



FACULTAD DE MATEMÁTICAS
ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

TRABAJO FIN DE GRADO

Comparación de curvas de supervivencia con datos censurados

ELENA DE LAS HERAS JIMÉNEZ
DOBLE GRADO EN MATEMÁTICAS Y ESTADÍSTICA

Dirigido por:
D. Juan Manuel Muñoz Pichardo

Sevilla, Julio de 2021

Índice general

Resumen	IV
Abstract	V
Índice de Figuras	VI
Índice de Tablas	VIII
1. Introducción	1
1.1. Breve referencia histórica	2
1.2. Estructura del trabajo	2
2. Conceptos básicos	4
2.1. Función de supervivencia	4
2.1.1. Estimador de Kaplan-Meier	5
2.1.2. Estimador actuarial	6
2.1.3. Estimador de Turnbull	9
2.1.4. Estimador de Fleming-Harrington	11
2.2. Tasas de fallo	12
2.2.1. Tasa de fallo instantánea o función de riesgo	12
2.2.2. Tasa de fallo acumulado	12
2.3. Tipo de datos censurados	13
2.3.1. Censura de tipo I	14
2.3.2. Censura de tipo II	14
2.3.3. Censura aleatoria	14
2.3.3.1. Censura a la derecha	14
2.3.3.2. Censura a la izquierda	15
2.3.3.3. Datos doblemente censurados	16
2.3.4. Datos censurados por intervalo	16
2.4. Teoría de proceso de conteo (<i>counting process theory</i>)	16
2.5. Test de permutación	17

2.6.	Teoría de Martingala	18
2.7.	Algoritmo EM	19
2.8.	Variables aleatorias intercambiables	19
2.9.	Método de Montecarlo	20
3.	Comparación de curvas de supervivencia	22
3.1.	Técnicas generales	22
3.1.1.	El test <i>log-rank</i> ponderado	22
3.1.2.	Prueba de tendencia (<i>trend-testing</i>)	24
3.2.	Técnicas específicas para datos censurados	26
3.2.1.	Datos censurados a la derecha	27
3.2.1.1.	Enfoque mediante el proceso de conteo	27
3.2.1.2.	Enfoque de permutación	28
3.2.2.	Datos censurados por intervalos	29
3.2.2.1.	Enfoque de permutación	29
3.2.2.2.	Enfoque basado en un vector de puntuación	31
4.	Comparación de curvas de supervivencia en R	33
4.1.	Librerías para el análisis de supervivencia	33
4.1.1.	Librería <i>survival</i>	33
4.1.1.1.	Estimador de Kaplan-Meier	34
4.1.1.2.	Estimador de Fleming-Harrington	38
4.1.1.3.	Estimador de la tasa de fallo acumulado	42
4.1.1.4.	Comparación de curvas de supervivencia	46
4.1.1.4.1.	Comparaciones múltiples	49
4.1.2.	Librería <i>survMisc</i>	51
4.1.2.1.	Comparación de curvas de supervivencia	51
4.1.3.	Librería <i>KMsurv</i>	58
4.1.3.1.	Estimador actuarial	58
4.1.4.	Librería <i>gte</i>	63
4.1.4.1.	Estimador de Turnbull	63
4.1.5.	Librería <i>coin</i>	66
4.1.5.1.	Comparación de curvas de supervivencia	66
4.2.	Librería <i>FHtest</i>	71
4.2.1.	Comparación de curvas de supervivencia	72

4.2.1.1.	Datos censurados a la derecha	72
4.2.1.1.1.	Enfoque mediante el proceso de conteo	72
4.2.1.1.2.	Enfoque de permutación	80
4.2.1.2.	Datos censurados por intervalos	83
4.2.1.2.1.	Enfoque de permutación	84
4.2.1.2.2.	Enfoque basado en un vector de puntuación	92
Conclusión		99
Bibliografía		105

Resumen

El análisis de supervivencia es una rama de la Estadística que tiene como objetivo analizar o modelizar el tiempo que tarda en ocurrir uno o más eventos de interés como la muerte de un paciente con cáncer o los fallos en los motores de unas máquinas de una industria textil. Para conseguir dicho objetivo, existe una extensa gama de herramientas entre las que se encuentra la comparación de curvas de supervivencia. El propósito de este trabajo es describir la teoría y la metodología de diversas técnicas de comparación de curvas de supervivencia con datos censurados. Además, con objeto de ilustrar la utilidad de estas técnicas se usarán librerías de R que contienen funciones que llevan a cabo dichas técnicas y se aplicarán sobre datos reales.

Palabras clave: análisis de supervivencia, datos censurados por intervalos, datos censurados a la derecha, función de supervivencia, prueba de tendencia, test *log-rank* ponderado, test de permutación, procesos de conteo.

Abstract

Survival analysis is a branch of Statistics that aims to analyse or model the time it takes for one or more events of interest to occur, such as the death of a cancer patient or the failure of machine motors in a textile industry. In order to achieve this goal, a wide range of tools are available, including the comparison of survival curves. The purpose of this dissertation is to describe the theory and methodology of tests for comparing survival curves with censored data. Furthermore, in order to illustrate the usefulness of these techniques, R libraries containing functions that perform these techniques will be used and applied on real data.

Keywords: survival analysis, interval-censored data, right-censored data, survival function, trend-testing, weighted log-rank test, permutation tests, counting processes.

Índice de figuras

2.1. Ejemplo de funciones de supervivencia	4
2.2. Estimador de Kaplan-Meier de la función de supervivencia	6
2.3. Estimador actuarial de la función de supervivencia	7
2.4. Estimador de Turnbull de la función de supervivencia	11
2.5. Estimador de Fleming-Harrington de la función de supervivencia	12
2.6. Ejemplo de funciones de riesgo	13
4.1. Estimador de Kaplan-Meier para los datos <i>AML</i>	38
4.2. Estimador de Fleming-Harrington para los datos <i>oropharynx</i>	42
4.3. Estimador de la tasa de fallo acumulado para los datos <i>rats</i>	45
4.4. Comparación de curvas de supervivencia con <i>ggsurvplot</i>	49
4.5. Prueba de tendencia con la función <i>comp</i>	53
4.6. Prueba de tendencia de manera gráfica	57
4.7. Estimador actuarial para los datos <i>Nurshome</i>	62
4.8. Estimador de Turnbull para los datos <i>simul</i>	66
4.9. Curvas de supervivencia de los pacientes de <i>lung</i>	69
4.10. Curvas de supervivencia de los pacientes de <i>lung</i> por género	71
4.11. Curvas de supervivencia de los pacientes de <i>ovarian</i>	76
4.12. Curvas de supervivencia de los pacientes de <i>bmt</i>	79
4.13. Curvas de supervivencia de los pacientes de <i>bcos</i>	88
4.14. Curvas de supervivencia de los pacientes de <i>illust3</i>	91
4.15. Curvas de supervivencia de los usuarios de <i>duser</i> por género	95
4.16. Curvas de supervivencia asociadas a cada período de <i>duser</i>	97

Índice de tablas

2.1. Datos usados para el estimador actuarial	7
2.2. Tabla de vida o estimador actuarial	8
3.1. Ejemplos de funciones pesos $\omega(t_i)$	24
4.1. Datos <i>AML</i>	35
4.2. Datos <i>AML</i> implementados en R	36
4.3. Estimador de Kaplan-Meier para el grupo <i>Maintained</i>	37
4.4. Estimador de Kaplan-Meier para el grupo <i>Nonmaintained</i>	37
4.5. Datos del fichero de <i>oropharynx</i>	39
4.6. Estimador de Fleming-Harrington para el grupo de pacientes con radioterapia sola (estándar)	41
4.7. Estimador de Fleming-Harrington para el grupo de pacientes con radioterapia junto a un agente quimioterapéutico (test)	41
4.8. Datos <i>rats</i> del paquete <i>survival</i>	43
4.9. Estimación de la función de riesgo acumulado para el grupo tratado con fármaco	44
4.10. Estimación de la función de riesgo acumulado para el grupo de control	44
4.11. Datos <i>colon</i> del paquete <i>survival</i>	47
4.12. Datos <i>larynx</i> del paquete <i>KMsurv</i>	53
4.13. Primera tabla de la salida <i>comp</i>	54
4.14. Segunda tabla de la salida <i>comp</i>	55
4.15. Datos <i>Nurshome</i>	59
4.16. Estimador actuarial del conjunto de datos <i>Nurshome</i>	61
4.17. Datos <i>simul</i> del paquete <i>gte</i>	64
4.18. Estimador de Turnbull del conjunto de datos <i>simul</i>	65
4.19. Datos <i>lung</i> del paquete <i>survival</i>	68
4.20. Datos <i>ovarian</i> del paquete <i>survival</i>	74

4.21. Datos <i>bmt</i> del paquete <i>KMsurv</i>	77
4.22. Datos <i>bcos</i> del paquete <i>interval</i>	86
4.23. Datos <i>illust3</i> del paquete <i>FHtest</i>	89
4.24. Datos <i>duser</i> del paquete <i>FHtest</i>	93

Capítulo 1

Introducción

El análisis de supervivencia tiene como objetivo principal analizar y modelizar el tiempo que transcurre hasta que se produce un determinado suceso en el que se está interesado. Este tipo de análisis está presente en muchas disciplinas tales como la medicina, la ingeniería o la sociología. Si se piensa en la medicina, esta técnica puede ser de interés para estudiar el tiempo que transcurre desde que a un paciente se le diagnostica un cáncer hasta que fallece o el tiempo que una persona tarda en recuperarse tras infectarse de la Covid-19. Por otro lado, en el ámbito de la ingeniería puede interesar el estudio del tiempo que tarda el motor de una máquina en dejar de funcionar. Por último, en estudios sociológicos puede tener interés el tiempo que tarda un trabajador en ascender en la empresa o el tiempo que tarda una persona en volver a delinquir tras salir de la cárcel.

En dicho análisis, la variable objetivo o de interés es el tiempo de vida y debe indicarse dónde se inicia el estudio de dicha variable para poder realizar un estudio riguroso en el que no necesariamente un individuo entra en el estudio en el punto de inicio considerado. De esta forma, la variable en estudio solo avanza en una dirección y nunca retrocede, por lo que se usarán técnicas no paramétricas o paramétricas sesgadas para modelar el tiempo que pasa hasta que el evento de interés ocurra. Cabe destacar también que esta técnica permite incorporar variables que pueden influir en la variable objetivo tiempo de vida, por lo que puede resultar interesante estudiar modelos de regresión para predecir el tiempo de fallo a través del resto de variables del estudio.

Otro de los aspectos interesantes que contempla el análisis de supervivencia es la existencia de lo que se conoce como datos censurados. Este tipo de datos incompletos son aquellos en los que se desconoce la variable de interés, ya sea porque el individuo ha abandonado el estudio o porque el estudio finaliza antes de que se produzca el suceso o evento de interés. Estos datos censurados deben ser tratados de manera específica para que no sean ignorados y aporten información al estudio que se esté realizando.

Por último, cabe destacar que en el análisis de supervivencia uno de los problemas estadísticos más importantes es la comparación de curvas de supervivencia. Dicho problema consiste en el estudio de la igualdad de las curvas de supervivencia de dos o más poblaciones distintas. En el caso de que dichas curvas no pudieran considerarse iguales, es interesante conocer a qué podría deberse dicha diferencia. Actualmente, por ejemplo, esta comparación puede ser de interés para observar el efecto que tiene una vacuna concreta destinada para el coronavirus en grupos de personas de distinta edad o para el estudio de su efectividad.

1.1. Breve referencia histórica

Los antecedentes más lejanos del análisis de supervivencia se pueden situar en la elaboración de tablas de mortalidad realizadas por el astrónomo Edmund Halley¹. Dichas tablas fueron elaboradas a partir del registro de funerales y nacimientos de la ciudad polaca de Breslavia. Otras fuentes sitúan que los orígenes de este análisis se encuentran en los estudios que realizó el estadístico John Graunt² en 1662 en su libro *Natural and Political Observations Made Upon the Bills of Mortality*.

Sin embargo, el concepto de análisis de supervivencia que se conoce hoy en día tiene su origen en la ingeniería y surgió con el principal objetivo de analizar la duración y fiabilidad de los diferentes elementos que forman una máquina. Además, la Segunda Guerra Mundial aceleró el desarrollo de ciertas técnicas del análisis de supervivencia para aplicarlas a la industria militar.

Cabe destacar también que durante décadas, el análisis de supervivencia estuvo ligado únicamente a la investigación de las tasas de mortalidad, pero desde los años 70 se ha extendido a la rama de la Ciencia de la Salud. El principal motivo de esta extensión es que permite generalizar el análisis de respuestas binarias y estudiar para cada sujeto una duración distinta del período de observación.

Actualmente, el análisis de supervivencia y la comparación de curvas de supervivencia son una de las herramientas esenciales en muchas ramas de la medicina como la epidemiología o la oncología y en la psicología. También es muy útil en la industria y en la ingeniería.

1.2. Estructura del trabajo

El presente trabajo se divide en cuatro capítulos con objeto de estructurar adecuadamente la información recogida. En este [primer capítulo](#) se ha realizado una pequeña introducción al tema del análisis de supervivencia y se han añadido brevemente algunas referencias históricas. Por otro lado, en el [Capítulo 2](#) se definirán conceptos fundamentales para comprender las técnicas de comparación de curvas de supervivencia que se detallarán en el [Capítulo 3](#). Tal y como se ha mencionado, en el [Capítulo 3](#) se describirán distintas técnicas de comparación de curvas de supervivencia. Se partirá de dos técnicas generales como son los test *log-rank* ponderados y las pruebas de tendencia; y seguidamente, se analizarán técnicas específicas para trabajar con datos censurados a la derecha y por intervalos. Cabe destacar, que en cada una de las técnicas en estudio se analizará de manera detallada la forma del estadístico involucrado en la comparación de las curvas de supervivencia. Por último, con el [Capítulo 4](#) se pretende ilustrar la utilidad de las técnicas definidas en el capítulo anterior con datos reales en R. Para ello, en primer lugar se irán enumerando y explicando algunos paquetes especializados en el análisis de supervivencia. A continuación, se elegirán aquellas funciones que son de interés para el estudio que se está realizando y se adjuntará una pequeña descripción de los argumentos involucrados en cada una de las funciones. Para finalizar e ilustrar el funcionamiento de

¹Astrónomo, matemático y físico inglés (1656 - 1742) conocido por el cálculo de la órbita del cometa que lleva su nombre y por elaborar el primer catálogo de estrellas del cielo austral.

²Estadístico inglés (1620 - 1674) que es considerado el primer demógrafo y epidemiólogo de la historia.

cada una de las funciones, estas se aplicarán a conjuntos de datos reales que en su mayoría se encuentran disponibles en distintos paquetes de R.

Capítulo 2

Conceptos básicos

En este capítulo se irán exponiendo los conceptos básicos necesarios para llevar a cabo un análisis de supervivencia y el posterior estudio de comparación de curvas de supervivencia. Téngase en cuenta que T es la variable aleatoria continua y no negativa que representa el tiempo de vida de un individuo en una población concreta. Se denota así que su función de densidad será $f(t)$ y su función de distribución será $F(t)$.

2.1. Función de supervivencia

La función de supervivencia $S(t)$ en t representa la probabilidad de que la variable aleatoria T supere el tiempo t :

$$S(t) = P[T > t] = 1 - F(t) = \int_t^{+\infty} f(x) dx$$

Además, se cumple que dicha función es monótona decreciente ya que:

$$S(0) = 1 \quad \lim_{t \rightarrow +\infty} S(t) = 0$$

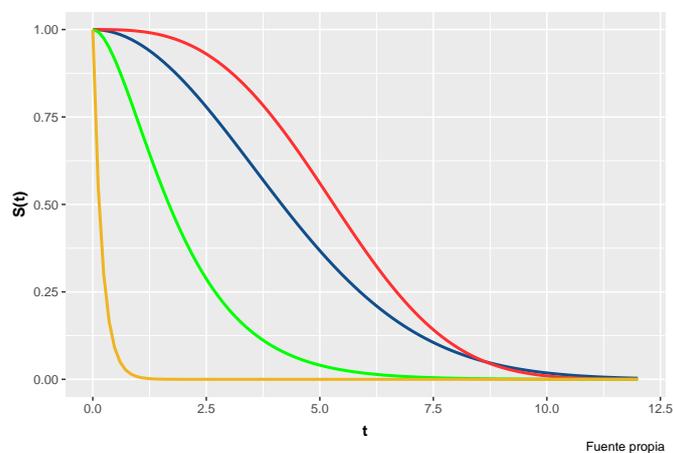


Figura 2.1: Ejemplo de funciones de supervivencia

La función anterior normalmente se estima mediante procedimientos no paramétricos entre los que destaca el método de Kaplan-Meier y el método actuarial que se suele aplicar cuando el tamaño de la muestra es grande, ya que agrupa los tiempos de supervivencia en intervalos. En ambos procedimientos, se parte de una muestra de n observaciones caracterizadas por la dupla (t_i, δ_i) donde t_i representa el tiempo de supervivencia o de censura del individuo i -ésimo y δ_i , una variable binaria que toma el valor 0 si el dato es censurado en el i -ésimo instante y 1 en caso de que el dato sea observado en dicho instante.

2.1.1. Estimador de Kaplan-Meier

Según Doménech (1992) este método fue desarrollado en 1958 por Kaplan y Meier (véase Kaplan y Meier (1958)) y su objetivo es el cálculo de la probabilidad de supervivencia que tiene un individuo de la muestra en el instante t_i , $i = 1, \dots, n$. 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} :

$$S(t_i) = S(t_{i-1}) \cdot S(t_i | t_{i-1})$$

La probabilidad condicionada de sobrevivir al instante t_i si se ha sobrevivido hasta el instante t_{i-1} puede ser estimada por el cociente $\frac{d_i}{r_i}$ ya que después de transcurrir el tiempo t_{i-1} quedan r_i individuos vivos y no censurados, es decir, individuos en riesgo; y d_i defunciones en el tiempo t_i . Por lo tanto, es lógico considerar que la probabilidad condicionada de sobrevivir al instante t_i si se ha sobrevivido hasta el instante t_{i-1} es:

$$\hat{S}(t_i | t_{i-1}) = 1 - \frac{d_i}{r_i} = \frac{r_i - d_i}{r_i}$$

Teniendo en cuenta las consideraciones anteriores, se obtiene que el estimador de Kaplan-Meier de la función de supervivencia será:

$$\hat{S}(t) = \prod_{i=1}^k \frac{r_i - d_i}{r_i}$$

donde k representa el k -ésimo instante de ocurrencia del evento o suceso en estudio.

El error estándar de la estimación de la función de supervivencia se puede calcular con la aproximación de Greenwood (1926):

$$SE(\hat{S}(t_i)) \approx \hat{S}(t_i) \sqrt{\sum_{j=1}^i \frac{d_j}{(r_j - d_j) r_j}}$$

Además un intervalo de confianza con nivel de confianza de $1 - \alpha$ para dicha estimación será:

$$\hat{S}(t_i) \pm z_{\alpha/2} SE(\hat{S}(t_i))$$

Este método es apropiado cuando los datos del estudio están censurados por la derecha y se trata de un estimador no paramétrico.

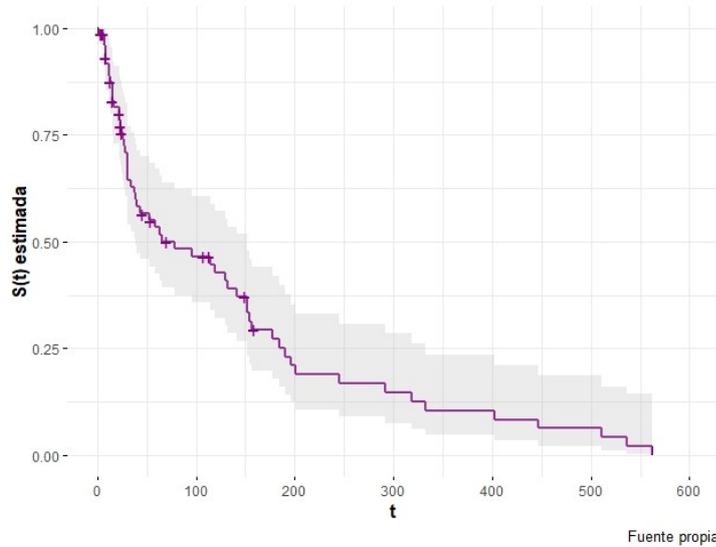


Figura 2.2: Estimación de la función de supervivencia mediante el método de Kaplan-Meier de los datos *kidney* (catéter renal) del paquete *survival* (Therneau (2021)) realizada con R

2.1.2. Estimador actuarial

Según los autores Doménech (1992) y Lee y Wang (2003), el estimador actuarial o método actuarial también conocido como el método de la tabla de vida fue desarrollado por Berkson y Gage (1950). Posteriormente, en los años 1958 y 1969 surgieron nuevas aportaciones realizadas por Cutler y Ederer (1958) y por Gehan (1969), respectivamente. Es destacable que este método es aproximadamente equivalente al estimador de Kaplan-Meier cuando se trabaja con muestras de gran tamaño. La principal diferencia entre ambos métodos es que en este caso, se requiere que los datos vengan agrupados en intervalos de tiempo, condición que conlleva a que las estimaciones serán poco precisas cuando el número de individuos en estudio no es suficientemente elevado.

Es aconsejable el uso de este estimador para muestras grandes y se caracteriza por la construcción de una tabla que resume muy bien la evolución del tiempo de supervivencia de los individuos estudiados. Para la construcción de dicha tabla se necesita crear intervalos de tiempo $[t_i, t_{i+1}]$ (normalmente de la misma amplitud) y para cada uno de ellos se debe considerar la siguiente información:

- N_i : individuos vivos al inicio del intervalo.
- d_i : individuos que mueren en el intervalo.
- $w_i + 1_i$: individuos censurados en el intervalo.
- $r_i = N_i - \frac{w_i + 1_i}{2}$: individuos expuestos al riesgo en el intervalo.
- $\frac{d_i}{r_i}$: probabilidad de morir en el intervalo $[t_i, t_{i+1}]$ condicionada a estar vivo al inicio del mismo.
- $\hat{S}(t_{i+1} | t_i) = 1 - \frac{d_i}{r_i} = \frac{r_i - d_i}{r_i}$: probabilidad condicionada de sobrevivir hasta el final del intervalo.

- $\hat{S}(t_{i+1}) = \hat{S}(t_i) \hat{S}(t_{i+1} | t_i) = \prod_{j=0}^i \frac{r_j - d_j}{r_j}$: probabilidad de estar vivo en el instante t_{i+1} , i.e., al inicio del siguiente intervalo.

Al igual que ocurría con el estimador de Kaplan-Meier, debe hallarse un intervalo de confianza a partir del error estándar que se calculará tal y como se ha expuesto en la [Sección 2.1.1](#) mediante la aproximación de Greenwood.

A continuación se adjuntará una representación de la función de supervivencia estimada mediante el método actuarial de los siguientes datos:

	nsubs	nlost	nevent
0-2	927	2	77
2-3	848	3	71
3-5	774	6	119
5-7	649	9	75
7-11	565	7	109
11-17	449	5	148
17-25	296	3	107
25-37	186	0	74
37-53	112	0	85
53-NA	27	0	27

Tabla 2.1: Datos extraídos del ejemplo que aparece en la ayuda de R de la orden `lifetab()` del paquete `KMsurv` (Klein y Moeschberger (2012))

La estimación gráfica siguiente se ha llevado a cabo a partir de la [Tabla 2.2](#) que muestra la información necesaria para la construcción de una tabla de vida.

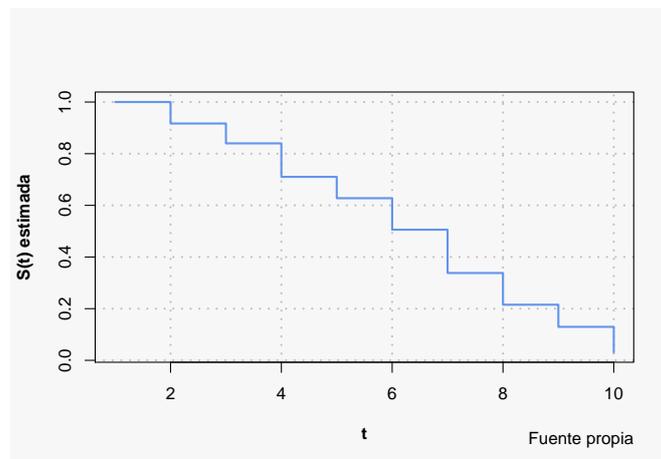


Figura 2.3: Estimación de la función de supervivencia mediante el método actuarial realizada con R

Tabla 2.2: Tabla de vida o estimador actuarial

	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-2	927	2	926.0	77	1.0000000	0.0415767	0.0433803	0.0000000	0.0045368	0.0049390
2-3	848	3	846.5	71	0.9168467	0.0769003	0.0875462	0.0090737	0.0087684	0.0103799
3-5	774	6	771.0	119	0.8399463	0.0648208	0.0836261	0.0120584	0.0055430	0.0076391
5-7	649	9	644.5	75	0.7103048	0.0413288	0.0617792	0.0149472	0.0045695	0.0071200
7-11	565	7	561.5	109	0.6276472	0.0304602	0.0537475	0.0159669	0.0027313	0.0051183
11-17	449	5	446.5	148	0.5058065	0.0279430	0.0662192	0.0165929	0.0020898	0.0053347
17-25	296	3	294.5	107	0.3381483	0.0153573	0.0554979	0.0158122	0.0013853	0.0052313
25-37	186	0	186.0	74	0.2152897	0.0071377	0.0413870	0.0138261	0.0007904	0.0046604
37-53	112	0	112.0	85	0.1296368	0.0061491	0.0764388	0.0113583	0.0006305	0.0065601
53-NA	27	0	27.0	27	0.0312517	NA	NA	0.0059119	NA	NA

Lo que representan cada una de las columnas de la [tabla anterior](#) es (véase [Klein y Moeschberger \(2012\)](#)):

<code>nsubs</code>	número de sujetos que entran en los intervalos y que no han experimentado el suceso en estudio
<code>nlost</code>	número de individuos censurados
<code>nrisk</code>	número de individuos en riesgo de sufrir el suceso en estudio
<code>nevent</code>	número de individuos que experimentaron el suceso
<code>surv</code>	función de supervivencia estimada al principio de los intervalos
<code>pdf</code>	función de densidad estimada en el punto medio de los intervalos
<code>hazard</code>	función de riesgo estimada en el punto medio de los intervalos
<code>se.surv</code>	desviación estándar de la función de supervivencia estimada al principio de los intervalos
<code>se.pdf</code>	desviación estándar de la función de densidad estimada en el punto medio de los intervalos
<code>se.hazard</code>	desviación estándar de la función de riesgo estimada en el punto medio de los intervalos

2.1.3. Estimador de Turnbull

Cuando se trabaja con [datos censurados por intervalos](#), el método de Kaplan-Meier no se puede aplicar. Ante esta situación, [Turnbull \(1974, 1976\)](#) (véase [Jaramillo y Lopera \(2017\)](#)) desarrolló un algoritmo iterativo que realiza una estimación no paramétrica de máxima verosimilitud de la función de supervivencia de los datos.

Sea una muestra de intervalos observados independientes $[L_i, R_i]$, $i = 1, \dots, n$ en los cuales se producen los tiempos de fallos T_1, \dots, T_n (variables aleatorias independientes). En estas condiciones, la función de verosimilitud que debe ser maximizada es:

$$L(S) = \prod_{i=1}^n [S(R_i^+) - S(L_i^-)] \quad (2.1)$$

donde $S(R_i^+)$ es el valor de la función de supervivencia justo después del extremo derecho R_i y $S(L_i^-)$ es el valor de la función de supervivencia justo antes del extremo izquierdo L_i de cada uno de los i intervalos observados.

Para la resolución del problema (2.1) es necesario definir un conjunto A que contenga los extremos izquierdos y otro conjunto B con los extremos derechos de la muestra inicial de intervalos observados:

$$\begin{aligned} A &= \{L_i, i = 1, \dots, n\} \\ B &= \{R_i, i = 1, \dots, n\} \end{aligned}$$

Una vez definidos los conjuntos anteriores, se pueden considerar nuevos intervalos de tiempo $[a_j, b_j]$, $j = 1, \dots, m$ con $m \leq n$, $a_j \in A$ y $b_j \in B$ de tal forma que una función que maximice (2.1) es constante entre los intervalos $[a_j, b_j]$ e indefinida dentro de ellos, lo

que implica que $\hat{P}[T \in (b_{j-1}, a_j)] = 0$ para cualquier j . Denótese los incrementos de S dentro de los intervalos $[a_j, b_j]$ por s_j , $j = 1, \dots, m$. Por tanto, (2.1) debe ser maximizada como función de s_1, \dots, s_m sujeto a que:

$$\begin{cases} s_j \geq 0 \\ s_m = 1 - \sum_{j=1}^{m-1} s_j \end{cases}$$

Para resolver el problema de optimización, Turnbull propone el uso del algoritmo de auto-consistencia. Sean $\alpha_{ij} = I_{\{[a_j, b_j] \in [L_i, R_i]\}}$, $i = 1, \dots, n$, $j = 1, \dots, m$ las variables indicadoras cuyo objetivo es informar de si el intervalo $[a_j, b_j]$ se encuentra dentro del intervalo $[L_i, R_i]$ o no. Entonces, la probabilidad de que T_i esté dentro del intervalo $[a_j, b_j]$ dado un vector $s = (s_1, s_2, \dots, s_m)'$ viene dado por:

$$\mu_{ij}(s) = \frac{\alpha_{ij} s_j}{\sum_{k=1}^m \alpha_{ik} s_k} \quad (2.2)$$

debido a que la estimación de S es constante fuera de los intervalos $[a_j, b_j]$. Además, la proporción de observaciones en $[a_j, b_j]$ es igual a:

$$\pi_j(s) = \frac{1}{n} \sum_{i=1}^n \mu_{ij}(s) \quad (2.3)$$

y se dice que un vector $s = (s_1, \dots, s_m)'$ es auto-consistente, si $s_j = \pi_j(s)$, $j = 1, \dots, m$.

Una vez definidas las herramientas necesarias, el algoritmo de auto-consistencia de Turnbull para calcular el estimador no paramétrico de $S(t)$ se puede implementar siguiendo los siguientes pasos:

1. Obtención de las estimaciones iniciales de s ; por ejemplo, $s_j^{(0)} = \frac{1}{m}$, $j = 1, \dots, m$.
2. Para $i = 1, \dots, n$, $j = 1, \dots, m$, hallar $\mu_{ij}(s^{(0)})$ conforme a (2.2), y luego $\pi_j(s^{(0)})$ según (2.3).
3. Obtención de estimaciones mejoradas para s calculando $s_j^{(1)} = \pi_j(s^{(0)})$.
4. Volver al paso 2, reemplazando $s^{(0)}$ por $s^{(1)}$ y continuar hasta lograr la convergencia.

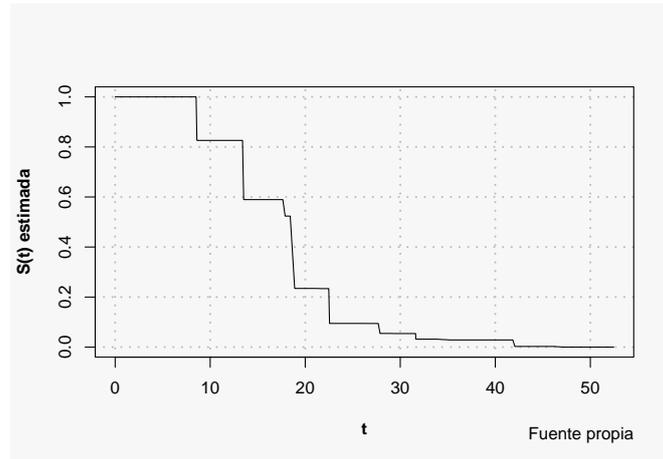


Figura 2.4: Estimación de la función de supervivencia mediante el estimador de Turnbull de los datos *simul* del paquete *gte* (Dehghan et al. (2015)) realizada con R

2.1.4. Estimador de Fleming-Harrington

El estimador de Fleming-Harrington (véase Fleming y Harrington (1984)) se trata de un estimador no paramétrico de la función de supervivencia y puede obtenerse a partir del estimador de la función de riesgo acumulado o tasa de fallo acumulado de Nelson y Aalen (véanse Ramírez-Montoya (2015), Nelson (1969) y Aalen y Johansen (1978)), el cual se puede calcular de la siguiente manera:

$$\hat{H}(t) = \sum_{i=1}^n \frac{d_i}{r_i}$$

representando d_i al número de muertes (o de ocurrencia del evento de interés) en el momento t_i y r_i al número de individuos en riesgo en el instante t_i .

El estimador también se puede calcular bajo el enfoque de procesos de conteo (véase la Sección 2.4) y toma la siguiente forma:

$$\hat{H}(t) = \sum_{i=1}^n \int_0^t \frac{dN_i(s)}{r(s)} ds$$

Este estimador suele ser útil cuando los tamaños de la muestra en estudio son pequeños.

Se deduce así que el estimador de Fleming-Harrington es:

$$\hat{S}(t) = e^{-\hat{H}(t)}$$

Además, una forma de calcular la varianza de la estimación de la función de supervivencia $\hat{S}(t)$ sería:

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{i=1}^n \frac{d_i}{r_i^2}$$

Asimismo, se podría calcular el error cuadrático medio del estimador en estudio:

$$ECM(\hat{S}(t_i)) = E[\hat{S}(t_i) - S(t_i)]^2 = [\hat{S}(t_i) - S(t_i)]^2 + Var[\hat{S}(t_i)]$$

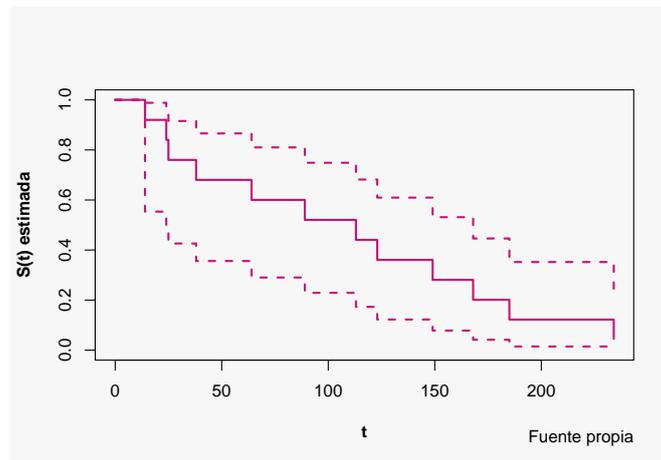


Figura 2.5: Estimación de la función de supervivencia mediante el estimador de Fleming-Harrington de los datos *NursingHome* (Pešta (2021)) realizada con R

2.2. Tasas de fallo

En análisis de supervivencia se conoce como tasa de fallo o función de riesgo a aquella función que tiene la capacidad de medir la probabilidad de que a un individuo le ocurra el suceso o evento en estudio a lo largo de un tiempo predeterminado.

2.2.1. Tasa de fallo instantánea o función de riesgo

La función de riesgo $h(t)$ (también conocida como tasa de fallo instantánea o *hazard function*) se define como la tasa de riesgo infinitesimal en $T = t$ condicionada a que el individuo ha sobrevivido al instante t :

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P[t < T \leq t + \Delta t | T > t]}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t [1 - F(t)]} = \frac{f(t)}{S(t)}$$

Equivalentemente, la expresión de dicha función es:

$$h(t) = -\frac{1}{S(t)} \frac{d}{dt} S(t) = -\frac{d}{dt} \ln [S(t)]$$

Se trata de una medida del potencial de fallo en el instante t , depende de la distribución de la variable T y puede tomar cualquier valor no negativo.

2.2.2. Tasa de fallo acumulado

La tasa de fallo acumulado $H(t)$ se define como la acumulación de la tasa de fallo instantánea hasta el instante t . Por tanto, para su cálculo basta considerar la integral de

la función de riesgo tal y como se indica a continuación:

$$H(t) = \int_0^t h(x) dx$$

Además, la tasa de fallo acumulado se relaciona directamente con la función de supervivencia mediante las siguientes expresiones:

$$H(t) = -\ln[S(t)] \Rightarrow S(t) = e^{-H(t)}$$

Para obtener un estimador de la tasa de fallo acumulado se puede trabajar de dos maneras distintas:

1. Estimación usando el método de máxima verosimilitud: este método consiste en tomar el menos logaritmo de la función de supervivencia.

$$\hat{H}(t) = -\log(\hat{S}(t))$$

2. Estimación de Nelson-Aalen: consultar la [Sección 2.1.4](#).

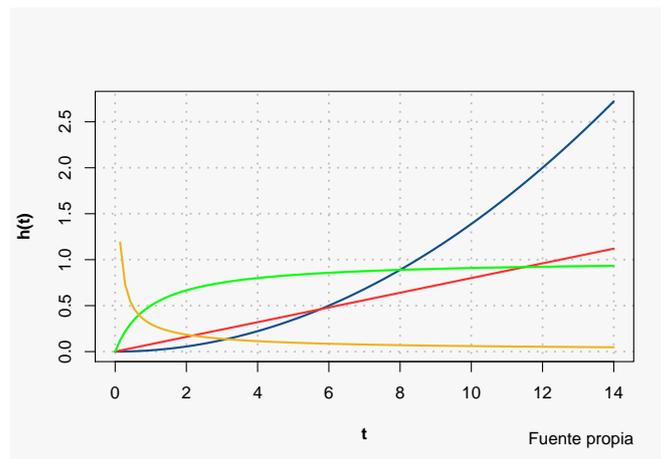


Figura 2.6: Ejemplo de funciones de riesgo

2.3. Tipo de datos censurados

Cuando se realiza un análisis de supervivencia a un conjunto de individuos, puede ocurrir que el suceso de interés no se observe durante el tiempo de estudio o de supervivencia. Esto podría suceder por varios motivos entre los que se encuentra que el suceso se haya producido en un instante anterior al momento en el que comienza el estudio, que el suceso ocurra después de que termine el estudio o simplemente porque un individuo decide o se ve obligado a abandonar el estudio. En todos esos casos, la información que aporta el individuo al estudio es parcial y dichos datos se conocen como datos censurados (*censored data*). En la práctica existen varios modelos de censura y a continuación, se presentan algunos de ellos (véanse [Gijbels \(2010\)](#), [Fahrmeir y Tutz \(2013\)](#) y [Cox y Oakes \(1984\)](#)).

2.3.1. Censura de tipo I

Este tipo de censura se produce cuando el experimento comienza en un momento determinado $t = 0$ para un conjunto de individuos concreto y se registra el tiempo de fallo de cada uno de ellos. El estudio termina en un tiempo prefijado $t = t_c$. Si T fuera el tiempo de fallo y n el número de individuos en estudio entonces las realizaciones de la variable observada serían:

$$Y_i = \begin{cases} T_i & \text{si } T_i \leq t_c \\ t_c & \text{si } T_i > t_c \end{cases}$$

Normalmente, se considera una variable cuya función es indicar si el tiempo es observado o censurado:

$$\delta_i = \begin{cases} 1 & \text{si } T_i \leq t_c \quad (\text{observado}) \\ 0 & \text{si } T_i > t_c \quad (\text{censurado}) \end{cases}$$

En este caso, el tiempo de censura t_c es fijo y el número de tiempos de fallo observados es una variable aleatoria que toma valores en el conjunto $\{0, 1, \dots, n\}$.

2.3.2. Censura de tipo II

La censura de tipo II se produce cuando el experimento se interrumpe cuando ha fallado una fracción prefijada r/n de individuos en estudio, siendo r un número entero tal que $1 \leq r \leq n$. Esta situación se suele dar por ejemplo en pruebas de fiabilidad cuando, para un determinado lote de productos o piezas, el experimento se detiene en cuanto ha fallado una fracción determinada de las n unidades. Si los tiempos de fallos son T_1, \dots, T_n y se considera la muestra ordenada $T_{(1)} \leq \dots \leq T_{(n)}$, entonces las n realizaciones de la variable observada serán:

$$Y_i = \begin{cases} T_{(i)} & \text{si } i \leq r \\ T_{(r)} & \text{si } i > r \end{cases}$$

Existen así $n - r$ observaciones de las que solo se conoce que su tiempo de fallo es mayor que $T_{(r)}$ y que se consideran observaciones censuradas. De esta forma, el número de tiempos observados r es fijo, mientras que el punto final T_r es aleatorio.

2.3.3. Censura aleatoria

2.3.3.1. Censura a la derecha

Este tipo de censura suele darse en estudios de medicina y se caracteriza por la existencia de individuos que no experimentan el suceso de interés antes de que acabe el tiempo de seguimiento prefijado. Considérese un estudio clínico con n pacientes ($i = 1, \dots, n$), cuyo objetivo es el estudio de la eficacia de un nuevo tratamiento médico y supóngase que el seguimiento de los pacientes será durante un período de tiempo fijado de antemano. Cada paciente recibió el tratamiento y sea la cura del paciente el evento de interés del estudio. Denótese:

E_i : momento en el que el i -ésimo paciente entra en el estudio

D_i : momento en el que el i -ésimo paciente se cura (evento de interés)

Siguiendo dicha notación, se tendrá entonces que el tiempo de supervivencia del i -ésimo individuo será:

$$T_i = D_i - E_i$$

Sin embargo, existen situaciones en las que los tiempos de supervivencia de todos los individuos no pueden ser observados y la información que se puede recabar de dichos pacientes es parcial. Esta censura se puede deber a alguna de las siguientes causas:

- Fin del estudio: termina el ensayo y no se ha observado el evento de interés.
- Abandono del individuo: el individuo deja de participar en el estudio.
- Salida forzosa: los efectos del tratamiento obligan al paciente a abandonar el estudio.

En los individuos en los que se produce alguna de las situaciones anteriores se observa que $C_i = T_i - E_i$, siendo $T_i > C_i$ y C_i el tiempo de censura. Se dice así que el tiempo de supervivencia T_i está censurado por el tiempo de censura C_i .

Una modelización estadística del problema contará con la participación de dos variables aleatorias: $T \equiv$ tiempo de supervivencia con función de distribución F y función de supervivencia S_f y $C \equiv$ tiempo de censura con función de distribución G y función de supervivencia S_g . De esta forma, en cada individuo se observa la dupla (Y_i, δ_i) siendo

$$Y_i = \min\{T_i, C_i\}; \quad \delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i \quad (\text{observado}) \\ 0 & \text{si } T_i > C_i \quad (\text{censurado}) \end{cases}$$

Normalmente se suele suponer independencia entre las variables T_i y C_i . No obstante, esta suposición es violada cuando el individuo debe salir forzosamente del estudio.

2.3.3.2. Censura a la izquierda

La censura a la izquierda se produce cuando solo se conoce que el tiempo de supervivencia de algún individuo es anterior a un momento determinado. Por ejemplo, este tipo de censura puede darse cuando el tiempo final del estudio ocurre antes del momento que se considera como origen del tiempo de estudio. Considérese un estudio pediátrico en el que se desea estudiar la edad en la que los bebés son capaces de dar sus primeros pasos. En este caso, es claro que el suceso de interés es empezar a caminar de manera independiente y la muestra de los bebés puede ser tomada de una guardería cualquiera. Un dato será censurado a la izquierda si al inicio del estudio el niño ya es capaz de caminar solo. Estas observaciones se caracterizan por la siguiente propiedad $T_i \leq C_i$ donde C_i será la edad del bebé al inicio del estudio.

En un problema con censura a la izquierda interviene un tiempo de censura aleatorio C_i asociado a cada tiempo de supervivencia T_i y cada una de las observaciones viene determinada por el par (Y_i, δ_i) donde:

$$Y_i = \max\{T_i, C_i\}; \quad \delta_i = \begin{cases} 1 & \text{si } T_i \geq C_i \quad (\text{observado}) \\ 0 & \text{si } T_i < C_i \quad (\text{censurado}) \end{cases}$$

Al igual que ocurría con la censura aleatoria a la derecha, una suposición que se suele hacer en este tipo de censura es considerar que las variables T_i y C_i son independientes.

2.3.3.3. Datos doblemente censurados

Los datos doblemente censurados se producen cuando existen a la vez tiempos de censura a la derecha y a la izquierda asociados a cada uno de los tiempos de supervivencia T_i . Cabe señalar que frecuentemente en estudios donde existe censura por la izquierda, también existe censura por la derecha. En estos casos, las observaciones vienen determinadas por la dupla (Y_i, δ_i) donde:

$$Y_i = \min\{\max(T_i, C_{l,i}), C_{r,i}\}; \quad \delta_i = \begin{cases} 1 & \text{si } C_{l,i} \leq T_i \leq C_{r,i} \quad (\text{observado}) \\ 0 & \text{si } C_{r,i} < T_i \quad (\text{censurado por la derecha}) \\ -1 & \text{si } T_i < C_{l,i} \quad (\text{censurado por la izquierda}) \end{cases}$$

En otras palabras, el tiempo de supervivencia real solo se observa cuando supera el tiempo de censura izquierdo $C_{l,i}$ (tiempo antes del cual algunos individuos presentan el evento de interés) y a la misma vez no supera el tiempo de censura derecho $C_{r,i}$ (tiempo después del cual algunos individuos presentan el evento de interés). Obsérvese que fijando $C_l = -\infty$ y $C_r = +\infty$, el caso de la censura doblemente aleatoria incluye como casos especiales la censura aleatoria a la derecha y la censura aleatoria a la izquierda, respectivamente.

2.3.4. Datos censurados por intervalo

La censura por intervalo o censura intermitente se produce cuando el individuo que está participando en el ensayo experimenta el punto final del estudio dentro de un intervalo determinado. Este tipo de censura suele darse cuando los individuos en estudio no son monitorizados de manera continua sino que solo se observan en momentos discretos de tiempo. Además, la censura a la izquierda es un caso especial de este tipo de censura, debido a que el tiempo final se producirá en el intervalo posterior a la entrada del individuo en el estudio y antes de que la investigación llegue a su fin.

Considérese el individuo i -ésimo del estudio que no ha presentado el evento de interés antes del tiempo L_i a pesar de que en el siguiente tiempo de observación R_i el individuo ya presentó el evento en estudio. En ese caso, se dice que T_i es censurado en el intervalo $(L_i, R_i]$, i.e., $L_i < T_i \leq R_i$.

2.4. Teoría de proceso de conteo (*counting process theory*)

Tal y como se enuncia en [Andersen et al. \(1985\)](#), un proceso de conteo multivariante $N = \{N_1(t), N_2(t), \dots, N_k(t)\}$, $t \geq 0$ es un proceso estocástico con k componentes que puede considerarse como un recuento de sucesos a lo largo del tiempo t siendo k los sucesos posibles que pueden ocurrir. De esta forma, $N_h(t)$ representará el número de sucesos de tipo h que se producen en el intervalo de tiempo $[0, t]$. Se supondrá que cada componente del proceso $N_h(t)$ tiene saltos de tamaño 1. Además, no se permite que se produzcan

eventos múltiples, i.e., no hay dos procesos que posean saltos en el mismo instante de tiempo. Nótese también que los sucesos suelen corresponderse con transiciones, para un individuo o grupo de individuos, entre los distintos estados del proceso estocástico en estudio.

La evolución del proceso N anterior se rige por un proceso de intensidad $\lambda = \{\lambda_1(t), \dots, \lambda_k(t)\}$, $t \in [0, 1]$ dado. Sea I_{dt} un intervalo de tiempo pequeño de longitud dt , entonces $\lambda_h(t) dt$ es la probabilidad de que N_h salte en I_{dt} condicionada a lo que ha ocurrido justo antes del instante t . Por otro lado, si se denota por $dN_h(t)$ el incremento de N_h sobre I_{dt} y sea \mathcal{F}_{t-} todo lo que ha sucedido hasta el instante t sin incluirlo, entonces:

$$\lambda_h(t) dt = P[dN_h(t) = 1 | \mathcal{F}_{t-}]$$

Téngase en cuenta que \mathcal{F}_{t-} incluye la trayectoria completa de N en el intervalo $[0, t)$, así como todos los demás acontecimientos incluidos ímplicita o explícitamente antes del instante t . Como consecuencia de ello, $\mathcal{F}_s \subseteq \mathcal{F}_t$ siempre que $s \leq t$, lo que refleja el hecho de que a medida que pasa el tiempo, se aprende cada vez más sobre el proceso.

Cabe destacar que el enfoque basado en procesos de conteo ha aumentado las posibilidades de los modelos de supervivencia.

2.5. Test de permutación

Siguiendo a [Clipson et al. \(2003\)](#), un test de permutación, test de aleatorización o test exacto es un tipo de test estadístico en el que la distribución del estadístico en estudio bajo la hipótesis nula se obtiene calculando todos los posibles estadísticos tras realizar permutaciones de los datos observados. El procedimiento no depende de suposiciones realizadas sobre la distribución de probabilidad subyacente. Dicha característica permite que sea una técnica que se pueda utilizar en un amplio abanico de estudios. En concreto, si la hipótesis de interés no tiene asociado un estadístico simple, un test de permutación puede ser una buena opción.

Si se considerase el conjunto completo de valores de la realización muestral y se formarían todas las posibles configuraciones de permutaciones de dicha realización, se obtendría un estimador del estadístico asociado al test para cada una de las muestras permutadas construidas. Sin necesidad de hacer ninguna suposición sobre la distribución de la variable aleatoria correspondiente al estadístico del test, se puede considerar el conjunto de estimadores calculados como una realización de una muestra aleatoria de esa distribución bajo la hipótesis nula. La significación empírica del valor correspondiente a la configuración observada podría entonces calcularse simplemente como el rango del valor observado en el conjunto de todos los valores.

Los test de permutación pueden utilizarse para analizar diseños desequilibrados y para combinar pruebas dependientes en mezclas de datos categóricos, ordinales y métricos. También pueden utilizarse para analizar datos cualitativos que han sido convertidos en números. Además, pueden ser ideales para analizar datos cuantificados que no satisfacen los supuestos estadísticos subyacentes a las pruebas paramétricas tradicionales (por ejemplo, pruebas t-student o ANOVA).

2.6. Teoría de Martingala

Considerando a [Lai y Xing \(2008\)](#), se puede definir la teoría de Martingala tal y como se indica a continuación. Sea un proceso estocástico $X = (S_t, \mathcal{F}_t)$, $t \in T \subseteq [0, \infty)$ definido en un espacio de probabilidad (Ω, \mathcal{F}, P) con una familia no decreciente de σ -álgebras $(\mathcal{F}_t)_{t \in T}$, $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$, $s \leq t$ donde \mathcal{F}_t denota al conjunto que contiene toda la información de los eventos que han ocurrido hasta el instante t . En tal caso, si se cumple que $E|S_t| < +\infty$, S_t es \mathcal{F}_{t-} medible y $E(S_t | \mathcal{F}_{t-1}) = S_{t-1}$ con probabilidad 1, entonces se dice que la secuencia de las variables S_t es martingala.

De forma más general, se denomina matriz de diferencia martingala al conjunto $\{x_{n,m}, \mathcal{F}_{n,m}, 1 \leq m \leq n\}$ siempre y cuando se cumpla que $E(x_{n,m} | \mathcal{F}_{n,m-1}) = 0$ para $1 \leq m \leq n$, donde $\mathcal{F}_{n,k}$ es el conjunto que contiene toda la información de los eventos en el momento k . Cabe destacar que el primer índice n puede considerarse como el tamaño de la muestra y $x_{n,m}$ como la m -ésima observación de la muestra.

Teorema central del límite de Martingala

Supóngase que $\{x_{n,m}, \mathcal{F}_{n,m}, 1 \leq m \leq n\}$ es una matriz de diferencia martingala. Sea $S_{n,k} = \sum_{m=1}^k x_{n,m}$, $V_{n,k} = \sum_{m=1}^k E(x_{n,m}^2 | \mathcal{F}_{n,m-1})$ y supóngase que a medida que $n \rightarrow +\infty$ se cumple que:

(I) $V_{n,[nt]} \rightarrow t$ en probabilidad para cada $0 \leq t \leq 1$, y

(II) $\sum_{m=1}^n E(x_{n,m}^2 I_{\{|x_{n,m}| > \delta\}} | \mathcal{F}_{n,m-1}) \rightarrow 0$ en probabilidad para todo $\delta > 0$.

Entonces $\{S_{n,[nt]}, 0 \leq t \leq 1\}$ converge en distribución al movimiento browniano³ cuando $n \rightarrow +\infty$.

El teorema tal y como se recoge en el apéndice de [Lai y Xing \(2008\)](#) implica que, cuando $n \rightarrow +\infty$, $S_{n,n}$ tiene como distribución límite a la distribución $N(0, 1)$. Además, la condición (II) suele denominarse como la condición de Lindeberg. Cuando $x_{n,1}, \dots, x_{n,n}$ son variables aleatorias independientes de media 0, tales que $\sum_{m=1}^n E(x_{n,m}^2) = 1$, la condición (II) se reduce a la condición clásica de Lindeberg

$$\lim_{n \rightarrow +\infty} \sum_{m=1}^n E(x_{n,m}^2 I_{\{|x_{n,m}| > \delta\}}) = 0 \text{ para todo } \delta > 0,$$

la cual es suficiente y necesaria para que $S_{n,n}$ tenga una distribución normal como límite.

³Sea \mathcal{F}_t la σ -álgebra generada por el proceso estocástico $\{W_s : s \leq t\}$. Entonces un movimiento browniano W_t verifica las siguientes propiedades:

- (i) $W_t - W_s$ es independiente de \mathcal{F}_s si $0 \leq s < t$
- (ii) $E(W_t | \mathcal{F}_s) = W_s$ siempre que $0 \leq s \leq t$

2.7. Algoritmo EM

Según [Gentle et al. \(2012\)](#), el algoritmo de esperanza-maximización (EM) es una técnica que se utiliza para el cálculo iterativo de estimaciones de máxima verosimilitud, sobre todo cuando se trabaja con datos incompletos. Se basa en la idea de resolver una sucesión de problemas más simples que se obtienen aumentando las variables observadas originales (datos incompletos) con un conjunto de variables adicionales que son inobservables o no disponibles para el usuario. Estos datos adicionales se conocen como datos ausentes en el marco EM. Además, este algoritmo está estrechamente relacionado con el enfoque “ad hoc” de estimación con datos ausentes, en el que los parámetros se estiman después de rellenar los valores iniciales de los datos que faltan. Estos últimos se actualizan mediante sus valores predichos utilizando las estimaciones iniciales de los parámetros. A continuación, se vuelven a estimar los parámetros, y así sucesivamente, procediendo de forma iterativa hasta la convergencia. En cada iteración del algoritmo EM hay dos pasos, uno de ellos denominado como paso de expectativa (o paso E) y el otro conocido como paso de maximización (o paso M).

El algoritmo EM tiene una serie de propiedades muy buenas, como su estabilidad numérica, convergencia global y simplicidad de implementación. Sin embargo, el algoritmo EM no está exento de limitaciones. En su forma básica, el algoritmo EM carece de un procedimiento incorporado para calcular la matriz de covarianza de los parámetros y a veces es muy lento para converger. Además, algunos problemas complejos con datos incompletos conducen a pasos E y M intratables. El algoritmo EM puede ser tratado como un procedimiento de optimización y tiene muchas similitudes con otros algoritmos como los de Montecarlo a través de cadenas de Markov. Se trata de un modelo muy útil en los campos relacionados con las ciencias biomédicas, ciencias de la salud y la bioinformática.

2.8. Variables aleatorias intercambiables

Tal y como se puede consultar en [Fristedt y Gray \(2013\)](#), una secuencia de variables aleatorias (finita o infinita) X_1, X_2, \dots se dice que es intercambiable si la distribución de probabilidad conjunta es invariante bajo permutaciones finitas $X_{\sigma(1)}, X_{\sigma(2)}, \dots$ para toda $\sigma \in S(n)$ con $n \in \mathbb{N}$. En particular, una sucesión de variables idénticamente distribuidas es intercambiable. En otras palabras, el concepto de intercambiabilidad generaliza el concepto de ser idénticamente distribuidas.

Teorema central del límite para variables aleatorias intercambiables

Si $\{X_n, n \geq 1\}$ es una secuencia de variables aleatorias intercambiables definida sobre (Ω, \mathcal{F}, P) con

$$\mathcal{L}\left(\frac{1}{b_n}\left(\sum_{i=1}^n X_i - a_n\right)\right) \rightarrow N(0, 1), \quad (2.4)$$

para algunas constantes a_n, b_n donde $0 < b_n \rightarrow +\infty$ y con \mathcal{L} se denota a la distribución, entonces existe una secuencia positiva $\varepsilon_n \downarrow 0$ tal que

$$nP^\omega\{|\xi_1| > \varepsilon_n b_n \xrightarrow{P} 0\} \quad (2.5)$$

donde P^ω es una distribución condicionada definida en $\mathcal{G} \subset \mathcal{F}$ tal que para cada $\omega \in \Omega$ las coordenadas de las variables aleatorias $\{\xi_n \equiv \xi_n^\omega, n \geq 1\}$ del espacio de probabilidad $(\mathbb{R}^\infty, \mathcal{B}^\infty, P^\omega)$ son independientes e idénticamente distribuidas.

Además, o bien $\frac{b_n}{n^{1/2}}$ varía lentamente con

$$\frac{v_{n,\omega}}{b_n} \xrightarrow{P} 1, \quad \frac{a_{n,\omega} - a_n}{b_n} \xrightarrow{P} 0 \quad (2.6)$$

o $\frac{b_n}{n}$ varía lentamente y

$$\frac{v_{n,\omega}}{b_n} \xrightarrow{P} 0, \quad \mathcal{L}\left(\frac{a_{n,\omega} - a_n}{b_n}\right) \rightarrow N(0, 1) \quad (2.7)$$

siendo

$$\begin{cases} a_{n,\omega} = E^\omega(T_{n,n,\omega}), & n \geq 1 \\ v_{n,\omega}^2 = E^\omega(T_{n,n,\omega} - a_{n,\omega})^2, & n \geq 1 \\ T_{j,n,\omega} = \sum_{i=1}^j \xi_{i,n,\omega} \equiv \sum_{i=1}^j \xi_i^\omega I_{[|\xi_i^\omega| \leq \varepsilon_n b_n]}, & 1 \leq j \leq n \end{cases}$$

Inversamente, si existen secuencias $a_n, 0 < b_n \rightarrow +\infty, \varepsilon_n \downarrow 0$, tales que (2.5) y (2.6) o (2.7) se cumplen, entonces se obtiene (2.4).

Este resultado y su demostración se pueden consultar en el artículo [Klass et al. \(1987\)](#), incluido como *theorem 2*.

2.9. Método de Montecarlo

Los métodos de Montecarlo constituyen el área o el campo de las matemáticas experimentales que se ocupa de los experimentos sobre eventos aleatorios. Dichos métodos pueden ser de tipo probabilístico o determinista. Además, según [Cleophas y Zwinderman \(2012\)](#) el método de Montecarlo es un enfoque alternativo y fiable para el análisis de datos. Generalmente, con dicho método se obtienen intervalos de confianza sin necesidad de conocer la distribución que siguen los datos y el tiempo que necesita para realizar los cálculos es menor que con los métodos exactos. Sin embargo, el método no es muy aconsejable cuando las muestras son pequeñas.

El objetivo general de la simulación de Montecarlo es estimar alguna característica de una variable aleatoria X como podría ser una función de distribución y al igual que en la inferencia estadística, el método de Montecarlo puede formularse como estimación de una integral definida de la forma

$$\theta = \int_D f(x) dx$$

en un dominio D y una función f que satisface ciertas condiciones de optimalidad. Si la función f se descompone para tener un factor que sea una función de densidad de probabilidad:

$$f(x) = g(x)p(x)$$

donde

$$\int_D p(x) dx = 1$$

y $p(x) \geq 0$ entonces la integral θ es la esperanza de la función g de la variable aleatoria con densidad de probabilidad p , es decir:

$$\theta = E(g(X)) = \int_D g(x) p(x) dx$$

Además, dada una muestra aleatoria x_1, \dots, x_m con distribución de densidad p , una estimación de θ es:

$$\hat{\theta} = \frac{\sum_{i=1}^m g(x_i)}{m}$$

Se ha resuelto así el problema estimando un valor esperado mediante el uso de una muestra simulada de la distribución de la variable aleatoria en estudio. Esta técnica se utiliza en muchos ámbitos de la estadística y tal como especifica [Gentle \(2009\)](#) el método posee tres pasos:

1. Descomponer la función de interés para incluir una función de densidad de probabilidad como factor.
2. Identificar el valor esperado.
3. Utilizar una muestra (simulada o no) para estimar el valor esperado.

Los métodos de Montecarlo pueden considerarse procedimientos estocásticos de integración y consisten en “inventar un juego aleatorio” tal que el valor esperado de una variable aleatoria sea exactamente igual al parámetro que se quiere calcular. El promedio de las muestras de Montecarlo repetidas e independientes de ese “juego” converge (en el sentido teórico de la medida) a la solución deseada.

Capítulo 3

Comparación de curvas de supervivencia

El objetivo principal de la comparación de curvas de supervivencia es estudiar si dos o más curvas de supervivencia de distintas poblaciones son iguales o difieren a lo largo del tiempo en estudio. Para dicho análisis, se debe plantear el siguiente contraste de hipótesis:

$$\begin{aligned} H_0 : S_1(t) = \cdots = S_k(t) & \quad \text{para todo } t \in T \\ H_1 : S_i(t) \neq S_j(t) & \quad \text{para algún } i \neq j, t \in T \end{aligned} \quad (3.1)$$

donde $k = 1, \dots, M$ siendo M el número de poblaciones en estudio.

También puede ser de interés el siguiente contraste:

$$\begin{aligned} H_0 : S_1(t) = \cdots = S_k(t) & \quad \text{para todo } t \in T \\ H_1 : S_1(t) \leq \cdots \leq S_k(t) & \quad \text{para todo } t \in T \end{aligned} \quad (3.2)$$

con al menos una desigualdad estricta. De manera similar, puede considerarse como hipótesis alternativa la siguiente:

$$H_1 : S_1(t) \geq \cdots \geq S_k(t) \quad \text{para todo } t \in T$$

3.1. Técnicas generales

3.1.1. El test *log-rank* ponderado

Tal y como se especifica en [Kleinbaum y Klein \(2012\)](#), una manera de abordar el contraste (3.1) es mediante el test *log-rank* ponderado. Los estadísticos asociados a dicha prueba se encargan de calcular la diferencia entre los fallos observados o reales y los esperados a lo largo del tiempo de estudio. Además, con el fin de que el test posea una mayor potencia y los estadísticos sean robustos, surgen distintos pesos o ponderaciones. Cabe destacar que si se toma una ponderación incorrecta, la potencia del test puede verse afectada de manera negativa. Las diferentes ponderaciones o pesos dan lugar a distintos test caracterizados por tener unas propiedades específicas.

A continuación, se detallan los k estadísticos en los que se basa el test. Suponiendo $t_1 \leq \cdots \leq t_n$ los tiempos observados, d_i , $i = 1, \dots, n$ el número de fallos o muertes (según

sea la naturaleza de los datos) en el instante t_i y r_i , $i = 1, \dots, n$ el número de individuos en riesgo en el instante t_i , se tiene que:

$$Z_g^{LR} = \sum_{i=1}^n \omega(t_i) r_{ig} \left(\frac{d_{ig}}{r_{ig}} - \frac{d_i}{r_i} \right), \quad g = 1, \dots, k,$$

donde d_{ig} es el número de fallos o muertes en el grupo o población g e instante t_i , r_{ig} el número de individuos en riesgo en el grupo o población g e instante t_i y $\omega(t_i)$ la función peso. Asimismo, se cumplen las siguientes relaciones:

$$d_i = \sum_{g=1}^k d_{ig}$$

$$r_i = \sum_{g=1}^k r_{ig}$$

Además, si se considera un vector al que se llamará Z^{LR} que contenga los k estadísticos dados por cada uno de los Z_g^{LR} y sea Σ^{LR} la matriz de varianza-covarianza del vector Z^{LR} , entonces el estadístico asociado al test vendrá dado por:

$$Q^{LR} = (Z^{LR})' (\Sigma^{LR})^{-1} Z^{LR} \quad (3.3)$$

donde $(\Sigma^{LR})^{-1}$ es la inversa generalizada⁴ de Σ^{LR} y los elementos de la matriz de varianza-covarianza del vector Z^{LR} son:

$$\sigma_{gj}^{LR} = \sum_{i=1}^n \omega(t_i)^2 \left(\frac{r_{ig}}{r_i} \right) \left(\delta_{gj} - \frac{r_{ij}}{r_i} \right) \left(\frac{r_i - d_i}{r_i - 1} \right) d_i$$

siendo

$$\delta_{gj} = \begin{cases} 1 & \text{si } g = j \\ 0 & \text{si } g \neq j \end{cases}$$

y $g, j = 1, \dots, k$.

El estadístico Q^{LR} sigue aproximadamente una distribución χ^2 con $k-1$ grados de libertad. A continuación, se enumeran algunas de las funciones peso más habituales. Sin embargo, el test que se suele utilizar más a menudo es aquel que usa pesos iguales y se conoce como el test *log-rank*.

En la siguiente tabla se pueden consultar algunos de los pesos $\omega(t_i)$ más utilizados a la hora de aplicar el test *log-rank* ponderado. Nótese que en dicha tabla se usará la siguiente notación:

$$\hat{S}(t) = \prod_{i=1}^m 1 - \frac{d_i}{r_i}$$

$$\tilde{S}(t) = \prod_{i=1}^m 1 - \frac{d_i}{r_i + 1}$$

donde m representa el m -ésimo instante de ocurrencia del evento en estudio.

⁴Sea X una matriz $n \times m$, entonces se dice que la matriz X^- de dimensión $m \times n$ es una inversa generalizada de X si se cumple que $XX^-X = X$.

Test	Peso
Log-rank	1
Gehan-Breslow	r_i
Tarone-Ware	$\sqrt{r_i}$
Peto-Peto	$\tilde{S}(t_i)$
Peto-Peto modificado	$\tilde{S}(t_i)r_i/(r_i + 1)$
Fleming-Harrington (0, 1)	$1 - \hat{S}(t_{i-1})$
Fleming-Harrington (1, 0)	$\hat{S}(t_{i-1})$
Fleming-Harrington (1, 1)	$(1 - \hat{S}(t_{i-1}))\hat{S}(t_{i-1})$
Fleming-Harrington (0.5, 2)	$(1 - \hat{S}(t_{i-1}))^2\sqrt{\hat{S}(t_{i-1})}$

Tabla 3.1: Ejemplos de funciones pesos $\omega(t_i)$

3.1.2. Prueba de tendencia (*trend-testing*)

En muchos experimentos y estudios cuando se desea comparar la eficacia de varios tratamientos o modalidades, es necesaria la existencia de pruebas sensibles. Un enfoque clásico de análisis es el análisis de la varianza (ANOVA) y las pruebas chi-cuadrado de grupos múltiples. Sin embargo, un enfoque más sensible puede ser el análisis de tendencias, entendiéndose por tendencia cualquier asociación entre el orden del tratamiento y la magnitud de la respuesta. Se suele usar cuando se trabaja con modelos de regresión y cuando se comparan tres o más curvas de supervivencia en las que las columnas están en un orden natural (edades o estadios del cáncer⁵). Además, en esencia, el test comprueba si existe una tendencia lineal entre el orden de las columnas y la mediana de supervivencia⁶.

Tal y como especifican Kirkwood y Sterne (2010), una prueba de tendencia es una aproximación (basada en una prueba de puntuación⁷) a un test de razón de verosimilitud en el que la hipótesis nula es el coeficiente de regresión para un efecto lineal igual a cero.

Este tipo de análisis es adecuado para datos binarios mediante la prueba χ^2 , aunque se pueden obtener resultados similares a partir de un modelo de regresión logística. Sin embargo, en general el test de tendencia suele ser más sensible que las pruebas estándar. Cuando se trabaja con dos tratamientos, la prueba de tendencia y la prueba estándar obtienen resultados idénticos ya que en cada caso, el estadístico correspondiente tiene los mismos grados de libertad. Sin embargo, con tres o más tratamientos, los grados de

⁵La determinación del estadio de un cáncer ayuda a definir la ubicación del cáncer, si se ha diseminado o hacia dónde, y si está afectando a otras zonas del cuerpo.

⁶Concepto que se suele usar en medicina y que denota el tiempo que pasa desde la fecha del diagnóstico o el comienzo del tratamiento de una enfermedad, como el cáncer, durante el cual la mitad de los pacientes con la enfermedad siguen vivos. En un ensayo clínico, se mide la mediana de supervivencia como una forma de determinar la eficacia de un tratamiento nuevo.

⁷La prueba de puntuación, *score test* o prueba del multiplicador de Lagrange es un test utilizado para comprobar si una restricción impuesta a un modelo estimado por máxima verosimilitud es violada por los datos. Dicha prueba evalúa las restricciones de los parámetros estadísticos basándose en el gradiente de la función de verosimilitud evaluada en el valor del parámetro hipotético bajo la hipótesis nula. Además, la distribución asintótica bajo la hipótesis nula es una distribución χ^2 .

libertad del estadístico asociado al análisis estándar aumentan rápidamente, mientras que con el análisis de tendencias no lo hacen, dando lugar a p-valores más pequeños.

No obstante, se deben tener en cuenta las limitaciones que presenta el análisis de tendencias. En primer lugar, si no existe tendencia en los datos, el método de análisis estándar puede ser más sensible que la prueba de tendencia. Como consecuencia, se deben realizar pruebas estándar además de las pruebas de tendencia. En segundo lugar, las pruebas de tendencia suponen una tendencia lineal en la respuesta de los datos con los tratamientos posteriores. Esto significa que, con datos continuos, las medias de los grupos de tratamiento aumentan linealmente, y con datos binarios, las probabilidades de respuesta aumentan exponencialmente, i.e., los logaritmos de las probabilidades aumentan linealmente. El efecto lineal es una suposición simplificadora que debe comprobarse. Dicha comprobación es fácil cuando se tienen como máximo tres categorías. De esta forma, se puede concluir que las pruebas de tendencia pueden ser más sensibles que los métodos estándar de comparación de tratamientos.

Las pruebas de tendencia están diseñadas para detectar diferencias ordenadas en las curvas de supervivencia, i.e., para al menos un grupo:

$$S_1(t) \geq S_2(t) \geq S_3(t) \geq \dots \geq S_k(t) \quad t \leq \tau$$

con al menos una desigualdad estricta y donde τ es la mayor $t \in T$ en la que todos los grupos involucrados en el estudio tienen al menos un sujeto en riesgo y $k = 1, \dots, M$ con M el número de poblaciones en estudio. Además, la hipótesis nula del test es:

$$S_1(t) = S_2(t) = \dots = S_k(t) \quad t \leq \tau$$

Según [Dardis \(2018\)](#) las puntuaciones utilizadas para construir la prueba suelen ser $s = 1, 2, \dots, k$, pero pueden darse como un vector \mathbf{s} que represente una característica numérica del grupo correspondiente. Además, la prueba se basa en los siguientes cálculos:

$$Z_g^{TT}(t_i) = \sum_{t_i \leq \tau} \omega(t_i) \left[e_{ig} - r_{ig} \frac{e_i}{r_i} \right], \quad g = 1, 2, \dots, k,$$

donde $\omega(t_i)$ es la función peso (véase la [tabla de pesos 3.1](#)), r_i es el número de individuos en riesgo en el instante t_i , r_{ig} es el número de individuos en riesgo en el grupo o población g e instante t_i , e_i es el número de eventos (por tiempo) y e_{ig} es el número de eventos (por tiempo) en el grupo o población g .

De esta forma, el estadístico asociado al test viene definido por:

$$Z^{TT} = \frac{\sum_{g=1}^k s_g Z_g^{TT}(\tau)}{\sqrt{\sum_{g=1}^k \sum_{j=1}^k s_g s_j \sigma_{gj}^{TT}}} \quad (3.4)$$

donde σ_{gj}^{TT} es el elemento apropiado de la matriz de varianza-covarianza asociada a dicho estadístico y s_g y s_j son los elementos g -ésimo y j -ésimo del vector \mathbf{s} respectivamente. Además, cuando la hipótesis nula del contraste se rechaza, se cumple que el estadístico Z^{TT} es mayor que el α -ésimo percentil de una distribución normal.

A continuación, se especifican los elementos de la matriz de varianza-covarianza de Z^{TT} para al menos $k \geq 2$ poblaciones:

$$\sigma_{gj}^{TT} = \begin{cases} \sum_{t_i \leq \tau} \frac{r_{ig} (r_i - r_{ig}) e_i (r_i - e_i)}{r_i^2 (r_i - 1)} & g, j = 1, \dots, k \text{ y } g = j, \\ - \sum_{t_i \leq \tau} \frac{r_{ig} r_{ij} e_i (r_i - e_i)}{r_i^2 (r_i - 1)} & g, j = 1, \dots, k \text{ con } g \neq j. \end{cases}$$

3.2. Técnicas específicas para datos censurados

Tal y como comentan Calle et al. (2009), el comportamiento asintótico de los estadísticos asociados a los test basados en datos censurados por intervalos es más difícil de estudiar que en el caso de la censura a la derecha porque por ejemplo, la teoría del proceso de recuento no puede ser aplicada. Esta dificultad se afronta con métodos de remuestreo o con métodos de inferencia estándar que requieren supuestos adicionales. Entre los métodos de remuestreo se encuentran los procedimientos de permutación, imputación múltiple y procedimientos bootstrap.

A continuación, se enumerarán diferentes técnicas para la comparación de curvas de supervivencia de dos o más poblaciones diferentes con la presencia de datos censurados a la derecha (RC \equiv *right censored data*) o por intervalo (IC \equiv *interval censoring* o *interval-censored data*) (véase Oller y Langohr (2017)). Para cada caso, se describirán los estadísticos de prueba de la clase de Fleming-Harrington $G^{\rho, \lambda}$, concepto que introdujeron Fleming y Harrington (1982) y que se trata de un subconjunto de lo que se conoce como la prueba *log-rank* ponderada (Sección 3.1.1).

Observación: No se tratarán técnicas para datos censurados a la izquierda debido a que dichos datos pueden transformarse fácilmente en datos censurados a la derecha. En el libro de Goel y Klein (1992) se determina que para obtener datos censurados a la derecha partiendo de una muestra de datos con censura a la izquierda basta con multiplicar cada uno de los datos por -1 . De esta forma, pueden extrapolarse las técnicas que se detallan a continuación a los datos censurados a la izquierda. ▲

Cuando se trata con muestras de dos o más poblaciones diferentes, puede ser de interés comparar las curvas de supervivencia asociadas a cada una de las poblaciones a través de los siguientes contrastes:

- Test con 2 poblaciones (bilateral):

$$\begin{aligned} H_0 : S_1(t) &= S_2(t) \quad \forall t \in T \\ H_1 : S_1(t) &\neq S_2(t) \quad \text{para algún } t \in T \end{aligned} \tag{3.5}$$

También puede contemplarse el contraste en el que la hipótesis alternativa es unilateral.

- Test con k poblaciones:

$$\begin{aligned} H_0 : S_1(t) &= \dots = S_k(t) \quad \forall t \in T \\ H_1 : \exists i, j, t &\text{ tal que } S_i(t) \neq S_j(t) \end{aligned} \tag{3.6}$$

- Prueba de tendencia con hipótesis alternativa creciente:

$$\begin{aligned} H_0 : S_1(t) &= \cdots = S_k(t) \quad \forall t \in T \\ H_1 : S_1(t) &\leq \cdots \leq S_k(t) \quad \forall t \in T \end{aligned} \quad (3.7)$$

con al menos una desigualdad estricta. También puede contemplarse la hipótesis alternativa decreciente $H_1 : S_1(t) \geq \cdots \geq S_k(t) \quad \forall t \in T$ con al menos una desigualdad estricta.

Bajo condiciones generales, el estadístico básico de cada uno de los contrastes anteriores es:

$$Z = (Z_1, \dots, Z_k)'$$

el cual es asintóticamente normal de media cero y matriz de varianza-covarianza Σ . La forma explícita de Z y Σ se indica a continuación y depende del tipo de censura y de la aproximación asintótica. El estadístico de prueba asociado al contraste (3.5) es:

$$V_1 = \frac{Z_2}{\sqrt{\sigma_{22}}} \quad (3.8)$$

donde σ_{22} es el segundo elemento de la diagonal de Σ y la distribución asintótica de V_1 es la distribución normal estándar. Por otro lado, el estadístico asociado a la hipótesis del test con k poblaciones (3.6) es:

$$V_2 = Z' \Sigma^{-1} Z \quad (3.9)$$

donde Σ^{-1} es la inversa generalizada de Σ y la distribución asintótica de V_2 es una χ^2 con $k - 1$ grados de libertad. Por último, el test de tendencia (3.7) con la hipótesis alternativa creciente puede resolverse con el estadístico:

$$V_3 = \frac{a' Z}{\sqrt{a' \Sigma a}} \quad (3.10)$$

donde $a = (a_1, \dots, a_k)'$, $a_1 < \cdots < a_k$, son los valores de las covariables de cada una de las poblaciones en estudio. La distribución asintótica de V_3 es una distribución normal estándar. Un test unilateral por la izquierda puede mostrar evidencias de una tendencia creciente y uno por la derecha, de una tendencia decreciente.

3.2.1. Datos censurados a la derecha

3.2.1.1. Enfoque mediante el proceso de conteo

Considérese una muestra aleatoria de datos censurados a la derecha $[L_i, R_i]$, $i = 1, \dots, n$ con $L_i = R_i$ (datos no censurados) o $R_i = +\infty$ (datos censurados a la derecha). Sean $t_1 < \cdots < t_m$ los m tiempos no censurados y supóngase que existen $n - m$ tiempos censurados a la derecha. En estas condiciones, las componentes del estadístico del test *log-rank* ponderado Z^{RC} son de la forma:

$$Z_g^{RC} = \sum_{q=1}^m \omega(t_q) \left(d_{qg} - \frac{r_{qg}}{r_q} d_q \right), \quad g = 1, \dots, k, \quad (3.11)$$

donde d_q representa al número de muertes en el instante t_q , r_q el número de individuos en riesgo en el instante t_q , es decir, el número de individuos vivos justo antes del instante t_q y $\omega(t_q)$ es el peso sugerido por Fleming-Harrington. Tanto d_{gg} como r_{gg} corresponden a las mismas definiciones pero referidas a las observaciones de la población g -ésima.

Las ponderaciones o pesos sugeridos por Fleming-Harrington son:

$$\omega(t_q) = \hat{S}(t_q^-)^\rho (1 - \hat{S}(t_q^-))^\lambda, \quad (3.12)$$

donde $\hat{S}(t_q^-)$ es el valor del [estimador de Kaplan-Meier](#) de la función de supervivencia justo antes del instante t_q . Cuando $(\rho, \lambda) = (0, 0)$ y $(\rho, \lambda) = (1, 0)$, el estadístico Z^{RC} corresponde a los estadísticos de prueba *log-rank* y Prentice-Wilcoxon, respectivamente. La elección de los pesos debe hacerse antes del examen de los datos teniendo en cuenta que deben proporcionar la mayor potencia estadística, que a su vez depende de cómo se cree que se viola la nulidad. No obstante, en los ensayos clínicos, el poder estadístico puede no ser la única consideración para determinar el esquema de ponderación, ya que también deben contemplarse los aspectos clínicos.

Asintóticamente, la distribución de (3.11) se obtiene mediante el [teorema central del límite de Martingala](#). Además, bajo la hipótesis nula, Z^{RC} sigue asintóticamente una distribución normal con media nula y matriz de varianza-covarianza $\Sigma^{RCCount.}$. Los elementos de dicha matriz vienen dados por:

$$\sigma_{gj}^{RCCount.} = \begin{cases} \sum_{q=1}^m \omega(t_q)^2 \frac{r_{qq}(r_q - r_{qq})}{r_q^2} \frac{(r_q - d_q) d_q}{r_q - 1} & g, j = 1, \dots, k \text{ y } g = j, \\ - \sum_{q=1}^m \omega(t_q)^2 \frac{r_{qq} r_{qj}}{r_q^2} \frac{(r_q - d_q) d_q}{r_q - 1} & g, j = 1, \dots, k \text{ con } g \neq j. \end{cases}$$

3.2.1.2. Enfoque de permutación

Según [Oller y Langohr \(2017\)](#), el estadístico (3.11) puede escribirse de forma lineal tal y como se indica a continuación:

$$Z^{RCPerm.} = \sum_{i=1}^n \mathbf{z}_i c_i \quad (3.13)$$

donde $\mathbf{z}_i = (Z_{1i}, \dots, Z_{ki})'$ es un vector de covariables indicadoras del grupo o población al que pertenece cada observación, i.e.,

$$Z_{gi} = \begin{cases} 1 & \text{si la } i\text{-ésima observación está en la } g\text{-ésima población} \\ 0 & \text{en caso contrario} \end{cases} \quad (3.14)$$

y c_i es un valor de puntuación asociado a cada una de las observaciones y toma la siguiente forma:

$$c_i = \begin{cases} \omega(t_q) - \sum_{s=1}^q \omega(t_s) \frac{d_s}{r_s} & \text{si } L_i = R_i = t_q, \\ - \sum_{s=1}^q \omega(t_s) \frac{d_s}{r_s} & \text{si } R_i = +\infty \text{ y } t_q \leq L_i < t_{q+1}. \end{cases} \quad (3.15)$$

El enfoque de permutación puede aplicarse fácilmente a esta forma lineal. Si la hipótesis nula es cierta y el proceso de censura subyacente es idéntico en todas las poblaciones, entonces las etiquetas de las puntuaciones c_i son intercambiables. La distribución de permutación de $Z^{RCPerm.}$ se obtiene entonces permutando las etiquetas y volviendo a calcular el estadístico de prueba para todas las posibles permutaciones. Además, los p-valores de la distribución de permutación pueden calcularse utilizando la enumeración completa, un algoritmo de red, un método de aproximación como el de Montecarlo o asintóticamente utilizando una versión del [teorema del límite central para variables aleatorias intercambiables](#). Nótese que este último método produce una aproximación a una distribución normal de media 0 y matriz de varianza-covarianza definida por:

$$\Sigma^{RCPerm.} = \frac{1}{n-1} \left(\sum_{i=1}^n c_i^2 \right) \left(\sum_{i=1}^n (\mathbf{z}_i \mathbf{z}_i' - \bar{\mathbf{z}}_i \bar{\mathbf{z}}_i') \right)$$

donde $\bar{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i$ es la media muestral.

3.2.2. Datos censurados por intervalos

3.2.2.1. Enfoque de permutación

Considérese una muestra aleatoria de datos censurados por intervalos $(L_i, R_i]$, dados $t_1 < \dots < t_m$ denótese los elementos ordenados $\{L_i, R_i; i = 1, \dots, n\}$ y sea $\hat{S}(t)$ la estimación no paramétrica de la función de supervivencia a través del [estimador de Turnbull](#). En dichas condiciones, las componentes del estadístico Z^{IC} son de la forma

$$Z_g^{IC} = \sum_{q=1}^m \omega^{IC}(t_q) \left(d_{qg}^{IC} - \frac{r_{qg}^{IC}}{r_q^{IC}} d_q^{IC} \right), \quad g = 1, \dots, k, \quad (3.16)$$

donde $d_q^{IC} = n(\hat{S}(t_q^-) - \hat{S}(t_q))$ representa el número estimado de muertes en el instante t_q y $r_q^{IC} = n\hat{S}(t_q^-)$ expresa el número estimado de individuos en riesgo en el instante t_q . Además, d_{qg}^{IC} y r_{qg}^{IC} se definen como:

$$d_{qg}^{IC} = n \left(\frac{1}{N_g} \sum_{j=1}^N Z_{gj} \hat{S}_j(t_q^-) - \frac{1}{N_g} \sum_{j=1}^N Z_{gj} \hat{S}_j(t_q) \right)$$

$$r_{qg}^{IC} = n \frac{1}{N_g} \sum_{j=1}^N Z_{gj} \hat{S}_j(t_q^-)$$

siendo N_g el tamaño de la muestra de la g -ésima población, Z_{gj} es la variable indicadora definida en (3.14), \hat{S}_j la estimación de la función de supervivencia del j -ésimo individuo usando el [algoritmo EM](#) y definida como la estimación de la función de supervivencia global de Turnbull $\hat{S}(t)$ truncada en el j -ésimo intervalo observado (véase [Gómez y Oller \(2008, 2012\)](#), i.e.,

$$\hat{S}_j(t) = P_{\hat{S}}[T \in (t, +\infty) | (L_i, R_i)] = \frac{\hat{S}(\max\{L_i, t\}) - \hat{S}(\max\{R_i, t\})}{\hat{S}(L_i) - \hat{S}(R_i)}$$

siendo $P_{\hat{S}}$ la probabilidad dada por la función de supervivencia $\hat{S}(t)$ y

$$N = \sum_{g=1}^k N_g.$$

Observación: La estimación de $\hat{S}_j(t)$ obtenida en el paso de maximización del algoritmo EM posee las siguientes propiedades:

1. $\hat{S}(t) = \frac{1}{n} \sum_{j=1}^n \hat{S}_j(t)$
2. $\hat{S}_j(t) = I_{\{t < t_j\}}$ cuando el j -ésimo individuo no está censurado y $T = t_j$. ▲

Por último, los pesos o ponderaciones propuestos por Oller y Langohr son:

$$\omega^{IC}(t_q) = \hat{S}(t_q^-) \frac{B(1 - \hat{S}(t_q); \lambda + 1, \rho) - B(1 - \hat{S}(t_q^-); \lambda + 1, \rho)}{\hat{S}(t_q^-) - \hat{S}(t_q)}, \quad (3.17)$$

donde $B(x; p, r) = \int_0^x t^{p-1} (1-t)^{r-1} dt$; $p, r \in \mathbb{C}$ tales que $\text{Re}(p) > 0$ y $\text{Re}(r) > 0$ denota a la función beta incompleta⁸. Cabe destacar que estas ponderaciones derivan del enfoque del vector de puntuación que se verá en la [Sección 3.2.2.2](#). Cuando la dupla (ρ, λ) toma los valores $(0, 0)$, el estadístico Z^{IC} se reduce al estadístico *log-rank* mientras que si toma los valores $(1, 0)$, el estadístico se reduce al estadístico de Wilcoxon-Peto. Además, como \hat{S} es asintóticamente continua en el instante t_q , la función peso $\omega^{IC}(t_q)$ se parece a $[\hat{S}(t_q^-)]^\rho [1 - \hat{S}(t_q^-)]^\lambda$ y como consecuencia, la interpretación de los parámetros λ y ρ reproducen la interpretación que posee la familia de datos censurados a la derecha.

Esta clase de estadísticos Z^{IC} puede escribirse en forma lineal tal y como se especifica a continuación:

$$Z^{ICPerm.} = \sum_{i=1}^n \mathbf{z}_i^{IC} c_i^{IC} \quad (3.18)$$

donde $\mathbf{z}_i^{IC} = (Z_{1i}^{IC}, \dots, Z_{Ki}^{IC})'$ es un vector de covariables indicadoras del grupo o población al que pertenece cada observación, i.e.,

$$Z_{gi}^{IC} = \begin{cases} 1 & \text{si la } i\text{-ésima observación está en la } g\text{-ésima población} \\ 0 & \text{en caso contrario} \end{cases} \quad (3.19)$$

y c_i^{IC} es un valor de puntuación asociado a cada una de las observaciones:

$$c_i^{IC} = \begin{cases} \frac{\hat{S}(R_i) B(1 - \hat{S}(R_i); \lambda + 1, \rho) - \hat{S}(L_i) B(1 - \hat{S}(L_i); \lambda + 1, \rho)}{\hat{S}(L_i) - \hat{S}(R_i)} & \text{si } \hat{S}(R_i) \neq 0, \\ -B(1 - \hat{S}(L_i); \lambda + 1, \rho) & \text{si } \hat{S}(R_i) = 0. \end{cases}$$

⁸La función beta incompleta es una generalización de la función beta. Además, coincide con la función beta cuando $x = 1$. Se define como $B(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt$ con $a, b \in \mathbb{C}$ tales que $\text{Re}(a) > 0$ y $\text{Re}(b) > 0$.

Nótese que la equivalencia entre la forma *log-rank* ponderada (3.16) y la forma lineal (3.18) se basa en las ponderaciones (3.17). Además, en el hipotético caso de que el tipo de censura sea el mismo en cada una de las poblaciones en estudio, la forma lineal $Z^{ICPerm.}$ dada en (3.18) permite el uso de una distribución de permutación con media 0 y matriz de varianza-covarianza definida de la siguiente forma:

$$\Sigma^{ICPerm.} = \frac{1}{n-1} \left(\sum_{i=1}^n (c_i^{IC})^2 \right) \left(\sum_{i=1}^n (\mathbf{z}_i^{IC} (\mathbf{z}_i^{IC})' - \bar{\mathbf{z}}_i^{IC} (\bar{\mathbf{z}}_i^{IC})') \right)$$

donde $\bar{\mathbf{z}}^{IC} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^{IC}$ es la media muestral.

Observación: Para la construcción del estadístico $Z^{ICPerm.}$ se ha considerado que los intervalos observados son semiabiertos aunque con pequeñas modificaciones, se puede hacer extensible a intervalos cerrados o abiertos. Además, dicha metodología puede aplicarse a los datos censurados a la derecha como un caso particular de censura por intervalos. En ese caso, los resultados obtenidos por las formas lineales $Z^{RCPerm.}$ dada por (3.13) y $Z^{ICPerm.}$ dada por (3.18) son parecidas pero no idénticas. ▲

3.2.2.2. Enfoque basado en un vector de puntuación

Cuando se restringe la clase de estadísticos Z^{IC} de Fleming-Harrington a la subclase $\lambda = 0$ (suele denominarse como familia G^ρ), cada estadístico del test dado por el valor de ρ puede considerarse como una prueba de puntuación eficiente asociado a un modelo lineal. Para tratar este tipo de enfoque es necesario que los datos censurados se muestren como datos discretos o agrupados en intervalos de tiempo continuo, i.e., $L, R \in \{t_0, t_1, \dots, t_m\}$ donde $0 = t_0 < t_1 < \dots < t_m < t_{m+1} = +\infty$ y además, se asume un modelo lineal de la forma:

$$g(T_i) = -(\mathbf{z}_i^{IC})' \boldsymbol{\beta} + \varepsilon_i$$

siendo g una función creciente desconocida, $\boldsymbol{\beta}$ un vector de constantes desconocidas y ε_i una variable que tiene asociada como función de supervivencia lo siguiente:

$$S_\varepsilon(t) = \left(1 + \rho e^t\right)^{-\frac{1}{\rho}}$$

En tales condiciones, se tiene que el estadístico asociado al contraste es:

$$Z^{ICVec.} = \left[\frac{\partial \log(L(\boldsymbol{\beta}, \boldsymbol{\theta}))}{\partial \boldsymbol{\beta}} \right]_{\boldsymbol{\beta}=\mathbf{0}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0} \quad (3.20)$$

donde $L(\boldsymbol{\beta}, \boldsymbol{\theta})$ es la función de verosimilitud y $\boldsymbol{\theta} = (\theta_q)_{q=1}^m$ son parámetros molestos o de ruido⁹ tales que $\theta_q = g(t_q)$.

Considerando datos discretos censurados por intervalos y suponiendo cierta la hipótesis nula del contraste de igualdad de curvas de supervivencia, el comportamiento asintótico del vector de puntuación $Z^{ICVec.}$ procede de la teoría de máxima verosimilitud. Usando

⁹Los parámetros molestos o de ruido (*nuisance parameters*) son parámetros de un modelo estadístico que no tienen interés científico alguno en el estudio realizado, pero cuyos valores son necesarios para hacer inferencia sobre otros parámetros de interés.

este enfoque es necesario que ninguno de los valores estimados del vector $\boldsymbol{\theta}$ se aproxime a la frontera del espacio que contiene a dichos parámetros, en otras palabras, al espacio definido por $1 > \hat{S}(t_1) > \cdots > \hat{S}(t_m) > 0$. En dicho caso, la distribución del estadístico es asintóticamente normal de media 0 y matriz de varianza-covarianza $\Sigma^{ICVec.}$ especificada por [Gómez y Oller \(2008\)](#) y que se muestra a continuación:

$$\Sigma^{ICVec.} = - \left[\frac{\partial^2 \log(L(S))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} - \left(\frac{\partial^2 \log(L(S))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}'} \right) \left(\frac{\partial^2 \log(L(S))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \left(\frac{\partial^2 \log(L(S))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}'} \right) \right]_{\boldsymbol{\beta}=\mathbf{0}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0}$$

donde

$$L(S) = \prod_{i=1}^n \{S(L_i) - S(R_i)\}$$

con $S(L_i)$ el valor de la función de supervivencia en el extremo izquierdo L_i , $S(R_i)$ el valor de la función de supervivencia en el extremo derecho R_i y para cualesquiera parámetros ϕ_x e ϕ_y arbitrarios se cumple:

$$\frac{\partial^2 \log(L(S))}{\partial \phi_x \partial \phi_y} = \sum_{i=1}^n \frac{1}{\Delta_i(S)} \left[\left(\frac{\partial^2 \Delta_i(S)}{\partial \phi_x \partial \phi_y} \right) - \frac{1}{\Delta_i(S)} \left(\frac{\partial \Delta_i(S)}{\partial \phi_x} \right) \left(\frac{\partial \Delta_i(S)}{\partial \phi_y} \right) \right]$$

con

$$\Delta_i(S) = S(R_i | \mathbf{z}_i^{IC}) - S(L_i | \mathbf{z}_i^{IC})$$

donde \mathbf{z}_i^{IC} es el vector de covariables indicadoras de la población a la que pertenece cada observación y cuyos elementos están definidos en [\(3.19\)](#). Además, se cumple que:

$$\begin{aligned} \left[\frac{\partial S(t_j | \mathbf{z}_i^{IC})}{\partial \beta_x} \right]_{\boldsymbol{\beta}=\mathbf{0}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= -\frac{1}{\rho} \{ \hat{S}(t_j) \} \{ 1 - (\hat{S}(t_j))^\rho \} Z_{ix}^{IC} \\ \left[\frac{\partial S(t_j | \mathbf{z}_i^{IC})}{\partial \theta_x} \right]_{\boldsymbol{\beta}=\mathbf{0}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= -\frac{1}{\rho} \{ \hat{S}(t_j) \} \{ 1 - (\hat{S}(t_j))^\rho \} I_{\{j=x\}} \\ \left[\frac{\partial^2 S(t_j | \mathbf{z}_i^{IC})}{\partial \beta_x \partial \beta_y} \right]_{\boldsymbol{\beta}=\mathbf{0}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \frac{1}{\rho^2} \{ \hat{S}(t_j) \} \{ 1 - (\hat{S}(t_j))^\rho \} \{ 1 - (\rho + 1)(\hat{S}(t_j))^\rho \} Z_{ix}^{IC} Z_{iy}^{IC} \\ \left[\frac{\partial^2 S(t_j | \mathbf{z}_i^{IC})}{\partial \beta_x \partial \theta_y} \right]_{\boldsymbol{\beta}=\mathbf{0}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \frac{1}{\rho^2} \{ \hat{S}(t_j) \} \{ 1 - (\hat{S}(t_j))^\rho \} \{ 1 - (\rho + 1)(\hat{S}(t_j))^\rho \} Z_{ix}^{IC} I_{\{j=y\}} \\ \left[\frac{\partial^2 S(t_j | \mathbf{z}_i^{IC})}{\partial \theta_x \partial \theta_y} \right]_{\boldsymbol{\beta}=\mathbf{0}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \frac{1}{\rho^2} \{ \hat{S}(t_j) \} \{ 1 - (\hat{S}(t_j))^\rho \} \{ 1 - (\rho + 1)(\hat{S}(t_j))^\rho \} I_{\{j=x=y\}} \end{aligned}$$

Observación: el estadístico Z^{RC} y su correspondiente matriz de varianza-covarianza $\Sigma^{RCCount.}$ pueden obtenerse a partir del enfoque del vector de puntuación cuando $\lambda = 0$. En dicho caso, en vez de utilizar una función de verosimilitud completa, se usa una función de verosimilitud parcial. Debido a esta apreciación, la combinación Z^{RC} y $\Sigma^{RCCount.}$ o la combinación $Z^{ICVec.}$ y $\Sigma^{ICVec.}$ pueden utilizarse para el análisis de datos censurados a la derecha a pesar de que los resultados obtenidos difieran un poco entre ellos.

▲

Capítulo 4

Comparación de curvas de supervivencia en R

En este capítulo, se detallará cómo se pueden usar las técnicas vistas en el capítulo anterior en el software libre R. Para ello, se irán exponiendo las librerías o paquetes imprescindibles para realizar dicho estudio y las funciones o comandos necesarios para obtener las conclusiones deseadas. Para mostrar cada una de las técnicas se irán poniendo ejemplos con datos disponibles en los propios paquetes o con datos extraídos de la asignatura Estadística Computacional II. Nótese que se usará un nivel de significación del 5% en todos los contrastes a realizar.

4.1. Librerías para el análisis de supervivencia

A continuación, se recogen brevemente distintas librerías de R que contienen funciones básicas para llevar a cabo un análisis de supervivencia. Para cada una de las librerías en estudio, se detallará la estructura y la salida de cada una de las funciones más relevantes.

4.1.1. Librería *survival*

La librería *survival* ([Therneau \(2021\)](#)) contiene funciones esenciales para el análisis de supervivencia y trabaja con objetos de tipo *Surv*. Este tipo de objetos combina la información temporal de cada individuo u observación con información de censura y se pueden crear con la función *Surv(time, time2, event, type, origin)* cuyos argumentos se definen como:

- *time*: en el caso de que los datos estén censurados a la derecha, en este argumento se anotará el tiempo de seguimiento. En cambio, cuando la censura sea por intervalos, se introducirá el instante inicial del intervalo.
- *event*: indica el estado de la observación o individuo en estudio, $0 = vivo$ y $1 = muerto$. Sin embargo, para datos censurados por intervalos el estado puede ser $0 = censura a la derecha$, $1 = evento observado$, $2 = censurado a la izquierda$ o $3 = censurado por intervalos$.

- *time2*: este argumento solo se anota cuando los datos están censurados por intervalos y en él se debe especificar el instante final del intervalo. Los intervalos se asumen semiabiertos de la forma $(inicio, fin]$. Además, para datos de procesos de conteo el argumento *event* indica si se produjo el evento al final del intervalo.
- *type*: determina el tipo de censura y admite los siguientes valores "right", "left", "counting", "interval", "interval2" o "mstate". Consultar Therneau (2021) para una explicación más detallada de cada uno de los posibles valores del argumento en estudio.
- *origin*: útil cuando se trabaja con datos de procesos de conteo e indica la función de riesgo al comienzo del estudio.

Una vez que se tienen definidos los datos como tipo *Surv*, ya se puede empezar a realizar un análisis de supervivencia de manera correcta. En los próximos apartados se irán detallando funciones de la librería *survival* necesarias para obtener estimaciones no paramétricas de la función de supervivencia o la función de riesgo acumulado. Antes de ello, es imprescindible cargar la librería *survival*:

```
library(survival)
```

4.1.1.1. Estimador de Kaplan-Meier

Para calcular un estimador de la función de supervivencia a través del [procedimiento de Kaplan-Meier](#) se usará la función *survfit()*. Esta función *survfit(formula, data, type)* es capaz de crear la curva de supervivencia asociada a un conjunto de datos a partir de una fórmula. Los argumentos de dicha función se detallan a continuación:

- *formula*: debe ser un objeto de la forma $y \sim x$ donde *y* será un objeto de supervivencia *Surv* que actúa como variable respuesta y *x* será el nombre de las covariables (puede tomar el valor 1 en el caso de que no existan covariables en el conjunto de datos en estudio).
- *data*: objeto *data frame* que contiene los datos del estudio.
- *type*: indica el tipo de estimador y puede tomar el valor "kaplan-meier" para obtener el estimador de Kaplan-Meier o en su defecto "fleming-harrington" para la obtención del estimador de Fleming-Harrington.

Para mostrar una aplicación de dicha función, se usará un conjunto de datos llamado *AML* del paquete en estudio y cuya recogida de datos fue realizada por Embury et al. (1977) para un estudio de la Universidad de Stanford. Dichos datos contienen los resultados de un ensayo clínico cuyo objetivo es evaluar la eficacia del uso de la quimioterapia como tratamiento de mantenimiento para la leucemia mieloide aguda. Los pacientes pertenecientes a la muestra han sido asignados de manera aleatoria a dos grupos tras alcanzar un estado de remisión de la enfermedad a través del tratamiento de quimioterapia. Al primer grupo de pacientes se le mantuvo el tratamiento de quimioterapia como mantenimiento y al segundo grupo (llamado de control) no se le aplicó ningún tratamiento. Con

este estudio se pretendía contrastar si la quimioterapia como tratamiento de mantenimiento prolonga o no el tiempo hasta la recaída del paciente.

En la [Tabla 4.1](#) se recogen los datos del estudio especificado. Nótese que el signo + especifica que el dato correspondiente está censurado:

Grupo	Tiempos de remisión en semanas
Mantenimiento	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
No mantenimiento	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Tabla 4.1: Datos *AML*

Para poder estimar la función de supervivencia es necesario implementar los datos en R en formato *data.frame*. A continuación, se muestran en forma de tabla los datos *AML* extraídos del conjunto de datos *cancer* del paquete *survival*:

```
data(cancer, package = "survival")
```

<i>Time</i>	<i>Status</i>	<i>Group</i>
9	1	Maintained
13	1	Maintained
13	0	Maintained
18	1	Maintained
23	1	Maintained
28	0	Maintained
31	1	Maintained
34	1	Maintained
45	0	Maintained
48	1	Maintained
161	0	Maintained
5	1	Nonmaintained
5	1	Nonmaintained
8	1	Nonmaintained
8	1	Nonmaintained
12	1	Nonmaintained
16	0	Nonmaintained
23	1	Nonmaintained
27	1	Nonmaintained
30	1	Nonmaintained
33	1	Nonmaintained

(continúa)

<i>Time</i>	<i>Status</i>	<i>Group</i>
43	1	Nonmaintained
45	1	Nonmaintained

Tabla 4.2: Datos *AML* implementados en R

Una vez extraídos los datos, se procede a la estimación de la función de supervivencia por el método de Kaplan-Meier para el grupo de pacientes de Mantenimiento (*Maintained*) y para el grupo de pacientes de No Mantenimiento (*Nonmaintained*):

```
KM.grupos <- survfit(Surv(time, status) ~ group, type = "kaplan-meier",
                    data = aml, conf.type = "plain")
```

Se ha añadido el argumento *conf.type = "plain"* para que la estimación del intervalo de confianza se realice a partir de la aproximación de Greenwood del error estándar.

La estimación de la función de supervivencia se puede consultar con la orden *summary()* y la información que devuelve es la siguiente:

- *time*: instante de la observación.
- *n.risk*: número de individuos en riesgo.
- *n.event*: número de individuos que presentaron el evento.
- *survival*: estimación de la función de supervivencia.
- *std.err*: desviación estándar de la estimación realizada.
- *lower 95 % CI* y *upper 95 % CI*: extremo inferior y superior del intervalo de confianza de la estimación al 95 %.

En las tablas 4.3 y 4.4 se recogen las estimaciones obtenidas para cada uno de las poblaciones de pacientes del conjunto de datos *AML*:

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>std.err</i>	<i>lower 95 % CI</i>	<i>upper 95 % CI</i>
9	11	1	0.909	0.0867	0.7392	1.000
13	10	1	0.818	0.1163	0.5903	1.000
13	10	0	0.818	0.1163	0.5903	1.000
18	8	1	0.716	0.1397	0.4422	0.990
23	7	1	0.614	0.1526	0.3145	0.913
28	6	0	0.614	0.1526	0.3145	0.913
31	5	1	0.491	0.1642	0.1691	0.813
34	4	1	0.368	0.1627	0.0494	0.687

(continúa)

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>std.err</i>	<i>lower 95 % CI</i>	<i>upper 95 % CI</i>
45	3	0	0.368	0.1627	0.0494	0.687
48	2	1	0.184	0.1535	0.0000	0.485
161	1	0	0.184	0.1535	0.0000	0.485

Tabla 4.3: Estimador de Kaplan-Meier para el grupo de pacientes *Maintained*

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>std.err</i>	<i>lower 95 % CI</i>	<i>upper 95 % CI</i>
5	12	2	0.8333	0.1076	0.6225	1.000
5	12	0	0.8333	0.1076	0.6225	1.000
8	10	2	0.6667	0.1361	0.3999	0.933
8	10	0	0.6667	0.1361	0.3999	0.933
12	8	1	0.5833	0.1423	0.3044	0.862
16	7	0	0.5833	0.1423	0.3044	0.862
23	6	1	0.4861	0.1481	0.1958	0.776
27	5	1	0.3889	0.1470	0.1008	0.677
30	4	1	0.2917	0.1387	0.0198	0.564
33	3	1	0.1944	0.1219	0.0000	0.433
43	2	1	0.0972	0.0919	0.0000	0.277
45	1	1	0.0000	NaN	NaN	NaN

Tabla 4.4: Estimador de Kaplan-Meier para el grupo de pacientes *Nonmaintained*

Se añade una gráfica en la que se puede observar la estimación de la función de supervivencia para cada una de las poblaciones o grupos de pacientes junto con el intervalo de confianza obtenido al 95 % para dicha estimación. La mencionada gráfica se ha realizado con la ayuda del paquete *survminer* (Biecek et al. (2021)) y la función *ggsurvplot()* que posee:

```
library(survminer)
ggsurvplot(
  KM.grupos,
  data = aml,
  xlab = "t (semanas)",
  ylab = "S(t) estimada",
  lab = c(10, 10, 7),
  size = 1, legend.title = "Grupos",
  palette =
    c("chartreuse", "cyan4"),
  conf.int = TRUE,
  legend.labs =
    c("Maintained", "Nonmaintained"),
```

```

ggtheme = theme_grey(),
caption = "Fuente propia",
font.caption = c(10, "italic", "black"),
font.x = c(12, "bold.italic", "black"),
font.y = c(12, "bold.italic", "black")
)

```

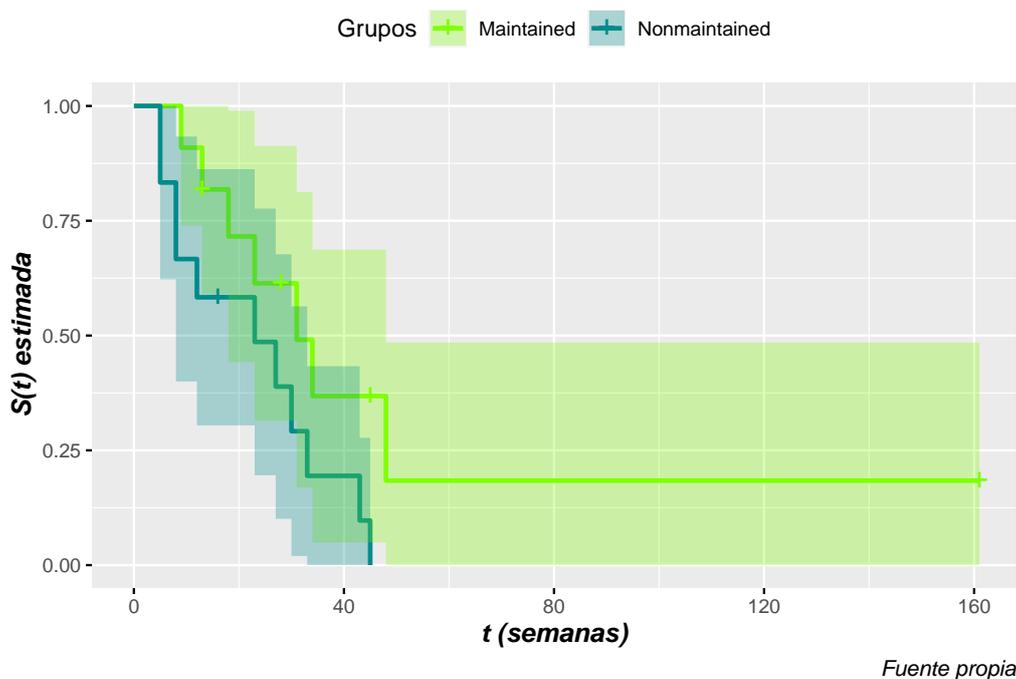


Figura 4.1: Estimador de Kaplan-Meier para cada uno de los grupos de pacientes del conjunto de datos *AML*

4.1.1.2. Estimador de Fleming-Harrington

Para obtener una estimación de la función de supervivencia a través del [procedimiento de Fleming-Harrington](#) se necesita, al igual que para el estimador de Kaplan-Meier, la función `survfit()` pero con el argumento `type = "fleming-harrington"`.

Se añade un ejemplo realizado con el conjunto de datos *oropharynx* ([Kalbfleisch y Prentice \(2002\)](#)). Dicho conjunto proporciona parte de los datos de un ensayo clínico realizado por *Radiation Therapy Oncology Group* de Estados Unidos. El estudio completo incluía pacientes con carcinoma epidermoide en 15 localizaciones de la boca y la garganta, y en él participaron 16 instituciones. Sin embargo, en el fichero de datos con el que se trabajará solo se incluyen 5 localizaciones distintas de las 6 instituciones más importantes. Además, los pacientes que entraron en la muestra fueron asignados aleatoriamente a uno de los dos grupos de tratamiento, radioterapia sola (estándar) o radioterapia junto a un agente quimioterapéutico (test). De esta forma, uno de los objetivos del estudio consistía en la comparación de las dos políticas de tratamiento con respecto a la supervivencia de los pacientes. La existencia de datos censurados se debe principalmente a que algunos

pacientes abandonaron el seguimiento debido a un traslado o a que sobrevivieron hasta la finalización del estudio.

Es destacable el hecho de que en los datos se incluyeron mediciones de covariables muy relacionadas con el análisis de supervivencia. Entre ellas se encuentra la variable *Condition* que proporciona una medida de la capacidad funcional del paciente en el momento del diagnóstico y la variable *Grade* que ofrece el grado en que la célula tumoral se asemeja a la célula huésped.

En primer lugar, se leerá el fichero que contiene dichos datos y se mostrará la cabecera del objeto *oropharynx* que será un *data.frame* que contendrá la información necesaria para realizar posteriormente una estimación de la función de supervivencia:

```
oropharynx <- read.table("oropharynx.txt", header = T)
```

<i>Inst</i>	<i>Sex</i>	<i>TrtGrp</i>	<i>Grade</i>	<i>Age</i>	<i>Cond</i>	<i>Site</i>	<i>T</i>	<i>N</i>	<i>entrydate</i>	<i>Status</i>	<i>Time</i>
2	2	1	1	51	1	2	3	1	2468	1	631
2	1	2	1	65	1	4	2	3	2968	1	270
2	1	1	2	64	2	1	3	3	3368	1	327
2	1	1	1	73	1	1	4	0	5768	1	243
5	1	2	2	64	1	1	4	3	9568	1	916
4	1	2	1	61	1	2	3	0	10668	0	1823
4	1	1	2	65	1	2	4	3	10768	1	637
4	1	2	3	84	1	4	1	3	12068	1	235
6	1	1	2	54	2	1	3	3	13368	1	255

Tabla 4.5: Primeros datos del fichero *oropharynx*

El significado de las variables en estudio son:

- *Inst*: indica la institución a la que pertenece cada uno de los pacientes. Los posibles valores que puede tomar dicha variable son 1, 2, 3, 4, 5 o 6.
- *Sex*: variable sexo que puede tomar los valores 1 = hombre o 2 = mujer.
- *TrtGrp*: grupo de tratamiento que puede tomar los valores 1 = estándar o 2 = test.
- *Grade*: grado en que la célula tumoral se asemeja a la célula huésped y puede tomar los valores 1 = bien diferenciado, 2 = moderadamente diferenciado o 3 = pobremente diferenciado.
- *Age*: edad del paciente en el momento del diagnóstico.
- *Condition*: medida de la capacidad funcional del paciente en el momento del diagnóstico y puede tomar los valores 1 = ninguna incapacidad, 2 = trabajo restringido, 3 = requiere ayuda con los cuidados personales o 4 = confinado en cama.

- *Site*: localización del carcinoma 1 = arco facial, 2 = fosa amigdalina, 3 = pilar posterior, 4 = lengua faríngea o 5 = pared posterior.
- *T*: 1 = primaria, 2 cm o menos, 2 = primarios, 2 a 4 cm, 3 = primaria, más de 4 cm o 4 = tumor invasivo masivo.
- *N*: 0 = no metástasis, 1 = un solo nodo de 3 cm o menos, 2 = un solo nodo de 3 cm o más o 3 = múltiples nodos.
- *Entrydate*: día del año y año, *dddy*.
- *Status*: estado de la censura (0 = censurado o 1 = muerte).
- *Time*: tiempo en días desde el primer día del diagnóstico.

Una vez presentadas cada una de las variables, se procede a estimar la función de supervivencia con respecto a la variable *TrtGrp*:

```
FH.oropharynx <- survfit(Surv(Time, Status) ~ TrtGrp,
                        type = "fleming-harrington",
                        data = oropharynx,
                        conf.type = "plain")
```

Al igual que en la sección anterior, se adjuntan dos tablas (Tabla 4.6 y Tabla 4.7) en las que se pueden consultar las estimaciones obtenidas para cada uno de los tratamientos que se contemplan en el conjunto de datos *oropharynx*:

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>std.err</i>	<i>lower 95 % CI</i>	<i>upper 95 % CI</i>
10	100	0	1.000	0.0000	1.000	1.000
50	100	0	1.000	0.0000	1.000	1.000
100	94	5	0.950	0.0218	0.907	0.993
150	88	6	0.890	0.0313	0.828	0.951
200	82	6	0.830	0.0377	0.756	0.903
250	76	6	0.769	0.0422	0.686	0.852
300	70	6	0.709	0.0455	0.620	0.798
400	58	12	0.588	0.0493	0.492	0.685
500	50	7	0.517	0.0502	0.418	0.615
600	43	7	0.445	0.0500	0.347	0.543
700	38	4	0.404	0.0494	0.307	0.501
800	28	7	0.326	0.0479	0.232	0.420
900	25	0	0.326	0.0479	0.232	0.420
1000	16	3	0.286	0.0473	0.193	0.379

(continúa)

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>std.err</i>	<i>lower 95 % CI</i>	<i>upper 95 % CI</i>
1250	10	2	0.248	0.0481	0.153	0.342
1575	1	2	0.117	0.0692	0.000	0.253

Tabla 4.6: Estimador de Fleming-Harrington para el grupo de pacientes con radioterapia sola (estándar)

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>std.err</i>	<i>lower 95 % CI</i>	<i>upper 95 % CI</i>
10	95	0	1.000	0.0000	1.000	1.000
50	91	4	0.958	0.0204	0.918	0.998
100	90	1	0.948	0.0227	0.903	0.992
150	84	6	0.885	0.0326	0.821	0.949
200	76	7	0.812	0.0400	0.733	0.890
250	68	8	0.727	0.0457	0.637	0.816
300	58	9	0.631	0.0496	0.534	0.728
400	46	12	0.502	0.0516	0.401	0.603
500	38	7	0.425	0.0512	0.325	0.526
600	26	9	0.323	0.0490	0.227	0.419
700	21	3	0.285	0.0479	0.191	0.379
800	18	1	0.270	0.0476	0.177	0.363
900	16	1	0.255	0.0472	0.163	0.348
1000	13	1	0.239	0.0470	0.147	0.331
1250	10	0	0.239	0.0470	0.147	0.331
1575	3	0	0.239	0.0470	0.147	0.331

Tabla 4.7: Estimador de Fleming-Harrington para el grupo de pacientes con radioterapia junto a un agente quimioterapéutico (test)

Se adjunta también una gráfica en la que se pueden observar las estimaciones de Fleming-Harrington de las funciones de supervivencia asociadas a las poblaciones de pacientes tratadas con radioterapia sola o con radioterapia junto a un agente quimioterapéutico:

```
ggsurvplot(
  FH.oropharynx,
  data = oropharynx,
  xlab = "t (días)",
  ylab = "S(t) estimada",
  lab = c(10, 10, 7),
  size = 1, legend.title = "Tratamientos",
  palette =
    c("darkslategray4", "darkorange"),
```

```

conf.int = TRUE,
legend.labs =
  c("Estándar", "Test"),
ggtheme = theme_grey(),
caption = "Fuente propia",
font.caption = c(10, "italic", "black"),
font.x = c(12, "bold.italic", "black"),
font.y = c(12, "bold.italic", "black")
)

```

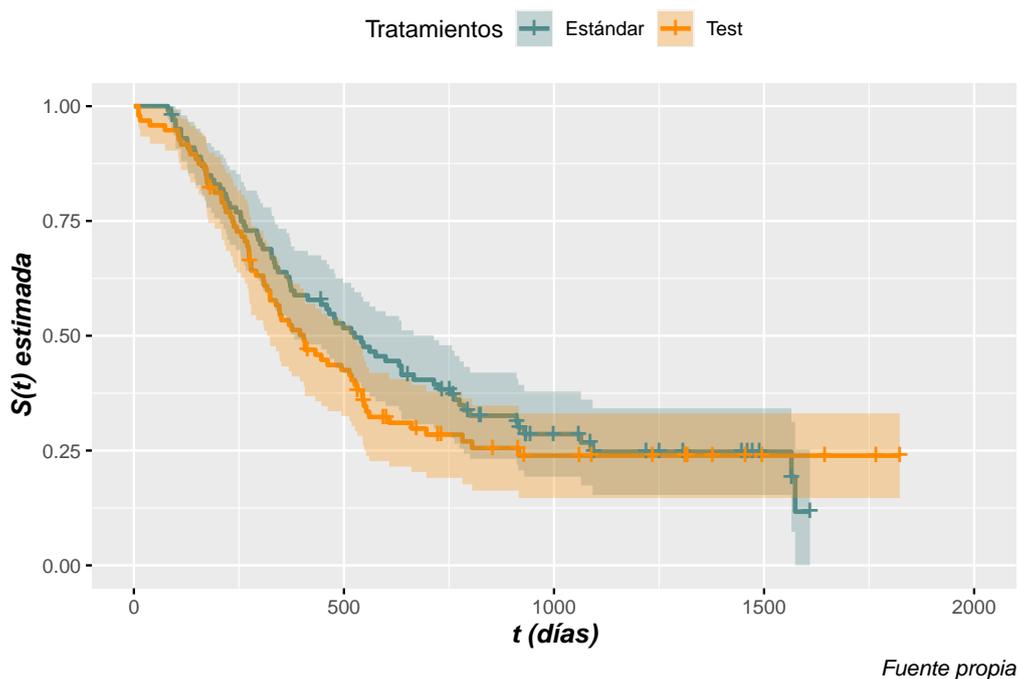


Figura 4.2: Estimador de Fleming-Harrington para cada uno de los tratamientos estudiados en el conjunto de pacientes disponibles en *oropharynx*

4.1.1.3. Estimador de la tasa de fallo acumulado

Un estimador de la función de riesgo acumulado puede obtenerse a partir de un estimador de la función de supervivencia. Se mostrará un ejemplo con el uso del conjunto de datos *rats* del paquete *survival*. Dicho conjunto procede del estudio realizado por [Mantel et al. \(1977\)](#) y en él se recoge información sobre la aplicación de dos tratamientos a 300 ratas. Se partió de 100 camadas de ratas y de manera aleatoria se eligieron 3 ratas de cada una de las camadas, una de las cuales fue tratada con un fármaco. Posteriormente, se observó la incidencia de tumores en cada uno de los individuos de la muestra.

Las variables involucradas en la investigación fueron:

- *litter*: camada a la que pertenece la rata y toma los valores del 1 al 100.

- *rx*: tratamiento administrado (1 = fármaco o 0 = control).
- *time*: tiempo hasta que aparece un tumor o el último seguimiento realizado.
- *status*: estado del evento (1 = tumor o 0 = censura).
- *sex*: sexo del animal (*m* = macho o *f* = hembra).

En la [Tabla 4.8](#) se pueden consultar algunos de los datos del conjunto *rats* con el que se trabajará:

<i>litter</i>	<i>rx</i>	<i>time</i>	<i>status</i>	<i>sex</i>
1	1	101	0	f
2	1	91	0	m
3	1	104	0	f
4	0	104	0	m
5	1	104	0	f
6	1	98	0	m
7	1	77	0	f
8	0	76	0	m
9	0	104	0	f
10	1	104	0	m
11	1	88	1	f

Tabla 4.8: Cabecera de los datos *rats* del paquete *survival*

Una vez expuestos los datos, se procede a la obtención del estimador deseado. En primer lugar, se debe estimar la función de supervivencia asociada a cada una de las poblaciones y luego, basta considerar la suma acumulada del cociente entre las variables *n.event* (individuos que presentaron el evento) y *n.risk* (individuos en riesgo):

```
ajuste.rats <- survfit(Surv(time, status) ~ rx, data = rats)
```

Se adjuntan dos tablas en las que se pueden observar el resultado de la estimación para cada uno de los tratamientos en estudio:

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>hazard</i>
34	99	1	0.990	0.0101
39	98	1	0.980	0.0203
45	97	1	0.970	0.0306
67	89	1	0.959	0.0419
70	86	1	0.948	0.0535
72	85	1	0.937	0.0652

(continúa)

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>hazard</i>
73	84	1	0.925	0.0771
80	78	2	0.902	0.1028
86	72	1	0.889	0.1167
88	67	1	0.876	0.1316
89	64	2	0.848	0.1629
92	54	1	0.833	0.1814
94	50	1	0.816	0.2014
96	47	1	0.799	0.2226
102	43	1	0.780	0.2459
103	41	3	0.723	0.3191
104	38	1	0.704	0.3454

Tabla 4.9: Estimación de la función de riesgo acumulado para el grupo tratado con fármaco

<i>time</i>	<i>n.risk</i>	<i>n.event</i>	<i>survival</i>	<i>hazard</i>
40	198	1	0.995	0.0051
49	196	1	0.990	0.0102
50	195	1	0.985	0.0153
54	191	1	0.980	0.0205
55	188	1	0.974	0.0258
64	184	1	0.969	0.0313
66	182	1	0.964	0.0368
68	181	1	0.958	0.0423
71	176	1	0.953	0.0480
73	173	1	0.948	0.0538
75	168	1	0.942	0.0597
77	164	1	0.936	0.0658
78	158	1	0.930	0.0721
79	156	1	0.924	0.0785
81	149	2	0.912	0.0920
84	142	2	0.899	0.1060
96	111	1	0.891	0.1151
101	98	1	0.882	0.1253
102	96	1	0.873	0.1357

Tabla 4.10: Estimación de la función de riesgo acumulado para el grupo de control

Por último, se añade una gráfica en la que se muestra la estimación. Para ello, basta añadir el argumento $fun = "cumhaz"$ en la función $ggsurvplot()$ con la que se ha trabajado en secciones anteriores:

```
ggsurvplot(
  ajuste.rats,
  fun = "cumhaz",
  xlab = "t",
  ylab = "H(t) estimada",
  lab = c(10, 10, 7),
  size = 1, legend.title = "Tratamientos",
  palette =
    c("burlywood3", "lightslateblue"),
  conf.int = TRUE,
  legend.labs =
    c("Fármaco", "Control"),
  ggtheme = theme_grey(),
  caption = "Fuente propia",
  font.caption = c(10, "italic", "black"),
  font.x = c(12, "bold.italic", "black"),
  font.y = c(12, "bold.italic", "black")
)
```

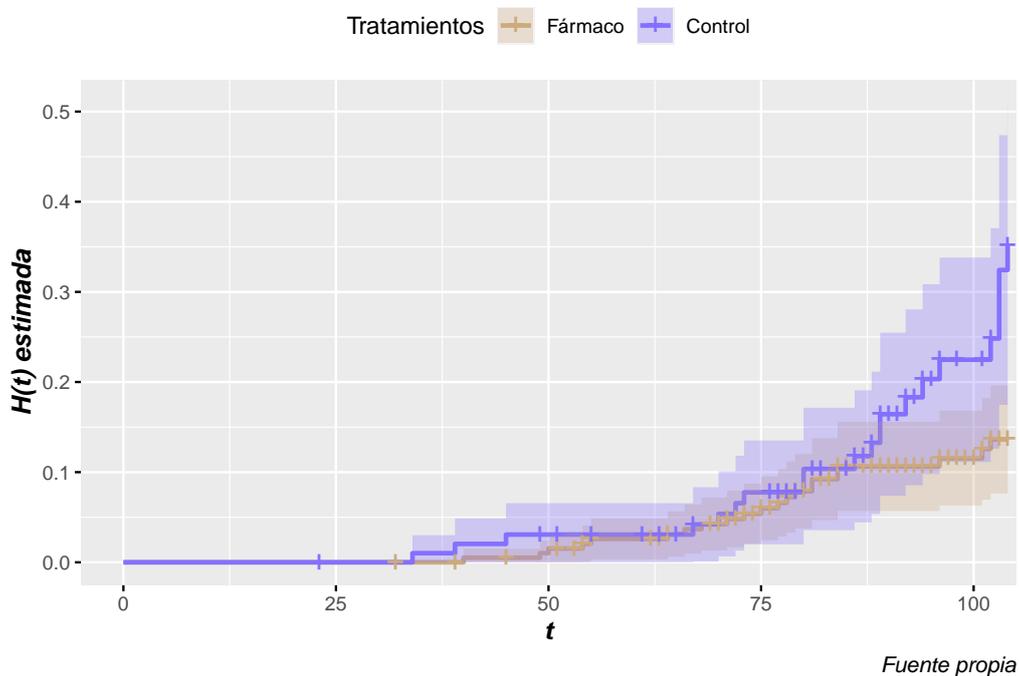


Figura 4.3: Estimador de la tasa de fallo acumulado para cada uno de los tratamientos estudiados en el conjunto de ratas disponibles en *rats* del paquete *survival*

4.1.1.4. Comparación de curvas de supervivencia

Uno de los objetivos del análisis de supervivencia es el estudio de la igualdad de dos o más curvas de supervivencia asociadas a distintas poblaciones en estudio (véase el contraste (3.1)). En el paquete *survival*, se encuentra la función `survdiff(formula, data, subset, na.action, rho)`, la cual es muy útil para dicho objetivo, ya que comprueba si existen diferencias significativas entre dos o más curvas de supervivencia con el uso de la familia G^ρ definida por Fleming y Harrington (1982). Además, la función en cuestión puede comparar una sola curva de supervivencia con una curva alternativa y conocida.

Tal y como se ha especificado, esta función `survdiff()` implementa la familia de test G^ρ de Harrington y Fleming con pesos en cada instante de fallo de $\{S_{KM}(t)\}^\rho$ siendo $S_{KM}(t)$ el [estimador de Kaplan-Meier](#) de la función de supervivencia. Los argumentos de la función en estudio son:

- *formula*: objeto fórmula del tipo *Surv* y de la forma $Surv(time, status) \sim predictor$.
- *data*: objeto *data.frame* en el cual se encuentran los datos.
- *subset*: argumento que indica qué subconjunto de las filas de datos debe utilizarse en el ajuste (vector lógico, vector numérico o vector de caracteres). Si no se indica nada, se incluyen todas las observaciones.
- *na.action*: argumento para filtrar los datos perdidos.
- *rho*: valor numérico que controla el tipo de test a aplicar. Si toma el valor 0 (por defecto) se usará el test *log-rank* y cuando toma el valor 1 se usará la modificación de Peto-Peto de la prueba de Gehan-Breslow, i.e., test *log-rank* ponderado con los pesos de Peto-Peto modificados (véase la [tabla de pesos 3.1](#)).

A continuación, se aplicará la función al conjunto de datos *colon* ([Laurie et al. \(1989\)](#)) del paquete *survival*. Los datos pertenecen a uno de los primeros ensayos que se realizaron sobre la aplicación a pacientes con cáncer de colon de distintos tratamientos. A una parte de la muestra de datos se le aplicó levamisol (compuesto de baja toxicidad que se usaba para tratar infecciones en animales y representado por *Lev*), a otra parte se le aplicó una combinación de levamisol y 5-FU (agente quimioterapéutico moderadamente tóxico y representado por *Lev+5FU*) y a la última parte de la muestra no se le aplicó ningún tratamiento (representado por *Obs*). Además, para cada paciente existen dos registros, uno para la recurrencia y otro para la muerte. Solo se trabajará con los datos de recurrencia y con las variables *rx* (tratamiento), *time* (días hasta que se produce el evento o la censura) y *status* (estado de la censura).

```
colon.rec <- colon[colon$etype == 1,]
```

<i>time</i>	<i>status</i>	<i>rx</i>
968	1	Lev+5FU
3087	0	Lev+5FU

(continúa)

<i>time</i>	<i>status</i>	<i>rx</i>
542	1	Obs
245	1	Lev+5FU
523	1	Obs
904	1	Lev+5FU
229	1	Lev
3192	0	Obs
3173	0	Lev
3308	0	Lev+5FU
2908	0	Lev
3309	0	Lev+5FU
1130	1	Obs
2231	1	Lev

Tabla 4.11: Muestra de datos del conjunto *colon* contenido en el paquete *survival*

El objetivo que se persigue es estudiar el contraste siguiente:

$$\begin{aligned} H_0 : S_{Obs}(t) = S_{Lev}(t) = S_{Lev+5FU}(t) & \quad \text{para todo } t \in T \\ H_1 : S_i(t) \neq S_j(t) & \quad \text{para algún } i \neq j, t \in T \end{aligned} \quad (4.1)$$

donde $i, j = Obs, Lev, Lev + 5FU$.

A continuación, se resuelve dicho contraste en R para $\rho = 0$ (test *log-rank* ponderado con pesos iguales):

```
survdif(Surv(time, status) ~ rx, data = colon.rec, rho = 0)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ rx, data = colon.rec,
##          rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=Obs     315      177      151      4.62      6.82
## rx=Lev     310      172      149      3.69      5.41
## rx=Lev+5FU 304      119      169     14.69     23.04
##
## Chisq= 23.1 on 2 degrees of freedom, p= 1e-05
```

Antes de concluir si el contraste es significativo o no, se detalla el significado de cada una de las componentes de la salida obtenida:

- N : número de individuos u observaciones existentes en cada grupo o población en estudio.

- *Observed*: número ponderado de eventos observados en cada uno de los grupos en estudio.
- *Expected*: número ponderado y esperado de eventos en cada uno de los grupos en estudio.
- $(O - E)^2/E$: representa el cálculo del estadístico χ^2 para cada uno de los grupos en estudio.
- $(O - E)^2/V$: valor numérico que se calcula elevando al cuadrado la diferencia de las componentes *Observed* y *Expected* y dividiendo entre la varianza.
- *chisq*: valor del estadístico χ^2 del test junto con los grados de libertad asociados.
- *var*: matriz de varianza del test.
- *p*: p-valor asociado al contraste.

Se concluye así que el contraste (4.1) realizado tiene asociado un p-valor de 0.00001 , por lo que, existen evidencias significativas en contra de la hipótesis nula o igualdad de las curvas de supervivencia de los tres tratamientos (*Obs*, *Lev* y *Lev+5FU*) en estudio.

En el caso de que se quisiera obtener un gráfico en el que se pudiera consultar la estimación de la curva de supervivencia asociada a cada tratamiento junto con el p-valor vinculado al contraste de hipótesis, se usará la función *ggsurvplot()* del paquete *survminer* utilizada en apartados anteriores. En este caso, a dicha función es necesario añadirle el argumento *pval = TRUE* y *log.rank.weights* para indicarle el tipo de peso que se quiere usar para el test *log-rank* ponderado. Además, con el argumento *risk.table = TRUE* se adjunta una tabla en la que se puede consultar el número de individuos en riesgo en cada uno de los instantes considerados. Para una consulta más detallada sobre los posibles argumentos que admite la función *ggsurvplot()* véase el apartado *ggsurvplot_arguments* del manual de [Biecek et al. \(2021\)](#).

A continuación, se muestra la gráfica asociada al contraste (4.1) en el que se puede visualizar el p-valor obtenido y la tabla de individuos en riesgo:

```
fit <- survfit(Surv(time, status) ~ rx, data = colon.rec,
              type = "kaplan-meier", conf.type = "plain")
ggsurvplot(
  fit,
  data = colon.rec,
  size = 1,
  conf.int = TRUE,
  pval = TRUE,
  legend.labs =
    c("Obs", "Lev", "Lev+5FU"),
  ylim = c(0.1, 1),
  xlim = c(0, 3250),
  lab = c(10, 10, 7),
  xlab = "t (días)",
  ylab = "S(t) estimada",
```

```

log.rank.weights = "1",
risk.table = TRUE,
risk.table.col = "strata",
risk.table.height = 0.32,
risk.table.title = "Número de individuos en riesgo",
ggtheme = theme_grey(),
tables.theme = theme_bw(),
legend.title = "Tratamientos"
)

```

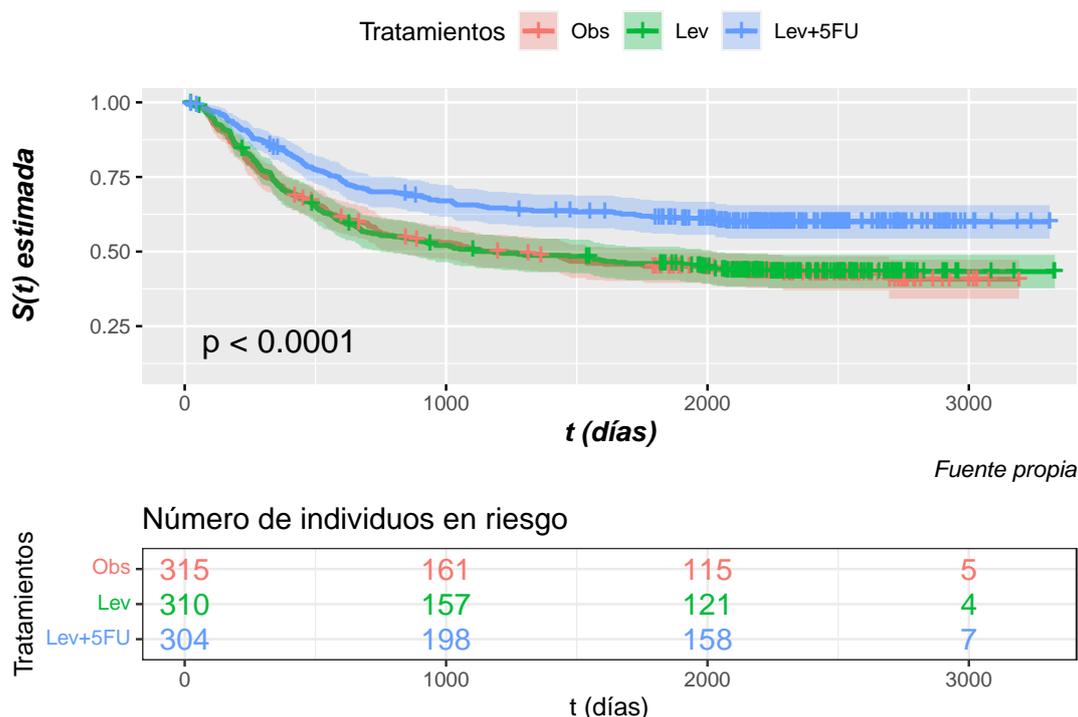


Figura 4.4: Comparación de curvas de supervivencia de manera gráfica con la ayuda del paquete *survminer*

A la vista de la gráfica obtenida, parece que se está rechazando la hipótesis nula debido a que el comportamiento de la curva de supervivencia asociada al tratamiento *Lev+5FU* es bastante dispar al resto. En contraposición, las otras dos curvas obtenidas son prácticamente iguales y casi no se pueden distinguir.

4.1.1.4.1. Comparaciones múltiples

Cuando se están comparando tres o más curvas de supervivencia y se rechaza la hipótesis nula del contraste (3.1) puede ser útil aplicar técnicas de comparaciones múltiples. Estas técnicas ayudan a averiguar la razón por la que se está rechazando la igualdad de las curvas de supervivencia. En el paquete *survminer*, existe la función *pairwise_survdiff()* capaz de realizar el citado análisis de comparaciones múltiples. La estructura de dicha función es *pairwise_survdiff(formula, data, p.adjust.method, na.action, rho)* donde:

- *formula*: objeto *Surv* de la forma $Surv(\text{tiempo}, \text{censura}) \sim \text{predictor}$.
- *data*: objeto *data.frame* en el cual se encuentran los datos.
- *p.adjust.method*: método para ajustar los p-valores. Los posibles valores pueden ser "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr" o "none" (si no se desea ajustar el p-valor).
- *na.action*: argumento para filtrar los datos perdidos.
- *rho*: valor numérico que controla el tipo de test a aplicar. Al igual que para la función *survdif()*, toma el valor 0 (por defecto) si se desea aplicar el test *log-rank* o el valor 1 si se desea utilizar la modificación de Peto-Peto de la prueba de Gehan-Breslow.

Con el código siguiente, se estudiará la razón por la que se ha rechazado la igualdad de curvas de supervivencia asociadas a los tratamientos del conjunto de datos *colon.rec*. Para el ajuste del p-valor, se usará la corrección de Bonferroni (véase [Dunn \(1961\)](#)) que es capaz de controlar el error de tipo I en cada uno de los $q = k \cdot (k - 1)/2$ contrastes a realizar siendo k el número de tratamientos en estudio. En general, cuando se aplica dicha corrección, el punto crítico que se tomará para cada uno de los contrastes es α/q :

```
pairwise_survdif(Surv(time, status) ~ rx, data = colon.rec,
                 p.adjust.method = "bonferroni", rho = 0)
```

```
##
## Pairwise comparisons using Log-Rank test
##
## data: colon.rec and rx
##
##      Obs      Lev
## Lev      1      -
## Lev+5FU 3.8e-05 7.7e-05
##
## P value adjustment method: bonferroni
```

En la salida obtenida, se muestran los tres p-valores asociados a cada uno de los contrastes realizados entre cada par de tratamientos posibles:

- El primer contraste considerado es aquel en el que se comparan las curvas de supervivencia de los tratamientos *Lev* y *Obs*. Este se puede enunciar tal y como sigue:

$$\begin{aligned} H_0 &: S_{Obs}(t) = S_{Lev}(t) \quad \text{para todo } t \in T \\ H_1 &: S_{Obs}(t) \neq S_{Lev}(t) \quad \text{para algún } t \in T \end{aligned} \quad (4.2)$$

En este caso, el p-valor asociado es 1, así que no existen evidencias en contra de la hipótesis nula, por lo que las curvas pueden considerarse iguales.

- El segundo contraste considera la igualdad de curvas de supervivencia asociadas a los tratamientos *Lev+5FU* y *Obs*:

$$\begin{aligned} H_0 &: S_{Obs}(t) = S_{Lev+5FU}(t) \quad \text{para todo } t \in T \\ H_1 &: S_{Obs}(t) \neq S_{Lev+5FU}(t) \quad \text{para algún } t \in T \end{aligned} \quad (4.3)$$

El p-valor asociado a dicho contraste es $3.8e-05$, por lo que existen suficientes evidencias en contra de la hipótesis nula.

- Por último, el tercer contraste a considerar es el que compara las curvas de supervivencia de los tratamientos Lev y $Lev+5FU$:

$$\begin{aligned} H_0 : S_{Lev}(t) &= S_{Lev+5FU}(t) \quad \text{para todo } t \in T \\ H_1 : S_{Lev}(t) &\neq S_{Lev+5FU}(t) \quad \text{para algún } t \in T \end{aligned} \quad (4.4)$$

Como el p-valor asociado a este contraste es $7.7e-05$, no se puede admitir la igualdad de las curvas de supervivencia.

Tras el análisis de los contrastes (4.2), (4.3) y (4.4), se puede concluir que efectivamente tal y como se había indicado anteriormente, se ha rechazado la hipótesis nula del contraste (4.1) porque no es posible considerar la igualdad de la curva de supervivencia asociada al tratamiento $Lev+5FU$ con el resto de curvas de supervivencia asociadas a los otros dos tratamientos en estudio.

4.1.2. Librería *survMisc*

La principal utilidad de la librería *survMisc* (Dardis (2018)) es abordar el problema de comparación de curvas de supervivencia cuando la hipótesis alternativa del contraste asociado contiene una relación creciente, decreciente o monótona, es decir, cuando H_1 es de una de las formas siguientes:

$$\begin{aligned} H_1 : S_1(t) &\geq \dots \geq S_k(t) \quad \text{para algún } t \in T \\ H_1 : S_1(t) &\leq \dots \leq S_k(t) \quad \text{para algún } t \in T \\ H_1 : S_1(t) &< \dots < S_k(t) \quad \text{para algún } t \in T \\ H_1 : S_1(t) &> \dots > S_k(t) \quad \text{para algún } t \in T \end{aligned}$$

y como hipótesis nula $H_0 : S_1(t) = \dots = S_k(t)$ para $k > 2$.

Tal y como se especificó en la Sección 3.1.2, lo más correcto es abordar los contrastes anteriores mediante una [prueba de tendencia](#). A continuación, se procede a cargar la librería con la que se trabajará para resolver los contrastes expuestos:

```
library(survMisc)
```

4.1.2.1. Comparación de curvas de supervivencia

En algunas ocasiones, puede ser de interés plantear contrastes de hipótesis en los que la hipótesis alternativa sea unilateral. Dicho planteamiento también es viable en el estudio de la comparación de curvas de supervivencia y para resolverlo en R, se usará la función `comp(x, p, q, scores)` del paquete *survMisc*. Esta función es capaz de resolver el contraste tanto mediante el [test log-rank ponderado](#) como con la [prueba de tendencia](#) y sus argumentos se definen de la siguiente forma:

- x : función `ten()` aplicada a un objeto de tipo *Surv*.
- p : valor de p del peso de Fleming-Harrington¹⁰ (puede ser ignorado).

¹⁰En este caso, se considera que los pesos de Fleming-Harrington son de la forma $(\hat{S}(t_{i-1}))^p(1 - \hat{S}(t_{i-1}))^q$.

- q : valor de q del peso de Fleming-Harrington (puede ser ignorado).
- $scores$: vector de pesos para la prueba de tendencia. Si no se indica, se toma por defecto el vector $1 : k$ donde k es el número de grupos o poblaciones en estudio.

Como resultado, la función proporciona dos tablas. En la primera de ellas, se pueden consultar diferentes test *log-rank* ponderados y en la segunda, pruebas de tendencia. La información más útil en general es la que aparece en la última columna de cada tabla. En ella se muestra el p-valor asociado a cada uno de los test realizados. Además, en cada una de las filas se indica el tipo de test aplicado.

Una aplicación de la función se realizará con los datos *larynx* del paquete *KMsurv* (Klein y Moeschberger (2012)):

```
data("larynx", package = "KMsurv")
```

Dicho conjunto de datos procede del libro Klein y Moeschberger (1997) y recoge información sobre 90 hombres que padecen cáncer de laringe. La información que contiene se divide en cinco columnas o variables:

- *stage*: estado de la enfermedad (1 = estado 1, 2 = estado 2, 3 = estado 3 o 4 = estado 4).
- *time*: tiempo hasta la muerte o hasta fin del estudio (en meses).
- *age*: edad del paciente en el momento del diagnóstico.
- *diagyr*: año en el que se ha diagnosticado el cáncer.
- *delta*: indicador de muerte, 0 = vivo o 1 = muerto.

En la siguiente tabla, se pueden consultar algunos de los datos del conjunto *larynx* con el que se trabajará:

<i>stage</i>	<i>time</i>	<i>age</i>	<i>diagyr</i>	<i>delta</i>
1	0.6	77	76	1
1	1.3	53	71	1
1	2.4	45	71	1
1	2.5	57	78	0
2	0.2	86	74	1
2	2.2	71	78	0
2	2.6	67	78	0
2	3.6	70	77	1
3	0.5	57	74	1
3	1	49	76	1

(continúa)

<i>stage</i>	<i>time</i>	<i>age</i>	<i>diagyr</i>	<i>delta</i>
3	3.7	52	77	0
3	4.8	54	76	0
4	0.1	65	72	1
4	2.3	62	71	1
4	2.9	74	78	0
4	4.3	48	76	0

Tabla 4.12: Muestra de datos del conjunto *larynx* contenido en el paquete *KMsurv*

Tomando como vector de pesos el que viene por defecto en la orden, $p, q = 1$ y creando un objeto *ten* se resolverá el problema de comparación de curvas de supervivencia asociado al estado de la enfermedad:

$$H_0 : S_1(t) = S_2(t) = S_3(t) = S_4(t) \quad \text{para todo } t \in T$$

$$H_1 : S_1(t) \neq S_2(t) \neq S_3(t) \neq S_4(t) \quad \text{para algún } t \in T$$

```
comp(ten(Surv(time, delta) ~ stage, data = larynx))
```

```

      chiSq df      pChisq
1      22.763  3 4.5252e-05 ***
n      23.177  3 3.7093e-05 ***
sqrtN  23.141  3 3.7746e-05 ***
S1     23.171  3 3.7199e-05 ***
S2     23.170  3 3.7213e-05 ***
FH_p=1_q=1 16.661  3 0.00082967 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
$tft
      Q      Var      Z      pNorm
1    -25.8061  48.1505 -3.7190 0.00020005 ***
n   -1939.0000 210644.6590 -4.2248 2.3919e-05 ***
sqrtN  -221.9185  2990.6169 -4.0580 4.9493e-05 ***
S1     -21.3895   26.8311 -4.1293 3.6380e-05 ***
S2     -21.0942   26.0075 -4.1363 3.5292e-05 ***
FH_p=1_q=1  -2.9412   1.5056 -2.3970 0.01653027 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$scores
[1] 1 2 3 4

```

Figura 4.5: Captura de la salida realizada en Rstudio Cloud¹¹

A continuación, se analizarán cada una de las tablas generadas y se explicará el significado de cada una de las filas y columnas:

¹¹Se ha adjuntado una captura de la salida debido a problemas generados por las versiones de R y del paquete en uso. Con la actualización de R a versiones superiores de la versión 4.0.0, la función *comp()* no funciona correctamente. En concreto, la última columna de los p-valores de cada una de las tablas desaparece y en su lugar aparecen números enteros que no tienen ningún significado.

	<i>chiSq</i>	<i>df</i>	<i>pChisq</i>
1	22.763	3	4.5252e-05 ***
n	23.177	3	3.7093e-05 ***
sqrtN	23.141	3	3.7746e-05 ***
S1	23.171	3	3.7199e-05 ***
S2	23.170	3	3.7213e-05 ***
FH_p=1_q=1	16.661	3	0.00082967 ***

Tabla 4.13: Primera tabla de la salida *comp* en la que se contemplan diferentes test *log-rank* ponderados

Los resultados de los test realizados en esta tabla por filas corresponden a la aplicación del test *log-rank* ponderado con pesos $\omega(t_i)$ distintos. Las explicaciones se basan en la notación tomada en la [Tabla 3.1](#) de la [Sección 3.1.1](#):

- 1: pesos iguales, que se conoce como test *log-rank*.
- n: r_i (número de individuos en riesgo) y se conoce como test de Gehan-Breslow o de Gehan-Wilcoxon.
- sqrtN: $\sqrt{r_i}$ y se conoce como test de Tarone-Ware.
- S1: $\prod_{i=1}^m 1 - \frac{d_i}{r_i + 1}$ y se conoce como test de Peto-Peto.
- S2: $\prod_{i=1}^m \left(1 - \frac{d_i}{r_i + 1}\right) \frac{r_i}{r_i + 1}$, conocido como test de Peto-Peto modificado por Andersen o modificación de Peto-Peto del test de Gehan-Breslow.
- FH: test de Fleming-Harrington con $(\hat{S}(t_{i-1}))^p(1-\hat{S}(t_{i-1}))^q$ siendo $\hat{S}(t_i) = \prod_{i=1}^m 1 - \frac{d_i}{r_i}$, en el que se pueden ir modificando los parámetros p y q según interese.

Además, en la columna *chiSq* se proporciona el valor del estadístico χ^2 al que se aproxima el estadístico (3.3) del test *log-rank* ponderado; en la columna siguiente (*df*) se muestran los $k - 1$ grados de libertad del estadístico χ^2 ; y en la última columna, se muestra el p-valor asociado al contraste. A la vista de los p-valores obtenidos, se puede concluir que en este caso, se rechaza por unanimidad la hipótesis nula de igualdad entre las curvas de supervivencia.

Observación: El hecho de que se generen los p-valores de varios test a la vez, garantiza que la conclusión extraída del contraste no es dudosa. A la hora de contrastar hipótesis siempre que sea posible, es aconsejable realizarlo por diversos métodos para cerciorarse de que las conclusiones alcanzadas son ciertas.

▲

Por otro lado, los resultados obtenidos por las pruebas de tendencia son:

	Q	Var	Z	$pNorm$
1	-25.8061	48.1505	-3.7190	0.00020005 ***
n	-1939.0000	210644.6590	-4.2248	2.3919e-05 ***
sqrtN	-221.9185	2990.6169	-4.0580	4.9493e-05 ***
S1	-21.3895	26.8311	-4.1293	3.6380e-05 ***
S2	-21.0942	26.0075	-4.1363	3.5292e-05 ***
FH_p=1_q=1	-2.9412	1.5056	-2.3970	0.01653027 *

Tabla 4.14: Segunda tabla de la salida *comp* en la que se contemplan diferentes pruebas de tendencia

En cada una de las filas se realiza una prueba de tendencia concreta con distintos pesos. Dichos pesos son los mismos que aparecen en la [Tabla 4.13](#) y que se han especificado antes. Por otro lado, en la columna Z se muestra el estadístico (3.4) de la prueba de tendencia y en la columna $pNorm$ se consulta el p-valor asociado al contraste. Nótese que el contraste que se está realizando es:

$$H_0 : S_1(t) = S_2(t) = S_3(t) = S_4(t) \quad t \leq \tau$$

$$H_1 : S_1(t) \geq S_2(t) \geq S_3(t) \geq S_4(t) \quad t \leq \tau$$

siendo τ la mayor $t \in T$ en la que todos los estados involucrados en el estudio tienen al menos un sujeto en riesgo. Si se observan los p-valores, todos ellos son menores que 0.05, así que claramente se rechaza la hipótesis nula.

Otra forma de obtener el p-valor asociado a la prueba de tendencia es de manera gráfica con la función *ggsurvplot()* de la librería *survminer* con la que ya se ha trabajado. En este caso, para visualizar el p-valor es necesario añadirle a la función los argumentos *pval = TRUE* y *test.for.trend = TRUE*. Además, con el argumento *log.rank.weights* se puede modificar el tipo de peso que se quiere usar en el test y con la orden *arrange_ggsurvplots()* se mostrarán varias gráficas en la misma figura. En las siguientes gráficas se muestran los mismos p-valores que se habían obtenido en la [Tabla 4.14](#):

```
# Ajuste de la función de supervivencia:
ajuste1 <- survfit(Surv(time, delta) ~ stage, data = larynx,
                  type = "kaplan-meier")
splots <- list()
# Prueba de tendencia con pesos iguales:
splots[[1]] <- ggsurvplot(
  ajuste1,
  data = larynx,
  pval = TRUE,
  test.for.trend = TRUE,
  pval.method = TRUE,
  log.rank.weights = "1"
)
```

```
# Test de Gehan-Breslow:
splots[[2]] <- ggsurvplot(
  ajustel,
  data = larynx,
  pval = TRUE,
  test.for.trend = TRUE,
  pval.method = TRUE,
  log.rank.weights = "n"
)
# Test de Tarone-Ware:
splots[[3]] <- ggsurvplot(
  ajustel,
  data = larynx,
  pval = TRUE,
  test.for.trend = TRUE,
  pval.method = TRUE,
  log.rank.weights = "sqrtN"
)
# Test de Peto-Peto:
splots[[4]] <- ggsurvplot(
  ajustel,
  data = larynx,
  pval = TRUE,
  test.for.trend = TRUE,
  pval.method = TRUE,
  log.rank.weights = "S1"
)
# Test de Peto-Peto modificado:
splots[[5]] <- ggsurvplot(
  ajustel,
  data = larynx,
  pval = TRUE,
  test.for.trend = TRUE,
  pval.method = TRUE,
  log.rank.weights = "S2"
)
# Test de Fleming-Harrington con  $p = 1$  y  $q = 1$ :
splots[[6]] <- ggsurvplot(
  ajustel,
  data = larynx,
  pval = TRUE,
  test.for.trend = TRUE,
  pval.method = TRUE,
  log.rank.weights = "FH_p=1_q=1"
)
arrange_ggsurvplots(splots, print = TRUE,
  ncol = 2, nrow = 3)
```

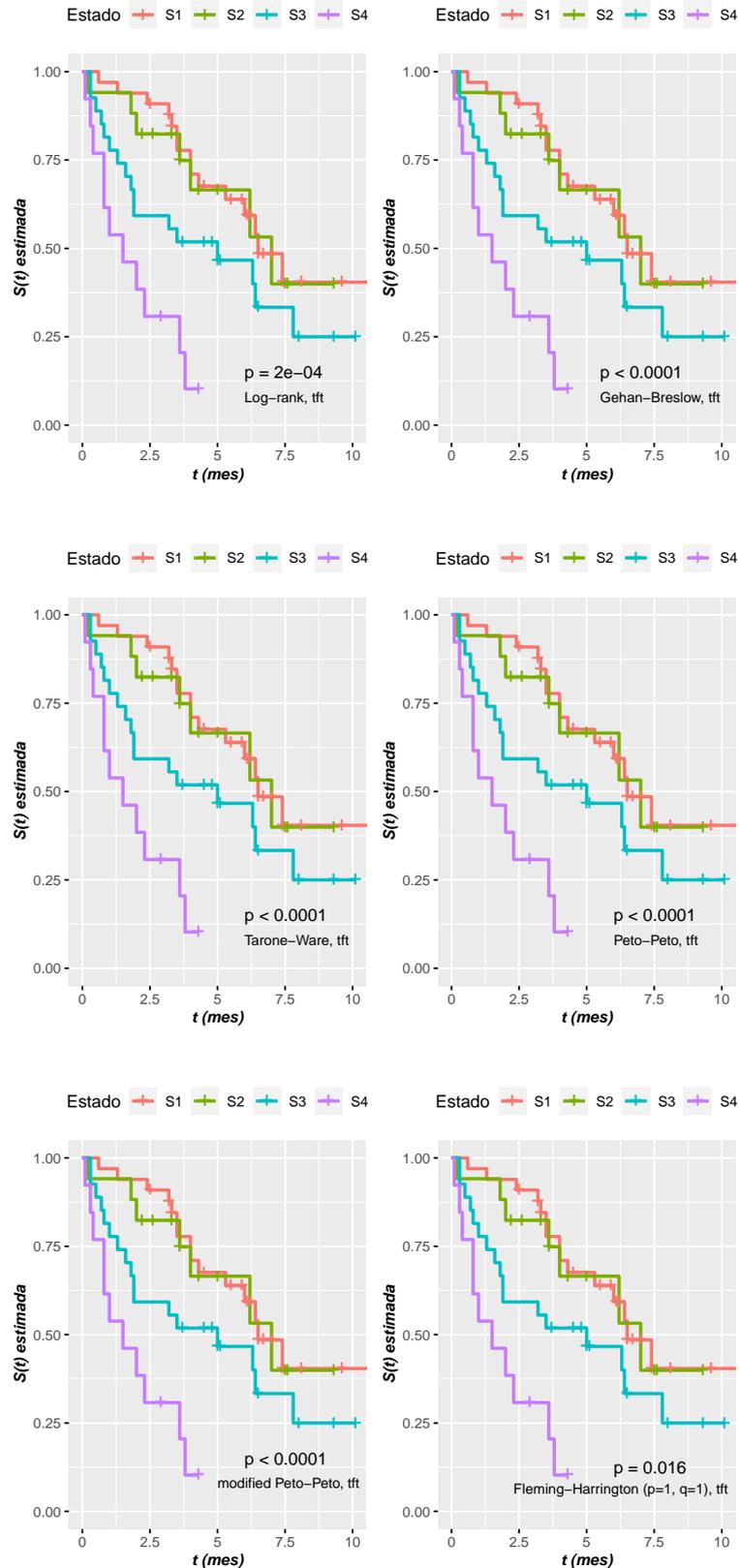


Figura 4.6: Obtención de los p-valores asociados a distintos pesos en pruebas de tendencia de manera gráfica con la librería *survminer*

4.1.3. Librería *KMsurv*

La librería *KMsurv* (Klein y Moeschberger (2012)) se caracteriza por poseer una gran selección de conjuntos de datos procedentes de estudios realizados y ejercicios propuestos en Klein y Moeschberger (1997). Además, incluye una función que es capaz de generar una tabla de vida o *estimador actuarial* a partir de un par de variables, el número de censuras y el número de eventos que se han producido en cada instante de tiempo.

```
library(KMsurv)
```

4.1.3.1. Estimador actuarial

Tal y como se especificó en la [Sección 2.1.2](#) del [Capítulo 2](#), una forma de calcular un estimador de la función de supervivencia cuando los tiempos vienen agrupados en intervalos es mediante el estimador actuarial o tabla de vida. En el paquete en estudio, existe la función *lifetab(tis, ninit, nlost, nevent)* capaz de calcular dicho estimador. A continuación, se especifica el significado de cada uno de los argumentos de la función anterior:

- *tis*: vector que contiene los instantes finales de los intervalos de tiempo involucrados en el estudio. Dicho vector debe contener un elemento más de los que poseen los vectores *nlost* y *nevent*.
- *ninit*: número de individuos que intervienen en el inicio del estudio.
- *nlost*: vector que contiene el número de individuos que abandonan el estudio o cuyos datos se pierden a lo largo del estudio.
- *nevent*: vector que posee el número de individuos que experimentaron el evento en estudio.

Además, como salida se obtiene una tabla con diez columnas en las que se puede consultar la estimación de la función de densidad, de la función de supervivencia o de la función de riesgo, entre otra información disponible. El significado de cada una de las columnas se puede consultar en la [página 9](#) del [Capítulo 2](#).

Se adjunta un ejemplo de aplicación de la función en cuestión. Para ello, se usarán los datos *nurshome* que se pueden consultar en el siguiente enlace <https://www2.karlin.mff.cuni.cz/~pesta/prednasky/NMFM404/Data/NursingHome.dat>. *The National Center for Health Services Research* estudió 36 residencias de ancianos con el objetivo de evaluar los efectos que tenían ciertos incentivos económicos en la duración de la estancia de cada uno de los residentes. Las residencias “incentivadas” recibían más dinero por los pacientes de *Medicaid* y bonificaciones por mejorar la salud de un paciente y por darles el alta. En el estudio participaron 1601 pacientes ingresados en dichas residencias entre el 1 de mayo de 1981 y el 30 de abril de 1982. Las variables que se consideraron para el estudio fueron:

- *los*: duración de la estancia en días.
- *age*: edad del residente.

- *rx*: indica el tipo de residencia. Si está incentivada toma el valor 1 y en caso contrario toma el valor 0.
- *gender*: especifica el género del residente. El valor 1 indica que el residente es hombre y el 0 que es mujer.
- *married*: indicador de si el residente estaba casado o no. Toma el valor 1 si está casado y el 0 en caso contrario.
- *health*: indicador del estado de salud del residente. Se indica a través de una escala numérica del 2 al 5 donde el 2 indica que el estado de salud es el mejor y 5 indica el peor estado de salud.
- *sensor*: indicador de la censura (1 = censurado o 0 = no censurado).

Se procede a la lectura de los datos y a la muestra de algunos de ellos:

```
nurshome <- read.table("http://www.karlin.mff.cuni.cz/~pesta/prednasky/
NFMF404/Data/NursingHome.dat",
header = FALSE, skip = 14,
col.names = c("los", "age", "rx",
"gender", "married",
"health", "sensor"))
```

<i>los</i>	<i>age</i>	<i>rx</i>	<i>gender</i>	<i>married</i>	<i>health</i>	<i>sensor</i>
864	94	1	1	0	2	1
18	86	1	1	0	4	0
1084	75	1	0	0	3	1
1092	77	1	0	1	2	1
23	86	1	0	0	4	0
971	82	1	1	1	4	1
176	93	0	0	0	5	0
396	86	0	0	1	3	1
533	89	0	0	0	5	1
25	65	0	0	0	4	0
2	84	0	1	0	5	0
38	95	0	1	0	3	0

Tabla 4.15: Muestra de datos del conjunto *Nurshome* localizado en <https://www2.karlin.mff.cuni.cz/~pesta/prednasky/NFMF404/Data/NursingHome.dat>

Para calcular un estimador actuarial, los datos se restringirán a aquellos en los que la variable *rx* toma el valor 1, es decir, aquellas residencias de ancianos que han sido incentivadas:

```
nurshome_inc <- nurshome[nurshome$rx == 1, ]
```

Además, como ya se ha indicado anteriormente, se necesita que los datos estén agrupados por intervalos, así que lo que se hará será agrupar la duración de la estancia de cada uno de los residentes en intervalos de 50 días. Para ello, se hará uso de la función *floor()* y se creará una nueva variable *int* que contendrá el extremo izquierdo del intervalo al que pertenece cada uno de los residentes o individuos de la muestra:

```
nurshome_inc$int <- floor(nurshome_inc$los/50)*50
```

A partir de esos extremos, se deben calcular los argumentos necesarios para introducir en la función *lifetab()*:

- Cálculo del vector *tis*, el cual debe contener el extremo derecho de cada uno de los intervalos considerados. Como el tiempo máximo que ha estado un paciente ingresado ha sido 1092, basta considerar el vector que comienza en el 0 y acaba en 1050 siendo los saltos de amplitud 50. El último intervalo a considerar será $(1050, +\infty)$:

```
extr_derecho <- c(seq(0, 1050, by = 50), Inf)
```

- Cálculo de *ninit*:

```
nTotal <- nrow(nurshome_inc)
```

- Cálculo de los vectores *nlost* y *nevent*. Se necesita una tabla de frecuencias en la que se contabilice para cada intervalo de tiempo considerado, el número de individuos censurados y completos. Para ello, se hará uso de la función *table()*:

```
tabla <- table(nurshome_inc$int, nurshome_inc$censor)
```

Una vez calculada la tabla, el vector que contabiliza el número de individuos censurados (*censura*) y no censurados (*evento*) en cada uno de los intervalos será:

```
censura <- tabla[, 2]
evento <- tabla[, 1]
```

Una vez que se tienen los datos preparados, se procede a calcular la tabla de vida o estimador actuarial:

```
tabla_vida <- lifetab(tis = extr_derecho, ninit = nTotal,
                    nlost = censura, nevent = evento)
```

En la siguiente página se puede consultar el resultado obtenido del estimador calculado:

Tabla 4.16: Estimador actuarial del conjunto de datos *Nurshome*

	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-50	712	0	712.0	241	1.00000	0.00677	0.00815	0.00000	0.00035	0.00051
50-100	471	0	471.0	89	0.66152	0.00250	0.00417	0.01773	0.00025	0.00044
100-150	382	0	382.0	48	0.53652	0.00135	0.00268	0.01869	0.00019	0.00039
150-200	334	0	334.0	38	0.46910	0.00107	0.00241	0.01870	0.00017	0.00039
200-250	296	0	296.0	35	0.41573	0.00098	0.00251	0.01847	0.00016	0.00042
250-300	261	0	261.0	30	0.36657	0.00084	0.00244	0.01806	0.00015	0.00044
300-350	231	0	231.0	15	0.32444	0.00042	0.00134	0.01755	0.00011	0.00035
350-400	216	0	216.0	23	0.30337	0.00065	0.00225	0.01723	0.00013	0.00047
400-450	193	1	192.5	17	0.27107	0.00048	0.00185	0.01666	0.00011	0.00045
450-500	175	0	175.0	15	0.24713	0.00042	0.00179	0.01617	0.00011	0.00046
500-550	160	0	160.0	5	0.22595	0.00014	0.00063	0.01568	0.00006	0.00028
550-600	155	0	155.0	8	0.21889	0.00023	0.00106	0.01550	0.00008	0.00037
600-650	147	0	147.0	3	0.20759	0.00008	0.00041	0.01521	0.00005	0.00024
650-700	144	0	144.0	10	0.20335	0.00028	0.00144	0.01509	0.00009	0.00045
700-750	134	10	129.0	7	0.18923	0.00021	0.00112	0.01469	0.00008	0.00042
750-800	117	20	107.0	3	0.17896	0.00010	0.00057	0.01440	0.00006	0.00033
800-850	94	13	87.5	1	0.17394	0.00004	0.00023	0.01428	0.00004	0.00023
850-900	80	16	72.0	3	0.17196	0.00014	0.00085	0.01426	0.00008	0.00049
900-950	61	15	53.5	4	0.16479	0.00025	0.00155	0.01425	0.00012	0.00078
950-1000	42	15	34.5	0	0.15247	0.00000	0.00000	0.01446	NaN	NaN
1000-1050	27	14	20.0	0	0.15247	0.00000	0.00000	0.01446	NaN	NaN
1050-Inf	13	13	6.5	0	0.15247	NA	NA	0.01446	NA	NA

A partir de la [Tabla 4.16](#), se puede graficar la estimación de la función de supervivencia asociada a aquellas residencias incentivadas:

```
plot(tabla_vida$surv, type = "s", xlab = "t (días)",
     ylab = "S(t) estimada", col = "darkslateblue")
```

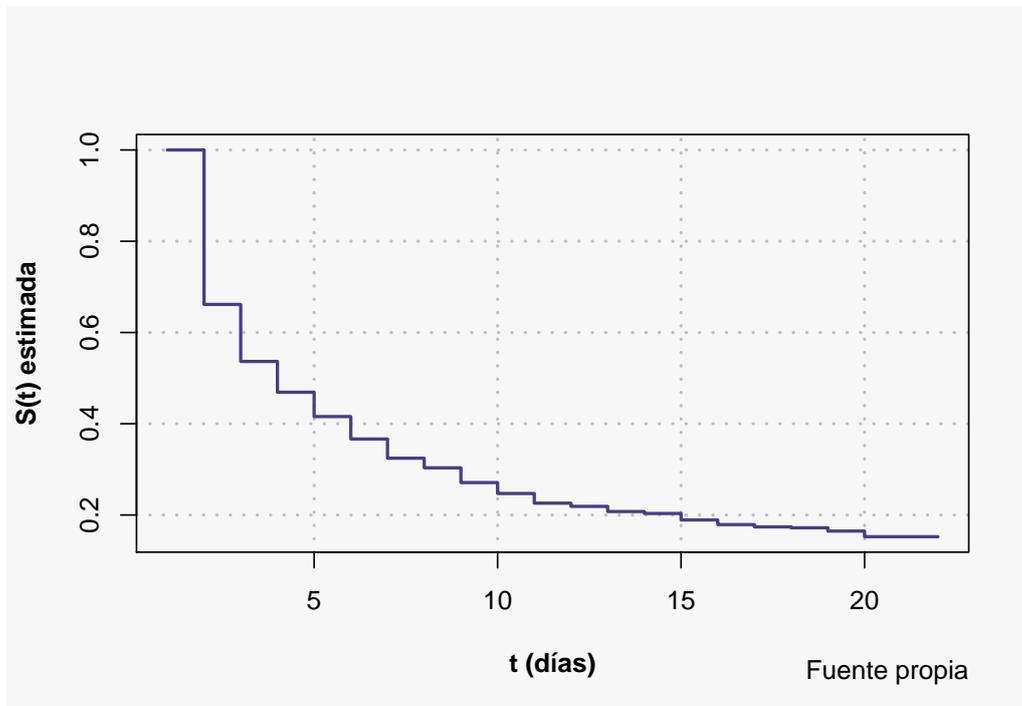


Figura 4.7: Estimador actuarial para el grupo de residencias de ancianos incentivadas del conjunto de datos *Nurshome*

Cabe destacar también, que en el paquete *biostat3* ([Tillander et al. \(2020\)](#)) existe otra función capaz de calcular la tabla de vida. Esta función se llama *lifetab2* (*formula*, *data*, *breaks*) y la principal diferencia que presenta con respecto a la función *lifetab()* del paquete *KMsurv* es que no es necesario preparar los datos en intervalos. Basta introducir como primer argumento un objeto *Surv* como *formula*; como segundo argumento, el *data.frame* de donde proceden los datos; y como último argumento, un vector con los extremos izquierdos de los intervalos a considerar. Con el siguiente código se obtiene la misma tabla de vida que la [Tabla 4.16](#). Nótese que se ha tenido que recodificar la variable *sensor* debido a que la función *Surv* considera como dato censurado aquellos que toman el 0 en la variable *sensor* en vez del valor 1 como considera el conjunto de datos *Nurshome*.

```
library(biostat3)
nurshome_inc$sensor <- car::Recode(nurshome_inc$sensor, "1='0'; 0='1'")
tabla_vida1 <- lifetab2(Surv(los, sensor) ~ 1, data = nurshome_inc,
                       breaks = seq(0, 1050, by = 50))
```

En este caso, para mostrar la gráfica del estimador de la función de supervivencia, basta realizar lo siguiente:

```
plot(tabla_vida1, type = "s")
```

4.1.4. Librería *gte*

La librería *gte* (Dehghan et al. (2015)) fue creada para obtener el estimador de Turnbull propuesto por Dehghan y Duchesne (2011) con una función capaz de realizar el cálculo de dicho estimador denominada *gte*. A continuación, se procede a cargar la librería:

```
library(gte)
```

4.1.4.1. Estimador de Turnbull

Tal y como se ha especificado, la función que calcula el estimador de Turnbull de la función de supervivencia es $gte(formula, data, z, itermax, tole)$. Los argumentos de dicha función son:

- *formula*: objeto de la forma $y \sim x$ donde y será un objeto de tipo *Surv* que actúa como variable respuesta y x será el nombre de las covariables.
- *data*: objeto *data.frame* que contiene el conjunto de datos en estudio.
- *z*: debe ser una matriz en la que cada fila contenga los valores de un vector de covariables con los que se calculan la función de supervivencia condicionada. En el caso de que exista una sola covariable, z será un vector.
- *itermax*: indica el número máximo de iteraciones que debe realizar el algoritmo (por defecto es 100000).
- *tole*: distancia máxima permitida entre iteraciones sucesivas antes de considerar la convergencia del método. Por defecto toma el valor 0.0005.

El estimador en cuestión es una generalización del estimador de Turnbull y consiste en una estimación no paramétrica de la función de supervivencia de la variable aleatoria T condicionada a un valor z_0 del vector de covariables Z , es decir, se pretende construir un estimador no paramétrico de $S(t | z_0) = P[T > t | Z = z_0]$.

Para ver el funcionamiento de la función, se aplicará al conjunto de datos *simul* del paquete en estudio. Dicho conjunto de datos contiene 100 observaciones censuradas por intervalos y tres variables:

- L : variable que contiene los extremos izquierdos de cada uno de los intervalos de censura.
- R : variable que posee cada uno de los extremos derechos de los intervalos de censura.

- Z : covariable continua.

Nótese que cuando $R = NA$, la observación en cuestión está censurada a la derecha y cuando $L = NA$, la observación está censurada por la izquierda. Además, si se cumple que $L = R$, entonces el instante en el que se ha producido el evento se conoce con exactitud.

A continuación, se leerá el conjunto de datos y se adjuntará una muestra de ellos:

```
data(simul)
```

L	R	Z
41.8099110	50.510700	18.122569
NA	36.926040	13.889860
2.8240208	37.949810	13.859484
NA	10.669690	5.353621
15.8330992	21.621300	17.423702
8.0705407	19.397390	5.345955
NA	22.475340	13.843842
46.2031487	NA	21.199203
51.6190973	NA	23.548849
12.7220895	28.013940	21.302348
17.4136333	25.578350	14.021004

Tabla 4.17: Muestra de datos del conjunto *simul* del paquete *gte*

Una vez mostrados los datos, se procede a generar el código necesario para obtener el estimador de Turnbull de los datos. En el objeto *Surv* que se debe introducir como primer argumento en la función *gte()* hay que especificar que el *type* es *interval2* debido a que los datos están censurados por intervalos y son indicados por el valor *NA*:

```
turnbull <- gte(Surv(L, R, type = "interval2") ~ Z,
               data = simul, z = c(15, 30, 45))
```

Como resultado se obtiene una tabla en la que se pueden consultar cada una de las estimaciones realizadas. Se adjunta una tabla del resultado obtenido:

	$survival/Z=z[1]$	$survival/Z=z[2]$	$survival/Z=z[3]$
[8.6042993, 8.6042993]	8.260578e-01	1.0000000	1.0000000
(10.0582938, 10.26988]	8.260488e-01	0.9999999	1.0000000
(10.5500817, 10.66969]	8.260488e-01	0.9999998	1.0000000
[13.5215962, 13.5215962]	5.895147e-01	0.9999678	1.0000000
(14.5131108, 14.53361]	5.895147e-01	0.9999367	1.0000000

(continúa)

	$survival/Z=z[1]$	$survival/Z=z[2]$	$survival/Z=z[3]$
(15.70474, 15.80604]	5.895147e-01	0.9999057	1.0000000
(16.1855673, 16.19669]	5.895061e-01	0.9999052	0.9999988
(16.6328335, 16.71368]	5.894748e-01	0.9999043	0.9999976
(17.6463082, 17.88326]	5.236541e-01	0.9999032	0.9999963
(18.4228483, 18.89019]	2.345770e-01	0.9999021	0.9999951
(21.1515274, 21.6213]	2.338500e-01	0.9998744	0.9999872
(22.3394385, 22.35767]	2.338496e-01	0.9854224	0.9999694
[22.5573127, 22.5573127]	9.464466e-02	0.9709315	0.9999517
(24.088661, 24.28747]	9.464466e-02	0.9593809	0.9999340
(24.4065347, 24.41195]	9.464466e-02	0.9475709	0.9999163
(25.870547, 26.93513]	9.445622e-02	0.9475709	0.9999163
(27.6882738, 27.87079]	5.469136e-02	0.9475709	0.9999163
(29.0303483, 29.33568]	5.422578e-02	0.4823702	0.6759858
(31.617589, 31.63937]	3.188399e-02	0.4823583	0.6669554
(33.8359598, 34.51381]	3.000955e-02	0.4823260	0.6579228
(34.7780274, 34.85963]	2.932396e-02	0.4784571	0.6481142
(34.9496932, 35.208]	2.864706e-02	0.2954358	0.4592661
(41.8460733, 42.05673]	2.675614e-03	0.2943977	0.4573820
(46.2031487, 47.13107]	9.146187e-05	0.2942248	0.4555470
(51.6190973, 52.4681]	0.000000e+00	0.0000000	0.0000000

Tabla 4.18: Estimador de Turnbull del conjunto *simul* del paquete *gte*

Una ventaja que presenta la función en estudio es la capacidad que tiene para calcular a la vez diferentes estimaciones condicionadas a diferentes valores de la covariable Z . Por ejemplo, en el caso expuesto se han calculado las siguientes estimaciones $P[T > t | Z = 15]$, $P[T > t | Z = 30]$ y $P[T > t | Z = 45]$ con una sola orden.

Se adjunta también una gráfica en la que se pueden observar cada una de las curvas de supervivencia estimadas:

```
plot(turnbull, xleg = "topright",
      shade = FALSE, xlab = "t",
      ylab = "S(t) estimada")
```

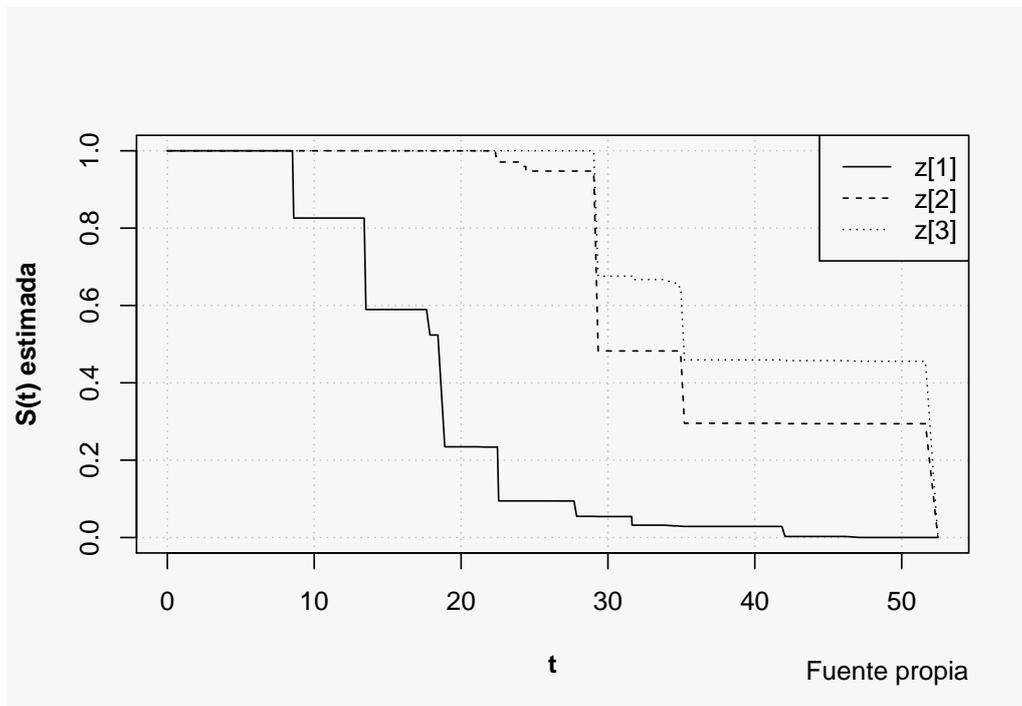


Figura 4.8: Estimador de Turnbull para los valores 15, 30 y 45 de la covariable Z del conjunto de datos *simul* del paquete *gte*

4.1.5. Librería *coin*

La librería *coin* (Hothorn et al. (2021)) implementa un marco general para los test de permutación y proporciona un amplio conjunto de funciones entre las que se encuentran pruebas de $k \geq 2$ muestras con la presencia de datos censurados. Además, las aproximaciones a la distribución nula exacta del contraste en estudio se pueden obtener de manera asintótica o mediante un remuestreo de Montecarlo.

```
library(coin)
```

4.1.5.1. Comparación de curvas de supervivencia

El test *log-rank ponderado* para $k \geq 2$ poblaciones o grupos está disponible en este paquete en la función `logrank_test(formula, data, type, rho, gamma, distribution, alternative)` cuyos argumentos son:

- *formula*: objeto de la forma $y \sim x$ donde y será un objeto de tipo *Surv* y x el nombre de las covariables en estudio. Es imprescindible que x sea de tipo factor.
- *data*: *data.frame* donde se encuentran los datos en estudio.
- *type*: indicador del tipo de test a realizar. Por defecto toma el valor "*logrank*" (pesos iguales) pero también puede aplicar los test de "*Gehan-Breslow*", "*Tarone-Ware*", "*Peto-Peto*" o "*Fleming-Harrington*".

- *rho*: dígito que indica el valor de la constante ρ del peso de Fleming-Harrington.
- *gamma*: dígito que indica el valor de la constante λ del peso de Fleming-Harrington.
- *distribution*: indicador del tipo de método usado para obtener el p-valor del test. Por defecto, el método que utiliza es el asintótico ("*asymptotic*") pero puede tomar el valor "*approximate*" (remuestreo por el método de Montecarlo) o "*exact*" (exacto).
- *alternative*: en el caso de que $k = 2$, se puede trabajar con desigualdades en las hipótesis. En esos casos, este argumento puede tomar el valor "*less*" o "*greater*".

Cuando se trabaja con $k = 2$, el contraste que se realiza por defecto es:

$$\begin{aligned} H_0 &: S_1(t) = S_2(t) \quad \text{para todo } t \in T \\ H_1 &: S_1(t) \neq S_2(t) \quad \text{para algún } t \in T \end{aligned}$$

o equivalentemente

$$\begin{aligned} H_0 &: \theta = 1 \\ H_1 &: \theta \neq 1 \end{aligned}$$

donde $\theta = \lambda_2/\lambda_1$ con λ_2 la tasa de riesgo asociada a la población o grupo 2 y λ_1 la tasa de riesgo asociada a la población o grupo 1. Sin embargo, si se indica que *alternative* = "*less*" el test que realizará será:

$$\begin{aligned} H_0 &: S_1(t) \geq S_2(t) \quad \text{para todo } t \in T \\ H_1 &: S_1(t) < S_2(t) \quad \text{para algún } t \in T \end{aligned}$$

o de forma equivalente

$$\begin{aligned} H_0 &: \theta \geq 1 \\ H_1 &: \theta < 1 \end{aligned}$$

y si *alternative* = "*greater*" entonces:

$$\begin{aligned} H_0 &: S_1(t) \leq S_2(t) \quad \text{para todo } t \in T \\ H_1 &: S_1(t) > S_2(t) \quad \text{para algún } t \in T \end{aligned}$$

o equivalentemente

$$\begin{aligned} H_0 &: \theta \leq 1 \\ H_1 &: \theta > 1 \end{aligned}$$

Como resultado, siempre se obtiene el valor del estadístico asociado al contraste y el p-valor correspondiente. Para mostrar el funcionamiento de la función, se usará el conjunto de datos *lung* (Loprinzi et al. (1994)) del paquete *survival* que contiene la supervivencia de 228 pacientes con cáncer de pulmón avanzado del *North Central Cancer Treatment Group*. Las variables que son más interesantes para el estudio a realizar son las siguientes:

- *time*: tiempo de supervivencia expresado en días.
- *status*: 1 = censurado o 2 = muerte del paciente.
- *sex*: sexo del enfermo (1 = hombre o 2 = mujer).

- *ph.ecog*: puntuación del ECOG¹² según el médico. Los valores posibles son 0 = asintomático, 1 = sintomático y capaz de realizar actividades cotidianas, 2 = en cama menos del 50% del día, 3 = en cama más del 50% del día o 4 = en cama.

Se adjunta una tabla con una muestra de los datos *lung*:

<i>time</i>	<i>status</i>	<i>sex</i>	<i>ph.ecog</i>
306	2	1	2
1010	1	1	1
361	2	2	3
81	2	2	1
965	1	2	2
196	1	1	2
583	2	1	2
252	1	2	1
1022	1	1	2
12	2	1	3
118	2	1	3
285	2	2	3

Tabla 4.19: Muestra de datos del conjunto *lung* del paquete *survival*

Se estudiará el contraste de igualdad de curvas de supervivencia de cada una de las puntuaciones del ECOG usando el test *log-rank* con pesos iguales:

```
logrank_test(Surv(time, status) ~ ph.ecog, data = lung, type = "logrank")
```

```
##
## Asymptotic K-Sample Logrank Test
##
## data: Surv(time, status) by ph.ecog (0, 1, 2, 3)
## chi-squared = 16.525, df = 3, p-value = 0.0008847
```

Se obtiene así que el p-valor asociado al contraste es significativo (0.0008847), así que no se puede admitir la igualdad de curvas de supervivencia. De igual forma, se estudiará el mismo contraste pero con los pesos de “*Gehan-Breslow*” y la obtención del p-valor a través del remuestreo de Montecarlo:

```
logrank_test(Surv(time, status) ~ ph.ecog, data = lung,
              type = "Gehan-Breslow", distribution = "approximate")
```

¹²ECOG alude a la escala diseñada por el *Eastern Cooperative Oncology Group* cuya función es valorar la calidad de vida del paciente. Se puntúa de 0 a 5 y es muy importante a la hora de aplicar un tratamiento al enfermo.

```
##
## Approximative K-Sample Gehan-Breslow Test
##
## data: Surv(time, status) by ph.ecog (0, 1, 2, 3)
## chi-squared = 20.086, p-value < 1e-04
```

De igual modo, se rechaza la hipótesis nula $H_0 : S_0(t) = S_1(t) = S_2(t) = S_3(t)$. Véase gráficamente que efectivamente no se puede aceptar la igualdad de las curvas de supervivencia asociadas a cada una de las puntuaciones del ECOG:

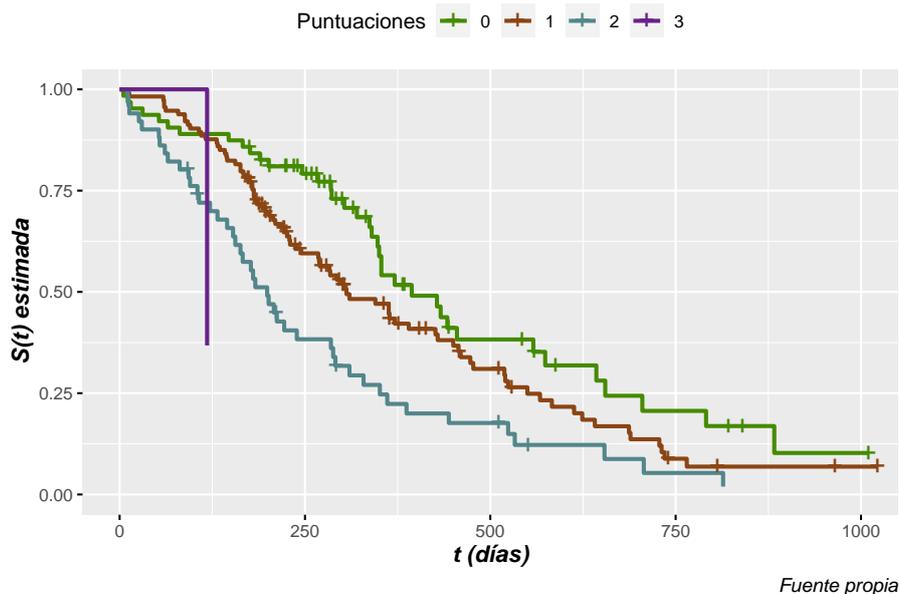


Figura 4.9: Curvas de supervivencia asociadas a cada una de las puntuaciones del ECOG del conjunto de datos *lung* del paquete *survival*

Por otro lado, se estudiarán contrastes de igualdad de curvas de supervivencia asociadas a la población de hombres y de mujeres. Dicho contraste se resolverá usando el test *log-rank* ponderado con pesos de Fleming-Harrington siendo $\rho = 1$ y $\lambda = 1$:

```
logrank_test(Surv(time, status) ~ sex, data = lung,
              type = "Fleming-Harrington", rho = 1,
              gamma = 1)
```

```
##
## Asymptotic Two-Sample Fleming-Harrington Test
##
## data: Surv(time, status) by sex (1, 2)
## Z = -2.8339, p-value = 0.004598
## alternative hypothesis: true theta is not equal to 1
```

El p-valor obtenido es 0.004598 por lo que existen evidencias en contra de la hipótesis nula. Por tanto, no se puede considerar la igualdad de curvas de supervivencia asociadas

a los pacientes masculinos y femeninos. A continuación, se estudiará la razón por la que son distintas dichas curvas. Se comenzará probando con la hipótesis nula

$$H_0 : S_{Masc.}(t) \geq S_{Fem.}(t) \text{ para todo } t \in T,$$

es decir, añadiendo el argumento *alternative* = "less":

```
logrank_test(Surv(time, status) ~ sex, data = lung,
              type = "Fleming-Harrington", rho = 1,
              gamma = 1, alternative = "less")
```

```
##
## Asymptotic Two-Sample Fleming-Harrington Test
##
## data: Surv(time, status) by sex (1, 2)
## Z = -2.8339, p-value = 0.002299
## alternative hypothesis: true theta is less than 1
```

En este caso, el p-valor sigue siendo menor que el nivel de significación considerado del 5%, así que no se puede admitir dicha H_0 . Por tanto, es lógico pensar que la curva de supervivencia asociada a los hombres es menor que la de las mujeres:

```
logrank_test(Surv(time, status) ~ sex, data = lung,
              type = "Fleming-Harrington", rho = 1,
              gamma = 1, alternative = "greater")
```

```
##
## Asymptotic Two-Sample Fleming-Harrington Test
##
## data: Surv(time, status) by sex (1, 2)
## Z = -2.8339, p-value = 0.9977
## alternative hypothesis: true theta is greater than 1
```

Tal y como se había indicado, en este caso se acepta la hipótesis nula

$$H_0 : S_{Masc.}(t) \leq S_{Fem.}(t) \text{ para todo } t \in T$$

pues el p-valor es muy cercano a 1. A continuación, se adjunta una gráfica de las estimaciones de las correspondientes funciones de supervivencia de hombres y mujeres pacientes de cáncer de pulmón:

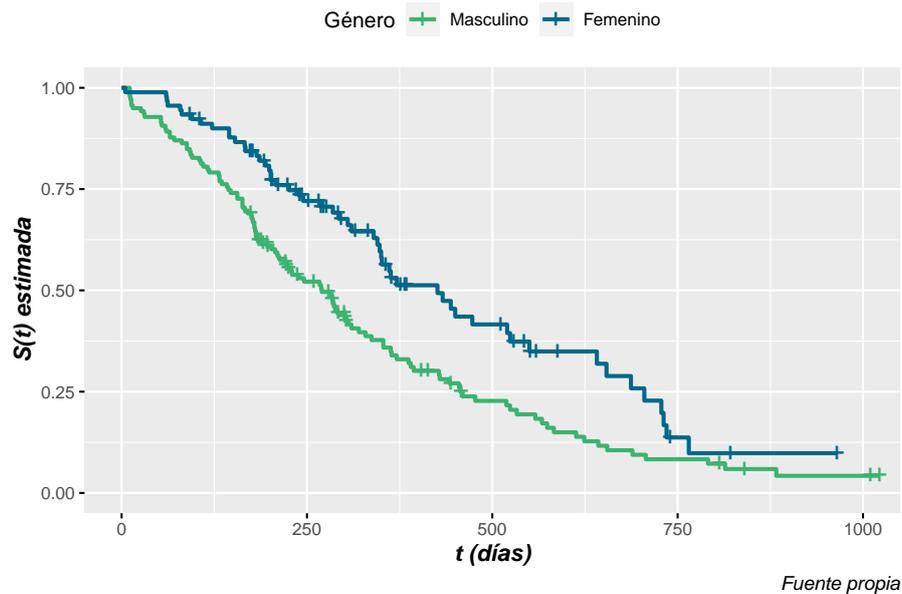


Figura 4.10: Curvas de supervivencia del conjunto de datos *lung* del paquete *survival* por género

Según la gráfica, el tiempo de supervivencia asociado a las mujeres es mayor durante todo el período en estudio aunque cabe destacar que al final del estudio, la diferencia entre ambas curvas de supervivencia es pequeña.

4.2. Librería *FHtest*

La librería *FHtest* (Oller y Langohr (2020)) fue implementada en el año 2017 por Ramon Oller y Klaus Langohr debido a la falta de funciones en R para test *log-rank* ponderados con datos censurados por intervalos basados en la clase de Fleming-Harrington. Además, incorpora funciones para trabajar con datos censurados por intervalos y datos censurados a la derecha. Por otro lado, proporciona un enfoque integrado para realizar pruebas de dos muestras, k muestras y de tendencia basadas en la teoría de proceso de conteo, teoría de verosimilitud o distribuciones de permutación.

Una peculiaridad que posee el paquete es que necesita un paquete externo llamado *Icens* (Bioconductor (2021b)) y que para ser instalado es conveniente consultar el siguiente enlace <http://www.bioconductor.org/packages/release/bioc/html/Icens.html>. Una forma de descargar dicho paquete es ejecutando las siguientes órdenes:

```
source("http://bioconductor.org/biocLite.R")
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install(version = "3.12")
BiocManager::install(c("GenomicFeatures", "AnnotationDbi"))
biocLite("Icens")
```

Una vez descargado el paquete necesario, se procede a cargar el paquete *FHtest*:

```
library(FHtest)
```

4.2.1. Comparación de curvas de supervivencia

La librería en estudio posee cuatro funciones destinadas a llevar a cabo las técnicas específicas de comparación de curvas de supervivencia con datos censurados por intervalos y a la derecha estudiadas en la [Sección 3.2](#) del [Capítulo 3](#). Las cuatro funciones permiten realizar pruebas o test de dos poblaciones o grupos, k poblaciones y de tendencia. Además, en los contrastes a realizar también se pueden contemplar el uso de hipótesis alternativas con desigualdades.

Debido a que dichas técnicas se basan en los pesos de Fleming-Harrington, la selección de los parámetros λ y ρ debe ser adecuada. Por ejemplo, si se está trabajando en ensayos clínicos y se quiere evaluar si el efecto de un tratamiento o terapia sobre la supervivencia es mayor en las primeras fases de la terapia, se deberá elegir $\lambda = 0$ y ρ de tal forma que cuánto mayor sea su valor, más se enfatizará en las diferencias tempranas. Sin embargo, en el caso de que existan evidencias médicas para creer que el efecto de la terapia está más pronunciado en la mitad o el final del período de seguimiento, se deberá tomar $\rho = \lambda > 0$ o $\rho = 0$, respectivamente. Además, cuanto más elevado sea el valor de λ más se enfatizará en las diferencias medias o tardías.

4.2.1.1. Datos censurados a la derecha

A la hora de abordar la comparación de curvas de supervivencia con datos censurados a la derecha, se puede utilizar la función *FHtestrcc()* o la función *FHtestrcp()*. Según el enfoque con el que se quiera resolver el contraste se utilizará una u otra función. La primera de ellas aplica un [enfoque basado en el proceso de conteo](#) mientras que la segunda aplica un [enfoque de permutación](#).

4.2.1.1.1. Enfoque mediante el proceso de conteo

La función asociada a este enfoque se llama *FHtestrcc* y tal y como se especificó en la [Sección 3.2.1.1](#), el método usa la familia de estadísticos $G^{\rho,\lambda}$ de Fleming-Harrington para detectar las diferencias entre dos o más curvas de supervivencia y se puede aplicar de dos formas distintas. La primera de ellas sigue el siguiente esquema *FHtestrcc(L, R, group, rho, lambda, alternative, method)* mientras que el segundo esquema es *FHtestrcc(formula, data, subset, na.action, rho, lambda, alternative, method)*. A continuación, se especifica el significado de cada uno de los argumentos:

- L : vector numérico que contiene los extremos izquierdos de los intervalos con los que se va a trabajar.
- R : vector numérico que contiene los extremos derechos de los intervalos de la muestra en estudio.

- *group*: vector que denota la variable para la que se desea efectuar el contraste. Además, determina el tipo de test a realizar. En el caso de que *group* sea de tipo factor con $k \geq 2$ posibles valores, entonces se llevará a cabo un test de k poblaciones. Sin embargo, si *group* es un vector numérico, entonces se llevará a cabo una prueba de tendencia.
- *rho*: valor de la constante ρ del peso de Fleming-Harrington (véase (3.12)). Por defecto, toma el valor 0.
- *lambda*: valor de la constante λ del peso de Fleming-Harrington. Por defecto, toma el valor 0.
- *alternative*: carácter que proporciona el tipo de hipótesis alternativa (H_1) que se quiere utilizar para los test con dos poblaciones o pruebas de tendencia. Los posibles valores son "*different*" (test bilateral), "*increasing*" (test unilateral con \leq) o "*decreasing*" (test unilateral con \geq).
- *method*: indicador del tipo de método usado para obtener el p-valor del test. Por defecto toma el valor "*pct*" y usa el teorema central del límite con permutaciones. Otros posibles valores pueden ser "*exact.mc*" que usa un método exacto vía Monte-carlo, "*exact.network*" que utiliza un método exacto a través de un algoritmo de red o "*exact.ce*" que a su vez usa un método exacto pero en este caso usando enumeración completa, lo que provoca que solo sea apropiada para muestras de tamaño pequeño. Nótese que los dos últimos posibles valores solo son compatibles con test de 2 poblaciones.
- *formula*: objeto de la forma $y \sim x$ donde y será un objeto de tipo *Surv* y x será la variable que contiene los grupos a los que pertenece cada individuo u observación en estudio.
- *data*: *data.frame* donde se encuentran los datos con los que se va a trabajar.
- *subset*: vector opcional que indica el subconjunto de individuos u observaciones que debe usar la función.
- *na.action*: función cuyo fin es indicar cómo se van a tratar aquellos valores perdidos *NA*.

Observación: cuando un tiempo i está censurado a la derecha, el elemento i -ésimo del vector R tomará el valor $+\infty$, de esta forma, el intervalo asociado a dicho tiempo será de la forma $(a, +\infty)$. Sin embargo, cuando el elemento i -ésimo del vector L y el correspondiente al vector R coinciden, se dice que se conoce de manera exacta el instante en el que se ha producido el evento en estudio y el intervalo es de la forma (a, a) . ▲

En la salida que proporciona la función se podrá consultar la siguiente información:

- *N*: número de observaciones existentes en cada una de las poblaciones o grupos en estudio.
- *Observed*: número ponderado de eventos observados en cada uno de los grupos en estudio.

- *Expected*: número ponderado de eventos esperados en cada uno de los grupos en estudio.
- *statistic*: estadístico asociado al contraste.
- *var*: matriz de varianzas del test.
- *pvalue*: p-valor asociado al contraste que se está realizando.

Además, se especifican los valores de los parámetros ρ y λ usados para los pesos de Fleming-Harrington, el tipo de test que se está aplicando y la descripción de la hipótesis alternativa del contraste. También se pueden observar las cantidades $(O - E)^2/E$ y $(O - E)^2/V$ necesarias para el cálculo del estadístico.

Para ilustrar su funcionamiento se mostrarán tres ejemplos, el primero de ellos consistirá en realizar un test de 2 poblaciones, el segundo constará de un contraste con $k > 2$ poblaciones y el tercero consistirá en una prueba de tendencia.

Test con 2 poblaciones

Se trabajará con el conjunto de datos *ovarian* (Edmonson et al. (1979)) disponible en el paquete *survival* y que contiene los tiempos de supervivencia de 26 pacientes con cáncer de ovario tratados cada uno de ellos con uno de los dos tratamientos disponibles con el fin de comparar la eficacia de ambos tratamientos. Entre las variables en estudio se encuentran:

- *ftime*: tiempo de supervivencia.
- *fustat*: estado de la censura (0 = censurado o 1 = no censurado).
- *rx*: tratamiento administrado al paciente. Si toma el valor 1, se hablará de *Tratamiento 1* y si toma el valor 2, se hablará entonces de *Tratamiento 2*.

A continuación, se muestran algunos de los datos del conjunto *ovarian*:

<i>ftime</i>	<i>fustat</i>	<i>rx</i>
59	1	1
115	1	1
464	1	2
563	1	2
770	0	2
1106	0	1
769	0	2
855	0	1

Tabla 4.20: Muestra de datos del conjunto *ovarian* del paquete *survival*

Para poder comparar la eficacia de los tratamientos se debe resolver el siguiente contraste:

$$\begin{aligned} H_0 : S_{Trat.1}(t) &= S_{Trat.2}(t) \quad \forall t > 0 \\ H_1 : S_{Trat.1}(t) &\neq S_{Trat.2}(t) \quad \text{para algún } t > 0 \end{aligned} \quad (4.5)$$

Una vez modelado el problema, se procede a resolverlo con la función en estudio mediante el test *log-rank* ponderado con pesos iguales (test *log-rank*):

```
FHtestrcc(Surv(futime, fustat) ~ rx, data = ovarian)
```

```
##
## Two-sample test for right-censored data
##
## Parameters: rho=0, lambda=0
## Distribution: counting process approach
##
## Data: Surv(futime, fustat) by rx
##
##      N Observed Expected   0-E (0-E)^2/E (0-E)^2/V
## rx=1 13         7      5.23  1.77    0.596    1.06
## rx=2 13         5      6.77 -1.77    0.461    1.06
##
## Statistic Z= -1, p-value= 0.303
## Alternative hypothesis: survival functions not equal
```

Analizando los resultados obtenidos del contraste, se concluye que debido a que el p-valor es *0.303*, no existen evidencias en contra de la igualdad de curvas de supervivencia. Para resolver el contraste sin necesidad de incluir una *formula* en la función es necesario que en el vector *L* se encuentren los tiempos observados mientras que *R* debe contener los tiempos de supervivencia no censurados o el $+\infty$ en caso de que el tiempo esté censurado a la derecha. Siguiendo las indicaciones expuestas para datos no censurados y censurados a la derecha la orden sería:

```
with(ovarian, FHtestrcc(L = futime, R = futime / fustat,
                        group = rx))
```

```
##
## Two-sample test for right-censored data
##
## Parameters: rho=0, lambda=0
## Distribution: counting process approach
##
## Data: {futime,futime/fustat} by rx
##
##      N Observed Expected   0-E (0-E)^2/E (0-E)^2/V
## rx=1 13         7      5.23  1.77    0.596    1.06
## rx=2 13         5      6.77 -1.77    0.461    1.06
##
## Statistic Z= -1, p-value= 0.303
## Alternative hypothesis: survival functions not equal
```

Tal y como era de esperar, el resultado obtenido es el mismo. La prueba expuesta utiliza pesos iguales en todos los sumandos de (3.11), pero en algunas ocasiones puede ser de interés detectar diferencias entre las curvas de supervivencia a corto o largo plazo. En el caso de que se esté interesado en las diferencias a largo plazo, es conveniente dar más peso a las últimas diferencias entre el número de eventos observados y esperados. Para ello, puede tomarse por ejemplo $\lambda = 1$ y $\rho = 0$:

```
FHtestrcc(Surv(futime, fustat) ~ rx, data = ovarian,
          rho = 0, lambda = 1)
```

```
##
## Two-sample test for right-censored data
##
## Parameters: rho=0, lambda=1
## Distribution: counting process approach
##
## Data: Surv(futime, fustat) by rx
##
##      N Observed Expected      O-E (O-E)^2/E (O-E)^2/V
## rx=1 13      1.11      1.11 -0.00447 1.80e-05 0.000102
## rx=2 13      1.50      1.49  0.00447 1.34e-05 0.000102
##
## Statistic Z= 0, p-value= 0.992
## Alternative hypothesis: survival functions not equal
```

En este caso, el estadístico de prueba es el recogido en la ecuación (3.8) y el p-valor es cercano a 1 por lo que se puede concluir que las diferencias a largo plazo entre las curvas de supervivencia asociadas a cada uno de los tratamientos son bastantes pequeñas. Véase gráficamente dicha diferencia:

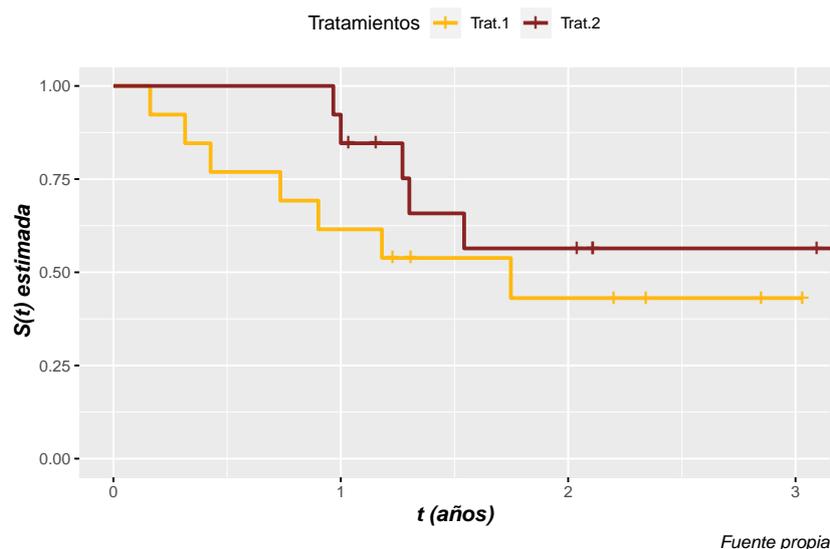


Figura 4.11: Curvas de supervivencia asociadas a cada uno de los tratamientos destinados a los pacientes de cáncer de ovario del conjunto de datos *ovarian* del paquete *survival*

Se observa así que existió una mejor supervivencia a corto plazo entre los pacientes con el tratamiento 2. Sin embargo, después de un año y medio, las diferencias entre las curvas de supervivencia disminuyeron.

Test con k poblaciones

Para ilustrar la prueba con $k > 2$ poblaciones se utilizará el conjunto de datos *bmt* (Klein y Moeschberger (1997)) del paquete *KMsurv*. El conjunto de datos procede de un estudio sobre la supervivencia de 137 pacientes con leucemia tras realizarle un trasplante de médula ósea a cada uno de ellos y está compuesto por 22 variables, entre las que se encuentran:

- *group*: grupo de enfermedad al que pertenece cada paciente. Puede tomar el valor 1 para referirse a la leucemia linfoblástica aguda (*ALL*), el valor 2 para hacer referencia a la leucemia mieloide aguda de bajo riesgo (*AML Low Risk*) o el valor 3 para referirse a la leucemia mieloide aguda de alto riesgo (*AML High Risk*).
- *t2*: tiempo de supervivencia de cada individuo libre de la enfermedad, es decir, el tiempo hasta la recaída, la muerte o el final del estudio.
- *d3*: indicador del tipo de censura. Si la variable toma el valor 1, se produce la muerte o la recaída del individuo en cuestión y si toma el valor 0, el tiempo de supervivencia estará censurado.

Se procede a la carga de los datos y se muestran algunos de ellos:

```
data("bmt", package = "KMsurv")
```

<i>group</i>	<i>t2</i>	<i>d3</i>
1	2081	0
1	1496	0
1	609	1
1	172	1
2	860	0
2	1324	0
2	390	1
2	79	1
3	2430	0
3	1345	0
3	100	1
3	2	1

Tabla 4.21: Muestra de datos del conjunto *bmt* del paquete *KMsurv*

Con el fin de que la interpretación de la salida sea más clara, se modificarán las etiquetas de la variable *group*:

```
bmt$group <- factor(bmt$group, labels = c("ALL", "AML low risk",
                                         "AML high risk"))
```

Una vez que se tienen los datos preparados, se pretende comparar si el tiempo de supervivencia hasta la recaída de los pacientes es igual para todos los grupos de enfermedades o no. En ese caso, el contraste que se debe resolver es:

$$\begin{aligned} H_0 : S_{ALL}(t) = S_{AML\ low\ risk}(t) = S_{AML\ high\ risk} & \quad \forall t > 0 \\ H_1 : S_i(t) \neq S_j(t) & \quad \text{para algún } t, i, j \end{aligned} \quad (4.6)$$

donde i, j toman como posibles valores *ALL*, *AML low risk* y *AML high risk*. Para resolver el contraste anterior, se elegirá $\rho = \lambda = 1$ de tal forma que se dará más importancia a las diferencias existentes entre las curvas en torno a la mediana global:

```
FHtestrcc(Surv(t2, d3) ~ group, data = bmt, rho = 1, lambda = 1)
```

```
##
## K-sample test for right-censored data
##
## Parameters: rho=1, lambda=1
## Distribution: counting process approach
##
## Data: Surv(t2, d3) by group
##
##           N Observed Expected   0-E (0-E)^2/E (0-E)^2/V
## group=ALL      38     4.55     3.79  0.769    0.156    1.02
## group=AML high risk 45     5.41     3.54  1.864    0.981    6.28
## group=AML low risk 54     4.87     7.50 -2.633    0.924    8.99
##
## Chisq= 9.9 on 2 degrees of freedom, p-value= 0.00697
## Alternative hypothesis: survival functions not equal
```

En la salida obtenida se observa que el estadístico de prueba (3.9) sigue asintóticamente una distribución χ^2 con $k - 1 = 2$ grados de libertad. Además, se muestra un p-valor muy pequeño, por lo que se debe rechazar la hipótesis nula. Así, se puede suponer que hay diferencias entre los tres grupos de enfermos de leucemia con respecto al tiempo de supervivencia libre hasta la recaída después de recibir un trasplante. Véanse dichas diferencias en la siguiente gráfica:

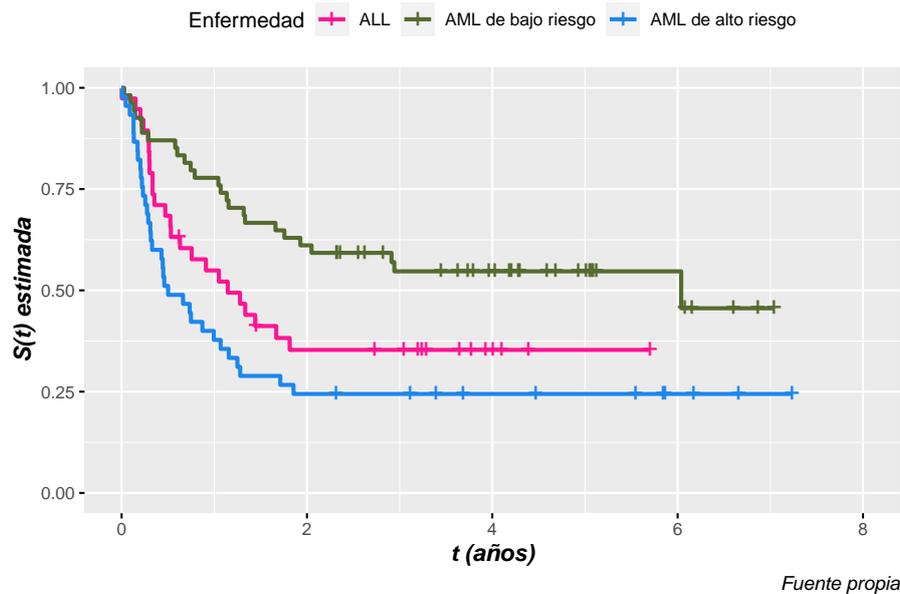


Figura 4.12: Curvas de supervivencia asociadas a cada uno de los pacientes de leucemia del conjunto de datos *bmt* del paquete *KMsurv*

En promedio, los tiempos de supervivencia libre de enfermedad fueron mayores entre los pacientes con *AML de bajo riesgo* y menores entre los pacientes con *AML de alto riesgo*.

Prueba de tendencia

Para poder aplicar una prueba de tendencia es imprescindible que la variable de agrupación *group* sea numérica. Este test puede ser más útil que la prueba con k poblaciones porque la hipótesis alternativa del contraste puede contener desigualdades.

Si se sigue trabajando con el conjunto de datos *bmt*, puede ser de interés resolver el siguiente contraste:

$$\begin{aligned}
 H_0 &: S_{ALL}(t) = S_{AML\ low\ risk}(t) = S_{AML\ high\ risk}(t) \quad \forall t > 0 \\
 H_1 &: S_{AML\ low\ risk}(t) \geq S_{ALL}(t) \geq S_{AML\ high\ risk}(t) \quad \forall t > 0
 \end{aligned}$$

con al menos una desigualdad estricta debido a que en la [Figura 4.12](#) se observa que la curva de supervivencia asociada al grupo *AML de bajo riesgo* está por encima de la de *ALL* y esta a su vez por encima de la curva del grupo *AML de alto riesgo*.

Como se necesita que la variable *group* sea numérica, se reconvierte a tipo *numeric* y se recodifica para que el valor 1 esté asociado al grupo *AML de bajo riesgo*; el 2 al grupo *ALL*; y el 3, al grupo *AML de alto riesgo*:

```
bmt$group <- as.numeric(relevel(bmt$group, ref = "AML low risk"))
```

Una vez preparados los datos, se procede a resolver el contraste. Para ello, se debe indicar que la hipótesis alternativa es decreciente (*alternative = "decreasing"*):

```

FHtestrcc(Surv(t2, d3) ~ group, data = bmt, rho = 1, lambda = 1,
           alternative = "decreasing")

##
## Trend FH test for right-censored data
##
## Parameters: rho=1, lambda=1
## Distribution: counting process approach
##
## Data: Surv(t2, d3) by group
##
##      N Observed Expected    O-E
## group=1 54      4.87      7.50 -2.633
## group=2 38      4.55      3.79  0.769
## group=3 45      5.41      3.54  1.864
##
## Statistic Z= 3.1, p-value= 0.000883
## Alternative hypothesis: decreasing survival functions (higher group implies earlier event times)

```

Nótese que en la salida de este test se suprimen las columnas $(O - E)^2/E$ y $(O - E)^2/V$ ya que estas no son necesarias para el cálculo del p-valor. La función ha utilizado el estadístico de prueba (3.10) y el p-valor obtenido es 0.000883 por lo que existen evidencias en contra de H_0 . Obsérvese que el signo negativo en la columna $O - E$ indica que existen menos eventos observados que esperados bajo la hipótesis nula, lo que quiere decir que la supervivencia es mejor de la esperada. Por lo tanto, el aumento de las diferencias $O - E$ del primer al último grupo avalan la hipótesis alternativa de funciones de supervivencia decrecientes.

4.2.1.1.2. Enfoque de permutación

La función asociada a este enfoque se llama *FHtestrcp* y utiliza la familia de estadísticos $G^{\rho, \lambda}$ para detectar las diferencias entre dos o más curvas de supervivencia. Al igual que la función estudiada en la sección anterior, admite dos tipos de esquemas. El primero de ellos es *FHtestrcp(L, R, group, rho, lambda, alternative, method, exact)* y el segundo *FHtestrcp(formula, data, subset, na.action, rho, lambda, alternative, method, exact)*. El significado de cada uno de los argumentos se puede consultar en la [descripción detallada de la función FHtestrcc\(\)](#) a falta del argumento *exact*, el cual puede tomar el valor *TRUE* o *FALSE* según se desee resolver el contraste de manera exacta o no. En el caso de que *method* sea no nulo, este argumento es ignorado. Otro posible argumento es *permcontrol* que se encarga de controlar aspectos relacionados con los test de permutación.

Al ejecutar la función se obtiene la siguiente información:

- *N*: número de observaciones en cada grupo de estudio.
- *O-E*: número ponderado de eventos observados menos el esperado en cada uno de los grupos en estudio.
- *var*: matriz de varianza del test.
- *statistic*: estadístico asociado al contraste.

- *pvalue*: p-valor asociado al contraste que se está realizando.

Además, se especifican los valores de los parámetros ρ y λ usados para los pesos de Fleming-Harrington, el tipo de test que se está aplicando y la descripción de la hipótesis alternativa del contraste. También se puede observar el intervalo de confianza del p-valor en el caso de que *method* = “*exact.mc*”. En contraposición con la salida de la función *FHtestrcc()*, no aparecen las cantidades $(O - E)^2/E$ y $(O - E)^2/V$ pues el p-valor se calcula en función de las permutaciones (véanse las ecuaciones (3.13) y (3.15)) y no en la distribución asintótica del proceso de recuento del estadístico (3.11).

Para mostrar el funcionamiento de esta función se usarán los mismos conjuntos de datos utilizados para el enfoque mediante el proceso de conteo.

Test con 2 poblaciones

Si se pretende resolver el contraste (4.5) según el enfoque de permutación con $\rho = 0$, $\lambda = 1$ y el método exacto a través de Montecarlo para el cálculo del p-valor, se obtendrá lo siguiente:

```
FHtestrcp(Surv(futime, fustat) ~ rx, data = ovarian,
          lambda = 1, exact = TRUE,
          permcontrol = permControl(p.conf.level = .95))

##
## Two-sample test for right-censored data
##
## Parameters: rho=0, lambda=1
## Distribution: permutation approach (exact method using Monte Carlo with 999 replications)
##
## Data: Surv(futime, fustat) by rx
##
##      N      O-E
## rx=1 13 -0.00447
## rx=2 13  0.00447
##
## Statistic= 0, p-value= 0.942
## 95 percent confidence interval on p-value:  0.878 1.000
## Alternative hypothesis: survival functions not equal
```

Observando el p-valor *0.942* y su intervalo de confianza (*0.878, 1*) se concluye que se debe aceptar la igualdad de las curvas de supervivencia asociadas a cada uno de los tratamientos tal y como se había obtenido en la sección anterior con la función *FHtestrcc()*.

Cabe destacar que tanto el número de réplicas de Montecarlo (999) como el nivel de confianza para la estimación del p-valor pueden modificarse cambiando los valores por defecto del argumento *permcontrol* (consultar la ayuda de [Fay \(2010\)](#)).

Se añade también un contraste unilateral que se resolverá haciendo hincapié en las diferencias entre las curvas en las fases tempranas de cada tratamiento. Para ello, se usará $\rho = 3$ y el método exacto a través de un algoritmo de red:

$$H_0 : S_{Trat.1}(t) = S_{Trat.2}(t) \quad \forall t > 0$$

$$H_1 : S_{Trat.1}(t) \leq S_{Trat.2}(t) \quad \forall t > 0$$

```

FHtestrcp(Surv(futime, fustat) ~ rx, data = ovarian,
           rho = 3, alternative = "increasing",
           method = "exact.network")

##
## Two-sample test for right-censored data
##
## Parameters: rho=3, lambda=0
## Distribution: permutation approach (exact method using a network algorithm)
##
## Data: Surv(futime, fustat) by rx
##
##      N    O-E
## rx=1 13  1.72
## rx=2 13 -1.72
##
## Statistic= -1.7, p-value= 0.0454
## Alternative hypothesis: increasing survival functions (group=2 has later event times)

```

El p-valor asociado al contraste es un tanto dudoso pues es muy cercano a 0.05 . Sin embargo, tomando un nivel de significación del 1% se concluye que efectivamente en las fases tempranas los tiempos de supervivencia asociados al tratamiento 1 son menores que los del tratamiento 2.

Test con k poblaciones

En el caso de que se quiera abordar el contraste (4.6) mediante el enfoque de permutación con los parámetros $\rho = \lambda = 1$, se obtendrá el siguiente resultado:

```

FHtestrcp(Surv(t2, d3) ~ group, data = bmt,
           rho = 1, lambda = 1)

##
## K-sample test for right-censored data
##
## Parameters: rho=1, lambda=1
## Distribution: permutation approach (asymptotic approximation using central limit theorem)
##
## Data: Surv(t2, d3) by group
##
##      N    O-E
## group=ALL      38  0.769
## group=AML high risk 45  1.864
## group=AML low risk 54 -2.633
##
## Chisq= 9.8 on 2 degrees of freedom, p-value= 0.00755
## Alternative hypothesis: survival functions not equal

```

El estadístico calculado se corresponde con (3.9) y el p-valor obtenido es 0.00755 por lo que existen evidencias en contra de la igualdad de curvas de supervivencia.

Prueba de tendencia

En el contraste que se ha resuelto se ha llegado a la conclusión de que las curvas de supervivencia asociadas a cada uno de los grupos de enfermedad *ALL*, *AML low risk* y *AML high risk* no pueden considerarse iguales. Además, en la [Figura 4.12](#) se detecta que en general los tiempos de supervivencia asociados a los pacientes con *AML high risk* son menores que los tiempos de *ALL* y estos a su vez son menores que los tiempos de los pacientes con *AML low risk*. Por lo tanto, se resolverá el siguiente contraste:

$$\begin{aligned} H_0 &: S_{ALL}(t) = S_{AML\ low\ risk}(t) = S_{AML\ high\ risk}(t) \quad \forall t > 0 \\ H_1 &: S_{AML\ high\ risk}(t) \leq S_{ALL}(t) \leq S_{AML\ low\ risk}(t) \quad \forall t > 0 \end{aligned}$$

con al menos una desigualdad estricta. Para ello, es necesario recodificar las etiquetas de la variable *group*:

```
bmt$group <- as.numeric(relevel(bmt$group, ref = "AML high risk"))
```

Una vez arreglada la variable *group*, se procede a resolver el contraste teniendo en cuenta que el argumento *alternative* debe tomar el valor *increasing*:

```
FHtestrcp(Surv(t2, d3) ~ group, data = bmt,
           rho = 1, lambda = 1, alternative = "increasing")
```

```
##
## Trend FH test for right-censored data
##
## Parameters: rho=1, lambda=1
## Distribution: permutation approach (asymptotic approximation using central limit theorem)
##
## Data: Surv(t2, d3) by group
##
##           N      O-E
## group=1 45  1.864
## group=2 38  0.769
## group=3 54 -2.633
##
## Statistic Z= -3, p-value= 0.00129
## Alternative hypothesis: increasing survival functions (higher group implies later event times)
```

El p-valor obtenido es *0.00129*, así que existen suficientes evidencias a favor de la hipótesis alternativa $S_{AML\ high\ risk}(t) \leq S_{ALL}(t) \leq S_{AML\ low\ risk}(t)$.

4.2.1.2. Datos censurados por intervalos

En el paquete *FHtest* existen dos funciones (*FHtesticp()* y *FHtestics()*) que son muy útiles cuando se trabaja con datos censurados por intervalos y se pretende resolver un contraste de igualdad de curvas de supervivencia. La primera de las funciones aborda el contraste mediante un [enfoque de permutación](#) mientras que la otra resuelve el contraste a través de un [enfoque basado en un vector de puntuación](#).

4.2.1.2.1. Enfoque de permutación

La función asociada a este enfoque se llama *FHtesticp*, usa la familia de estadísticos $G^{\rho,\lambda}$ para detectar las diferencias existentes entre dos o más curvas de supervivencia y admite dos esquemas para aplicarla. El primer esquema es *FHtesticp(L, R, group, rho, lambda, alternative, icFIT, icontrol, exact, method, Lin, Rin, permcontrol)* y el segundo es *FHtesticp(formula, data, subset, na.action, rho, lambda, alternative, icFIT, icontrol, exact, method, Lin, Rin, permcontrol)*. A continuación, se especifica la utilidad de cada uno de los argumentos:

- *L*: vector numérico que contiene los extremos izquierdos de los datos censurados por intervalos.
- *R*: vector numérico que contiene los extremos derechos de los datos censurados por intervalos.
- *group*: vector que denota la variable para la que se desea efectuar el contraste. Además, determina el tipo de test a realizar. En el caso de que *group* sea de tipo factor con $k \geq 2$ posibles valores, entonces se llevará a cabo un test de k poblaciones. Sin embargo, si *group* es un vector numérico, entonces se llevará a cabo una prueba de tendencia.
- *rho*: valor de la constante ρ del peso de Fleming-Harrington. Por defecto, toma el valor 0.
- *lambda*: valor de la constante λ del peso de Fleming-Harrington. Por defecto, toma el valor 0.
- *alternative*: carácter que proporciona el tipo de hipótesis alternativa (H_1) que se quiere utilizar para los test con dos poblaciones o de tendencia. Los posibles valores son "different" (test bilateral), "increasing" (test unilateral con \leq) o "decreasing" (test unilateral con \geq).
- *icFIT*: objeto *icfit*¹³ precalculado cuyo fin es aumentar la velocidad de cálculo.
- *icontrol*: lista de argumentos que se encargan de controlar el algoritmo de la estimación no paramétrica de máxima verosimilitud del objeto *icfit*.
- *exact*: valor lógico que indica si se desea resolver el test de manera exacta o no.
- *method*: indicador del tipo de método usado para obtener el p-valor del test. Los valores posibles son "pctl", "exact.network", "exact.ce" o "exact.mc" como en las funciones *FHtestrec()* y *FHtestrecp()*.
- *Lin*: vector lógico que tiene como finalidad indicar si *L* debe ser incluido o no en el intervalo. El valor por defecto es no incluir el extremo a menos que existan individuos en los que se cumpla que el elemento correspondiente del vector *L* y *R* coincidan.

¹³Este tipo de objeto se construye con la función *icfit()* del paquete *interval* (Fay (2020)) la cual calcula la estimación no paramétrica de la función de distribución mediante máxima verosimilitud usando un estimador auto-consistente y con la presencia de datos censurados por intervalos.

- *Rin*: vector lógico que tiene como finalidad indicar si R debe ser incluido o no en el intervalo. Nótese que R debe ser incluido en el intervalo cuando existan elementos de R que tomen el valor infinito.
- *formula*: objeto de la forma $y \sim x$ donde y será un objeto de tipo *Surv* y x será la variable que contiene los grupos a los que pertenece cada individuo u observación en estudio. Debido a que se está trabajando con datos censurados por intervalos, es necesario añadir el argumento *type* = "interval2" en el objeto *Surv*.
- *data*: *data.frame* donde se encuentran los datos con los que se va a trabajar.
- *subset*: vector opcional que indica el subconjunto de individuos u observaciones que debe usar la función.
- *na.action*: función cuyo fin es indicar cómo se van a tratar aquellos valores perdidos *NA*.
- *permcontrol*: lista de argumentos cuya función es controlar los test de permutación.

Al ejecutar la función en estudio se obtiene la siguiente información:

- *N*: número de observaciones en cada grupo de estudio.
- *O-E*: número ponderado de eventos observados menos el esperado en cada uno de los grupos en estudio.
- *var*: matriz de varianzas del test.
- *statistic*: estadístico asociado al contraste.
- *pvalue*: p-valor asociado al contraste que se está realizando.

También se especifican los valores de los parámetros ρ y λ usados para los pesos de Fleming-Harrington, el tipo de test que se está aplicando y la descripción de la hipótesis alternativa del contraste. Además, se puede observar el intervalo de confianza del p-valor en el caso de que *method* = "exact.mc".

Para mostrar el funcionamiento de esta función se trabajará con datos proporcionados por los paquetes *interval* y *FHtest* y se realizarán test con 2 y k poblaciones y una prueba de tendencia.

Test con 2 poblaciones

Se usará el conjunto de datos *bcos* disponible en el paquete *interval* (Fay (2020)) que contiene los tiempos de supervivencia censurados por intervalos de 94 pacientes con cáncer de mama sometidos a dos tipos de tratamiento: radiación (*Rad*) o radiación más quimioterapia (*RadChem*). Estos datos proceden de estudios realizados por Finkelstein y Wolfe (1985). El objetivo final es comparar la eficacia de ambos tratamientos en la cura del cáncer de mama. Las variables en estudio son:

- *left*: vector que contiene el extremo izquierdo de cada uno de los intervalos de censura expresado en meses.
- *right*: vector que contiene el extremo derecho de cada uno de los intervalos de censura expresado en meses.
- *treatment*: tratamiento con el que ha sido tratado cada paciente (*Rad* o *RadChem*).

Se cargan los datos y se adjunta una tabla en la que se pueden consultar algunas de las observaciones del conjunto de datos en estudio:

```
data("bcos", package = "interval")
```

<i>left</i>	<i>right</i>	<i>treatment</i>
45	Inf	Rad
0	7	Rad
25	37	Rad
17	Inf	Rad
4	8	RadChem
21	Inf	RadChem
16	60	RadChem
32	Inf	RadChem

Tabla 4.22: Muestra de datos del conjunto *bcos* del paquete *interval*

Con el objetivo de comparar la eficacia de los dos tratamientos en estudio, se deberá resolver el contraste que se define a continuación:

$$\begin{aligned}
 H_0 &: S_{Rad}(t) = S_{RadChem}(t) \quad \forall t > 0 \\
 H_1 &: S_{Rad}(t) \neq S_{RadChem}(t) \quad \text{para algún } t > 0
 \end{aligned}$$

Recuérdese que si se usa un objeto *Surv* para resolver el contraste, se debe añadir el argumento *type = "interval2"* para indicar que se está trabajando con datos censurados por intervalos. Además, en la función *FHtesticp()* es conveniente incorporar los argumentos *Lin* y *Rin* debido a que se incluyen observaciones en las que la variable *right* toma el valor *Inf*:

```
FHtesticp(Surv(left, right, type = "interval2") ~ treatment, data = bcos,
           Rin = TRUE, Lin = TRUE, lambda = 2, rho = 2)
```

```
##
## Two-sample test for interval-censored data
##
## Parameters: rho=2, lambda=2
## Distribution: permutation approach (asymptotic approximation using central limit theorem)
##
```

```
## Data: Surv(left, right, type = "interval2") by treatment
##
##           N    O-E
## treatment=Rad    46 -0.598
## treatment=RadChem 48  0.598
##
## Statistic Z= 3.9, p-value= 0.000113
## Alternative hypothesis: survival functions not equal
```

El p-valor asociado al contraste usando $\lambda = \rho = 2$ es 0.000113 , por lo que se concluye que la hipótesis de igualdad de curvas debe ser rechazada. Si se quisiera estudiar la razón por la que se está rechazando la H_0 , se podría resolver un contraste con hipótesis alternativa de tipo *decreasing*:

$$H_0 : S_{Rad}(t) = S_{RadChem}(t) \quad \forall t > 0$$

$$H_1 : S_{Rad}(t) \geq S_{RadChem}(t) \quad \forall t > 0$$

```
FHtesticp(Surv(left, right, type = "interval2") ~ treatment, data = bcos,
           Rin = TRUE, Lin = TRUE, alternative = "decreasing",
           lambda = 2, rho = 2)
```

```
##
## Two-sample test for interval-censored data
##
## Parameters: rho=2, lambda=2
## Distribution: permutation approach (asymptotic approximation using central limit theorem)
##
## Data: Surv(left, right, type = "interval2") by treatment
##
##           N    O-E
## treatment=Rad    46 -0.598
## treatment=RadChem 48  0.598
##
## Statistic Z= 3.9, p-value= 5.67e-05
## Alternative hypothesis: decreasing survival functions (group=RadChem has earlier event times)
```

En este contraste, el p-valor obtenido es mucho más pequeño que el nivel de significación prefijado, así que existen suficientes evidencias en contra de la hipótesis nula y por tanto, se puede concluir que en general los tiempos de supervivencia asociados al tratamiento de radioterapia son mayores que los asociados al tratamiento de radioterapia más quimioterapia, sobre todo en la mitad del período de seguimiento. Por consiguiente, da la impresión de que la efectividad del tratamiento en radioterapia es mayor. Véanse gráficamente las estimaciones de las curvas de supervivencia:

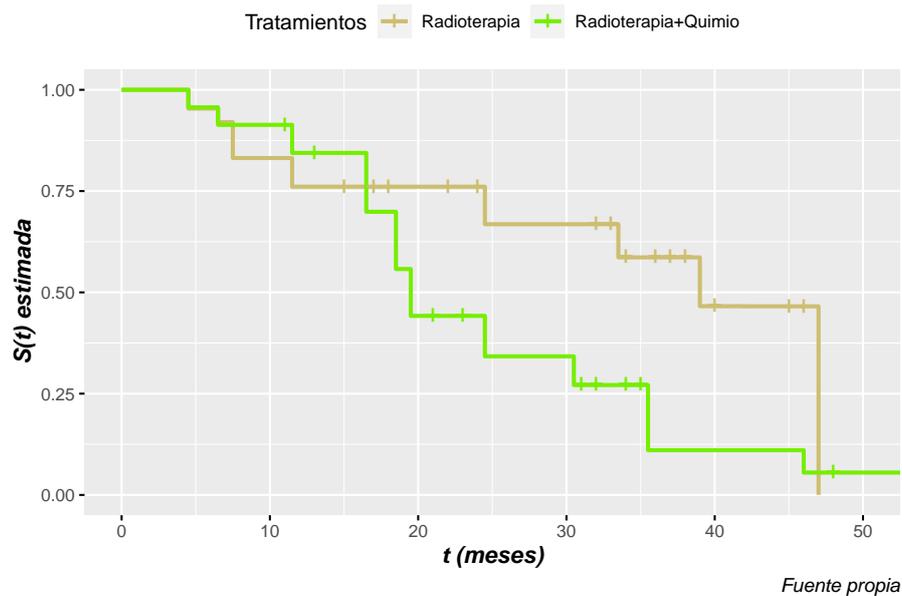


Figura 4.13: Curvas de supervivencia asociadas a cada uno de los tratamientos destinados a los pacientes de cáncer de mama del conjunto de datos *bcos* del paquete *interval*

Se observa así que aproximadamente en los primeros siete meses de estudio, las dos curvas de supervivencia son prácticamente iguales. Sin embargo, en los siguientes diez meses la curva de supervivencia asociada al tratamiento de radioterapia más quimioterapia se sitúa por encima de la asociada al tratamiento de radioterapia. Finalmente, a largo plazo se observa que la curva de supervivencia asociada al tratamiento de radioterapia se sitúa por encima de la otra curva, por lo que, en general, la efectividad del tratamiento de radioterapia es mayor en dicho período de tiempo.

Test con k poblaciones

El conjunto de datos *illust3* del paquete *FHtest* que contiene datos procedentes de un ensayo clínico sobre el sida se utilizará para la ilustración del test con k poblaciones y datos censurados por intervalos. El marco de datos anterior posee 1607 observaciones divididas en tres variables y como se ha especificado, procede de un ensayo clínico sobre el sida que comenzó a realizarse hacia el año 1987 por [Volberding et al. \(1995\)](#). Fue diseñado para estudiar los beneficios de la terapia con zidovudina en pacientes en la fase inicial de la infección por VIH y las tres variables en estudio son:

- *left*: extremo izquierdo de cada uno de los intervalos de tiempo en estudio.
- *right*: extremo derecho de cada uno de los intervalos de tiempo en estudio.
- *group*: grupo de tratamiento en el que el valor 1 hace referencia a terapia diferida (*deferred therapy*), el 2 hace referencia a dosis de 500 mg/día (*500 mg/day dosage*) y el 3 hace referencia a dosis de 1500 mg/día (*1500 mg/day dosage*).

La variable censurada por intervalos es el tiempo (en meses) desde que el paciente comienza el tratamiento hasta que muestra síntomas del sida o muere. A continuación, se cargan los datos y se muestran algunos de ellos:

```
data("illust3", package = "FHtest")
```

<i>left</i>	<i>right</i>	<i>group</i>
7	999	1
9	35	1
2	2	1
0	999	2
2	8	2
16	45	2
8	9	3
46	55	3
70	999	3

Tabla 4.23: Muestra de datos del conjunto *illust3* del paquete *FHtest*

En este conjunto de datos, existen tiempos censurados por intervalos, tiempos censurados a la derecha, tiempos perdidos y tiempos sin censura. Los tiempos sin censura son aquellos en los que la variable *left* y *right* coinciden; los tiempos perdidos se dan cuando la variable *left* toma el valor 0 y *right* toma el valor 999; los censurados a la derecha toman el valor 999 en la variable *right* y un número distinto de 0 en la variable *left* y por último, los censurados por intervalo son aquellos que no verifican ninguna de las condiciones anteriores. Se convertirán los valores 999 a *Inf* para distinguir de manera más sencilla los tiempos censurados:

```
library(dplyr)
illust3 <- illust3 %>%
  mutate(right = ifelse(right == 999, Inf, right))
```

Una vez presentados y arreglados los datos, se procede a resolver el siguiente contraste:

$$H_0 : S_1(t) = S_2(t) = S_3(t) \quad \forall t > 0$$

$$H_1 : S_1(t) \neq S_2(t) \neq S_3(t) \quad \text{para algún } t > 0$$

donde S_1 hace referencia a la función de supervivencia asociada a la terapia diferida; S_2 a la función de supervivencia asociada a la dosis de 500 mg/día; y S_3 , a la dosis de 1500 mg/día.

Dado que la variable *group* es numérica, se necesita recodificar la variable:

```
illust3$group <- factor(illust3$group, labels = c("Deferred therapy",
                                                "500 mg/day dosage",
                                                "1500 mg/day dosage"))
```

El contraste anterior se resolverá para los valores de $\rho = 1$ y $\lambda = 1$ de forma que se dé más importancia a las diferencias intermedias que a las diferencias a largo o corto plazo entre las curvas de supervivencia:

```
FHtesticp(Surv(left, right, type = "interval2") ~ group, data = illust3,
          rho = 1, lambda = 1, Lin = TRUE, Rin = TRUE)
```

```
##
## K-sample test for interval-censored data
##
## Parameters: rho=1, lambda=1
## Distribution: permutation approach (asymptotic approximation using central limit theorem)
##
## Data: Surv(left, right, type = "interval2") by group
##
##              N    O-E
## group=1500 mg/day dosage 528 -7.63
## group=500 mg/day dosage  538 -1.72
## group=Deferred therapy  541  9.35
##
## Chisq= 15.7 on 2 degrees of freedom, p-value= 4e-04
## Alternative hypothesis: survival functions not equal
```

El estadístico asociado a la prueba sigue asintóticamente una distribución χ^2 con 2 grados de libertad y el p-valor obtenido es 0.0004 , por lo que existen evidencias suficientes en contra de la igualdad de curvas de supervivencia. Se adjunta una gráfica de la estimación de las curvas de supervivencia asociadas a cada uno de los tratamientos en estudio. En ella se puede observar que aproximadamente durante el primer año de tratamiento, las curvas de supervivencia son prácticamente iguales. Sin embargo, posteriormente la curva asociada a la terapia diferida se sitúa por debajo de las otras dos, aunque la del grupo de las dosis de 1500 mg/día es algo mayor que la de 500 mg/día. Además, con la ayuda del contraste realizado se ha podido comprobar que realmente existen diferencias entre los tres tratamientos con respecto a la supervivencia de cada uno de los pacientes.

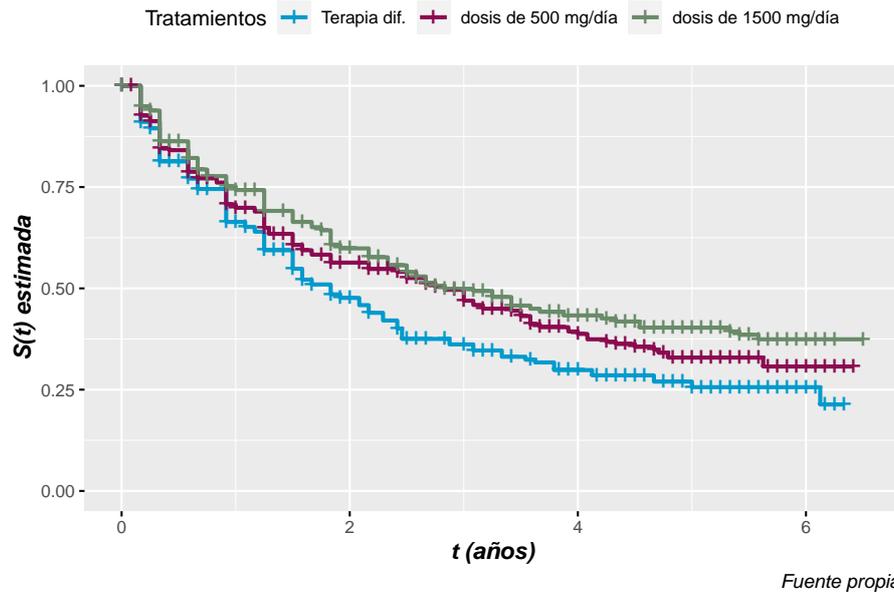


Figura 4.14: Curvas de supervivencia asociadas a cada uno de los tratamientos destinados a los pacientes de sida del conjunto de datos *illust3* del paquete *FHtest*

Prueba de tendencia

Con la ayuda de una prueba de tendencia se estudiará si puede considerarse que el tiempo de supervivencia asociado al tratamiento de la dosis de 1500 mg/día es superior al resto. Se considerará el siguiente contraste:

$$H_0 : S_1(t) = S_2(t) = S_3(t) \quad \forall t > 0$$

$$H_1 : S_1(t) \leq S_2(t) \leq S_3(t) \quad \forall t > 0$$

con al menos una desigualdad estricta. A la hora de resolver dicho contraste, se introducirá el argumento *alternative* = “*increasing*” y se convertirá a una variable numérica la variable de agrupación *group*:

```
FHtesticp(Surv(left, right, type = "interval2") ~ as.numeric(group),
          data = illust3, Lin = TRUE, Rin = TRUE)
```

```
##
## Trend FH test for interval-censored data
##
## Parameters: rho=0, lambda=0
## Distribution: permutation approach (asymptotic approximation using central limit theorem)
##
## Data: Surv(left, right, type = "interval2") by as.numeric(group)
##
##              N    O-E
## as.numeric(group)=1 541  51.4
## as.numeric(group)=2 538  -8.0
```

```
## as.numeric(group)=3 528 -43.4
##
## Statistic Z= -4.2, p-value= 3.15e-05
## Alternative hypothesis: survival functions not equal
```

Se ha obtenido un p-valor cercano a 0 , así que se puede suponer que efectivamente el tiempo de supervivencia de los pacientes con una dosis de 1500 mg/día es mayor que el tiempo de supervivencia asociado a los pacientes con una dosis de 500 mg/día y este tiempo es a su vez mayor que el asociado a la terapia diferida.

4.2.1.2.2. Enfoque basado en un vector de puntuación

La función asociada a este enfoque se llama *FHtestics*, utiliza la familia de estadísticos $G^{\rho, \lambda}$ con $\lambda = 0$ para encontrar las diferencias entre dos o más curvas de supervivencia y al igual que todas las funciones del paquete en estudio, su aplicación puede llevarse a cabo de dos formas distintas. El primer esquema de aplicación es *FHtestics(L, R, group, rho, alternative, tol, icFIT, icontrol, Lin, Rin)* y el segundo *FHtestics(formula, data, subset, na.action, rho, alternative, tol, icFIT, icontrol, Lin, Rin)*. El significado de cada uno de los argumentos se puede consultar en la [descripción detallada de la función *FHtesticp\(\)*](#) a falta del argumento *tol*, el cual denota la tolerancia permitida para el cálculo de la inversa de g .

En la salida que proporciona la función se podrá consultar la siguiente información:

- *N*: número de observaciones en cada grupo de estudio.
- *O-E*: número ponderado de eventos observados menos el esperado en cada uno de los grupos en estudio.
- *statistic*: estadístico asociado al contraste.
- *pvalue*: p-valor asociado al contraste que se está realizando.
- *var*: matriz de varianza del test.
- *d2L.dB2*: segunda derivada de la log-verosimilitud con respecto a β .
- *d2L.dgam2*: segunda derivada de la log-verosimilitud con respecto a θ .
- *d2L.dBdgam*: segunda derivada de la log-verosimilitud con respecto a β y θ .

Además, se especifica el valor de ρ usado en los pesos de Fleming-Harrington, el tipo de test aplicado y la descripción de la hipótesis alternativa del contraste. Por último, en el caso de que se utilice el método *exact.mc*, se puede consultar un intervalo de confianza del p-valor.

En las siguientes páginas se mostrarán ejemplos en los que se usará la función en estudio, entre ellos, habrá test con $k \geq 2$ poblaciones y una prueba de tendencia.

Test con 2 poblaciones

Se trabajará con el conjunto de datos *duser* (Gómez et al. (2000)) que contiene 940 observaciones y está disponible en el paquete *FHtest*. Dicho conjunto de datos contiene los tiempos desde que un usuario usa por primera vez una serie de drogas inyectables hasta la seroconversión del VIH. Los datos proceden de la unidad de desintoxicación del hospital universitario *Germans Trias i Pujol* de Badalona y las variables en estudio son:

- *left*: extremo izquierdo de cada uno de los intervalos en estudio expresado en meses.
- *right*: extremo derecho de cada uno de los intervalos en estudio expresado en meses.
- *zper*: período del calendario que se denota por un valor del 1 al 4. El primer período va desde el año 1972 al 1980, el segundo desde 1981 a 1985, el tercero desde 1986 hasta 1991 y el último desde 1992 hasta 1997.
- *zgen*: género del usuario (0 = Hombre y 1 = Mujer)
- *age*: edad a la que comienza el usuario a consumir drogas inyectables.

La variable censurada por intervalos es el tiempo (en meses) desde el inicio del consumo de drogas inyectables hasta la seroconversión, síntoma que puede utilizarse como indicador de la infección del sida. A continuación, se cargan los datos y se muestran algunos de ellos:

```
data("duser", package = "FHtest")
```

<i>left</i>	<i>right</i>	<i>zper</i>	<i>zgen</i>	<i>age</i>
201	9999	1	0	20
0	211	1	1	12
83	9999	2	0	35
0	106	2	1	17
47	61	2	0	17
19	28	3	0	25
35	9999	3	1	14
18	23	4	0	49
24	9999	4	1	31

Tabla 4.24: Muestra de datos del conjunto *duser* del paquete *FHtest*

Los tiempos sin censura se caracterizan por la coincidencia de las variables *right* y *left*; los tiempos censurados a la izquierda son aquellos en los que la variable *left* toma el valor 0 y los tiempos censurados a la derecha son aquellos en los que la variable *right* toma el valor 9999. Por último, aquellos tiempos que no verifiquen ninguna de las condiciones expuestas se considerarán que están censurados por intervalos. A continuación, se convertirán los valores 9999 a *Inf* para facilitar el estudio:

```
duser <- duser %>%
  mutate(right = ifelse(right == 9999, Inf, right))
```

Una vez preparados los datos, se procede a comparar los tiempos de supervivencia según el género del usuario, es decir, se debe resolver el siguiente contraste:

$$H_0 : S_{Hombre}(t) = S_{Mujer}(t) \quad \forall t > 0$$

$$H_1 : S_{Hombre}(t) \neq S_{Mujer}(t) \quad \text{para algún } t > 0$$

Se tomará como ρ el valor 1 para dar más importancia a las diferencias tempranas y se añadirán los parámetros *Lin* y *Rin* para considerar intervalos de tiempo cerrados. Además, se aumentará el número máximo de iteraciones que utiliza el programa internamente para estimar las funciones de supervivencia por el método EM debido a que surgen problemas de convergencia cuando se mantiene el valor por defecto de 10000 iteraciones. Para ello, se añadirá el argumento *icontrol*:

```
FHtestics(Surv(left, right, type = "interval2") ~ zgen, data = duser,
  rho = 1, Rin = TRUE, Lin = TRUE,
  icontrol = icfitControl(maxit = 50000))
```

```
##
## Two-sample test for interval-censored data
##
## Parameters: rho=1, lambda=0
## Distribution: score vector approach
##
## Data: Surv(left, right, type = "interval2") by zgen
##
##          N    O-E
## zgen=0 759 -12.4
## zgen=1 181  12.4
##
## Statistic Z= 2.3, p-value= 0.0214
## Alternative hypothesis: survival functions not equal
```

El p-valor obtenido es 0.0214 , por lo que existen evidencias en contra de la hipótesis nula y por tanto, no se puede considerar que los tiempos de supervivencia hasta la seroconversión sean iguales para los hombres y las mujeres. Ante el resultado obtenido, podría plantearse resolver un contraste unilateral como el siguiente:

$$H_0 : S_{Hombre}(t) = S_{Mujer}(t) \quad \forall t > 0$$

$$H_1 : S_{Hombre}(t) \geq S_{Mujer}(t) \quad \forall t > 0$$

El contraste anterior se resuelve de la siguiente forma:

```
FHtestics(Surv(left, right, type = "interval2") ~ zgen, data = duser,
  Rin = TRUE, Lin = TRUE, rho = 1, alternative = "decreasing")
```

```
##
## Two-sample test for interval-censored data
##
## Parameters: rho=1, lambda=0
## Distribution: score vector approach
##
## Data: Surv(left, right, type = "interval2") by zgen
##
##          N    O-E
## zgen=0 759 -12.4
## zgen=1 181  12.4
##
## Statistic Z= 2.3, p-value= 0.0107
## Alternative hypothesis: decreasing survival functions (group=1 has earlier event times)
```

Dado que el p-valor es significativo, se concluye que se rechaza la hipótesis nula y que por tanto puede considerarse que la curva de supervivencia de los hombres es mayor que la de las mujeres. Se adjunta una gráfica de las estimaciones de las correspondientes funciones de supervivencia de hombres y mujeres consumidores de drogas:

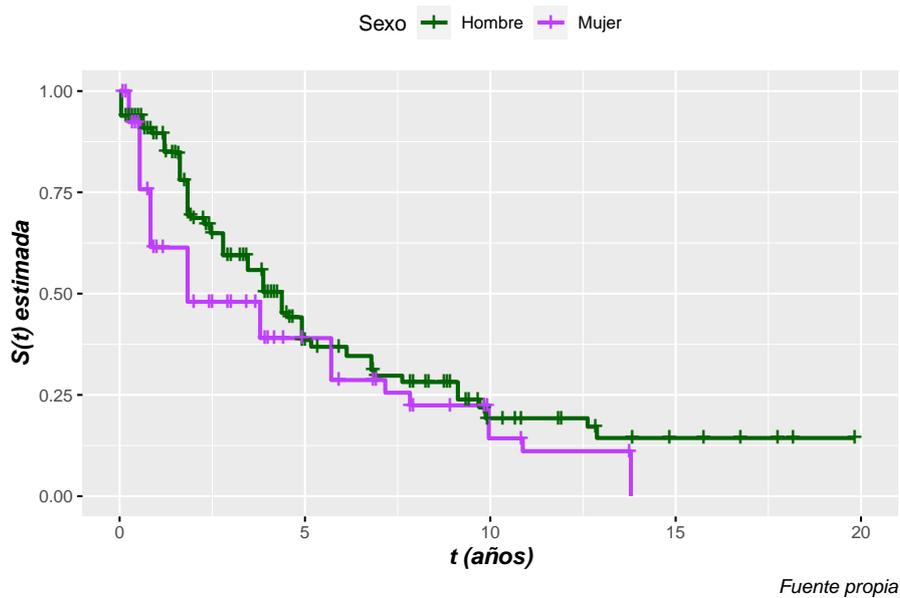


Figura 4.15: Curvas de supervivencia del conjunto de datos *duser* del paquete *FHtest* por género

Según la gráfica, la probabilidad de seroconversión durante los tres o cuatro primeros meses es mayor entre las mujeres. Después, la función de supervivencia asociada a los hombres se sitúa por encima de la asociada a las mujeres a excepción de que, al comienzo del quinto año, la curva que se sitúa por encima es la asociada a las mujeres. Es destacable también que después de 13 años, no se observa más seroconversión entre los consumidores femeninos.

Test con k poblaciones

Para ilustrar la prueba con $k > 2$ poblaciones o grupos se usará el conjunto de datos *duser* que se ha utilizado en el apartado anterior. En este caso, la variable de agrupación que se tomará es *zper* y el estudio se centrará en los consumidores de drogas masculinos. En concreto, se resolverá el siguiente contraste:

$$H_0 : S_{P1,H}(t) = S_{P2,H}(t) = S_{P3,H} = S_{P4,H} \quad \forall t > 0$$

$$H_1 : S_{P1,H}(t) \neq S_{P2,H}(t) \neq S_{P3,H} \neq S_{P4,H} \quad \text{para algún } t > 0$$

A la hora de aplicar dicho contraste en R, debe recordarse que la variable de agrupación no puede ser numérica y para trabajar solo con los datos de usuarios masculinos se añade el argumento *subset* junto con la condición *zgen == 0*:

```
FHtestics(Surv(left, right, type = "interval2") ~ as.character(zper),
          data = duser, subset = (zgen == 0), Lin = TRUE, Rin = TRUE,
          rho = 5)
```

```
##
## K-sample test for interval-censored data
##
## Parameters: rho=5, lambda=0
## Distribution: score vector approach
##
## Data: Surv(left, right, type = "interval2") by as.character(zper)
##
##              N      O-E
## as.character(zper)=1 146 -0.524
## as.character(zper)=2 300  2.632
## as.character(zper)=3 240 -0.997
## as.character(zper)=4  73 -1.111
##
## Chisq= 1.8 on 3 degrees of freedom, p-value= 0.605
## Alternative hypothesis: survival functions not equal
```

Tomando como ρ el valor 5, se obtiene un p-valor de *0.605*, por lo que no existen evidencias suficientes en contra de H_0 . De esta forma, puede admitirse que las curvas de supervivencia asociadas a cada uno de los períodos son iguales, sobre todo en las fases tempranas del estudio.

A continuación, se adjunta una gráfica de las estimaciones de las funciones de supervivencia asociadas a cada uno de los períodos en estudio para la población masculina:

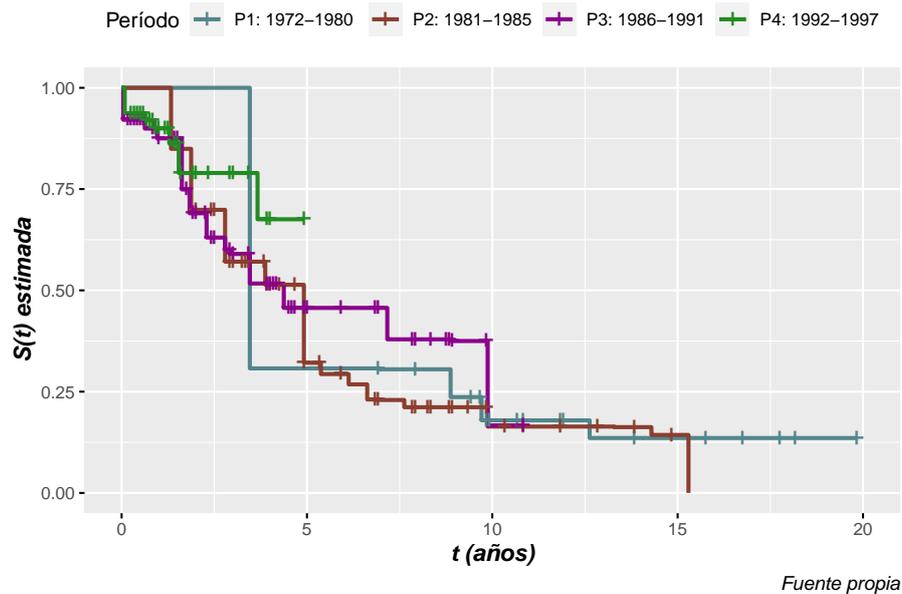


Figura 4.16: Curvas de supervivencia asociadas a cada período de tiempo del conjunto de datos *duser* del paquete *FHtest*

Se observa así que aproximadamente en el primer año de consumo, las funciones de supervivencia de los dos primeros períodos son indistinguibles. Lo mismo ocurre con los otros dos períodos del calendario. Es destacable que en el período 4 no se observa más seroconversión a partir del quinto año. Además, a largo plazo las curvas de supervivencia de los períodos 1, 2 y 3 son prácticamente iguales. Por último, en el tramo comprendido entre los 5 y 10 años, la función asociada al período 3 es mayor que la del período 1 y esta a su vez, es mayor que la del período 2.

Prueba de tendencia

Para mostrar un ejemplo de prueba de tendencia se usarán también los datos *duser*. En este caso, se tomará como variable de agrupación la variable *zper* eliminando el primero de los períodos disponibles. Además, el estudio se centrará en los datos pertenecientes a usuarios masculinos. Partiendo de los datos seleccionados, se resolverá el siguiente contraste:

$$\begin{aligned} H_0 &: S_{P2,H}(t) = S_{P3,H} = S_{P4,H} \quad \forall t > 0 \\ H_1 &: S_{P2,H}(t) \leq S_{P3,H} \leq S_{P4,H} \quad \forall t > 0 \end{aligned}$$

con al menos una desigualdad estricta.

Al igual que en el test realizado con k poblaciones, se debe añadir el argumento *subset* junto con las condiciones impuestas de los períodos y el género:

```
FHtestics(Surv(left, right, type = "interval2") ~ zper,
  data = duser, subset = (zper > 1) & (zgen == 0),
  Lin = TRUE, Rin = TRUE, alternative = "increasing")
```

```
##
## Trend FH test for interval-censored data
##
## Parameters: rho=0, lambda=0
## Distribution: score vector approach
##
## Data: Surv(left, right, type = "interval2") by zper
##
##           N      O-E
## zper=2 300  19.45
## zper=3 240 -15.26
## zper=4  73  -4.19
##
## Statistic Z= -2.5, p-value= 0.00616
## Alternative hypothesis: increasing survival functions (higher group implies later event times)
```

El p-valor obtenido es 0.00616 , por lo que existen suficientes evidencias en contra de la igualdad de curvas de supervivencia. Además, podría suponerse que en los primeros meses se cumple que la curva asociada al segundo período es menor que la curva asociada al tercer período, y esta a su vez, menor que la del cuarto período dado que se está trabajando con la familia de estadísticos $G^{\rho,\lambda}$ con $\lambda = \rho = 0$.

Tras exponer ejemplos de uso de todas las funciones disponibles en el paquete *FHtest*, se ha demostrado que dicho paquete es una herramienta útil y fácil de usar que integra y mejora métodos de otros paquetes con datos censurados a la derecha y por intervalos.

Conclusión

Tras realizar en el [Capítulo 1](#) una breve introducción histórica al tema del análisis de supervivencia, en el [Capítulo 2](#) se recogen definiciones básicas y necesarias para comprender las técnicas de comparación de curvas de supervivencia que se desarrollan en el [Capítulo 3](#). Además, en el [Capítulo 3](#) se analizan tanto técnicas generales como específicas para la existencia de datos censurados a la derecha y por intervalos. Tal y como se especificó en dicho capítulo, estas técnicas pueden extrapolarse también a otras censuras tras realizar una transformación sobre los datos originales que se quieran estudiar y analizar. De esta forma, se ha querido evidenciar la importancia de dichas técnicas para extraer toda la información que encierran los datos censurados y que no se debe eludir. Cabe resaltar también, el hecho de que el problema de la comparación de curvas de supervivencia con datos censurados se puede abordar mediante distintos enfoques según el tipo de censura que posean los datos.

Por último, en el [Capítulo 4](#) se enumeran distintas librerías o paquetes de R que son útiles para realizar un completo análisis de supervivencia. Se parte del paquete más general *survival* con el que se puede obtener entre otras utilidades, la estimación no paramétrica de la función de supervivencia asociada a un conjunto de datos y a continuación, se van presentando otros paquetes interesantes para un análisis más exhaustivo de conceptos y técnicas expuestas en el trabajo. Además, es destacable la función *ggsurvplot()* del paquete *survminer* que es capaz de mostrar gráficas más elaboradas como la [Figura 4.4](#) en la que se puede consultar tanto el p-valor asociado al test de comparación de curvas de supervivencia como una tabla con los individuos en riesgo. Para finalizar el capítulo, se presenta uno de los paquetes más eficientes para trabajar con datos censurados a la derecha y por intervalos. Dicho paquete se llama *FHtest* y fue desarrollado por [Oller y Langohr \(2020\)](#). Además, a lo largo del [Capítulo 4](#) se van utilizando algunas funciones de cada uno de los paquetes en estudio y se van enumerando y explicando conjuntos de datos que han sido utilizados en estudios reales, sobre todo en el ámbito de la medicina.

De esta forma, se evidencia la cantidad de herramientas disponibles en R para el análisis de supervivencia y la utilidad de todas ellas para abordar el objetivo de comparar curvas de supervivencia. Téngase siempre presente que según la tipología del estudio que se quiera realizar y la naturaleza de los datos implicados en el mismo, el investigador se decantará por un recurso concreto de R.

Bibliografía

- O.O. Aalen y S. Johansen. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978.
- A. Allignol y A. Latouche. *CRAN Task View: Survival Analysis*. R Foundation for Statistical Computing, Version 2021-04-29. Disponible en: <https://cran.r-project.org/web/views/Survival.html>.
- P.K. Andersen, Ø. Borgan, N.L. Hjort, E. Arjas, J. Stene, y O. Aalen. Counting Process Models for Life History Data: A Review [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 97–158, 1985.
- J. Berkson y R.R. Gage. Calculation of Survival Rates for Cancer. *Proceedings of Staff Meetings, Mayo Clinic*, (25):270–286, 1950.
- K.J. Berry, J.E. Johnston, y P.W. Mielke Jr. Permutation Methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6):527–542, 2011. doi: <https://doi.org/10.1002/wics.177>.
- P. Biecek, S. Fabian, A. Kassambara, y M. Kosinski. *Survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.9, 2021. Disponible en: <https://CRAN.R-project.org/package=survminer>.
- Bioconductor. "Using *Bioconductor* - Install R". Disponible en <https://bioconductor.org/install/>, 2021a.
- Bioconductor. "Software Packages - I cens". Disponible en <http://www.bioconductor.org/packages/release/bioc/html/Icens.html>, 2021b.
- Britannica, T. Editors of Encyclopaedia (2020, April 20). "John Graunt. Encyclopedia Britannica.". Disponible en <https://www.britannica.com/biography/John-Graunt>.
- M.L. Calle, G. Gómez, K. Langohr, y R. Oller. Tutorial on Methods for Interval-Censored Data and their Implementation in R. *Statistical Modelling*, 9(4):259–297, 2009. doi: <https://doi.org/c4mkev>.
- L. Camilleri. History of Survival Analysis. *Allied Newspapers Ltd.*, 2019.
- T.J. Cleophas y A.H. Zwinderman. *Statistics Applied to Clinical Studies*. Springer Science & Business Media, 2012.
- A. Clipson, R. Epstein, T. Hestenberg, S. Monaghan, y D.S. Moore. *Introduction to Statistical Learning*. W. H. Freeman and Company, New York, 2003.

-
- D.R. Cox y D. Oakes. *Analysis of Survival Data*, volume 21. CRC press, 1984.
- S.J. Cutler y F. Ederer. Maximum Utilization of the Life Tables Method in Analyzing Survival. *Journal of Chronic Diseases*, (8):699–712, 1958. doi: <https://doi.org/bz2h69>.
- C. Dardis. *survMisc: Miscellaneous Functions for Survival Data*. R package version 0.5.5, 2018. Disponible en: <https://CRAN.R-project.org/package=survMisc>.
- M.H. Dehghan y T. Duchesne. A Generalization of Turnbull's Estimator for Nonparametric Estimation of the Conditional Survival Function with Interval-Censored Data. *Lifetime data analysis*, 17(2):234–255, 2011. doi: <https://doi.org/bxjh2s>.
- M.H. Dehghan, T. Duchesne, y S. Baillargeon. *gte: Generalized Turnbull's Estimator*. R package version 1.2-2, 2015. Disponible en: <https://CRAN.R-project.org/package=gte>.
- Instituto Nacional del Cáncer. "Diccionario de Cáncer". Disponible en <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer>.
- J.M. Doménech. Una aplicación del análisis de la supervivencia en ciencias de la salud. *Anuario de psicología/The UB Journal of psychology*, (55):109–142, 1992.
- O.J. Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: <https://doi.org/gd85vm>.
- J.H. Edmonson, T.R. Fleming, D.G. Decker, G.D. Malkasian, E.O. Jorgensen, J.A. Jefferies, M.J. Webb, y L.K. Kvols. Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma versus Minimal Residual Disease. *Cancer Treat Rep*, 63(2):241–247, 1979.
- S.H. Embury, L. Elias, P.H. Heller, C.E. Hood, P.L. Greenberg, y S.L. Schrier. Remission Maintenance Therapy in Acute Myelogenous Leukemia. *Western Journal of Medicine*, 126(4):267, 1977.
- I. Ercan y P.G. Karadeniz. Examining Tests for Comparing Survival Curves with Right Censored Data. *Stat Transit*, 18(2):311–328, 2017. doi: <https://doi.org/gddx>.
- B.S. Everitt y A. Skrondal. *The Cambridge Dictionary of Statistics*. Cambridge University Press, Fourth edition, 2010.
- L. Fahrmeir y G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Science & Business Media, 2013.
- M.P. Fay. *perm: Exact or Asymptotic permutation tests*. R package version 1.0-0.0, 2010. Disponible en: <https://CRAN.R-project.org/package=perm>.
- M.P. Fay. *interval: Weighted Logrank Tests and NPMLE for Interval Censored Data*. R package version 1.1-0.7, 2020. Disponible en: <https://CRAN.R-project.org/package=interval>.
- H. Fehske, R. Schneider, y A. Weiße. *Computational Many-Particle Physics*, volume 739. Springer, 2007.
- P. Fernández. Análisis de supervivencia. *Unidad de Epidemiología Clínica*, 1995.
-

-
- T. Fernández, M. Ruiza, y E. Tamaro. "Biografía de Edmund Halley - Enciclopedia biográfica". Disponible en <https://www.biografiasyvidas.com/biografia/h/halley.html>, 2004.
- D.M. Finkelstein y R.A. Wolfe. A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data. *Biometrics*, pages 933–945, 1985. doi: <https://doi.org/crsfs5>.
- T.R. Fleming y D.P. Harrington. A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*, 69(3):553–566, 1982. doi: <https://doi.org/bxzdgg>.
- T.R. Fleming y D.P. Harrington. Nonparametric Estimation of the Survival Distribution in Censored Data. *Communications in Statistics-Theory and Methods*, 13(20):2469–2486, 1984. doi: <https://doi.org/csng36>.
- B.E. Fristedt y L.F. Gray. *A Modern Approach to Probability Theory*. Springer Science & Business Media, 2013.
- E.A. Gehan. Estimating Survival Functions from the Life Table. *Journal of Chronic Diseases*, 21(9–10):629–644, 1969. doi: <https://doi.org/dtzq3c>.
- J.E. Gentle. *Computational Statistics*, volume 308. Springer, 2009.
- J.E. Gentle, Y. Mori, y W.K. Härdle. *Handbook of Computational Statistics: Concepts and Methods*. Springer Science & Business Media, 2012.
- I. Gijbels. Censored Data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(2):178–188, 2010. doi: <https://doi.org/dt8zms>.
- P. Goel y J.P. Klein. *Survival Analysis: State of the Art*. Springer Science & Business Media, First edition, 1992.
- P. Good. *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Science & Business Media, 2013.
- M. Greenwood. The Errors of Sampling of the Survivorship Tables. Reports on Public Health and Medical Subjects. London. *His Majesty's Stationery Office*, (33), 1926. Appendix 1.
- G. Gómez y R. Oller. *A New Class of Rank Tests for Interval-Censored Data*. Harvard University Biostatistics Working Paper Series, 2008. Disponible en: <http://biostats.bepress.com/harvardbiostat/paper93/>.
- G. Gómez y R. Oller. A Generalized Fleming and Harrington's Class of Tests for Interval-Censored Data. *Canadian Journal of Statistics*, 40(3):501–516, 2012. doi: <https://doi.org/10.1002/cjs.11139>.
- G. Gómez, M.L. Calle, J.M. Egea, y R. Muga. Risk of HIV Infection as a Function of the Duration of Intravenous Drug Use: a Non-parametric Bayesian Approach. *Statistics in medicine*, 19(19):2641–2656, 2000.
- T. Hothorn, H. Winell, K. Hornik, M.A. van de Wiel, y A. Zeileis. *coin: Conditional Inference Procedures in a Permutation Test Framework*. R package version 1.4-1, 2021. Disponible en: <https://CRAN.R-project.org/package=coin>.
-

- M.C. Jaramillo y C.M. Lopera. Estudio del efecto de la imputación de fallas en la estimación de la curva de supervivencia bajo censura a intervalo. *Ciencia en Desarrollo*, 8(1):21–28, 2017.
- H.W. Jones. John Graunt and his Bills of Mortality. *Bulletin of the Medical Library Association*, 33(1):3–4, 1945.
- J.D. Kalbfleisch y R.L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Second edition, 2002.
- E.L. Kaplan y P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi: <https://doi.org/gdz3gq>.
- C. Kartsonaki. Survival Analysis. *Diagnostic Histopathology*, 22(7):263–270, 2016. ISSN 1756-2317. doi: <https://doi.org/gf7fph>. Mini-Symposium: Medical Statistics.
- B.R. Kirkwood y J.A.C. Sterne. *Essential Medical Statistics*. John Wiley & Sons, 2010.
- M. Klass, H. Teicher, et al. The Central Limit Theorem for Exchangeable Random Variables without Moments. *The Annals of Probability*, 15(1):138–153, 1987.
- J.P. Klein y M.L. Moeschberger. *Survival Analysis Techniques for Censored and Truncated Data*. Springer, 1997.
- J.P. Klein y M.L. Moeschberger. *KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis*. R package version 0.1-5, 2012. Disponible en: <https://CRAN.R-project.org/package=KMsurv>.
- D.G. Kleinbaum y M. Klein. *Survival Analysis: A Self-Learning Text*. Springer, Third edition, 2012.
- A. Knezevic y S. Patil. "Combination Weighted Log-rank Tests for Survival Analysis with Non-Proportional Hazards". Disponible en <https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/5062-2020.pdf>, 2020.
- T.L. Lai y H. Xing. *Statistical Models and Methods for Financial Markets*. Springer, 2008.
- J.A. Laurie, C.G. Moertel, T.R. Fleming, H.S. Wieand, J.E. Leigh, J. Rubin, G.W. McCormack, J.B. Gerstner, J.E. Krook, y J. Malliard. Surgical Adjuvant Therapy of Large-Bowel Carcinoma: an Evaluation of Levamisole and the Combination of Levamisole and Fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. *Journal of Clinical Oncology*, 7(10):1447–1456, 1989.
- E.T. Lee y J. Wang. *Statistical Methods for Survival Data Analysis*, volume 476. John Wiley & Sons, 2003.
- C.L. Loprinzi, J.A. Laurie, H.S. Wieand, J.E. Krook, P.J. Novotny, J.W. Kugler, J. Bartel, M. Law, M. Bateman, y N.E. Klatt. Prospective Evaluation of Prognostic Variables from Patient-Completed Questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, 12(3):601–607, 1994.

-
- P.L. Luque. Escribir un Trabajo Fin de Estudios con R Markdown. Disponible en <http://destio.us.es/calvo>, 2017.
- P.L. Luque. Cómo crear Tablas de información en R Markdown. Disponible en <http://destio.us.es/calvo>, 2019.
- M. Maathuis. *MLEcens: Computation of the MLE for bivariate (interval) censored data*. R package version 0.1-4, 2013. Disponible en: <https://CRAN.R-project.org/package=MLEcens>.
- N. Mantel, N.R. Bohidar, y J.L. Ciminera. Mantel-Haenszel Analyses of Litter-Matched Time-to-Response Data, with Modifications for Recovery of Interlitter Information. *Cancer Research*, 37(11):3863–3868, 1977.
- C.M. Martínez. Bonferroni's Method to Compare Survival Curves with Recurrent Events. *Ingeniería Industrial. Actualidad y Nuevas Tendencias*, 3(10):105–114, 2013.
- D.F. Moore. *Applied Survival Analysis Using R*. Springer, 2016.
- W. Nelson. Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1(1):27–52, 1969. doi: <https://doi.org/gfgm>.
- R. Oller y K. Langohr. FHtest: An R Package for the Comparison of Survival Curves with Censored Data. *Journal of Statistical Software*, 81(15):1–25, 2017. doi: <https://doi.org/fz35>.
- R. Oller y K. Langohr. *FHtest: Tests for Right and Interval-Censored Survival Data Based on the Fleming-Harrington Class*. R package version 1.5., Barcelona, España, 2020. Disponible en: <https://CRAN.R-project.org/package=FHtest>.
- M. Pešta. "Survival Data Analysis". Disponible en <https://www2.karlin.mff.cuni.cz/~pesta/NMFM404/survival.html>, 2021.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. Disponible en: <https://www.R-project.org/>.
- J. Ramírez-Montoya. Comparación de estimadores no paramétricos frente a los paramétricos para la función de confiabilidad. *Ingeniería Investigación y Desarrollo*, 15(2): 15–24, 2015.
- P. Rebas. Conceptos básicos del análisis de supervivencia. *Cirugía española*, 78(4): 222–230, 2005. ISSN 0009-739X. doi: <https://doi.org/c87m3p>.
- B. Ripley, B. Venables, D.M. Bates, K. Hornik, A. Gebhardt, y D. Firth. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-54, 2021. Disponible en: <https://CRAN.R-project.org/package=MASS>.
- RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015. Disponible en: <http://www.rstudio.com/>.
- P. Sedgwick. Kaplan-Meier Survival Analysis: Types of Censored Observations. *Bmj*, 347, 2013. doi: <https://doi.org/gh7rp5>.
-

- A.N. Shiryaev. "Martingale. Encyclopedia of Mathematics.". Disponible en <http://encyclopediaofmath.org/index.php?title=Martingale&oldid=49256>.
- NCSS Statistical Software. "Kaplan-Meier Curves (Logrank Tests)". Disponible en <https://www.ncss.com/software/ncss/ncss-documentation/>.
- M. Tableman y J.S. Kim. *Survival Analysis Using S: Analysis of Time-to-Event Data*. CRC press, 2003.
- T.M. Therneau. *A Package for Survival Analysis in R*. R package version 3.2-11, 2021. Disponible en: <https://CRAN.R-project.org/package=survival>.
- A. Tillander, A. Karlsson, J. Zetterqvist, P. Strom, B. Delcoigne, y M. Clements. *biostat3: Utility Functions, Datasets and Extended Examples for Survival Analysis*. R package version 0.1.5, 2020. Disponible en: <https://CRAN.R-project.org/package=biostat3>.
- B.W. Turnbull. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association*, 69(345):169–173, 1974. doi: <https://doi.org/gg2nc6>.
- B.W. Turnbull. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295, 1976. doi: <https://doi.org/ggqzts>.
- P.A. Volberding, S.W. Lagakos, J.M. Grimes, D.S. Stein, J. Rooney, T. Meng, M.A. Fischl, A.C. Collier, J.P. Phair, M.S. Hirsch, et al. A Comparison of Immediate with Deferred Zidovudine Therapy for Asymptomatic HIV-Infected Adults with CD4 Cell Counts of 500 or More per Cubic Millimeter. *New England Journal of Medicine*, 333(7):401–407, 1995. doi: <https://doi.org/dk6b89>.
- H. Wickham y Rstudio. *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.3.1, 2021. Disponible en: <https://CRAN.R-project.org/package=tidyverse>.
- H. Wickham, W. Chang, L. Henry, T.L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani, y D. Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.3, 2020. Disponible en: <https://CRAN.R-project.org/package=ggplot2>.