$$
- Una variable dicotómica para distinguir los eventos de las censuras:
$$\Delta_i = I(T_i \leq C_i) = \begin{cases} 1, & \text{si } T_i \leq C_i \quad \text{(observado)}, \\ 0, & \text{si } T_i > C_i \quad \text{(censurado)}. \end{cases}$$

Según este esquema, los individuos podrán abandonar el estudio en cualquier momento, y es esta razón la que justifica que los tiempos de censura, C_i, de cada individuo sean considerados como variables aleatorias. La incorporación de los individuos al estudio también podría ser escalonada.

Al igual que en el modelo de censura de tipo I generalizada, también es factible, y a veces conveniente, representar los datos mediante un Diagrama de Lexis.

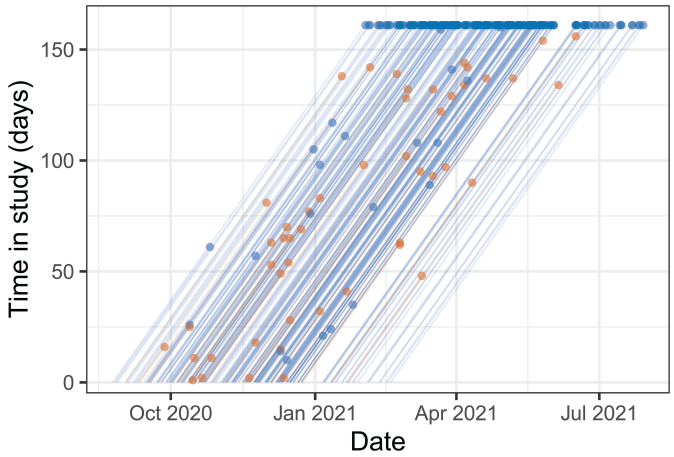


Figura 3.5: Esquema censura aleatoria mediante Diagrama de Lexis. Fuente: [37]

[6]No es un término muy utilizado, pues en comparación con los tipos I y II, que son mecanismos más estructurados, la censura aleatoria abarca cualquier censura que no se ajuste a estos dos modelos. Otros autores, en cambio, hablan de censura de tipo III para referirse al caso en el que los individuos entran al estudio, de duración preestablecida, en diferentes tiempos (lo que hemos llamado censura de tipo I generalizada).

En la Figura 3.5 se muestra un Diagrama de Lexis procedente de un estudio realizado en Países Bajos en el año 2020 que tenía como objetivo analizar la entrada del virus SARS-CoV-2 en los hogares neerlandeses durante un periodo de observación de 161 días. La longitud de las líneas representa el tiempo desde la entrada en el estudio hasta la infección del primer miembro de la familia (puntos naranjas), hasta completar la duración del estudio (puntos azules situados en la parte superior), o bien hasta el abandono del estudio (puntos azules situados por debajo de los 161 días).

Funciones asociadas a la supervivencia en presencia de censura aleatoria

Al haber introducido la notación (Y, Δ), es natural preguntarse qué variaciones provoca en el conjunto de funciones definidas en el Capítulo 1.

Respecto a la función de supervivencia de Y, esta será el producto de las funciones de supervivencia de T y de C (ambas continuas e independientes). Veamos:

$$S_Y(t) = P(Y > t) = P\big(\min(T, C) > t\big) = P(T > t)P(C > t) = S_T(t)\, S_C(t). \tag{3.1}$$

Respecto a la función de distribución de Y, usando (3.1), se tiene que

$$\begin{aligned} F_Y(t) &= 1 - S_T(t)\, S_C(t) = 1 - (1 - F_T(t))\,(1 - F_C(t)) \\ &= F_T(t) + F_C(t) - F_T(t)\, F_C(t). \end{aligned} \tag{3.2}$$

Derivando (3.2) se obtiene la función de densidad de Y, a saber,

$$\begin{aligned} f_Y(t) &= f_T(t) + f_C(t) - f_T(t)\, F_C(t) - f_C(t)\, F_T(t) \\ &= f_T(t)\,(1 - F_C(t)) + f_C(t)\,(1 - F_T(t)) = f_T(t)\, S_C(t) + f_C(t)\, S_T(t). \end{aligned} \tag{3.3}$$

Ahora, usando la relación (1.3), junto con (3.1) y (3.3), obtenemos la función de riesgo de Y,

$$h_Y(t) = \frac{f_T(t)\, S_C(t) + f_C(t)\, S_T(t)}{S_T(t)\, S_C(t)} = h_T(t) + h_C(t). \tag{3.4}$$

Por último, usando (3.4), por linealidad de la integral se obtiene la función de riesgo acumulado de Y,

$$H_Y(t) = H_T(t) + H_C(t).$$

Muestras simple y múltiplemente censuradas

Dependiendo de si existe un valor de censura único o no, las muestras generadas por los experimentos pueden ser simple o múltiplemente censuradas.

Una muestra se dice **simplemente censurada** si solo existe un nivel de censura, es decir, tiene un valor de censura común para todas las observaciones censuradas. Esto requiere que todos los individuos comiencen el ensayo al mismo tiempo. En los ensayos industriales son más habituales este tipo de muestras.

En cambio, una muestra se dice **múltiplemente censurada** si existen distintos niveles de censura. Esto sucede en los ensayos médicos ya que, además de establecerse una limitación temporal, es habitual que los individuos se incorporen al ensayo en distintos instantes de tiempo y que se produzcan abandonos, que dan lugar a observaciones censuradas, pues sólo se conoce que el evento no se había observado hasta el momento del abandono (véase Ejemplo 3.1).

Aunque históricamente fue importante distinguir entre censura simple y múltiple debido a las limitaciones de las técnicas analíticas disponibles, hoy en día, con herramientas estadísticas más avanzadas, esta distinción es en muchos casos prescindible. Las técnicas modernas permiten una mayor flexibilidad y capacidad para manejar datos censurados de diversas maneras sin necesidad de una categorización tan estricta.

Ejemplo 3.1. Se considera un ensayo clínico de un año de duración y con un periodo de reclutamiento de tres meses. De los ocho pacientes que entraron al estudio, dos de ellos lo abandonaron (B y E), cuatro experimentaron el evento (A,D,G y H) y dos seguían vivos al finalizar el estudio (C y F) (Figura 3.6).

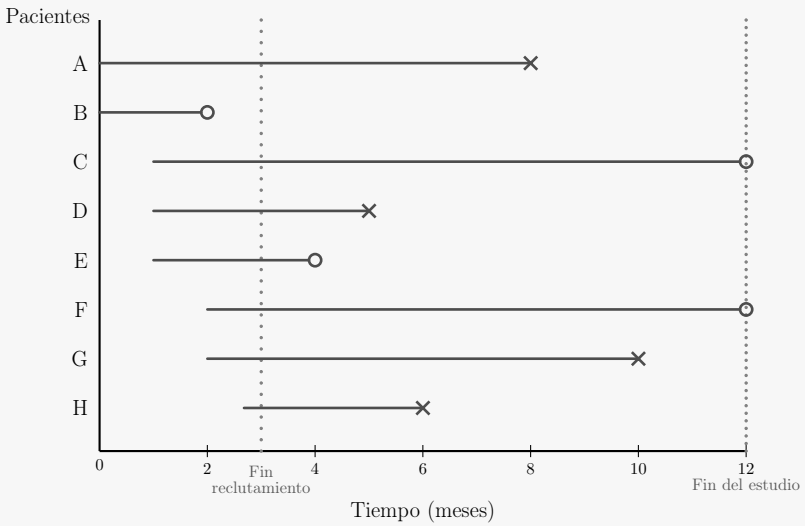


Figura 3.6: Tiempos de supervivencia en un ensayo clínico (I)

Debido a que la magnitud de interés es el tiempo de supervivencia T ($t_{evento} - t_{entrada}$), el instante en el que se comenzó a medir cada observación no suele ser relevante; por lo tanto, las observaciones suelen representarse con el mismo origen, como se muestra en la Figura 3.7.

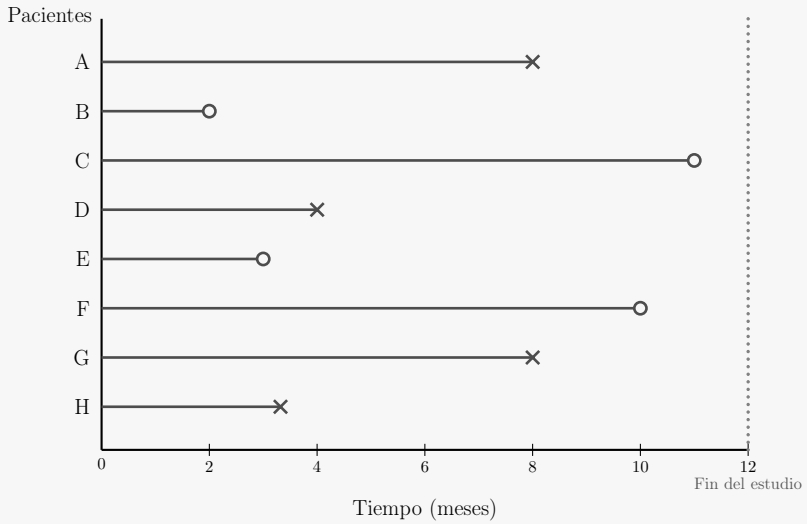


Figura 3.7: Tiempos de supervivencia en un ensayo clínico (II)

Observación 3.4. Como se viene comentando, las observaciones del tipo *time-to-event* son pares (t_i, c_i), donde t_i denota el tiempo del i-ésimo individuo, y c_i indica si el evento o la censura ocurre en ese tiempo. También se comentó que es habitual medir el tiempo t_i desde un punto de partida común (véase Ejemplo 3.1). Esta técnica, conocida por algunos autores como *alignment*, puede ser inapropiada en algunos contextos. Por ejemplo, cuando la base de datos procede de historias clínicas (en inglés: *electronic health record* o EHR), se debe tener en cuenta que los pacientes ingresan a esta base de datos en diferentes momentos en relación con su estado de salud, y por tanto, alinear a todos los pacientes desde un inicio común podría no reflejar adecuadamente la progresión de la enfermedad [31] (véase Ejemplo 3.2). Esta limitación dentro del análisis de supervivencia fue observada por [26], quien por primera vez propuso[7] un enfoque centrado en el evento (*alignment by failure*) donde el tiempo se mide hacia atrás desde el evento de interés en lugar de hacia adelante desde un punto de inicio común. Este enfoque garantiza que todos los pacientes compartan características similares en el momento en el que se produce el evento.

Ejemplo 3.2. Un estudio está evaluando la recuperación postoperatoria de pacientes que han tenido cirugías de reemplazo de cadera. Si se decide alinear a todos los pacientes desde el día de la cirugía, se podrían obtener resultados engañosos, pues es razonable pensar que los pacientes más jóvenes y saludables se recuperen más rápido que los pacientes mayores. En cambio, alinear desde el alta hospitalaria permitiría comparar de manera más justa el tiempo de rehabilitación.

[7] La solución propuesta por [26] fue probada solo en un conjunto de datos pequeño y no se aplicaba directamente a las observaciones censuradas. Tres años después [31] ampliaría esta técnica para censuras por intervalo.

3.1.2. Hipótesis sobre el comportamiento de la censura

A continuación se añaden unas notas sobre la censura independiente y la censura no informativa, que no son clasificaciones de las censuras, sino suposiciones que se pueden hacer acerca de estas.

Censura independiente

Más allá del modelo estadístico que explica el vector aleatorio (Y, Δ), no debemos perder de vista que nuestro interés se centra en la distribución de T, que puede ser descrita, como vimos, por su función de supervivencia o de riesgo, entre otras.

Definición 3.1 (Censura independiente)**.** Sea T una variable aleatoria continua[8] denotando el tiempo de supervivencia. La variable de censura C se dice que satisface la condición de **censura independiente** si y solo si

$$\lim_{\Delta t \to 0} \frac{P\big(t < T \leq t + \Delta t \,\big|\, T > t\big)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P\big(t < T \leq t + \Delta t \,\big|\, Y > t\big)}{\Delta t}, \quad c.t.p.$$

Notar que en la definición no se exige la independencia estocástica entre T y C (véase Figura 3.8).

Cuando las variables T y C son independientes (en sentido estocástico), hemos visto que la función de supervivencia de Y está relacionada con la de T mediante la expresión (3.1), siempre que se conozca la distribución de C. Ahora bien, en general, se tiene que

$$S_Y(t) = S_T(t)\, P(C > t) \leq S_T(t).$$

Cuando se dispone de información completa (no hay censuras), la función de supervivencia de Y se corresponde con la de T, pues $Y = \text{mín}(T, C) = T$, es decir,

$$P(T > t \mid T \leq C) = \frac{P(t < T \leq C)}{P(T \leq C)} = P(t < T) = S_T(t).$$

Supongamos ahora que la variable aleatoria T es continua y consideremos su función de riesgo $h(t)$. Se cumple entonces que, por la independencia[9] entre T y C,

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0^+} \frac{P\big(t < T \leq t + \Delta t \,\big|\, T > t\big)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P\big(t < T \leq t + \Delta t \,\big|\, T > t, C > t\big)}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{P\big(t < T \leq t + \Delta t \,\big|\, Y > t\big)}{\Delta t}. \end{aligned} \tag{3.5}$$

[8] Una generalización de la condición de censura independiente para distribuciones arbitrarias del tiempo de supervivencia se puede encontrar en [21].

[9] La independencia (en sentido estocástico) entre T y C es una **condición suficiente, pero no necesaria** para que ocurra (3.5) (veáse [21], pág. 14). Por ejemplo, en el modelo de censura de tipo II, C_i no es independiente de T_i, pero se cumple (3.5), pues: $T_i, T_{(r)} \geq t \Rightarrow T_i \geq t$.

Observación 3.5. En términos más intuitivos, la expresión (3.5) demuestra que la tasa de riesgo de un sujeto que permanece en el estudio (individuo 'en riesgo') coincide con la tasa de riesgo de la población, es decir, si se verifica (3.5), los individuos bajo seguimiento que no hayan sido censurados serán representativos de la población objeto de estudio (véase [17]).

Por tanto, según (3.5), la función de riesgo de T puede ser recuperada de los datos censurados cuando observamos, para los individuos que permanecen en riesgo, que el evento de interés ocurra en un determinado instante de tiempo. Esta es la razón por la que la función de riesgo es una herramienta tan apropiada en el análisis de datos de supervivencia.

Censura no informativa

La **censura no informativa**[10](o 'no pronóstica') se da cuando el punto en el que se censura no guarda ninguna información acerca de la verosimilitud que el individuo tiene de experimentar el evento (su tiempo de supervivencia) [22]. En la Figura 3.8 se ilustra la relación entre censura independiente y la censura no informativa.

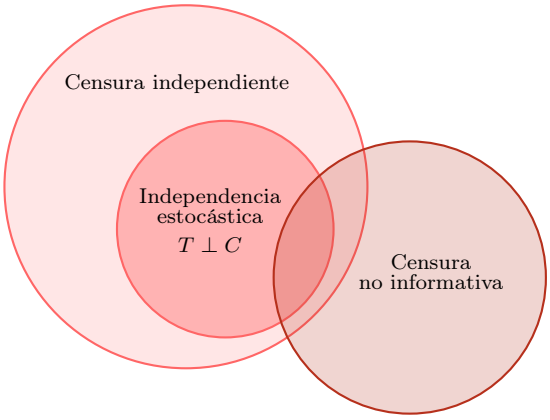


Figura 3.8: [25] Diagrama de Venn que ilustra la relación que existe entre los distintos tipos de hipótesis de censura. Independencia estocástica (entre T y C) implica censura independiente, pero no necesariamente ocurre el recíproco. Las tres hipótesis pueden ser ciertas a la vez, pero no por definición.

Observación 3.6. Dada la variable aleatoria $Y = \text{mín}(T, C)$ y una observación arbitraria suya, y, la condición

$$P(t \leq T < t + \Delta t \mid C = y, T > y) = P(t \leq T < t + \Delta t \mid T > y) \tag{3.6}$$

implica la no informatividad[11] de C en el sentido de que la información que proporciona una

[10] El primer modelo de censura no informativa fue introducido por Williams y Lagakos en 1977 [38].

[11] También se dice que el mecanismo de censura actúa de manera no informativa o no informativamente sobre los tiempos de vida.

observación censurada en y es tan solo que su tiempo de supervivencia supera el valor y. La condición (3.6) es un tipo particular del modelo MCAR mencionado en la introducción del capítulo (véase [28]).

En el contexto de ensayos clínicos, no es lo mismo que un paciente abandone el estudio debido a que cambie de hospital o que lo abandone debido a efectos adversos provocados por el tratamiento. Parece razonable que haya que tratar estas dos situaciones de manera distinta [35]. Los individuos no observados o perdidos durante el periodo de estudio (*loss to follow-up*), se clasifican según sean (véase [15]):

- *Withdrawal*: paciente que es obligado a abandonar el estudio por la presentación o aparición de determinadas circunstancias especificadas en el protocolo (por ejemplo, reacciones adversas graves), pero que es factible su seguimiento.
- *Drop-out*: paciente incluido en un ensayo clínico que no quiere o no puede continuar en el estudio ni el seguimiento (por ejemplo, no vuelve a las revisiones o abandona).

Ejemplo 3.3. En oncología, una terapia adyuvante es un tratamiento que tiene como objetivo eliminar las células tumorales que pudieran haber quedado tras la cirugía. Considérese un estudio clínico en el que se pretende evaluar la eficacia del fármaco *nivolumab* como terapia adyuvante en pacientes con cáncer de pulmón. Para ello se establecen dos grupos de pacientes: los que reciben el fármaco (grupo experimental) y los que reciben una solución salina (grupo de control), ambos administrados por vía intravenosa cada dos semanas durante un año. Supongamos que un paciente abandona el estudio a los seis meses debido a posibles efectos adversos que pudiera estar experimentando. Este individuo es menos probable que sobreviva que otro que continúe en el estudio pasados estos seis meses, y su censura es, por tanto, informativa. Asumir censura no informativa derivaría en una valoración sesgada del tratamiento.

3.2. Censura por la izquierda

La **censura por la izquierda** ocurre cuando solo se tiene conocimiento de que los eventos han ocurrido antes de un cierto momento. En particular, si ese momento es el inicio del estudio, los individuos que hayan presentado el evento antes de ingresar al mismo serán censurados por la izquierda; para ellos se sabe que su tiempo de supervivencia (no observado) es menor que el tiempo de censura observado (véase Ejemplo 3.4).

Ejemplo 3.4. [8] En un estudio pediátrico se desea estudiar la edad en la que los bebés son capaces de dar sus primeros pasos. Para ello se recoge una muestra de una guardería seleccionada al azar. Teniendo en cuenta que el evento de interés se corresponde con: 'empezar a caminar de manera autónoma', un dato será censurado por la izquierda si al inicio del estudio el infante ya es capaz de caminar por cuenta propia. Estas observaciones se caracterizan por la propiedad: $T_i \leq C_i$, donde C_i será la edad del bebé al inicio del estudio.

3.3. Censura por intervalo

La **censura por intervalo** normalmente aparece junto con evaluaciones periódicas en las que el evento de interés pudo ya haber ocurrido o no. Denotemos el proceso de observación por $\mathcal{Y} = (Y_0, Y_1, Y_2, \ldots)$ donde $Y_0 = 0 < Y_1 < Y_1 < \cdots$ son tiempos aleatorios de medición. El modelo supone la existencia de algún $j > 0$ tal que[12] $T \in (Y_j, Y_{j+1}]$, por lo que el valor exacto de T se encuentra acotado por un intervalo, pero continúa desconocido.

Más formalmente, observamos un vector aleatorio (L, R) tal que $T \in (L, R]$ con probabilidad uno de ocurrir. Las variables L y R son tiempos sucesivos ($L = Y_j$ y $R = Y_{j+1}$) observados, respectivamente, antes y después de la ocurrencia del evento de interés [28].

Ejemplo 3.5. Supongamos que una paciente es llamada regularmente para que le practiquen una biopsia como medida preventiva del cáncer de mama, por intervalos de dos años (edad $60, 62, 64, \ldots$), y se le detecta un tumor a los 64 años. Evidentemente, la paciente no desarrolló el tumor el día del examen médico; la información de la que disponemos es que la aparición del tumor tuvo lugar en algún momento entre los 62 y los 64 años.

Supongamos ahora que solo se observa si ocurre $T < 1$ o no. Esta situación es un caso particular de censura por intervalo, en la que solo sabemos si una observación está dentro del intervalo $(0, 1]$ o bien está en $(1, \infty)$. Este tipo de observaciones se dicen **doblemente censuradas**.

Ejemplo 3.6. Considérese un estudio en el que se pretende conocer el hábito tabáquico de cierta población. El objetivo es estimar la edad media de inicio en el consumo diario de tabaco. Una vez extraída una muestra, se encuesta a cada individuo sólo una vez y no se le sigue más en el tiempo. Los datos recogidos estarían doblemente censurados. Pongamos por caso que dos de los encuestados, i, j, tienen, respectivamente, 15 y 20 años de edad; dos posibles observaciones serían: $t_i \in (15, \infty)$ y $t_j \in (0, 20]$.

[12] La elección de intervalos semiabiertos se debe a que son más comunes en la práctica y reflejan de forma más natural situaciones en las que los sujetos son inspeccionados intermitentemente.

3.4. Truncamiento

El truncamiento tiene lugar cuando sólo se observan aquellos sujetos que experimentan el evento dentro de una ventana observacional (U, V), del resto no se realiza ningún seguimiento y, por tanto, no se obtiene información sobre ellos. El truncamiento 'oculta' a los individuos, por lo que el investigador no es consciente de su existencia. Por tanto, en este modelo, sólo una parte de la población es observada; la fracción de la población sin observar es completamente desconocida (no hay información parcial).

El **truncamiento por la izquierda** aparece cuando los sujetos han estado en riesgo antes de entrar en el estudio (véase Ejemplo 3.7). Formalmente, las observaciones truncadas por la izquierda son observaciones t_i tales que $t_i < U$.

El tipo de truncamiento por la izquierda más común ocurre cuando los individuos entran al estudio en diferentes edades y son seguidos desde que entran hasta que experimentan el evento, o bien sean censurados por la derecha. Por esta razón, a este modelo se le conoce por 'entrada tardía' (*delayed entry*).

Ejemplo 3.7. Considérese un estudio demográfico en el que se pretende analizar la mortalidad infantil en regiones con malnutrición. Los individuos que no sean estudiados desde su nacimiento estarán truncados por la izquierda en un tiempo correspondiente al momento en el que entren al estudio.

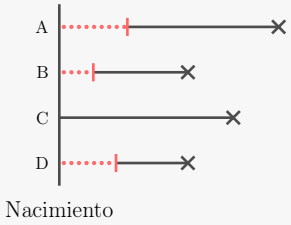


Figura 3.9: Modelo de entrada tardía

En la Figura 3.9 se muestran las observaciones para cuatro individuos. El sujeto C fue seguido desde su nacimiento, mientras que los sujetos A, B y D fueron incorporados más tarde al estudio. La raya roja representa el truncamiento por la izquierda; la línea punteada representa el periodo en el que los individuos permanecieron a riesgo y sin observar.

En la Figura 3.9 todos los individuos empiezan a estar en riesgo simultáneamente, esto no va a ser siempre así. Por ejemplo, en un estudio en el que se pretenda estudiar el tiempo hasta la aparición del Síndrome de la Inmunodeficiencia Adquirida (AIDS), los pacientes empiezan a estar en riesgo en el momento en el que la infección del Virus de la Inmunodeficiencia Humana (VIH) tiene lugar (véase Figura 3.10).

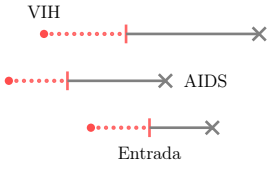


Figura 3.10: Esquema truncamiento por la izquierda

El **truncamiento por la derecha** ocurre cuando sólo los individuos que han experimentado el evento de interés son incluidos en el muestra (se observan). Formalmente, las observaciones truncadas por la izquierda son aquellas tales que $t_i > V$.

Una muestra truncada por la derecha se puede encontrar en investigaciones médicas en las que haya un periodo de latencia (o de incubación) de una enfermedad. Este modelo adquiere particular relevancia en estudios de seropositividad.

Ejemplo 3.8. El periodo de latencia del síndrome de la inmunodeficiencia adquirida se define como el tiempo desde la infección del virus VIH hasta el diagnóstico de la enfermedad. La transfusión de sangre contaminada es una vía de transmisión del VIH; si una persona recibe una transfusión de sangre y meses después se le diagnostica la enfermedad, se puede seguir la pista del momento de la infección[a]. El número de individuos infectados es desconocido y solo se dispone de información de aquellos individuos que se infectaron y desarrollaron la enfermedad dentro de cierto periodo de tiempo (por ejemplo, entre 1985 y 1990). Los individuos que aún estén por desarrollar la enfermedad son desconocidos por el investigador y no se incluyen en la muestra.

[a]Consultar https://cran.r-project.org/web/packages/coxrt/vignettes/coxrt-vignette.html

También puede existir truncamiento por la izquierda y por la derecha a la vez (véase Ejemplo 3.9)

Ejemplo 3.9. Considérese que se está realizando un estudio para conocer la edad en la que los niños contraen paperas (enfermedad que en la mayoría de los casos ocurre una vez en la vida) en cierta región. Para ello, fueron revisados los informes de un colegio de la zona. Los datos incluyen los seis años de educación primaria más los primeros dos de instituto. En la Figura 3.11 se muestran cinco observaciones procedentes del estudio. Se destaca la última de ellas (sujeto E), que empezó el colegio con siete años.

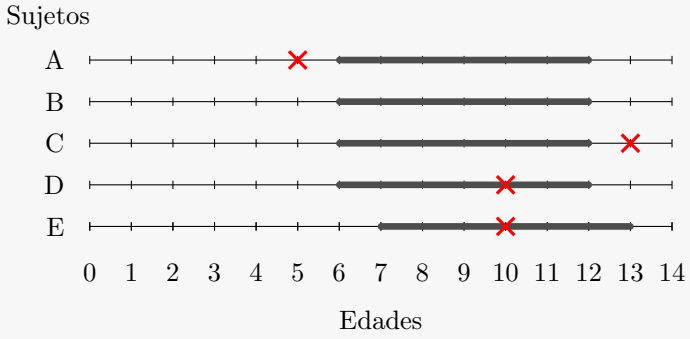


Figura 3.11: Esquema truncamiento por la izquierda y derecha

Denotemos por X la edad en la que se contraen paperas. Los tres primeros individuos (A, B y C) no fueron observados; mientras que los últimos dos (D y E) producen observaciones truncadas por la izquierda y por la derecha: $(X = 10 \mid 6 < X \leq 12)$, $(X = 10 \mid 7 < X \leq 13)$.

Capítulo 4

Verosimilitud con información incompleta

La función de verosimilitud es una función dependiente de uno o varios parámetros de una distribución mediante la cual pueden realizarse inferencias estadísticas (estimación y contraste de hipótesis) acerca del valor paramétrico desconocido sobre la base de la muestra de observaciones disponibles [9].

Definición 4.1. Sea $(x_1, \ldots, x_n)$ una muestra de observaciones independientes a partir de una variable aleatoria X cuya distribución depende de $\theta \in \Theta$ (siendo Θ el espacio paramétrico o conjunto de valores que puede tomar θ) y sea $f(x;\theta)$ la función de probabilidad/densidad de X. La **función de verosimilitud** de $(x_1, \ldots, x_n)$ se define como la aplicación $L(\cdot\,; x_1, \ldots, x_n) : \Theta \to [0, \infty)$ que a cada $\theta \in \Theta$ le asocia el valor

$$L(\theta) \coloneqq L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdots f(x_n; \theta).$$

Observación 4.1. La función de verosimilitud, si bien coincide en valor con la función de probabilidad/densidad conjunta de la muestra aleatoria simple $(X_1, \ldots, X_n)$, no se debe interpretar como tal, sino como una función únicamente dependiente de θ, dada la muestra $(x_1, \ldots, x_n)$.

Intuitivamente, de acuerdo con los datos observados, cada valor de los parámetros nos dará una puntuación a la que llamamos verosimilitud, medida que refleja cuán creíbles son dichos valores.

> **Ejemplo 4.1.** Si comparamos la función de verosimilitud en dos valores de Θ, digamos θ_1 y θ_2, y encontramos $L(\theta_1) > L(\theta_2)$, entonces la muestra que hemos observado es **más verosímil** que ocurra si $\theta = \theta_1$ que si $\theta = \theta_2$. En otras palabras, θ_1 es un valor más plausible para θ que θ_2.

De acuerdo con el tipo de observación en el contexto del análisis de supervivencia, se tienen las siguientes contribuciones (suponiendo Censura no informativa) a la función de verosimilitud (véase [20]):

Tipo de observación	T	L_i
Completa	$T = t_i$	$f(t_i;\theta)$
Censurada por la derecha	$T > t_i$	$S(t_i;\theta)$
Censurada por la izquierda	$T < t_i$	$F(t_i;\theta)$
Censurada por intervalo	$l_i < T < r_i$	$F(r_i;\theta) - F(l_i;\theta)$
Truncada por la izquierda	$T = t_i \mid T > u$	$f(t_i;\theta)/S(u;\theta)$
Truncada por la derecha	$T = t_i \mid T < v$	$f(t_i;\theta)/F(v;\theta)$

Tabla 4.1: Contribución a la función de verosimilitud según el tipo de observación

4.1. Verosimilitud para observaciones censuradas por la derecha

Por ser la más habitual en la práctica, estudiaremos en detalle la construcción de la función de verosimilitud cuando las observaciones están censuradas por la derecha.

4.1.1. Verosimilitud en presencia de censura aleatoria

Sean $(T_1, C_1), \ldots, (T_n, C_n)$ vectores aleatorios independientes dos a dos, que por estar en presencia de censura aleatoria, también se tiene la independencia de sus componentes marginales, T_i y C_i. Denotemos, como siempre, $Y_i = \text{mín}(T_i, C_i)$ a la variable observada y $\Delta_i = I(T_i \leq C_i)$ al indicador del evento. Los datos consistirán en los pares independientes $(Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)$.

Supongamos que las variables aleatorias $T_1, \ldots, T_n$ están idénticamente distribuidas con función de distribución $F(y;\boldsymbol{\theta})$, función de supervivencia $S(y;\boldsymbol{\theta})$ y función de densidad $f(y;\boldsymbol{\theta})$, donde $\boldsymbol{\theta} \in \Theta^p$ es un vector de parámetros p-dimensional. Por simplicidad de notación, escribiremos a partir de ahora las funciones sin el parámetro $\boldsymbol{\theta}$.

Denotemos por $G_i(y)$ a la función de supervivencia de la variable de censura C_i y por $g_i(y)$ a su densidad. No vamos a suponer que los tiempos de censura estén igualmente distribuidos[1], pero sí supondremos independencia entre las variables T_i y C_i.

Supongamos que las variables aleatorias T_i y C_i son continuas[2]. Así, Y_i también será continua.

[1] Basta considerar un ensayo clínico en el que se produzcan censuras por diferentes motivos, por ejemplo: por cambio de hospital, por finalización del estudio y por retirada debido a toxicidad en la quimioterapia.

[2] No es cierto que los tiempos de censura C_i vayan a ser siempre variables continuas. A menudo son mixturas finitas de distribuciones discretas y continuas. Para una demostración más general consultar [21], pág. 16.

Por la suposición de independencia, la función de verosimilitud coincide con el producto de las funciones de verosimilitud de cada observación por separado, $L_i(\boldsymbol{\theta})$, y por tanto,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}).$$

Teniendo en cuenta que el vector aleatorio (Y_i, Δ_i) es mixto (Y_i es una variable aleatoria continua y Δ_i, discreta), la verosimilitud $L_i(\boldsymbol{\theta})$ para la observación (y_i, δ_i) vendría dada por:

$$f_{Y_i}(y_i \mid \delta_i = 1)P(\delta_i = 1) = f_{Y_i}(y_i \mid T_i \leq C_i)P(T_i \leq C_i),$$

si corresponde a la observación de un evento.

Reescribiendo $f_{Y_i}(y_i \mid T_i \leq C_i)$ en términos de la otra distribución condicionada,

$$f_{Y_i}(y_i \mid T_i \leq C_i) = f_{T_i}(y_i \mid T_i \leq C_i) = f_{T_i}(y_i)\,\frac{P(T_i \leq C_i \mid T_i = y_i)}{P(T_i \leq C_i)},$$

se sigue que

$$\begin{aligned} f_{Y_i}(y_i \mid T_i \leq C_i)P(T_i \leq C_i) &= f_{T_i}(y_i)P(T_i \leq C_i \mid T_i = y_i) \\ &= f(y_i)P(y_i \leq C_i) = f(y_i)G_i(y_i). \end{aligned}$$

En cambio, si lo que se observa es una censura, la contribución de (y_i, δ_i) a la verosimilitud se corresponde con

$$f_{Y_i}(y_i \mid \delta_i = 0)P(\delta_i = 0) = f_{Y_i}(y_i \mid T_i > C_i)P(T_i > C_i)$$

De nuevo, reescribiendo $f_{Y_i}(y_i \mid T_i > C_i)$ en términos de la otra distribución condicionada,

$$f_{Y_i}(y_i \mid T_i > C_i) = f_{C_i}(y_i \mid T_i > C_i) = f_{C_i}(y_i)P(T_i > C_i \mid C_i = y_i)/P(T_i > C_i),$$

se sigue que

$$\begin{aligned} f_{Y_i}(y_i \mid T_i > C_i)P(T_i > C_i) &= f_{C_i}(y_i)P(T_i > C_i \mid C_i = y_i) \\ &= g_i(y_i)P(T_i > y_i) = g_i(y_i)S(y_i). \end{aligned}$$

En resumen:

- Para $\delta = 0$ observamos un tiempo censurado y, y sabemos que $C = y$ y que $T > y$. Su densidad será $g_i(y)\,S(y)$.

- Para $\delta = 1$ observamos un tiempo de supervivencia y, y sabemos que $T = y$ y que $C \geq y$. Su densidad será $f(y)\,G_i(y)$.

Por tanto, la contribución de un individuo a la verosimilitud vendrá dada por la siguiente expresión:

$$\big(f(y)\,G_i(y)\big)^{\delta}\,\big(g_i(y)\,S(y)\big)^{1-\delta},$$

y la verosimilitud de $\boldsymbol{\theta}$, dada la muestra de observaciones $(y_1,\delta_1),\ldots,(y_n,\delta_n)$, será:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\Big\{f(y_i)^{\delta_i}S(y_i)^{1-\delta_i}\Big\}\prod_{i=1}^{n}\Big\{G_i(y_i)^{\delta_i}g_i(y_i)^{1-\delta_i}\Big\}. \tag{4.1}$$

Si C_i actúa no informativamente sobre T_i y ambas variables son independientes, entonces la estimación de $\boldsymbol{\theta}$ a través de la función de verosimilitud no dependerá[3] de C_i y, por tanto, los factores $G_i(y_i)^{\delta_i}$ y $g_i(y_i)^{1-\delta_i}$ no serán informativos para hacer inferencia sobre la función de supervivencia. Podremos trabajar entonces con el primer producto de (4.1), y se tendrá:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n}\Big\{f(y_i)^{\delta_i}S(y_i)^{1-\delta_i}\Big\} = \prod_{i:\delta_i=1} f(y_i)\prod_{i:\delta_i=0} S(y_i). \tag{4.2}$$

Equivalentemente, dado que $f(y) = h(y)\,S(y)$, también:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n}\Big\{h(y_i)^{\delta_i}S(y_i)\Big\}.$$

Los resultados típicos de la teoría de estimación máximo-verosímil pueden ser aplicados para obtener el estimador máximo-verosímil de θ, así como su distribución asintótica. Sin embargo, para la mayoría de distribuciones del tiempo de supervivencia, la función *score* (derivada de la función log-verosimilitud con respecto a θ) y la cantidad de información de Fisher no son fáciles de calcular.

Ejemplo 4.2. Supongamos que se observan $(Y_1,\Delta_1),\ldots,(Y_n,\Delta_n)$ tales que $Y_i = \text{mín}(T_i, C_i)$, donde C_i es un tiempo de censura aleatorio y $T_i \sim \text{Exp}(\lambda)$, con C_i y T_i independientes. La variable Δ_i se define tal que $\Delta_i = 1$ si $T_i \leq C_i$ y $\Delta_i = 0$ en otro caso. Dada la realización muestral $(y_1,\delta_1),\ldots,(y_n,\delta_n)$, encuentra una expresión para el estimador máximo-verosímil de λ que tenga en cuenta las observaciones censuradas.

Solución:

Denotemos por $\mathcal{U} \coloneqq \{i : \delta_i = 1\}$ al conjunto de índices de las observaciones sin censurar, y por $\mathcal{C} \coloneqq \{i : \delta_i = 0\}$ a los índices de las observaciones censuradas. Tenemos (véase Modelo Exponencial):

$$f(y_i;\lambda) = \lambda\exp\left(-\lambda\,y_i\right) \quad \text{y} \quad S(y_i;\lambda) = \exp\left(-\lambda\,y_i\right).$$

[3] En otras palabras, el mecanismo de censura no contiene ninguna información sobre el parámetro de interés $\boldsymbol{\theta}$, digamos que $C_i \sim \mathcal{D}_i(\boldsymbol{\eta_i})$, con $\boldsymbol{\eta_i}$ un vector de parámetros diferente de $\boldsymbol{\theta}$.

Teniendo en cuenta (4.2), se sigue que

$$L(\lambda; (y_1, \delta_1), \ldots, (y_n, \delta_n)) \propto \prod_{i \in \mathcal{U}} \lambda \exp(-\lambda y_i) \prod_{i \in \mathcal{C}} \exp(-\lambda y_i)$$
$$= \lambda^{n_U} \exp\left(-\lambda \sum_{i \in \mathcal{U}} y_i\right) \exp\left(-\lambda \sum_{i \in \mathcal{C}} y_i\right),$$

donde n_U denota el número de observaciones sin censurar.

La función log-verosimilitud es, por tanto,

$$\ell(\lambda; (y_1, \delta_1), \ldots, (y_n, \delta_n)) = n_U \log \lambda - \lambda \sum_{i \in \mathcal{U}} y_i - \lambda \sum_{i \in \mathcal{C}} y_i.$$

Derivando esta función con respecto a λ e igulando a cero, se obtiene el estimador máximo-verosímil de λ. Así,

$$\frac{\partial}{\partial \lambda} \ell(\lambda; (y_1, \delta_1), \ldots, (y_n, \delta_n)) = \frac{n_U}{\lambda} - \sum_{i \in \mathcal{U}} y_i - \sum_{i \in \mathcal{C}} y_i,$$

de donde se sigue finalmente, por los criterios de la primera y segunda derivada, que

$$\widehat{\lambda}_{\text{MV}} = \frac{n_U}{\sum_{i \in \mathcal{U}} y_i + \sum_{i \in \mathcal{C}} y_i} = n_U \left(\sum_{i=1}^{n} y_i\right)^{-1}.$$

4.1.2. Verosimilitud en presencia de censuras de tipo I y II

Consideremos $T_1, \ldots, T_n$ tiempos de supervivencia independientes e idénticamente distribuidos, cada uno asociado a un tiempo de censura $C_i > 0$ (fijo o aleatorio) y con $T_i \sim \mathcal{D}(\boldsymbol{\theta})$, donde $\boldsymbol{\theta} \in \Theta^p$. Consideremos una realización muestral $\{(y_i, \delta_i)\}_{i=1}^n$ de $(Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)$. De nuevo, por simplicidad de notación, no escribiremos el parámetro $\boldsymbol{\theta}$.

Censura de tipo I

En el caso de Censura de tipo I, donde $C_i = c$ es un valor fijo para todo $i = 1, \ldots, n$, la probabilidad de que un individuo sea censurado y el tiempo observado sea c es:

$$P(Y_i = c, \delta_i = 0) = P(Y_i = c \mid \delta_i = 0) P(\delta_i = 0)$$
$$= P(\delta_i = 0) = P(T_i > c) = S(c),$$

mientras que si un individuo experimenta el evento de interés durante la observación y en el instante y_i, se tiene:

$$f_{Y_i}(y_i \mid \delta_i = 1)\, P(\delta_i = 1) = f_{T_i}(y_i \mid T_i \leq c)\, P(T_i \leq c) = f(y_i).$$

Por tanto, la contribución de cada observación (y_i, δ_i) a la función de verosimilitud es: $f(y_i)^{\delta_i}\, S(y_i)^{1-\delta_i}$. Así,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i)^{\delta_i}\, S(y_i)^{1-\delta_i} = \prod_{i:\delta=1} f(t_i) \prod_{i:\delta=0} S(c_i). \tag{4.3}$$

Nótese que (4.3) no tiene la misma forma que en la Definición 4.1, forma que hubiera quedado reflejada si se hubieran observado todos los eventos.

Censura de tipo II

Supongamos ahora que de los mismos tiempos de supervivencia $T_1, \ldots, T_n$, se observan los r tiempos de fallo más pequeños, dejando a $n - r$ tiempos censurados por la derecha.

De la teoría de estadísticos de orden (véase, por ejemplo, [14]) se tiene que la función de densidad conjunta de $T_{(1)}, \ldots, T_{(r)}$ es:

$$f_{(T_{(1)},\ldots,T_{(r)})}(t_1, \ldots, t_r) = \frac{n!}{(n-r)!} \left(\prod_{i=1}^{r} f(t_i) \right) \left[1 - F(t_{(r)})\right]^{n-r} I(t_1 < \cdots < t_n).$$

Así, dadas las realizaciones muestrales, la expresión anterior coincidirá en valor con la verosimilitud de $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) = \frac{n!}{(n-r)!} \left(\prod_{i=1}^{r} f(t_i) \right) \left[S(t_{(r)})\right]^{n-r} I(t_1 < \cdots < t_n). \tag{4.4}$$

El producto $\prod_{i=1}^{r} f(t_i)$ corresponde con los r eventos observados, mientras que $\left[S(t_{(r)})\right]^{n-r}$ constituye la aportación de las observaciones censuradas en el r-ésimo fallo.

Observación 4.2. Las verosimilitudes (4.3) y (4.4) son proporcionales, pues son idénticas en forma, salvo por el factor $n!/(n-r)!$

Capítulo 5

Técnicas no paramétricas

En este capítulo se introducen las técnicas no paramétricas más habituales que se utilizan para analizar estadísticamente datos que no se ajustan a ninguna distribución conocida (exponencial, Weibull, etc.).

5.1. Introducción

En general, las técnicas no paramétricas en análisis de supervivencia son útiles para hacer un análisis preliminar, que nos puede ayudar, si procede, a elegir un modelo paramétrico plausible. En el caso de no poder elegir un modelo, las técnicas no paramétricas nos proporcionan herramientas para estudiar los datos de tipo tiempo de vida, lo que nos permitirá, entre otras cosas, aproximar la función de supervivencia a partir de los datos o contrastar si las funciones de supervivencia de varios grupos de la población coinciden.

Los datos relacionados con el comportamiento humano no suelen seguir las mismas pautas, por lo que normalmente es inviable ajustar un modelo; mientras que en estudios de fiabilidad y control de calidad, suele ser factible encontrar un modelo adecuado.

Comenzaremos estudiando la estimación no paramétrica de la función de supervivencia, $S(t)$, usando datos no censurados, para realizar inferencia en modelos de tiempo de vida.

Sean $T_1, T_2, \ldots, T_n$ variables aleatorias denotando los tiempos de supervivencia, independientes e idénticamente distribuidas. En el modelo sin censuras, donde todos los eventos son observados, podemos hacer una encuesta a n individuos y obtener la muestra de observaciones $(t_1, t_2, \ldots, t_n)$ y la muestra ordenada $(t_{(1)}, t_{(2)}, \ldots, t_{(n)})$.

Teniendo presente el teorema de Glivenko-Cantelli[1], consideremos como estimador de la función de supervivencia:

$$\widehat{S}(t) = 1 - \widehat{F}_n(t),$$

donde $\widehat{F}_n(t)$ es la función de distribución empírica (Figura 5.1).

Formalmente, la función de distribución empírica, fijado $t \in \mathbb{R}$, es una variable aleatoria que a cada $\omega \in \Omega$ le hace corresponder el valor:

$$\widehat{F}_n(t)(\omega) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}\big(T_i(\omega)\big),$$

o equivalentemente,

$$\widehat{F}_n(t)(\omega) = \frac{1}{n} \sum_{i=1}^{n} I_{\{T_i(\omega) \leq t\}}.$$

Se puede probar fácilmente que la función de distribución empírica es un estimador insesgado de la función de distribución $F(t)$. Además, fijado un $\omega \in \Omega$, es $\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(t_i \leq t)$, esto es, una función escalonada como muestra la Figura 5.1.

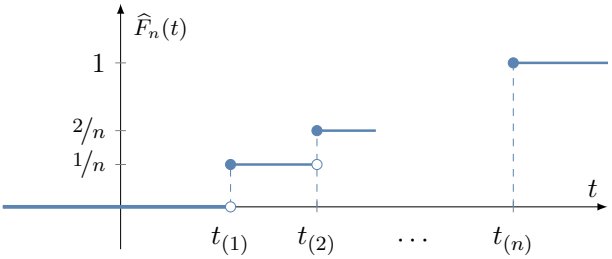


Figura 5.1: Función de distribución empírica

Por tanto, consideramos $\widehat{S}(t)$ como

$$\widehat{S}(t_{(i)}) = 1 - p_i,$$

donde $p_i = i/n$ (proporción de datos menores o iguales que $t_{(i)}$). En la literatura especializada, se recomienda tomar una versión corregida[2] de p_i, a saber,

$$p_i = \frac{i - a}{n - 2a + 1}, \quad i = 1, 2, \ldots, n, \tag{5.1}$$

[1]El teorema de Glivenko-Cantelli asegura que la función de distribución empírica se acerca cada vez más a la función de distribución teórica a medida que el tamaño de la muestra aumenta. Más adelante, Kolmogorov probó que se acercaba (convergía) uniformemente.

[2]La expresión (5.1) es la expresión general para la obtención de los puntos p_i, llamados puntos de posición gráfica (*plotting positions*). La expresión con $a = 1/2$ fue propuesta por Hazen en 1930 y es la que se utiliza habitualmente en el cálculo de cuantiles muestrales. Otra propuesta que se elige con frecuencia es $a = 3/8$.

fijado $a \in (0,1)$. Se llama entonces **estimador natural** de la función de supervivencia a

$$\widehat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{T_i > t\}} = \frac{\text{nº de individuos con tiempo de supervivencia} > t}{n}.$$

Si la muestra contiene observaciones censuradas, este no es un estimador adecuado porque trata las censuras como instantes de fallo y es probable que infraestime la función de supervivencia.

5.2. Método Kaplan-Meier

Este método fue desarrollado en 1958 por Kaplan y Meier, y tiene como objetivo calcular la probabilidad de supervivencia de un individuo de la muestra en un instante $t_i, i = 1, \ldots, n$. Para ello, se realiza el producto de la probabilidad de supervivencia en el instante t_{i-1} por la probabilidad condicionada de sobrevivir al instante t_i si se ha sobrevivido hasta el instante t_{i-1}, es decir,

$$S(t_i) = S(t_{i-1})\, P(T > t_i \mid T > t_{i-1}).$$

Consideremos la muestra $(Y_1, \Delta_1), (Y_2, \Delta_2), \ldots, (Y_n, \Delta_n)$, donde Y_i es la variable tiempo observado y Δ_i es el indicador de no censura. Supongamos que contiene n_s observaciones sin censurar y n_c observaciones censuradas. Denotemos las observaciones sin censurar (ya ordenadas de menor a mayor) como (véase [8]):

$$t'_1 \leq t'_2 \leq \cdots \leq t'_{n_s}.$$

Tomemos $t'_0 = 0$ y notemos que para una observación arbitraria y sin censurar t'_k, podemos escribir

$$\begin{aligned}
S(t'_k) &= P(T > t'_k) \\
&= P(T > t'_k \mid T > t'_{k-1})\, P(T > t'_{k-1}) \\
&= P(T > t'_k \mid T > t'_{k-1})\, P(T > t'_{k-1} \mid T > t'_{k-2})\, P(T > t'_{k-2}) \\
&\vdots \\
&= \left[\prod_{j=1}^{k} P(T > t'_j \mid T > t'_{j-1})\right] P(T > t'_0) \\
&= \prod_{j=1}^{k} P(T > t'_j \mid T > t'_{j-1}).
\end{aligned}$$

Cada una de las probabilidades

$$P(T > t'_j \mid T > t'_{j-1}) = 1 - P(T \leq t'_j \mid T > t'_{j-1})$$

puede ser estimada mediante

$$1 - \frac{1}{\#\{i : Y_i \geq t'_j\}},$$

donde $\{i : Y_i > t'_j\}$ es el conjunto de todos los individuos que han sobrevivido al tiempo t'_{j-1} y que, por tanto, están a riesgo en t'_j. Así, una estimación de la función de supervivencia en t'_k es:

$$\widehat{S}(t'_k) = \prod_{j=1}^{k} \left[1 - \frac{1}{\#\{i : Y_i \geq t'_j\}}\right] = \prod_{j:y_j \leq t'_j} \left[1 - \frac{1}{\#\{i : Y_i \geq y_j\}}\right]^{\delta_j}.$$

En todos los demás puntos $t \in [t'_k, t'_{k+1})$, el estimador toma el mismo valor. El estimador de Kaplan-Meier[3] viene dado por:

$$\widehat{S}_{\mathrm{KM}}(t) = \prod_{j:Y_j \leq t} \left[1 - \frac{1}{\#\{i : Y_i \geq Y_j\}}\right]^{\Delta_j}.$$

Denotando por $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ a los estadísticos de orden de las variables Y_i observadas y por $\Delta_{(i)}$ a la variable indicadora asociada con $Y_{(i)}$, el estimador de Kaplan-Meier se puede escribir como

$$\begin{aligned}
\widehat{S}_{\mathrm{KM}}(t) &= \prod_{j:Y_{(j)} \leq t} \left[1 - \frac{1}{\#\{i : Y_i \geq Y_{(j)}\}}\right]^{\Delta_{(j)}} \\
&= \prod_{j:Y_{(j)} \leq t} \left[1 - \frac{1}{n-j+1}\right]^{\Delta_{(j)}} \\
&= \prod_{j:Y_{(j)} \leq t} \left(\frac{n-j}{n-j+1}\right)^{\Delta_{(j)}},
\end{aligned}$$

o, equivalentemente,

$$\widehat{S}_{\mathrm{KM}}(t) = \prod_{j:Y_{(j)} \leq t} \left(1 - \frac{\Delta_{(j)}}{n-j+1}\right).$$

Discutiremos ahora algunas propiedades básicas del estimador de Kaplan-Meier. Supongamos que todas las observaciones están sin censurar. Es fácil ver que en este caso el estimador para F se reduce a la habitual función de distribución empírica: para cualquier $t \in \left[T_{(k)}, T_{(k+1)}\right)$, se tiene que

$$\widehat{F}_{\mathrm{KM}}(t) = 1 - \widehat{S}_{\mathrm{KM}}(t) = 1 - \prod_{j:Y_{(j)} \leq t} \left(\frac{n-j}{n-j+1}\right) = 1 - \prod_{j=1}^{k} \left(\frac{n-j}{n-j+1}\right) = 1 - \frac{n-k}{n} = \frac{k}{n}.$$

Cuando el momento en el que se observa el evento de interés coincide exactamente con otra observación que está censurada, el convenio sugiere considerar que ha ocurrido antes la observación

[3] También llamado estimador del **límite-producto**.

sin censurar. Cuando hay más observaciones repetidas, digamos que hay D_j veces en las que el evento de interés ha sido observado en el j-ésimo instante, entonces denotando los diferentes tiempos de observación por $Y'_{(1)} \leq Y'_{(2)} \leq \cdots \leq Y'_{(r)}$, y por $\Delta'_{(i)}$ la función indicador asociada, el estimador de Kaplan-Meier se puede escribir como:

$$\widehat{S}_{\mathrm{KM}}(t) = \prod_{j:Y'_{(j)} \leq t} \left(1 - \frac{D_j}{N_j}\right)^{\Delta'_{(j)}},$$

donde N_j denota el número de individuos que permanecen a riesgo en el instante $Y'_{(j)}$.

Dada una muestra de observaciones, la expresión habitual de la estimación de Kaplan-Meier es:

$$\widehat{S}_{\mathrm{KM}}(t) = \prod_{\{i:t_i \leq t\}} \left(1 - \frac{d_i}{n_i}\right), \tag{5.2}$$

donde d_i es el número de eventos en t_i, y n_i es el número de individuos a riesgo antes de t_i.

Es importante notar que a la hora de evaluar la estimación de Kaplan-Meier sólo los eventos que ocurran en los tiempos de observación $\{t_i\}_{i=1}^n$ son importantes. Por tanto, entre dos eventos, digamos en los instantes t_i y t_{i+1}, la estimación de la función de supervivencia es constante. Esto permite reescribir (5.2) de forma recursiva:

$$\widehat{S}_{\mathrm{KM}}(t_k) = \widehat{S}_{\mathrm{KM}}(t_{k-1})\, \frac{n_k - d_k}{n_k}.$$

Observación 5.1. El estimador de Kaplan-Meier es una función continua por la derecha[4] y escalonada que se mantiene constante cuando pasa por las censuras y cuyos saltos se sitúan en las observaciones sin censurar. Además, la altura de los escalones de $\widehat{S}_{\mathrm{KM}}$ siempre es aleatoria, haya observaciones repetidas o no[5].

Cuando el mayor de los tiempos observados en la muestra, $t_{(n)}$, corresponde con un evento, la estimación de Kaplan-Meier toma el valor 0 a partir de ese instante. Sin embargo, si $t_{(n)}$ es una observación censurada, $\widehat{S}_{\mathrm{KM}}(t)$ no está bien definida para $t > t_{(n)}$ porque no se puede saber cuándo este último individuo hubiera fallado de no haber sido censurado. Más formalmente, se tiene que $\widehat{S}_{\mathrm{KM}}(t) = \widehat{S}_{\mathrm{KM}}(t_{(n)}) \neq 0$ para todo $t > t_{(n)}$, produciendo un sesgo positivo, ya que $\lim_{t\to\infty} S(t) = 0$. Tampoco se puede tomar $\widehat{S}_{\mathrm{KM}}(t) = 0$ para $t > t_{(n)}$ porque produce un sesgo negativo.

La construcción del estimador de Kaplan-Meier tiene una interpretación intuitiva gracias al algoritmo de redistribución de masa (*redistribute-to-the-right algorithm*) de Efron (véase Ejemplo 5.1), quien en 1967 demostró que coincidía con el estimador.

[4] Cuando se toma $\{j : t_{(j)} < t\}$ entonces $\widehat{S}(t)$ es una función escalonada continua por la izquierda.

[5] Si no hay empates, conocemos la altura de los escalones de la función de distribución empírica, estos son: $1/n$.

Ejemplo 5.1. Considérense los siguientes datos de tiempo de vida:

$$2, 2, 3^+, 5, 5^+, 7, 9, 16, 16, 18^+.$$

1. Ordenar las observaciones de forma creciente, situando las censuras después de los eventos cuando se produzcan repeticiones.
2. Colocar un peso de $1/n$ en cada observación.
3. Empezando de izquierda a derecha, redistribuir el peso de cada censura equitativamente entre el resto de observaciones (censuradas y sin censurar).
4. Repetir el paso anterior hasta que todas las observaciones censuradas (excepto la observación más grande) no tengan peso.

Paso 1	2	2	3^+	5	5^+	7	9	16	16	18^+
Paso 2	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Paso 3	$\downarrow$	$\downarrow$	$\hookrightarrow$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$	$\frac{1}{70}$
Paso 4	$\downarrow$	$\downarrow$		$\downarrow$	$\hookrightarrow$	$\frac{1}{5}\frac{8}{70}$	$\frac{1}{5}\frac{8}{70}$	$\frac{1}{5}\frac{8}{70}$	$\frac{1}{5}\frac{8}{70}$	$\frac{1}{5}\frac{8}{70}$
Masa total	$\frac{2}{10}$		0	$\frac{8}{70}$	0	$\frac{24}{175}$	$\frac{24}{175}$	$\frac{48}{175}$		> 18

Figura 5.2: Algoritmo de redistribución de masa

Por consiguiente, la estimación de la supervivencia en el instante t vendrá dada por 1 menos la suma de los pesos de todos los instantes anteriores a t o iguales a t:

$$\widehat{S}_{\mathrm{KM}}(t) = \begin{cases} 1, & \text{si } 0 \leq t < 2, \\ 0.8, & \text{si } 2 \leq t < 5, \\ 0.69, & \text{si } 5 \leq t < 7, \\ 0.55, & \text{si } 7 \leq t < 9, \\ 0.41, & \text{si } 9 \leq t < 16, \\ 0.14, & \text{si } 16 \leq t \leq 18, \\ \text{no definido}, & \text{si } t > 18. \end{cases}$$

Observación 5.2. Desde un punto de vista teórico, si asumimos que el tiempo de supervivencia es continuo, no es posible que con probabilidad positiva se produzcan empates en tiempos de fallo. Sin embargo, en la práctica este hecho se puede dar debido a la forma en la que tomamos

las observaciones (tomando los tiempos de supervivencia como tiempos de observación). Una forma de romper esos empates podría consistir en considerar que realmente no han ocurrido en un tiempo t, sino que han ocurrido de manera secuencial en instantes de tiempo muy próximos (infinitesimalmente) al tiempo t. Este esquema evita los empates, pero es inmediato comprobar que el factor que contribuye a la estimación de la función de supervivencia en el tiempo t es el mismo[6] que el dado por Kaplan-Meier (véase [1]).

Proposición 5.1. *El estimador de Kaplan-Meier es el estimador no paramétrico máximo-verosímil de la función de supervivencia (véase [18], pág. 475).*

5.2.1. Incertidumbre en la estimación de Kaplan-Meier

El mejor estimador de la varianza del estimador de Kaplan-Meier de $S(t)$ para un tiempo t fijo es el que proporciona la Fórmula de Greenwood:

$$\widehat{\mathrm{Var}}\left(\widehat{S}_{\mathrm{KM}}(t)\right) = \widehat{S}^2_{\mathrm{KM}}(t) \sum_{\left\{i:T_{(i)}\leq t\right\}} \frac{D_i}{N_i\left(N_i - D_i\right)}. \tag{5.3}$$

Demostración. Usando el Método Delta[7] se tiene que

$$\mathrm{Var}[f(T)] \approx \left[f'\big(E[T]\big)\right]^2 \mathrm{Var}(T).$$

Así, tomando $f(x) = \log x$, se sigue que

$$\mathrm{Var}\left(\log T\right) \approx \left(\frac{1}{E[T]}\right)^2 \mathrm{Var}(T). \tag{5.4}$$

Habida cuenta de que el riesgo h_i es una proporción (cuando se discretiza la v.a. T), podemos obtener su varianza a través de la distribución binomial, de forma que si $\hat{p} = (\#\text{ eventos})/n$ es un estimador de p, entonces, $\mathrm{Var}(\hat{p}) = p\,(1-p)/n$. Así, para cada i, se tiene que

$$\mathrm{Var}\Big(\widehat{h}_i\Big) = \frac{h_i\,(1-h_i)}{N_i}.$$

[6] Supongamos que antes del tiempo t_j hay n_j individuos a riesgo, y en t_j se producen d_j eventos, que si los suponemos separados infinitesimalmente se tendría:

$$\left(1-\frac{1}{n_j}\right)\left(1-\frac{1}{n_j-1}\right)\cdots\left(1-\frac{1}{n_j-d_j+1}\right) = \left(\frac{n_j-1}{n_j}\right)\left(\frac{n_j-2}{n_j-1}\right)\cdots\left(\frac{n_j-d_j}{n_j-d_j+1}\right) = \frac{n_j-d_j}{n_j}.$$

[7] La aproximación por el Método Delta asume que $f(X) \approx f(c) + f'(c)(X-c)$ para una función $f(X)$ de una variable aleatoria, con c próximo a $E(X)$.

Usando que la función de riesgo es aproximadamente independiente[8] en tiempos de fallo distintos,

$$\begin{aligned}\operatorname{Var}\left(\log \widehat{S}_{\mathrm{KM}}(t)\right) &= \operatorname{Var}\left(\sum_{\{i:T_i\le t\}} \log\left(1-\frac{D_i}{N_i}\right)\right) = \sum_{\{i:T_i\le t\}} \operatorname{Var}\left(\log\left(1-\widehat{h}_i\right)\right)\\ &\approx \sum_{\{i:T_i\le t\}} \left(\frac{1}{1-h_i}\right)^2 \operatorname{Var}\left(\widehat{h}_i\right) = \sum_{\{i:T_i\le t\}} \left(\frac{1}{1-h_i}\right)^2 \frac{h_i\,(1-h_i)}{N_i},\end{aligned}$$

que se estima a través de

$$\sum_{\{i:T_i\le t\}} \left(\frac{1}{1-D_i/N_i}\right)^2 \frac{D_i/N_i\,(1-D_i/N_i)}{N_i} = \sum_{\{i:T_i\le t\}} \frac{D_i}{N_i\,(N_i-D_i)}.$$

Por último, usando (5.4) y que el estimador de Kaplan-Meier es insesgado, se sigue el resultado. □

Corolario 5.2. *En ausencia de censura, la expresión* (5.3) *se reduce a*

$$\widehat{S}_{\mathrm{KM}}(t)\,\frac{1-\widehat{S}_{\mathrm{KM}}(t)}{n}.$$

Demostración. En efecto, como $N_i - D_i = N_{i+1}$ y $\widehat{S}_{\mathrm{KM}}(t) = \frac{N_{s+1}}{n}$ con $s = \max\{i : T_i \le t\}$, se tiene

$$\begin{aligned}\sum_{\{i:T_i\le t\}} \frac{D_i}{N_i(N_i-D_i)} &= \sum_{\{i:T_i\le t\}} \left(\frac{1}{N_i-D_i}-\frac{1}{N_i}\right) = \sum_{\{i:T_i\le t\}} \left(\frac{1}{N_{i+1}}-\frac{1}{N_i}\right)\\ &= \frac{1}{N_{s+1}}-\frac{1}{n} = \frac{1}{N_{s+1}}\left(1-\widehat{S}_{\mathrm{KM}}(t)\right) = \frac{1}{n\,\widehat{S}_{\mathrm{KM}}(t)}\left(1-\widehat{S}_{\mathrm{KM}}(t)\right).\end{aligned}$$

□

Bajo la hipótesis de censura aleatoria, el estimador $\widehat{S}_{\mathrm{KM}}(t)$ es asintóticamente normal para un t fijo. Un intervalo de confianza para $S(t)$ a un nivel del $100(1-\alpha)\,\%$ para un valor concreto de t es:

$$\widehat{S}_{\mathrm{KM}}(t) \pm z_{1-\alpha/2}\, EE\left(\widehat{S}_{\mathrm{KM}}(t)\right),$$

donde $z_{1-\alpha/2}$ es el cuantil correspondiente de la distribución $\mathcal{N}(0,1)$ y EE denota el error estándar, es decir, $\sqrt{\operatorname{Var}/n}$, que se calcula utilizando (5.3).

Es importante notar que este estimador presenta problemas cuando la hipótesis de independencia entre los tiempos de supervivencia T y los tiempos de censura C no se verifica (véase Ejemplo 5.2).

[8] Se tiene que $\hat{h}_i, \hat{h}_j$ son asintóticamente independientes para todo $i \neq j$ (demostración disponible en [32]).

Ejemplo 5.2. Consideremos

$$(\log T, \log C) \sim \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right), \quad \text{donde } \rho \in \{-0.9, 0, 0.9\}. \tag{5.5}$$

Las distribuciones marginales de (5.5) son $\mathcal{N}(0,1)$ y, por tanto, T y $C \sim$ Log-Normal$(0,1)$. En la Figura 5.3 se representa el estimador de Kaplan-Meier (en línea continua naranja) junto con los intervalos de confianza para $S(t)$ al nivel 95 % (en líneas discontinuas naranjas) y la función de supervivencia (exacta) de T (en azul).

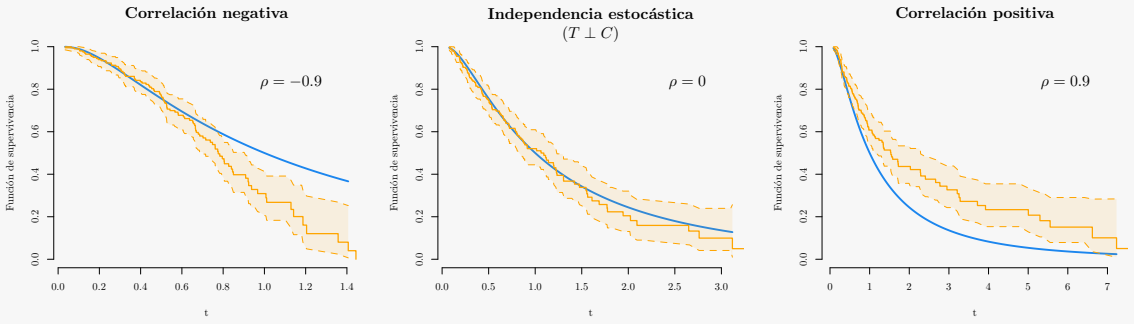


Figura 5.3: Estimador de Kaplan-Meier con distintas correlaciones

5.3. Método Nelson-Aalen

La relación entre las funciones de supervivencia y de riesgo en el caso continuo (1.6) no parece proporcionar una buena generalización del caso discreto (1.10). En la literatura especializada (véase [10]) ambas fórmulas pueden ser agrupadas e intuitivamente interpretables introduciendo el concepto de producto integral[9], que está definido como el límite de aproximaciones de productos finitos. Para una formalización adecuada se puede consultar [24].

Definición 5.1. Sea T una variable aleatoria denotando el tiempo de supervivencia. Se define el estimador de Nelson-Aalen para la función de riesgo acumulada como

$$\widehat{H}(t) = \sum_{\{i:T_{(i)} \le t\}} \frac{D_i}{N_i},$$

y su correspondiente estimación:

$$\sum_{\{i:t_{(i)} \le t\}} \frac{d_i}{n_i},$$

[9]Esta formalización es necesaria para no hacer suposiciones sobre la naturaleza de la variable aleatoria, pudiendo ser esta continua, discreta, o incluso una mixtura. Habitualmente se denota por $\prod$.

donde $t_{(i)}$ representa los tiempos observados ordenados, d_i el número de eventos ocurridos en el instante $t_{(i)}$, y n_i el número de individuos que estaban a riesgo antes de $t_{(i)}$.

A partir de la definición anterior, es posible obtener una estimación de la función de supervivencia a través de la expresión:

$$\widehat{S}_{\mathrm{NA}}(t) = \exp\left(-\widehat{H}(t)\right) = \exp\left(-\sum_{\{i:T_i\leq t\}} \frac{D_i}{N_i}\right).$$

Observación 5.3. Se puede probar que, en general, se cumple: $\widehat{S}_{\mathrm{KM}}(t) \leq \widehat{S}_{\mathrm{NA}}(t)$.

Razonando de la misma forma que en la demostración de la Fórmula de Greenwood, se puede probar (en el caso continuo):

$$\mathrm{Var}\left(\widehat{H}(t)\right) \approx \sum_{\{i:T_i\leq t\}} \frac{D_i\,(N_i - D_i)}{N_i^3}, \text{ y}$$

$$\mathrm{Var}\left(\widehat{S}_{\mathrm{NA}}(t)\right) = \mathrm{Var}\left(\exp(-\widehat{H}(t))\right) \approx \left(e^{-H(t)}\right)^2 \sum_{\{i:T_i\leq t\}} \frac{D_i\,(N_i - D_i)}{N_i^3} \approx \left(\widehat{S}(t)\right)^2 \sum_{\{i:T_i\leq t\}} \frac{D_i\,(N_i - D_i)}{N_i^3}.$$

Observación 5.4. El estimador de Kaplan-Meier también puede ser usado para estimar la función de riesgo acumulada, a saber:

$$\widehat{H}_{\mathrm{KM}}(t) = \sum_{\{i:T_{(i)}\leq t\}} -\log\left(1 - \frac{D_i}{N_i}\right).$$

Ambos estimadores de H están estrechamente ligados, ya que $\log(1-x) \approx -x$ para x pequeño.

5.4. Comparación de la supervivencia de dos o más grupos (Test Log-Rank)

Con frecuencia resulta interesante comparar distribuciones de supervivencia de dos o más grupos de pacientes (cohortes). Entre los distintos contrastes de hipótesis no paramétricos para comparar distribuciones de supervivencia, describiremos el test Log-Rank[10] pues es, en general[11], apropiado cuando se trabaja con datos censurados por la derecha. Cabe destacar el test de Breslow[12] y el

[10] También conocido como test de Mantel-Cox. Fue propuesto originalmente por Nathan Mantel en 1966.

[11] Es muy potente para calcular diferencias cuando los logaritmos de las funciones de supervivencias son proporcionales, pero presenta problemas para detectar las diferencias cuando las curvas de supervivencias se cruzan.

[12] Detecta diferencias cuando las curvas de supervivencia se cruzan, pero solamente al principio, por lo que no es recomendable para un estudio a largo plazo.

de Tarone-Ware, este último es un test intermedio a los otros dos. Las hipótesis planteadas serán:

$$H_0 : S_0(t) = S_1(t) \quad \text{frente a} \quad H_1 : S_0(t) \neq S_1(t).$$

No existe ningún test de la clase uniformemente más potente (UMP) para este tipo de constrastes.

El test Log-Rank es un método no paramétrico que compara las funciones de supervivencia de dos o más grupos de individuos y el marco de trabajo es el mismo que cuando calculamos el estimador de Kaplan-Meier. No ofrece ninguna información sobre la magnitud de las diferencias entre los grupos o un intervalo de confianza y para conocer este tipo de información se utiliza el cociente de riesgos que se explica en el Capítulo 6.

La idea en la que se basa este contraste es la misma que cuando se comparan dos distribuciones a través del test de independencia χ^2. En primer lugar, clasifiquemos los individuos según se haya observado el evento de interés o no (véase Tabla 5.1).

	Evento observado		
Grupo	Sí	No	Total
0	D_0	$N_0 - D_0$	N_0
1	D_1	$N_1 - D_1$	N_1
Total	D	$N - D$	N

Tabla 5.1: Tabla de contingencia para comparar dos grupos

Bajo la hipótesis de no asociación entre evento y grupo (establecida en la hipótesis nula H_0), se cumple[13]:

$$E(D_0) = \frac{N_0\, D}{N} \quad \text{y} \quad \text{Var}(D_0) = \frac{N_0\, N_1\, D\,(N - D)}{N^2\,(N-1)}.$$

Por tanto, bajo H_0, se cumple:

$$\chi^2_{\text{MC}} = \frac{(d_0 - n_0 d/n)^2}{\dfrac{n_0\, n_1\, d\,(n-d)}{n^2\,(n-1)}} \sim \chi^2_1. \tag{5.6}$$

La expresión (5.6) es el estadístico Log-Rank y es aproximadamente equivalente al estadístico del test χ^2 de Pearson para la independencia de dos grupos (véase Ejemplo 5.3):

$$\sum_i^2 \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_1,$$

donde O_i denotan las frecuencias observadas y E_i las frecuencias esperadas.

Observación 5.5. El estadístico χ^2 de Pearson se obtiene cuando los elementos de la Tabla 5.1

[13] Estas expresiones provienen de la distribución hipergeométrica.

están fijos (y por supuesto se observan todos los eventos), entonces la varianza se reemplaza por:

$$\mathrm{Var}(d_0) = \frac{n_0\, n_1\, d\,(n-d)}{n^3}.$$

Ejemplo 5.3. Considérese que se está llevando a cabo un estudio clínico para conocer el comportamiento de la toxicidad en la quimioterapia entre dos cohortes de pacientes. Los datos recogidos se muestran en la Tabla 5.2.

	Presenta toxicidad		
Grupo	Sí	No	Total
0	8 (5)	42 (45)	50
1	2 (5)	48 (45)	50
Total	10	90	100

Tabla 5.2: Datos recogidos del estudio. Las frecuencias esperadas se encuentran entre

El estadístico χ^2 de Pearson aplicado a la muestra de observaciones es, por tanto:

$$\sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{(8-5)^2}{5} + \cdots + \frac{(48-45)^2}{45} = 4.00.$$

Así, al ser $P\big(\chi^2_1 > 4\big) = 0.046$, se obtiene un p-valor menor que el nivel de significación habitual (5 %) y, por tanto, hay evidencias para rechazar la independencia entre el comportamiento de la toxicidad y la cohorte, en otras palabras, un grupo de pacientes es más susceptible a una mala evolución con el tratamiento quimioterápico.

Análogamente, se tiene un valor del estadístico Log-Rank $\chi^2_{\mathrm{MC}} = 3.96$ y un p-valor $= 0.047$.

Supongamos ahora que se observan los siguientes datos de supervivencia:

$$\textbf{Grupo 0}: (y_{01}, \delta_{01}), \ldots, (y_{0n_0}, \delta_{0n_0}) \quad \text{y} \quad \textbf{Grupo 1}: (y_{11}, \delta_{11}), \ldots, (y_{1n_1}, \delta_{1n_1})$$

Denotemos por d_i al número de eventos observados en el grupo $i \in \{0, 1\}$, esto es, $d_i = \sum_j^{n_i} \delta_{ij}$.

La formulación general del test Log-Rank para dos grupos consiste en considerar para cada tiempo de evento una tabla de contingencia (2×2), y comparar las frecuencias de los eventos entre los dos grupos, condicionado al número de individuos que estén a riesgo en ese momento y en cada grupo. Las tablas se combinan usando el contraste de Cochran-Mantel-Haenszel.

Denotemos por $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ a los tiempos (ordenados) de supervivencia observados. Se considera la tabla de contingencia en el j-ésimo tiempo de supervivencia (Tabla 5.3), donde d_{1j}

y d_{2j} denotan el número de eventos observados en el grupo 1 y 2, respectivamente, en el j-ésimo tiempo de supervivencia; r_{1j} y r_{2j} es el número de individuos a riesgo en ese instante.

	Evento observado		
Grupo	Sí	No	Total
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
2	d_{2j}	$r_{2j} - d_{2j}$	r_{2j}
Total	d_j	$r_j - d_j$	r_j

Tabla 5.3: Tabla de contingencia en el j-ésimo tiempo de supervivencia

Proposición 5.3. *La expresión general del estadístico Log-Rank es*[14]*:*

$$\chi^2_{logrank} = \frac{\left[\sum_{j=1}^{k}\left(d_{1j} - r_{1j}\,\frac{d_j}{r_j}\right)\right]^2}{\sum_{j=1}^{k}\frac{r_{2j}\,r_{1j}\,d_j\,(r_j - d_j)}{r_j^2\,(r_j-1)}} = \frac{\left[\sum_{j=1}^{k}\left(\frac{r_{1j}\,r_{2j}}{r_j}\,(\hat{h}_{2j} - \hat{h}_{1j})\right)\right]^2}{\sum_{j=1}^{k}\frac{r_{2j}\,r_{1j}\,d_j\,(r_j - d_j)}{r_j^2\,(r_j-1)}}. \tag{5.7}$$

Asumiendo que todas las tablas de contingencia son independientes, la distribución de este estadístico corresponde aproximadamente a la de una χ^2 con un grado de libertad.

Demostración. Cuando la hipótesis nula es cierta, es decir la función de supervivencia es igual en ambas poblaciones, la probabilidad condicionada de fallo en $t_{(j)}$ es igual para los dos grupos, por lo tanto, la distribución de probabilidad de (d_{1j}, d_{2j}) viene dada por:

$$\prod_{i=1}^{2}\left[\binom{n_{ij}}{d_{ij}}\,h_j^{d_{ij}}\,(1-h_i)^{n_j - d_{ij}}\right] = \prod_{i=1}^{2}\left[\binom{n_{ij}}{d_{ij}}\right]\,h_j^{d_{ij}}\,(1-h_i)^{n_j - d_{ij}}.$$

Bajo H_0 las funciones de supervivencia coinciden, por la que la función de riesgo es la misma en ambas poblaciones, y el fallo es, por tanto, independiente del grupo, lo que implica que los eventos esperados en el grupo 1 vendrán dados por la expresión:

$$e_{1j} = \frac{n_{1j}\,d_j}{n_j}.$$

Definimos los eventos observados menos los esperados como:

$$u_i := \sum_{j=1}^{k}(d_{ij} - e_{ij}).$$

[14]Notar que es indistinto el grupo que se elija para hacer la suma en el numerador, pues al estar elevado al cuadrado, el estadístico de contraste queda invariante.

Cuando k es suficientemente grande se sigue, por el Teorema del Límite Central, que

$$\frac{u_1}{\sqrt{v}} = \frac{\sum_{j=1}^{k}(d_{1j} - e_{1j})}{\sqrt{\sum_{j=1}^{k} v_j}} \sim \mathcal{N}(0,1),$$

donde $v_j = \mathrm{Var}(d_{1j})$, que usando la distribución hipergeométrica es igual a:

$$v_j = \frac{n_{1j}\, n_{2j}\, d_j\, (n_j - d_j)}{n_j^2\, (n_j - 1)}.$$

El estadístico Log-Rank se define, por tanto, como $u^2/v \sim \chi_1^2$. □

Observación 5.6. El numerador de la expresión (5.7) puede ser interpretado como $\sum_i (O_i - E_i)$, donde O_i es el número de eventos observados en el grupo i, y E_i es el número de eventos esperados, teniendo en cuenta el conjunto de individuos a riesgo. La frecuencia esperada será igual al número de eventos multiplicado por la proporción de individuos que están a riesgo en el grupo i.

Observación 5.7. Si no se producen empates de eventos, el estadístico Log-Rank adquiere la forma:

$$\chi^2_{logrank} = \frac{\left[\sum_{j=1}^{k} d_{1j} - r_{1j}/r_j\right]^2}{\sum_{j=1}^{k} r_2\, r_{1j}/r_j^2}.$$

Proposición 5.4. *El test de Log-Rank es el más potente cuando la odds ratio es constante sobre los intervalos. Es decir, es el más potente cuando se cumple la hipótesis de riesgos proporcionales.*

Observación 5.8. Tarone y Ware definieron una clase de test añadiendo pesos w_j. Los contrastes Log-Rank, Wilcoxon y Peto-Prentice Wilcoxon están incluidos como casos particulares (véase Tabla 5.4).

Contraste	Peso (w_j)
Log-Rank	1
Tarone-Ware	$\sqrt{r_j}$
Gehan-Breslow/Wilcoxon	r_j
Peto-Peto/Prentice	$n\widehat{S}(t_j)$
Fleming-Harrington	$\left[\widehat{S}(t_j)\right]^{\alpha}$

Tabla 5.4: Tipo de contraste según el peso w_j

Capítulo 6

Modelo de regresión de Cox

En este capítulo exploraremos la relación que hay entre la supervivencia y otras variables[1] y nos plantearemos cómo pueden estar influyendo en la supervivencia (qué impacto tienen). Para mayor grado de detalle, consúltese [4].

Cuando se utiliza este modelo para analizar la supervivencia de los individuos en un ensayo clínico, éste nos permite separar los efectos del tratamiento de los efectos de otras variables. El modelo puede utilizarse también si se conoce de antemano que hay otras variables, aparte del tratamiento, que están influyendo en la supervivencia y que no pueden ser controladas fácilmente en el ensayo clínico. Por tanto, utilizando este modelo se puede mejorar la estimación del efecto del tratamiento [30].

6.1. Formulación del modelo

El modelo de Cox expresa la función de riesgo $h(t)$ en función del tiempo t y de un conjunto de covariables[2], $\boldsymbol{X} = (X_1, \ldots, X_p)$, que definen al sujeto en estudio del siguiente modo:

$$h(t|\boldsymbol{X}) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j X_j\right),$$

donde $h_0(t)$ es la función de riesgo basal y corresponde al riesgo de un individuo que tiene como valor 0 en todas las variables explicativas, esto es,

$$h_0(t) = h(t|X_1 = 0, \ldots, X_p = 0).$$

[1] En los capítulos anteriores sólo estudiábamos la variable tiempo de supervivencia.
[2] También llamadas: variables explicativas, predictores, factores de riesgo o variables de confusión.

Una interpretación alternativa de la función de riesgo basal es aquella función 'básica' del modelo si este no incorporase factores de riesgo.

Como hipótesis de partida supondremos que los tiempos de supervivencia tienen distribuciones continuas, que están tomados de forma exacta y que no existe la posibilidad de empates. Para cada sujeto i, con $i = 1, \ldots, n$, conoceremos su tiempo de seguimiento y_i, su estado de fallo o censura δ_i, variable codificada con 1 si el dato no está censurado y con 0 si el dato sí lo está, y las covariables fijas $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$. Si incluimos el subíndice i para denotar a un sujeto determinado, el modelo se podría reescribir como:

$$h(t|\boldsymbol{X}_i) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j X_{ij}\right). \tag{6.1}$$

6.2. Hipótesis de riesgos proporcionales

El modelo de Cox permite analizar la relación entre los riesgos de muerte de dos individuos expuestos a factores de riesgo diferentes. Para ello, el modelo parte de una hipótesis fundamental, la de que los riesgos son proporcionales. Para comprender esta noción, definiremos previamente la denominada 'razón de riesgos' (*Hazard Ratio*, HR) entre dos sujetos con diferente vector de covariables $\boldsymbol{X} = (X_1, \ldots, X_p)$ y $\boldsymbol{X}^* = (X_1^*, \ldots, X_p^*)$ como:

$$HR = \frac{h(t|\boldsymbol{X}^*)}{h(t|\boldsymbol{X})}. \tag{6.2}$$

Al igual que se realiza con los *odds ratios*, típicamente se evalúa en el numerador el grupo de mayor riesgo (definido por $\boldsymbol{X}^*$, sin pérdida de generalidad), y en el denominador, el grupo de menor riesgo (definido por $\boldsymbol{X}$). En tal caso, el HR será mayor que 1, ya que $h(t|\boldsymbol{X}^*) > h(t|\boldsymbol{X})$ y cuantificará cuántas veces es mayor el riesgo de morir con perfil $\boldsymbol{X}^*$ que con $\boldsymbol{X}$.

Si sustituimos (6.2) en la expresión (6.1) obtenemos:

$$\begin{aligned}
\frac{h(t|\boldsymbol{X}^*)}{h(t|\boldsymbol{X})} &= \frac{h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}^*)}{h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})} \\
&= \frac{\exp\left(\sum_{j=1}^{p} \beta_j X_j^*\right)}{\exp\left(\sum_{j=1}^{p} \beta_j X_j\right)} \\
&= \exp\left((\boldsymbol{X}^* - \boldsymbol{X})'\boldsymbol{\beta}\right)
\end{aligned}$$

6.3. Función de verosimilitud parcial

Para la estimación de los coeficientes del modelo de Cox, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, se utiliza el método de máxima verosimilitud. Sin embargo, en lugar de utilizar la verosimilitud completa, se usa una verosimilitud parcial, que se concentra en los k tiempos de fallo observados. La función de verosimilitud parcial para el modelo de Cox se define como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left[\frac{h(t|\boldsymbol{X}_i)}{\sum_{l \in R(t_i)} h(t|\boldsymbol{X}_l)} \right]^{\delta_i},$$

donde $R(t_i)$ es el conjunto de sujetos que están en riesgo en el tiempo t_i, es decir, aquellos que aún no han experimentado el evento ni han sido censurados antes de t_i.

La función de verosimilitud parcial se maximiza con respecto a los coeficientes β_j para obtener las estimaciones $\hat{\beta}_j$. Esto se puede hacer utilizando métodos numéricos como el algoritmo de Newton-Raphson.

6.4. Evaluación de la hipótesis de riesgos proporcionales

Para evaluar la validez de la hipótesis de riesgos proporcionales en el modelo de Cox, se pueden utilizar varios métodos, tanto gráficos como analíticos.

6.4.1. Métodos gráficos

Uno de los métodos gráficos más comunes para evaluar la hipótesis de riesgos proporcionales es el gráfico de residuos de Schoenfeld. Este gráfico muestra los residuos de Schoenfeld en función del tiempo y ayuda a identificar cualquier tendencia que pueda indicar una violación de la hipótesis de riesgos proporcionales.

Otro tipo de residuos son los de Cox-Snell, útiles para evaluar el ajuste global del modelo planteado. Si el modelo dado por (6.1) es aplicable, entonces las estimaciones del tiempo de supervivencia del modelo planteado vienen dadas por un estimador de la función de supervivencia $\widehat{S}(t_{(i)})$ que debe ser muy similar al verdadero valor de $S(t_{(i)})$.

Es importante hacer notar que si el modelo apropiado se ajusta bien a los datos, entonces los residuos de Cox-Snell tendrán para cada variable predictora un valor exp(1), es decir, distribución exponencial con razón o tasa de riesgo igual a 1.

6.4.2. Gráfico log-log de la curva de supervivencia

Otro método gráfico es el gráfico log-log de la curva de supervivencia, que compara las curvas de supervivencia para diferentes niveles de las covariables. Si las curvas log-log son paralelas, esto sugiere que la hipótesis de riesgos proporcionales es razonable.

6.4.3. Contrastes de hipótesis

Existen test de hipótesis para contrastar:

$$H_0 : \text{los riesgos son proporcionales}$$
$$H_1 : \text{los riesgos no son proporcionales}$$

La idea del test es la siguiente: para un predictor en particular se cumplirá la hipótesis de proporcionalidad de los riesgos si los residuos de Schoenfeld de dicho predictor no están correlacionados con los tiempos de supervivencia. Gráficamente, si dibujamos los residuos de Schoenfeld del predictor, éstos serán horizontales si se cumple la hipótesis de proporcionalidad ya que en tal caso los residuos son independientes del tiempo [4].

6.5. Extensiones del modelo de Cox

En algunos casos, la hipótesis de riesgos proporcionales puede no ser válida para todas las covariables en el modelo. En tales situaciones, se puede utilizar un modelo de Cox estratificado. Este modelo permite que la función de riesgo basal varíe entre diferentes estratos definidos por una o más covariables, mientras se asume que los coeficientes de las otras covariables son constantes entre los estratos.

En el modelo de Cox estratificado, los sujetos en el estrato m-ésimo tienen una función de riesgo basal arbitraria, $h_{0m}(t)$. El efecto de otras covariables explicativas sobre la función de riesgo puede representarse por un modelo de riesgos proporcionales en ese estrato:

$$h_m(t|\boldsymbol{X}) = h_{0m}(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}), \quad m = 1, \ldots, s.$$

Por otro lado, el modelo de Cox también se puede extender para incluir covariables que varían en el tiempo. En este caso, el modelo se expresa como:

$$h(t|\boldsymbol{X}_i) = h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_{ij}(t)\right),$$

donde $X_{ij}(t)$ es el valor de la covariable j para el sujeto i en el tiempo t.

Capítulo 7

Aplicación a datos reales

En este capítulo, se aplica la metodología propuesta a un conjunto de datos real. El trabajo se ha realizado en colaboración con Rebeca Fernández González a través del servicio de Oncología médica del Hospital Universitario Central de Asturias.

El Cáncer de Pulmón de Células No Pequeñas (abreviadamente CPCNP) surge de las células epiteliales pulmonares desde los bronquios principales hasta los alvéolos terminales. El tipo histológico de CPCNP se correlaciona con el sitio de origen y refleja las variaciones epiteliales de las vías respiratorias desde los bronquios hasta los alvéolos. Por lo general, el carcinoma de células escamosas surge cerca de un bronquio principal. El adenocarcinoma y el carcinoma bronquioloalveolar, por lo común, surgen en el tejido pulmonar periférico.

A todos los pacientes analizados en el estudio se les ha aplicado la técnica NGS (secuenciación de última generación). Algunas de las razones por la que esta técnica es preferible frente a otras, como pueden ser RT-PCR (reacción en cadena de la polimerasa con transcripción inversa), IHQ (inmunohistoquímica) o FISH (hibridación in situ fluorescente), son la mayor variedad de mutaciones que proporciona o la menor cantidad de muestra que se necesita, entre otras.

El hecho de estudiar mutaciones tiene como objetivo aplicar un tratamiento personalizado (tratamiento 'diana'), es decir, tratar a los pacientes de forma individualizada en función de estos hallazgos moleculares.

No todas las mutaciones conducen a un tratamiento personalizado, pues para poder aplicarse se necesitan hacer muchos ensayos clínicos y pasar un largo proceso de inspección. Las mutaciones que permiten tratamiento diana son EGFR, ALK y ROS1, de acuerdo con la última guía clínica de la Sociedad Española de Oncología Médica para el CPCNP [23]. El resto de alteraciones o mutaciones para este tipo de cáncer se dicen no accionables.

7.1. Conjunto de datos

El conjunto de datos está formado por 93 individuos y 16 variables, descritas a continuación:

- Numéricas: edad (en años), TMB (carga mutacional tumoral; en nº mutaciones somáticas detectadas por Megabase de DNA tumoral), inicio (inicio del diagnóstico o del tratamiento; en días).
- Fechas: inicioDiana (fecha en la que se empezó el tratamiento personalizado), progresion (fecha de progresión), muerte (fecha de muerte), ultimo (última fecha conocida del individuo).
- Dicotómicas: sexo (masculino o femenino), karnofsky (índice de Karnofsky que permite cuantificar el estado funcional de los pacientes con cáncer; buen estado [Karnofsky $\geq 70\,\%$], mal estado [Karnofsky $< 70\,\%$]), histologia (tipos histológicos en los que se divide el CPCNP; escamoso, no escamoso [adenocarcinoma y carcinoma de células grandes]), diana (indicador de tratamiento personalizado; sí, no).
- Categóricas nominales no dicotómicas: fumador (hábito tabáquico; fumador, no fumador, ex fumador), tratamiento (tipo de tratamiento; quimioterapia, inmunoterapia, inmunoterapia añadida a la quimioterapia, tratamiento dirigido ITKs), lineaNGS (línea de tratamiento en la que se hizo NGS; línea 1, línea 2, línea 3); mutacion (tipo de mutación encontrada).

7.2. Objetivo del estudio

El objetivo del estudio es analizar de qué manera ha influido la implantación de la NGS frente a la utilización de otras técnicas diagnósticas en los resultados terapéuticos de los pacientes con CPCNP en estadios avanzados en el Hospital Universitario Central de Asturias (HUCA). En particular, el evento de interés en este estudio es el fallecimiento del paciente.

Se trata de un estudio epidemiológico observacional retrospectivo de cohortes de pacientes con diagnóstico de CPCNP estadio IV, diagnosticados y tratados en el HUCA, a los que se les hizo un estudio genético con NGS.

Se evaluaron 93 pacientes desde junio de 2019 hasta enero de 2024 con aplicación NGS. Las variantes moleculares (mutaciones) encontradas más frecuentemente fueron TP53 (40.9 %), KRAS G12C (25.8 %) y CDKN2A (21.5 %) (Figura B.3). Todas fueron más frecuentes en personas expuestas en algún momento al tabaco que en los nunca fumadores, a excepción de un caso. La mediana de supervivencia global (SG) para 68 casos evaluados y 25 censurados por no haber presentado el evento de interés fue de 16 meses. La ausencia de mutaciones registradas mediante NGS y el hecho de haber recibido quimioterapia en primera línea se asociaron a una mejor SG en el análisis univariante. En el multivariante, valores de TMB superiores a 10 mutaciones/mb se asociaron a mayor SG.

7.3. Análisis descriptivo de las variables

Empezaremos el estudio con un análisis descriptivo de las variables disponibles más relevantes a nivel clínico, con la ayuda del lenguaje de programación estadístico R (versión 4.2.2). Las variables numéricas se han descrito con la media, el mínimo y el máximo, mientras que de las variables categóricas se han aportado las frecuencias absolutas y las relativas en términos de porcentajes. Las tablas resumen se muestran al comienzo del Apéndice B. Cabe destacar que la primera tabla resume la muestra global y estratificada por sexo, mientras que la segunda estratifica a los pacientes según el hábito tabáquico.

Todos los datos de la tabla 1 recogen a los 93 pacientes de la muestra, de los cuales 50 (53.8 %) eran hombres y 43 mujeres (46.2 %). La mayoría tenían una edad comprendida entre los 60 y los 70 años, siendo la mediana de edad de 66 años. El 59 % eran exfumadores, el 31 % fumadores activos y solo el 9.7 % nunca fueron fumadores.

La gran mayoría (un 88 %) presentaban un buen estado general en el momento del diagnóstico (índice de Karnofsky mayor de 70 %) y solo el 12 % un mal estado general. En cuanto a la histología, hubo gran predominio del adenocarcinoma no escamoso (80 %), frente al 20 % de los escamosos. El tratamiento de primera línea que más pacientes recibieron fue quimioterapia junto con inmunoterapia (el 46 % de los pacientes), seguido de un 37 % que recibió inmunoterapia, un 14 % que recibió quimioterapia y un 3.4 % que recibió un tratamiento dirigido basado en ITK.

7.4. Análisis de supervivencia

Mediante el uso del lenguaje R, en particular de las funciones de las librerías survival y survminer, hemos representado gráficamente las funciones de supervivencia estimadas por Kaplan-Meier (5.2) y los intervalos de confianza para cada tiempo t con la Fórmula de Greenwood (5.3), a un nivel del 95 % (véase Apéndice B).

En primer lugar, se deben disponer los datos de forma adecuada para trabajar con la librería survival. Esto se hace creando un objeto en R de clase 'Surv', que establezca claramente qué observaciones se corresponden con eventos y cuáles con censuras. Esto se realiza mediante la función Surv() y necesita dos argumentos para muestras censuradas a la derecha: un vector que contiene todas las observaciones y_i, y un vector de estados que indica para cada observación si es un evento (1) o una censura (0).

En cuanto a resultados, de los 93 pacientes de la muestra a los que se les realizó la NGS, se evaluaron un total de 68 casos, mientras que 25 fueron censurados por no haber tenido lugar el evento en estudio o por no disponer de la información suficiente en su seguimiento. La mediana de SG obtenida fue de 16 meses (IC 95 % : $12 - 20$). La probabilidad de supervivencia a los 12 meses fue del 58 %, a los 24 meses del 26 % y finalmente a los 36 meses del 14 % (Figura B.4).

Dentro de los 93 pacientes de la muestra, 64 presentaban al menos una alteración molecular accionable y su mediana de SG tras analizar 44 casos y 20 censuras, fue de 12 meses (IC 95 % : 7.7 − 20). Por otro lado, de los 29 pacientes que no presentaban mutaciones o no se registraron en los informes pertinentes de la NGS, se analizaron 24 casos y 5 censuras, obteniéndose una mediana de SG de 20 meses (IC 95 % : 18 − 32) (Figura B.5). El p-valor resultante del test Log-Rank fue 0.0069, indicando diferencias estadísticamente significativas en la supervivencia entre los que presentan o no mutaciones.

Respecto a la línea en la que se realizó la NGS, también se aprecian diferencias estadísticamente significativas (p-valor = 0.0019). En aquellos pacientes a los que se les realizó en primera línea, su mediana de supervivencia fue de 12 meses, en segunda línea de 20 meses y en tercera línea de 28 meses (Figura B.6). Finalmente, en cuanto al TMB, el p-valor no es significativo, aunque más allá de los 5 meses de supervivencia, los que tuvieron un valor alto de TMB presentan una supervivencia estimada mayor, seguidos de los de TMB intermedio y, por último, los pacientes con TMB bajo, para los que se espera un peor pronóstico, con una mediana de supervivencia de 6.7 meses (Figura B.7).

Los pacientes con más de 65 años tuvieron una mediana de supervivencia de 14 meses (IC 95 % : 8.4 − 23), frente a los menores de 65, cuya mediana fue de 18 meses (IC 95 % : 12 − 23) (Figura B.8). Observamos que la supervivencia estimada para los mayores de 65 años no llega a tomar el valor cero en su último tiempo de observación; esto se debe a que el tiempo de supervivencia máximo corresponde con un dato censurado. En cuanto al sexo, la mediana de SG fue similar, siendo la de las mujeres de 16 meses (IC 95 % : 12 − 23) frente a 18 meses de los hombres (IC 95 % : 7.7 − 23) (Figura B.9). Aquellos pacientes que presentaban un buen estado general (índice de Karnofsky superior al 70 %) obtuvieron una mediana de SG de 16 meses, y aquellos con un mal estado general, de 9 meses, aunque cabe recordar que solo 11 pacientes se encontraban en mal estado (Figura B.10). En cuanto a la relación entre la supervivencia y el hábito tabáquico, los fumadores obtuvieron una mediana de SG de 9 meses, los exfumadores de 16 meses y los no fumadores de 20 meses (Figura B.11).

En relación con la histología, aquellos pacientes con carcinoma escamoso tuvieron una mediana de supervivencia de 20 meses, frente a los no escamosos, con 16 meses. Sin embargo, los intervalos de confianza para la estimación de supervivencia de los primeros son muy amplios, probablemente debido a que solo hay un 20 % de pacientes en esa situación y bastante variabilidad en sus tiempos de supervivencia (Figura B.12). Existen diferencias estadísticamente significativas (p-valor = 0.015) en la supervivencia según el tratamiento de primera línea recibido. Para aquellos pacientes que recibieron quimioterapia, la mediana de SG fue de 23 meses, seguida de 20 meses para los que recibieron inmunoterapia, 13 meses para quimioterapia junto con inmunoterapia y, finalmente, para los que recibieron ITK fue de 11 meses (Figura B.13).

Respecto al análisis multivariante, se propone en la Tabla 7.1 un posible modelo de Cox. Obtenemos 'ser fumador' como la única variable predictora que se asocia con la supervivencia a un

nivel de significación del 10 % (p-valor=0.088 $< \alpha = 0.10$). En este caso, $\exp(\beta_i) = 1.63$ si el individuo es fumador (con respecto a ser ex fumador, que es la categoría de referencia de esta variable), lo que significa que el riesgo aumenta un 63 %. En base al análisis gráfico de residuos de Schoenfeld (Figura B.14), se cumple la hipótesis de riesgos proporcionales para cada una de las variables (p-valores $> \alpha = 0.05$) y a nivel global (p-valor= 0.49 $> \alpha = 0.05$). Esto avala las conclusiones extraídas a partir del modelo a nivel inferencial.

term	estimate	std.error	statistic	p.value
sexo=="M"	0.1309801	0.2773961	0.4721772	0.6368003
fumador=="F"	0.4885193	0.2866997	1.7039411	0.0883921
fumador=="NF"	-0.0011123	0.4576090	-0.0024307	0.9980606
karnofsky=="ME"	0.3932336	0.3858355	1.0191742	0.3081203
histologia=="NE"	-0.1792888	0.3124473	-0.5738210	0.5660889
edad	0.0095998	0.0149931	0.6402828	0.5219888

Tabla 7.1: Análisis multivariante I

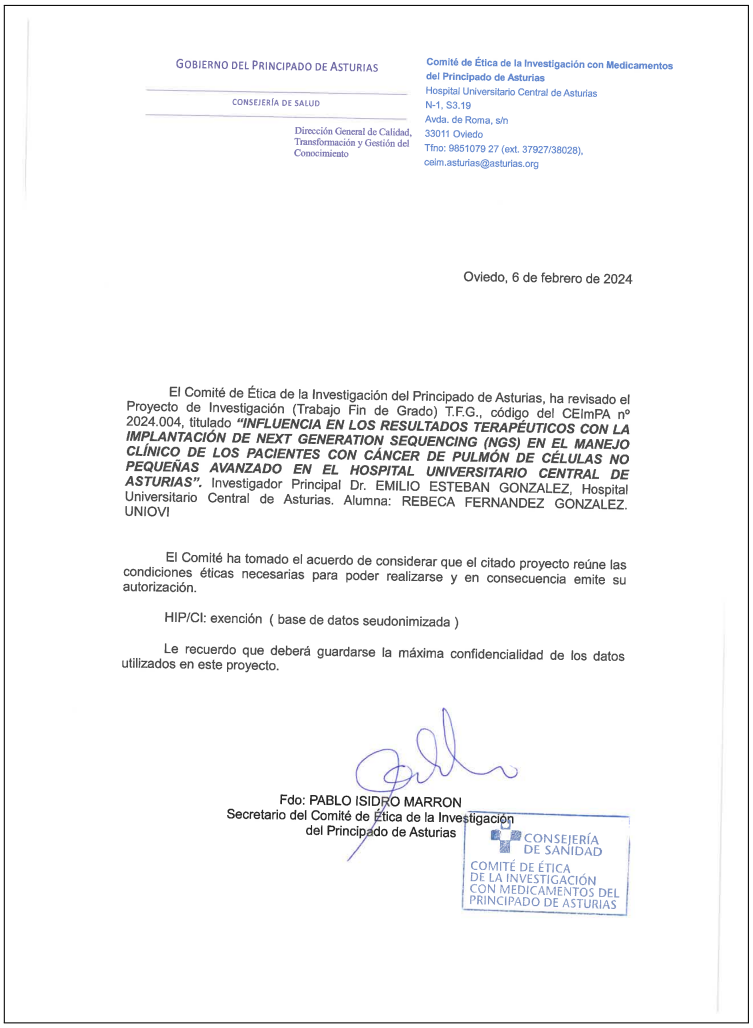
En cambio, si consideramos todas las variables como predictoras de la supervivencia (Tabla 7.2), la asociación de esta con ser fumador se 'diluye'. La única característica que presenta una asociación estadísticamente significativa con el riesgo de fallecimiento es el TMB (p-valor=0.043 $< \alpha = 0.05$). Esta asociación es negativa, pues al aumentar en una unidad el TMB se reduce el riesgo esperado un 8.29 %.

term	estimate	std.error	statistic	p.value
sexo=="M"	0.1761831	0.4631408	0.3804093	0.7036416
fumador=="F"	0.0070303	0.4699356	0.0149602	0.9880640
fumador=="NF"	-0.1228642	1.2643779	-0.0971736	0.9225885
karnofsky=="ME"	-0.2563881	1.0889357	-0.2354483	0.8138608
histologia=="NE"	-0.2382393	0.6270851	-0.3799155	0.7040082
edad	0.0119477	0.0249758	0.4783709	0.6323863
tratamiento==2	-0.9346545	1.6631095	-0.5619922	0.5741213
tratamiento==3	-0.2436459	1.6925044	-0.1439558	0.8855353
tratamiento==4	-0.8677134	1.7144051	-0.5061309	0.6127647
TMB	-0.0812014	0.0403102	-2.0144136	0.0439661

Tabla 7.2: Análisis multivariante II

Apéndice A

Autorización Comité de Ética



GOBIERNO DEL PRINCIPADO DE ASTURIAS

CONSEJERÍA DE SALUD

Dirección General de Calidad, Transformación y Gestión del Conocimiento

Comité de Ética de la Investigación con Medicamentos del Principado de Asturias
Hospital Universitario Central de Asturias
N-1, S3.19
Avda. de Roma, s/n
33011 Oviedo
Tfno: 9851079 27 (ext. 37927/38028),
ceim.asturias@asturias.org

Oviedo, 6 de febrero de 2024

El Comité de Ética de la Investigación del Principado de Asturias, ha revisado el Proyecto de Investigación (Trabajo Fin de Grado) T.F.G., código del CEImPA nº 2024.004, titulado ***"INFLUENCIA EN LOS RESULTADOS TERAPÉUTICOS CON LA IMPLANTACIÓN DE NEXT GENERATION SEQUENCING (NGS) EN EL MANEJO CLÍNICO DE LOS PACIENTES CON CÁNCER DE PULMÓN DE CÉLULAS NO PEQUEÑAS AVANZADO EN EL HOSPITAL UNIVERSITARIO CENTRAL DE ASTURIAS".*** Investigador Principal Dr. EMILIO ESTEBAN GONZALEZ, Hospital Universitario Central de Asturias. Alumna: REBECA FERNANDEZ GONZALEZ. UNIOVI

El Comité ha tomado el acuerdo de considerar que el citado proyecto reúne las condiciones éticas necesarias para poder realizarse y en consecuencia emite su autorización.

HIP/CI: exención (base de datos seudonimizada)

Le recuerdo que deberá guardarse la máxima confidencialidad de los datos utilizados en este proyecto.

Fdo: PABLO ISIDRO MARRON
Secretario del Comité de Ética de la Investigación
del Principado de Asturias

CONSEJERÍA DE SANIDAD
COMITÉ DE ÉTICA DE LA INVESTIGACIÓN CON MEDICAMENTOS DEL PRINCIPADO DE ASTURIAS

Figura A.1: Autorización del Comité de Ética de la Investigación con Medicamentos del Principado de Asturias

Bibliografía

[1] Aznar Gimeno, R. *Métodos estocásticos en medicina: Análisis de Supervivencia.* Trabajo Fin de Grado, Universidad de Zaragoza, 2015. Disponible en https://zaguan.unizar.es/record/31686.

[2] Balakrishnan, N. y Aggarwala, R. *Progressive Censoring.* Birkhäuser Boston, Boston, Massachusetts, 2000.

[3] Billingsley, P. *Probability and Measure*, capítulo 6, págs. 445–458. John Wiley & Sons, Nueva York, tercera edición, 1995.

[4] Boj del Val, E. El modelo de regresión de Cox. Informe técnico, Universidad de Barcelona, Departamento de Matemática Económica, Financiera y Actuarial de la Facultad de Economía y Empresa, Abril 2017. Disponible en http://hdl.handle.net/2445/49070.

[5] Bruce Rubin, D. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. URL https://doi.org/10.2307/2335739.

[6] Castro Cacabelos, M. *Imputación de datos faltantes en un modelo de tiempo de fallo acelerado.* Trabajo Fin de Máster, Universidad de Santiago de Compostela, Julio 2014. Disponible en http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_940.pdf.

[7] Fleming, T.R. y Harrington, D.P. *Counting Processes and Survival Analysis*, capítulo Apéndice A, págs. 320–322. John Wiley & Sons, 2005. Disponible en https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118150672.app1.

[8] Gijbels, I. Censored data. *WIREs Computational Statistics*, 2(2):178–188, 2010. URL https://doi.org/10.1002/wics.80.

[9] Gil Álvarez, M.Á., López Díaz, M., Montenegro Hermida, M.F., Sinova Fernández, B. y Terán Agraz, P.N. Probabilidades y Estadística. Informe técnico, Universidad de Oviedo, Departamento de Estadística e Investigación Operativa y D.M., 2010.

[10] Gill, R.D. y Johansen, S. A survey of product-integration with a view towards application in Survival Analysis. *The Annals of Statistics*, 18(4):1501–1555, 1990. URL https://doi.org/10.1214/aos/1176347865.

[11] Graunt, J. *Natural and Political Observations Mentioned in a following Index, and made upon the Bills of Mortality*. Impreso por Thomas Roycroft y por Thomas Dicas en St. Paul's Churchyard, Londres, segunda edición, 1662.

[12] Greenwood, M. Medical statistics from Graunt to Farr. *Biometrika*, 32(2):101–127, 1941. URL https://doi.org/10.2307/2332126.

[13] Gutiérrez Cabria, S. Origen y desarrollo de la estadística en los siglos XVII y XVIII. *Estadística Española*, Cuarto trimestre(97):19–32, 1982. Disponible en https://www.ine.es/art/ree/97_2.pdf.

[14] Herbert Aron, D. y Haikady Navada, N. *Order Statistics*, capítulo 2, págs. 9–32. Wiley Series in Probability and Statistics. John Wiley & Sons, Julio 2003.

[15] Hospital del Mar Research Institute, Barcelona. *Términos de Ensayos Clínicos*. Disponible en https://www.imim.cat/media/upload/arxius/terminologia.pdf.

[16] Kalbfleisch, J.D. y MacKay, R.J. On constant-sum models for censored survival data. *Biometrika*, 66(1):87–90, Abril 1979. URL https://doi.org/10.2307/2335246.

[17] Kalbfleisch, J.D. y Prentice, R.L. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, segunda edición, Agosto 2002.

[18] Kaplan, E.L. y Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. URL https://doi.org/10.2307/2281868.

[19] Keiding, N. Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 332(1627):487–509, 1990. URL https://doi.org/10.1098/rsta.1990.0128.

[20] Klein, J.P. y Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer, segunda edición, 2005. (3ª impresión corregida).

[21] Kulich, M. Censored data analysis. Informe técnico, Charles University, Department of Probability and Mathematical Statistics. Faculty of Mathematics and Physics, Octubre 2021. Disponible en https://www.karlin.mff.cuni.cz/~kulich/vyuka/cens/doc/cens_notes_ext_220102.pdf.

[22] Lagakos, S.W. General right censoring and its impact on the analysis of survival data. *Biometrics*, págs. 139–156, 1979. URL https://doi.org/10.2307/2529941.

[23] Majem, M., Juan, O., Insa, A., Reguart, N., Trigo, J. et al. *SEOM clinical guidelines for the treatment of non-small cell lung cancer*. Clinical Guides in Oncology, 2018.

[24] Matadamas Segura, M.A. *Inferencia para modelos de supervivencia de un solo evento y extensiones para modelos de riesgos competitivos.* Tesis, Universidad Autónoma Metropolitana, México, Abril 2010. Disponible en http://mat.izt.uam.mx/mat/documentos/produccion_academica/toda_la_produccion/Tesis%20dirigidas-14-66.pdf.

[25] Mattsson, A. *Censoring Bias in Oncology Clinical Trials.* Bachelor's theses in mathematical sciences, Lund University, Suecia, 2020. Disponible en https://www.lunduniversity.lu.se/lup/publication/9022651.

[26] McCullagh, P. y Dempsey, W. Survival models and health sequences. *Lifetime Data Analysis*, 24:550–584, Febrero 2018. URL https://doi.org/10.1007/s10985-018-9424-9.

[27] Muro Guerrero, J. *Análisis de Datos Censurados: Técnicas de Estimación e Inferencia No Paramétricas y Paramétricas.* Trabajo Fin de Grado, Universidad de Zaragoza, Septiembre 2019. Disponible en https://zaguan.unizar.es/record/87415.

[28] Oller, R. y Gómez Melis, G. Survival analysis with censored data: a further twist on ignorability conditions. *Statistics*, 57(6):1529–1550, Febrero 2024. URL https://doi.org/10.1080/02331888.2023.2283091.

[29] Pawitan, Y. *In all likelihood: statistical modelling and inference using likelihood*, capítulo 11, págs. 309–320. Oxford University Press, 2001.

[30] Pérez Fernández, S. *Estimación de la curva ROC acumulativa/dinámica.* Trabajo Fin de Máster, Universidad de Oviedo, Julio 2015. Disponible en https://digibuo.uniovi.es/dspace/handle/10651/32024.

[31] Ranganath, R., Perotte, A., Elhadad, N. y Blei, D. Deep survival analysis. En *Machine Learning for Healthcare Conference*, págs. 101–114. PMLR, 2016.

[32] Rodenkirchen, J. y Hoyer, A. An extension of Greenwood's formula to variances, Junio 2024. Apéndice 2 (página 10). Disponible en https://arxiv.org/abs/2406.07994.

[33] Ruiz Guzman, J. Historia de las estadísticas de salud. *Gaceta Médica Boliviana*, 29(2):72–77, 2006. Disponible en http://www.scielo.org.bo/pdf/gmb/v29n2/a15.pdf.

[34] Saran, J., Pushkarna, N. y Verma, K. Progressively Type-II right censored order statistics from odds generalized exponential-pareto distribution and related inference. *Applied Mathematics E-Notes*, 21(47):451–466, 2021. Disponible en https://www.emis.de/journals/AMEN/2021/AMEN-200610.pdf.

[35] Shih, W.J. Problems in dealing with missing data and informative censoring in clinical trials. *Current controlled trials in cardiovascular medicine*, 3(2):1–7, Enero 2002. URL https://doi.org/10.1186%2F1468-6708-3-4.

[36] Tapia Granados, J.A. Economía y mortalidad en las ciencias sociales: del renacimiento a las ideas sobre la transición demográfica. *Salud colectiva*, 1(3):285–308, 2005. Disponible en https://digitalrepository.unm.edu/lasm_cucs_es/140.

[37] van Boven. Michiel, van Dorp, C.H., Westerhof, I., Jaddoe, V., Heuvelman, V., Duijts, L. et al. Estimation of introduction and transmission rates of SARS-CoV-2 in a prospective household study. *PLOS Computational Biology*, 20(1):1–20, Enero 2024. URL https://doi.org/10.1371/journal.pcbi.1011832.

[38] Williams, J. y Lagakos, S.W. Models for censored survival analysis: Constant-sum and variable-sum model. *Biometrika*, 64(2):215–224, 1977. URL https://doi.org/10.2307/2335687.

[39] Zhou, M. y Jeong, J.H. Empirical likelihood ratio test for median and mean residual lifetime. *Statistics in medicine*, 30(2):152–159, Noviembre 2010. URL https://doi.org/10.1002/sim.4110.