Modelos de Supervivencia

Sofía Villers Gómez Carlos Fernando Vásquez Guerra
Luis Angel Ramirez Teodoro

Índice general

Prefacio

Primera edición del bookdown *Modelos de Supervivencia* para su uso en la materia Análisis de Supervivencia y Series de tiempo y sus relacionadas impartidas por los autores, así como para aquellos estudiantes que deseen adquirir el conocimiento pertinente de tal tópico.

Objetivos

- Otorgar un material electrónico de calidad con el contenido referente al Análisis de Supervivencia como un esfuerzo de los autores para lograr un proceso de aprendizaje autodidacta por parte del alumno y así optimizar el tiempo, tanto de los profesores, como el de los alumnos.
- Plasmar las bases teóricas de esta rama de la estadística con el uso de ejemplos y contenido visual para un mejor entendimiento de cada subtema que se trate.

Estructura

Este libro se compone de diez diferentes capítulos comenzando con los tipos de datos con el que se puede encontrar el lector en un estudio de supervivencia, dado esto, en el capítulo 3 se comienza con el estudio de las distintas funciones de supervivencia, así como las funciones que se deriven de esta, para los casos discreto y continuo. En el capítulo 4 se desglosa la teoría correspondiente a la obtención de parámetros poblacionales, como lo son la media y la varianza; para así en el capítulo 5 tratar modelos paramétricos comunes donde se asume una distribución para los datos a tratar y se obtienen las funciones vistas en los capítulos previos. El capítulo 6 abarca lo correspondiente a la obtención de la función de verosimilitud en datos censurados y truncados. Respecto a los capítulos 7 y 8 se da la teoría, junto a varios ejemplos, para el análisis no paramétrico en un análisis de supervivencia para que en el capítulo 9 se establezcan las pruebas de hipótesis pertinentes. Finalmente en el capítulo 10 se considera el modelo de riesgos proporcionales y se estudian las implicaciones sobre algunos tópicos de los capítulos anteriores.

Se recomienda que la consulta de los capítulos se realice de acuerdo al índice, ya que a medida que se avanza en índice, se asume el conocimiento de los capítulos previos.

Detalles técnicos

Para la creación de este material se hizo uso de varios sistemas de software como LaTeX y CSS para el diseño de ciertos elementos. Todos los cálculos y gráficas fue creado con el lenguaje de programación `R` ya sea con el uso del paquete `base` o algún otro de los paquetes que se mencionan a continuación.

```
R version 3.6.2 (2019-12-12)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Catalina 10.15.6

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
```

```
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] survival_3.1-8      kableExtra_1.2.1    knitr_1.28
 [4] actuar_2.3-3        futurevisions_0.1.1 shape_1.4.4
 [7] latex2exp_0.4.0     patchwork_1.0.0     devtools_2.2.2
[10] usethis_1.5.1       forcats_0.5.0       stringr_1.4.0
[13] dplyr_1.0.1         purrr_0.3.4         readr_1.3.1
[16] tidyr_1.1.0         tibble_3.0.3        ggplot2_3.3.2
[19] tidyverse_1.3.0

loaded via a namespace (and not attached):
 [1] httr_1.4.2        pkgload_1.1.0     splines_3.6.2     viridisLite_0.3.0
 [5] jsonlite_1.7.1    modelr_0.1.6      assertthat_0.2.1  cellranger_1.1.0
 [9] yaml_2.2.1        remotes_2.1.1     sessioninfo_1.1.1 lattice_0.20-40
[13] pillar_1.4.6      backports_1.1.10  glue_1.4.2        digest_0.6.25
[17] rvest_0.3.5       colorspace_1.4-1  Matrix_1.2-18     htmltools_0.4.0
[21] pkgconfig_2.0.3   broom_0.7.0       haven_2.2.0       bookdown_0.20
[25] webshot_0.5.2     scales_1.1.1      processx_3.4.4    generics_0.0.2
[29] ellipsis_0.3.1    withr_2.3.0       cli_2.0.2         magrittr_1.5
[33] crayon_1.3.4      readxl_1.3.1      memoise_1.1.0     evaluate_0.14
[37] ps_1.3.4          fs_1.5.0          fansi_0.4.1       xml2_1.2.2
[41] pkgbuild_1.1.0    tools_3.6.2       prettyunits_1.1.1 hms_0.5.3
[45] expint_0.1-6      lifecycle_0.2.0   munsell_0.5.0     reprex_0.3.0
[49] callr_3.4.4       compiler_3.6.2    rlang_0.4.7       grid_3.6.2
[53] rstudioapi_0.11   rmarkdown_2.3     testthat_2.3.2    gtable_0.3.0
[57] DBI_1.1.0         R6_2.4.1          lubridate_1.7.9   rprojroot_1.3-2
[61] desc_1.2.0        stringi_1.4.6     Rcpp_1.0.5        vctrs_0.3.4
[65] dbplyr_1.4.2      tidyselect_1.1.0  xfun_0.12
```

Este libro fue escrito con bookdown usando RStudio.

Esta versión fue escrita con:

```
Finding R package dependencies ... Done!
```

setting	value
version	R version 3.6.2 (2019-12-12)
os	macOS Catalina 10.15.6
system	x86_64, darwin15.6.0
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	America/Mexico_City
date	2020-11-02

Licencia

- **Remix**—remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- **Attribution**—You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **ShareAlike**—If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
- **No additional restrictions**—You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Parte I

Todo es cuestion de tiempo

Motivación

En los dos cursos anteriores de estadística se ha trabajado con datos (muestras aleatorias) exactos, es decir, datos que se conocen en su totalidad y que, con base en ellos, hemos realizado estimaciones sobre los parámetros poblacionales. Sin embargo, puede suceder que los datos se conozcan parcialmente, o bien, se desconozcan; esto suele suceder en estudios que tienen limitados recursos y tiempo para llevarse a cabo, por ejemplo: estudios sobre la efectividad de un tratamiento clínico, la reaparición de cierta enfermedad en pacientes, confiabilidad industrial, etcétera.

Lo anterior sugiere plantearnos algunas preguntas interesantes: *¿Cómo realizamos estimaciones con datos parciales?, ¿Se puede ajustar algún modelo paramétrico con datos parciales?, ¿Podemos comparar poblaciones con datos parciales?*. Éstas y otras cuestiones abordaremos a lo largo del curso de *Modelos de Supervivencia y Series de Tiempo*. En general, en la primer parte del curso (Análisis de supervivencia) se establecerán las bases teóricas para el tratamiento de datos parciales en modelos conocidos (paramétricos), y además estudiaremos modelos No paramétricos para este tipo de datos.

Por otro lado, responder preguntas como: *¿Cuál será el precio de las acciones de Facebook para el último bimestre del 2020?, ¿Cuál será el nivel de partículas contaminantes en la CDMX para noviembre de 2020?, ¿Cuál será la capacidad de un procesador intel para el año 2021?*, puede parecer, en primera instancia, una tarea complicada. Si bien no tenemos una "bola mágica" con la que podamos adivinar el futuro, disponemos de ciertos procesos estocásticos llamados *Series de Tiempo*, cuyo objetivo principal es el *pronóstico*; estos se abordarán en la segunda parte del curso, y por ende en un próximo bookdown.

Análisis de Supervivencia

El análisis de supervivencia se basa en el estudio del **tiempo**, en la ocurrencia de un **evento**. El término supervivencia se debe a que en las primeras aplicaciones de este método de análisis se utilizaba como **evento** la muerte de un paciente; tradicionalmente el análisis de supervivencia se ha asociado al análisis de datos en ensayos médicos.

El **tiempo de superviviencia o falla** se define como el tiempo transcurrido desde el estado inicial hasta la ocurrencia de un evento dado. Por ejemplo, en un estudio que consiste en observar la remisión de cierta enfermedad en pacientes, se puede definir el tiempo de falla como el tiempo en el que tarda en reaparecer la enfermedad en los pacientes. Otros ejemplos de tiempo de falla son: los tiempos que toman los individuos para completar tareas específicas en experimentación psicológica, tiempos en los que tardan ciertas máquinas industriales en descomponerse, la longitud de trayectorias sobre una placa fotográfica en física de partículas.

Para determinar el tiempo de falla de forma precisa, hay tres requerimientos: un tiempo de origen que debe ser definido, una escala para medir el paso del tiempo que debe ser congruente al problema y finalmente, el significado de falla que debe ser completamente claro. Frecuentemente la escala para medir el paso de tiempo es el tiempo reloj (tiempo real), sin embargo hay otras posibilidades, como el kilometraje en un auto o el uso operacional de un sistema. Evidentemente, se pide que los tiempos de falla sean **No negativos**.

Resulta de interés conocer (si es posible) la distribución de los tiempos de falla. En general los estudios que hemos ejemplificado anteriormente tienen limitaciones para llevarse a cabo, en consecuencia se establece algún periodo fijo de observación; es en este punto en el que pueden surgir *datos parcialmente conocidos*.

A continuación se muestra con ejemplos algunas de las diferentes formas como se pueden presentar los datos de supervivencia. Todos los ejemplos tienen un tiempo de origen, una escala de medición y una definición de falla.

1. Datos de supervivencia de pacientes psiquiatricos.

Género	Edad de admisión	Tiempo de seguimiento
Femenino	51	1
Femenino	58	1
Femenino	55	2
Masculino	21	30^+
Femenino	25	32
Masculino	19	28
Masculino	24	33^+

En esta forma de presentar los datos $^+$ significa que la observación es censurada, es decir no se observo más allá de los 30 o 33 años. Las observaciones censuradas se verán en la siguiente sección.

2. Tiempos de infección (en meses) de pacientes en diálisis con diferentes procedimientos de cateterización.

Colocación de catéter de forma quirúrgica

Tiempos de Infección: 1.5,3.5,4.5,4.5,5.5,8.5,8.5,9.5,10.5,11.5,15.5,16.5,18.5,23.5,26.5

Observaciones Censuradas:2.5,2.5,3.5,3.5,3.5,4.5,5.5,6.5,6.5,7.5,7.5,7.5,7.5,8.5,9.5

Colocación de catéter de forma percutánea

Tiempos de Infección: 0.5,0.5,0.5,0.5,0.5,0.5,2.5,2.5,3.5,6.5,15.5

Observaciones Censuradas:0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,1.5,1.5,1.5,1.5,2.5

3. Consumo de la marihuana en la preparatoria

Edad	Total de observaciones	Aún no la han probado	Empezaron a fumar a edad más temprana
10	4	0	0
11	12	0	0
12	19	2	0
13	24	15	1
14	20	24	2

Los ejemplos anteriores son extracciones de ejemplos en el libro (Klein and Moeschberger, 2006).

Capítulo 1

Datos Censurados

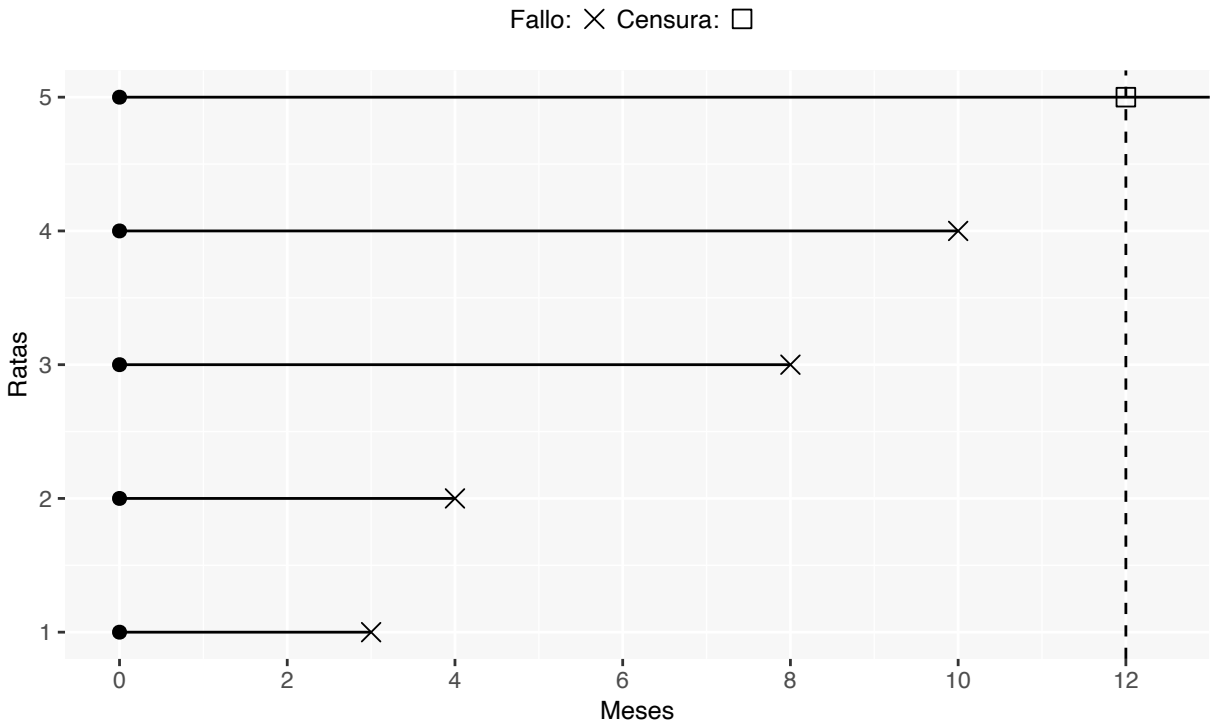
Ocurren cuando el tiempo de falla se conoce sólo en cierto periodo de tiempo. Por ejemplo, un estudio consiste en observar la muerte (tiempo de falla) de pacientes en un periodo establecido, algunos de ellos pueden estar vivos todavía al final del periodo de estudio. Los tiempos de supervivencia exactos de estos sujetos son desconocidos, pero tenemos información parcial. Estas son llamadas **observaciones censuradas**, y se denotan por +.

1.1. Censura por la Derecha Tipo I

Nos referimos a este tipo de censura cuando el estudio se termina a un tiempo fijo predeterminado (tiempo de censura) independientemente del tamaño de la muestra (número total de individuos en el estudio).

Ejemplo

Un investigador de la facultad de ciencias realiza un estudio en 5 ratas de laboratorio y ha determinado que la duración de este será de un año y medirá el tiempo (en meses) en que cada rata fallece . Al cabo del año se obtuvieron los siguientes resultados: $3, 4, 8, 10, 12^{+}$, esto quiere decir que una rata murió a los tres meses iniciado el estudio, otra lo hizo en 4 meses, etcétera; la quinta rata permaneció con vida hasta el final del estudio, es decir, no presentó la falla en el tiempo predeterminado de observación por lo tanto es un dato censurado.

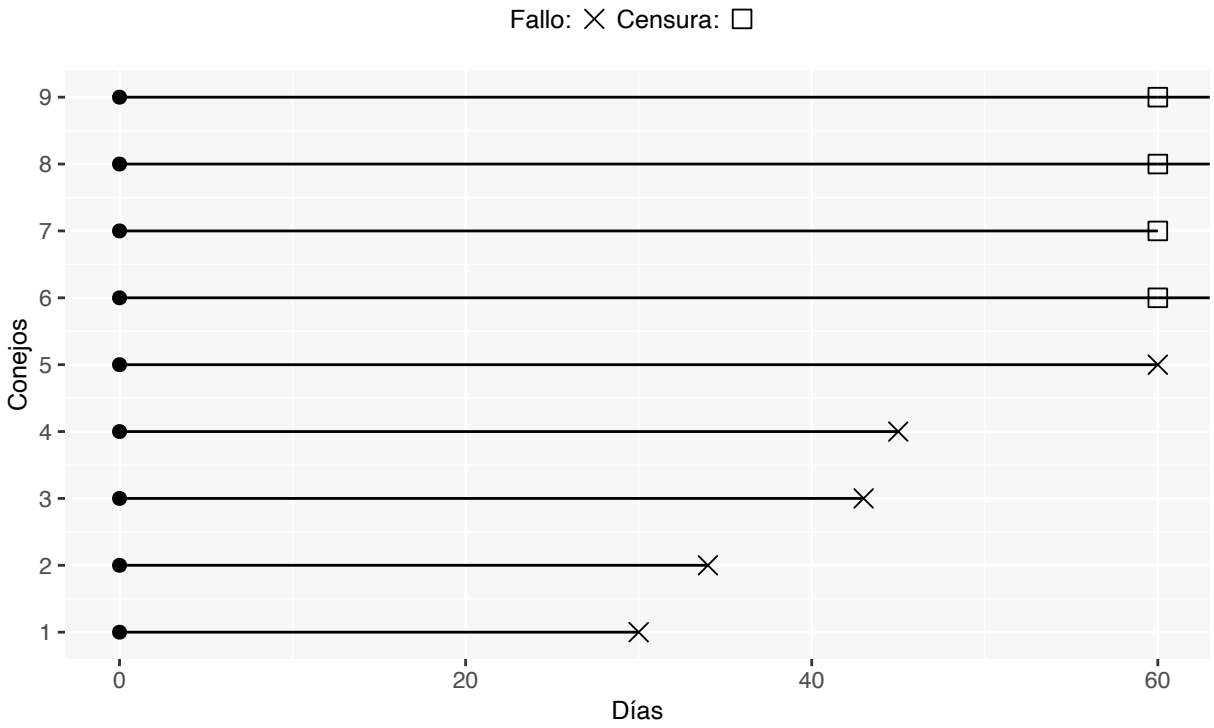


1.2. Censura por la Derecha Tipo II

Ocurre cuando el estudio continua hasta que se presenta la falla de los primeros $r < n$ individuos. Donde r es el número de individuos predeterminado a observar y n es el número total de individuos en el estudio. En este tipo de censura hay dependencia del tamaño de muestra y de las fallas que se observen.

Ejemplo

Las limitaciones económicas para la investigación científica han hecho que un especialista en cancerología tome la decisión de observar los tiempos (en días) de sólo 5 conejos hasta que desarrollen un tumor, de un total de 9. Al final se obtuvo lo siguiente: $30, 34, 43, 45, 60, 60^+, 60^+, 60^+, 60^+$, ¿Cómo interpretaría usted los resultados obtenidos por el especialista?



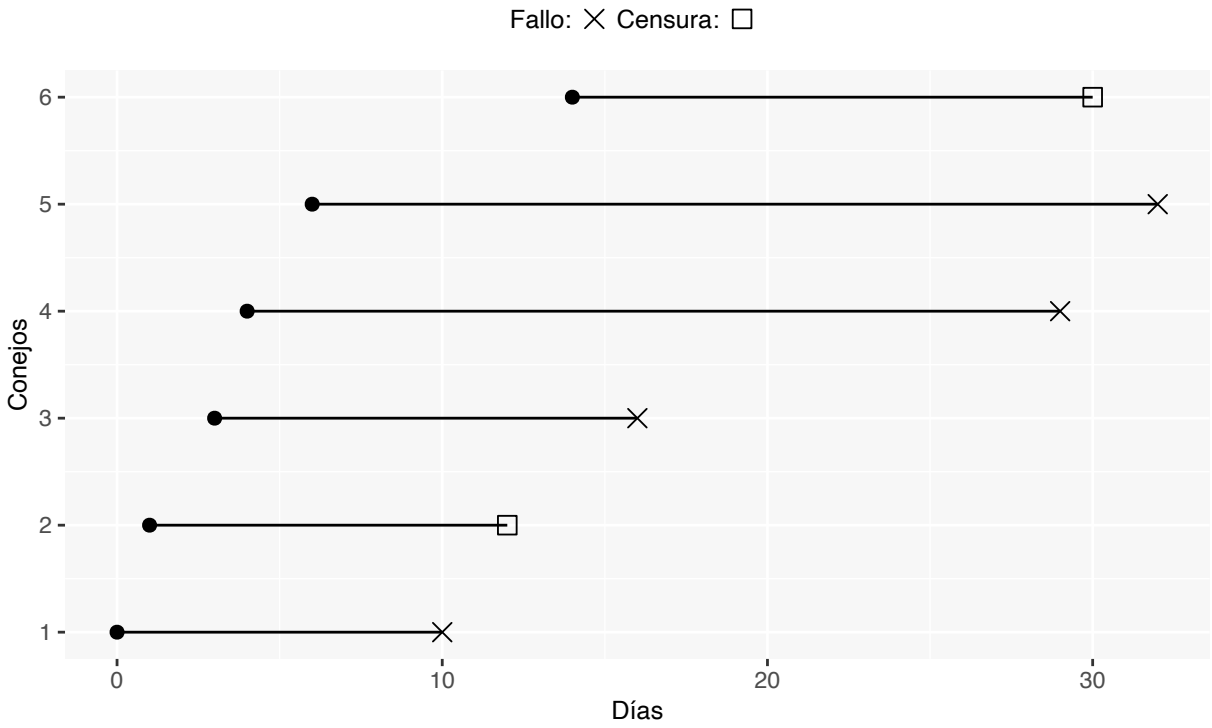
1.3. Censura por la Derecha Tipo III

Llamada también censura aleatoria. Se denomina así porque el tiempo de censura lo determina un fenómeno aleatorio, que tiene lugar durante la consecución del estudio e impide seguir con la observación del individuo hasta el tiempo final; es decir que T_i es una variable aleatoria. En general, la censura aleatoria surge cuando los individuos salen del estudio sin presentar la falla por razones **no** controladas por el investigador. Cuando sucede tal censura, se dice que se tiene una *análisis de riesgos competitivos.*

Finalmente, si el mecanismo de censura aleatoria es *dependiente* de los tiempos de falla, se dice que este es una *censura informativa*, ya que otorga información a los tiempos de falla; en caso contrario es *no informativa.*

Ejemplo

Algunos ejemplos son la migración a otra ciudad o los casos donde el paciente se retira del estudio o muere por alguna causa ajena al evento de interés. La siguiente gráfica muestra un ejemplo de datos con censura aleatoria.



1.4. Censura por la Izquierda

Las observaciones censuradas por la izquierda serán aquellas en las que el evento de interés ha tenido lugar antes del punto de inicio del estudio. Así, el tiempo de censura en este caso será el tiempo de inicio del estudio.

Ejemplo

Un ejemplo de este tipo de censura son los tiempos al primer uso de marihuana. Como se observa en el ejemplo en la introducción de este capítulo, el levantamiento de la información se hace sobre alumnos de preparatoria y hay casos en los que el inicio del consumo es a una edad previa al levantamiento de los datos. Entonces aquellos jóvenes cuyo primer uso de marihuana es a una edad previa a la registrada en ese momento, serán consideradas como observaciones censuradas por la izquierda.

Frecuentemente si un estudio tiene censura por la izquierda, entonces se tendrá doble censura, por lo que se denota como C_i al tiempo antes del cual algunos individuos presentan el evento de interés y C_d al tiempo después del cual algunos de los individuos presentan el evento de interés.

1.5. Censura por Intervalos

Ocurre cuando los individuos en el estudio son monitoreados intermitentemente en momentos discretos de tiempo, de modo que, es posible que el suceso de estudio haya tenido lugar en un tiempo entre dos de las mediciones.

Si el individuo i no ha presentado el evento de interés al fin del tiempo I_i pero a la siguiente observación D_i sí presentó el evento, entonces es una falla censurada en el intervalo (I_i, D_i).

Ejemplo

Un investigador lleva a cabo un estudio en ratas diseñado para evaluar los efectos de dietas ricas en vegetales en el riesgo de cáncer de mama. El tumor mamario es inducido con una única dosis de DMBA

al inicio del estudio. 6 semanas después de la administración del DMBA, cada rata es examinada una vez a la semana por 14 semanas y el tiempo en que el tumor es palpable es registrado. Si una rata presenta tumor en la quinta revisión, entonces lo que se sabe es que el tumor en la cuarta revisión no era palpable y por lo tanto este debió presentarse entre la cuarta y quinta revisión dando como resultado que esa observación esta censurada en el intervalo (semana 10, semana 11).

Capítulo 2

Datos Truncados

El truncamiento tiene lugar cuando sólo aquellos sujetos que manifiestan el evento dentro de una ventana observacional (U, V) son observados, del resto no se realiza ningún seguimiento y, por tanto, no se obtiene información sobre ellos (no hay información parcial). Las observaciones t_i tales que $t_i < U$, serán observaciones truncadas por la izquierda y aquellas que $t_i > V$ serán observaciones truncadas por la derecha.

Ejemplos de Datos Truncados

Deseamos medir la supervivencia de adultos mayores de 60 años. Entonces necesitamos que los individuos tengan al menos 60 años para que sean considerados en el estudio, por lo que el estudio está truncado por la izquierda. Otro ejemplo es si deseáramos medir la distancia de la tierra a las estrellas, este sería truncado por la derecha ya que no se puede ver más allá de un límite.

Es importante mencionar que la diferencia entre *truncamiento* y *censura* es la información parcial disponible. En datos censurados hay información parcial, mientras que en datos truncados no.

Ejemplos del caso continuo y caso discreto

Un número grande de individuos sanos fueron enrolados en un estudio que inicio el 1/01/1970. Los individuos fueron seguidos por 30 años para estudiar la edad a la que desarrollaron cáncer de mama. Fueron sometidos a exámenes clínicos cada 3 años. Menciona si hay censura o truncamiento en:

1. Individuo sano, enrolado a los 30 años, durante el periodo de estudio nunca desarrollo cáncer de mama.
2. Individuo sano, enrolado en el estudio con 40 años de edad, se le diagnosticó cáncer de mama en el quinto examen clínico.
3. Individuo sano, enrolado con 50 años de edad, murió a los 63 años por paro cardíaco.
4. Individuo con cáncer de piel en remisión.

Parte II

Estudio paramétrico

Capítulo 3

Funciones para el Análisis de Supervivencia

Dado que el análisis de supervivencia se basa en tiempos de falla, definiremos a continuación funciones importantes que se pueden asociar a estos. Evidentemente, estamos pensando que los tiempos de falla provienen de una variable aleatoria **No negativa**[1] T la cual llamaremos *variable aleatoria del tiempo de falla*, equivalentemente *longitud de tiempo de vida futura* o *tiempo de superviviencia*. Por simplicidad la llamaremos *tiempo de supervivencia*.

T es usualmente descrita o caracterizada por cuatro funciones:

- *1.- Función de supervivencia*
- *2.- Función de densidad de probabilidad*
- *3.- Función de riesgo*
- *4.- Función de riesgo acumulado*

Estas funciones son matemáticamente equivalentes; a partir de una se derivan las otras tres. Cabe destacar que, dependiendo el caso, T puede ser una variable aleatoria continua o bien, discreta.

3.1. Función de Supervivencia

Caso continuo

La función de supervivencia $S(t)$, tanto en el caso continuo como en el discreto, se define como la probabilidad de que un individuo sobreviva más allá del tiempo t. Para el caso continuo:

$$S(t) = \mathbb{P}(T > t) = 1 - F_T(t) = \int_t^\infty f_T(u)du$$

Si tomamos la igualdad $S(t) = 1 - F_T(t)$ y derivamos en ambos lados y multiplicamos por -1, obtenemos:

$$-\frac{d}{dt}S(t) = f_T(t)$$

Las propiedades de $S(t)$ son:

1.-Es monótona no creciente.

2.- $S(t) = 1$ para $t = 0$.

[1] Véase que al mencionar que la variable T es no negativa se da por entendido que dicha variable aleatoria en el espacio de probabilidad $(\Omega, S, \mathbb{P})$, donde Ω es el espacio muestral, S la σ-álgebra definida en Ω y $\mathbb{P}$ la medida de probabilidad correspondiente, asigna valores en $\mathbb{R}^{+\{0\}}$; es decir: $T : \Omega \to \mathbb{R}^{+\{0\}}$.

3.- $S(t) = 0$ cuando $t \to \infty$.

La función $S(t)$ es conocida también como la *tasa de superviviencia acumulativa*; en el contexto industrial se conoce como *función de confiabilidad*.

Por otro lado, es de suma importancia representar $S(t)$ gráficamente ya que de ella se puede obtener información interesante; por ejemplo el cálculo de diversos cuantiles (como el cuantil 50) nos permitirán hacer inferencias y comparar distribuciones de superviviencia de dos o más grupos de individuos. La gráfica de $S(t)$ es llamada **curva de supervivencia**, los gráficos de la figura 3.1 son ejemplos de esta.

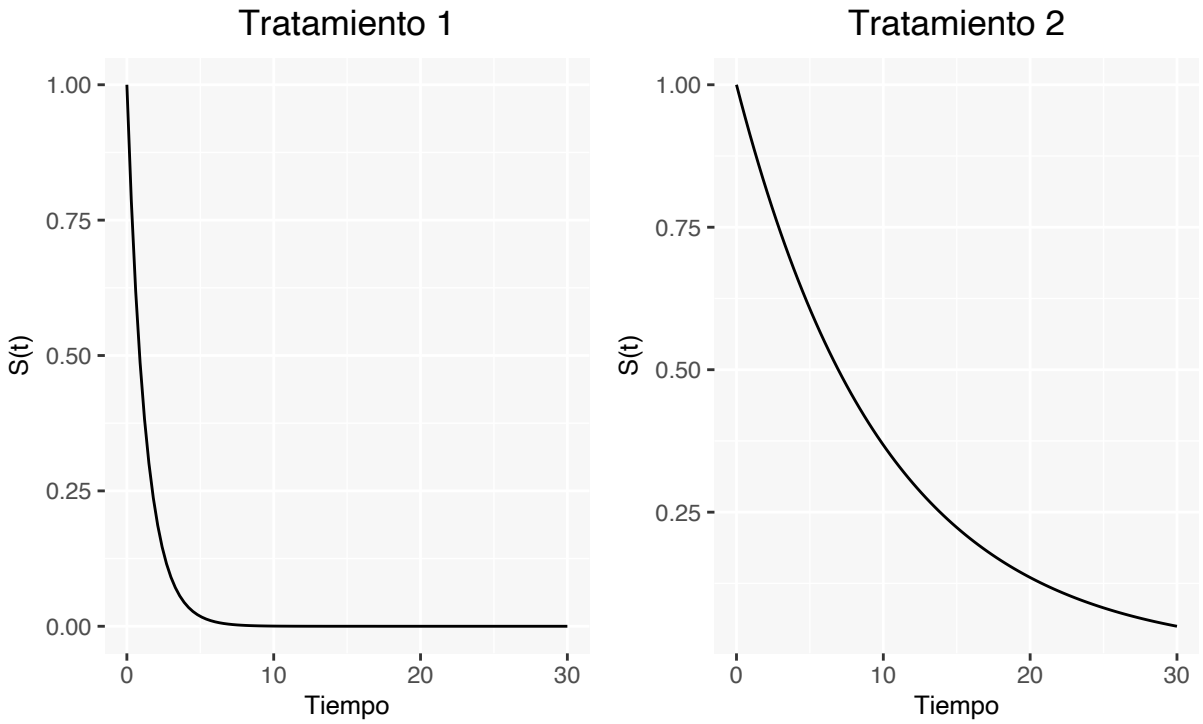


Figura 3.1: Ejemplos de la curva de supervivencia.

En la primer gráfica se observa una tasa de supervivencia baja, mientras que en la segunda gráfica se tiene una tasa alta. En este contexto, ¿qué podríamos decir de los pacientes con el tratamiento 1 *versus* los pacientes con el tratamiento 2?

Caso discreto

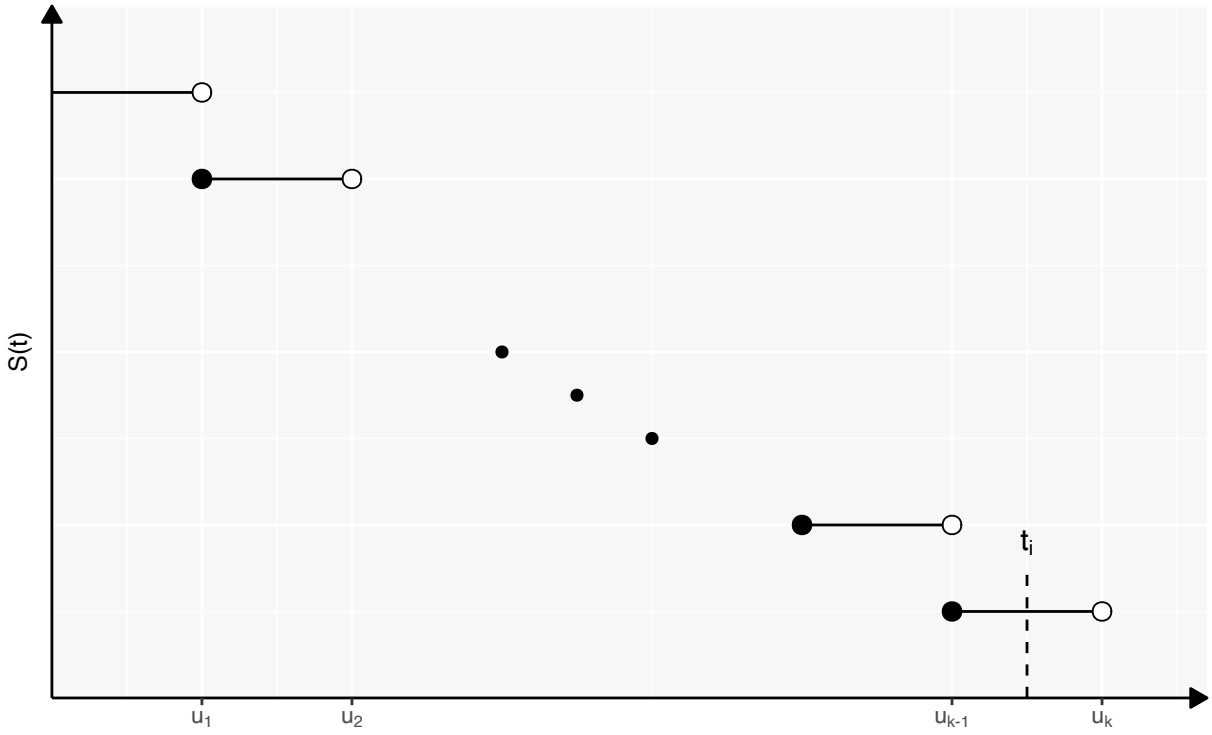
Si T es una variable aleatoria discreta que toma valores $0 < t_1 < t_2 < ...$ Entonces la función de probabilidad de T es:

$$f(t) = \begin{cases} \mathbb{P}(T = t_j) & \text{si } t = t_j, j = 1, 2, ... \\ 0 & \text{en otro caso} \end{cases}$$

Por lo que su función de supervivencia es:

$$S(t) = \mathbb{P}(T > t) = \sum_{t<t_j} f(t_j)$$

La siguiente es una representación gráfica de la función de supervivencia en el caso discreto



Al igual que el caso continuo, se tienen las propiedades:

1.-Es monótona no creciente

2.- $S(t) = 1$ para $t = 0$.

3.- $S(t) = 0$ cuando $t \to \infty$.

Ejemplos del caso continuo y caso discreto

Estos pueden verse al final de la sección 4

3.2. Función de Riesgo

Caso continuo

La función de riesgo $h(t)$ (hazard function), también llamada **tasa de falla condicional** (en el análisis de confiabilidad) o *tasa de mortalidad* (en demografía), se define como la probabilidad de falla durante un intervalo de tiempo muy pequeño suponiendo que el individuo ha sobrevivido hasta el inicio del intervalo[2]; en expresiones matemáticas es:

$$h(t) = \lim_{\alpha \to 0} \frac{1}{\alpha} \mathbb{P}(t < T \leq t + \alpha | T \geq t) = \lim_{\alpha \to 0} \frac{1}{\alpha} \frac{\mathbb{P}(T \leq t + \alpha) - \mathbb{P}(T < t)}{\mathbb{P}(T \geq t)} = \lim_{\alpha \to 0} \frac{1}{\alpha} \frac{F(t + \alpha) - F(t)}{S(t)} = \frac{f(t)}{S(t)}$$

Entonces

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt} log(S(t))$$

[2]Aunque en la definición de $h(t)$ se tenga explícitamente la palabra "probabilidad", hay que tener en claro que esta función no es una función de probabilidad, si no tal cual una **tasa**, ya que la acumulación de esta puede dar valores superiores a 1.

Si tomamos la igualdad

$$h(t) = -\frac{d}{dt} log(S(t))$$

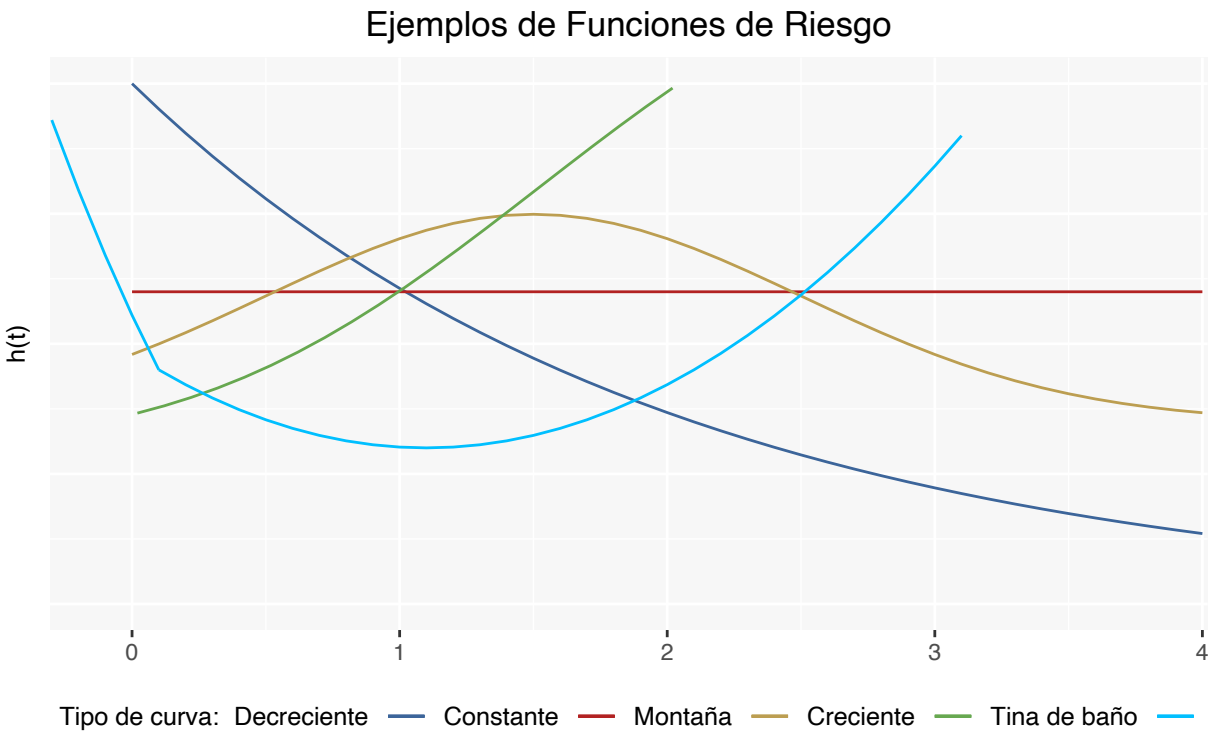
Y despejamos a $S(t)$ tenemos que:

$$S(t) = exp\left[-\int_0^t h(u)du\right] = exp[-H(t)]$$

Donde $H(t) = \int_0^t h(u)du$ es conocida como **la función acumulada de riesgo**, la cual veremos más adelante.

La función de riesgo juega un papel importante en el análisis de supervivencia. Describe la forma en que cambia la *tasa instantánea de muerte de un individuo al paso del tiempo* (constante, lineal, exponencial, etc.). El conocer $h(t)$ puede darnos alguna idea sobre la selección del modelo para la distribución del tiempo de supervivencia, por ejemplo, puede ser útil al considerar restricciones para modelos con funciones de riesgo no decrecientes o modelos con funciones de riesgo no crecientes.

No hay un comportamiento "habitual" en la gráfica de $h(t)$, es decir, $h(t)$ puede crecer, decrecer, ser constante o mostrar algo más complicado. El esquema siguiente muestra algunos ejemplos de la gráfica de $h(t)$:



La curva $h(t)$ en color azul (*llamada curva de tina de baño*) describe el proceso de la vida humana: al inicio existe mortalidad infantil y el riesgo de morir es alto, crece el individuo y el riesgo de morir se reduce y hasta cierto punto es constante, después viene el envejecimiento(hay deterioro) y entonces el riesgo de morir aumenta. Una función de riesgo creciente como la curva de color verde implica un envejecimiento natural. Una función decreciente como la curva color morado es menos común e indica rejuvenecimiento. Una función en forma de montaña como la curva color café, podría representar un comportamiento de muerte por enfermedad después de llevar un tratamiento para la misma enfermedad.

Caso discreto

En este caso, la función de riesgo proporciona la probabilidad condicional de falla al tiempo $t = u_k$, dado que el individuo estaba vivo antes de u_k.

Sea T una variable aleatoria discreta con soporte en $\{u_1, u_2, u_3, ...\}$. La función de riesgo al tiempo u_k se define como[3]

$$h(u_k) = \mathbb{P}(T = u_k | T \geq u_k) = \frac{\mathbb{P}(T = u_k)}{\mathbb{P}(T \geq u_k)} = \frac{f_T(u_k)}{S_T(u_{k-1})}$$

Observemos que:

$$f_T(u_k) = \mathbb{P}(T = u_k) = \mathbb{P}(T \geq u_k) - \mathbb{P}(T > u_k) = S(u_{k-1}) - S(u_k)$$

si dividimos entre $S(u_{k-1})$, tenemos entonces:

$$h(u_k) = \frac{S(u_{k-1}) - S(u_k)}{S(u_{k-1})} = 1 - \frac{S(u_k)}{S(u_{k-1})}$$

Por otro lado, para $S(t)$ se cumple que:

$$S(t) = \frac{S(t)}{1} = \frac{S(t)}{S(0)} = \frac{S(u_1)}{S(0)} \cdot \frac{S(u_2)}{S(u_1)} \cdot \frac{S(u_3)}{S(u_2)} \cdots \frac{S(u_k)}{S(u_{k-1})} \cdot \frac{S(t)}{S(u_k)}$$

Entonces, de la expresión anterior y ocupando que $h(u_k) = 1 - \frac{S(u_k)}{S(u_{k-1})}$ tenemos finalmente:

$$S(t) = \prod_{k:u_k \leq t} \frac{S(u_k)}{S(u_{k-1})} = \prod_{k:u_k \leq t} (1 - h(u_k))$$

Si queremos obtener $f(u_k)$, la podemos conocer partir de la función de riesgo:

$$f(u_k) = \frac{f(u_k)}{S(u_{k-1})} S(u_{k-1}) = h(u_k) S(u_{k-1}) = h(u_k) \prod_{j<k} \frac{S(u_j)}{S(u_{j-1})} = h(u_k) \prod_{j<k} (1 - h(u_j))$$

La expresión

$$S(t) = \prod_{k:u_k \leq t} (1 - h(u_k))$$

la ocuparemos más adelante.

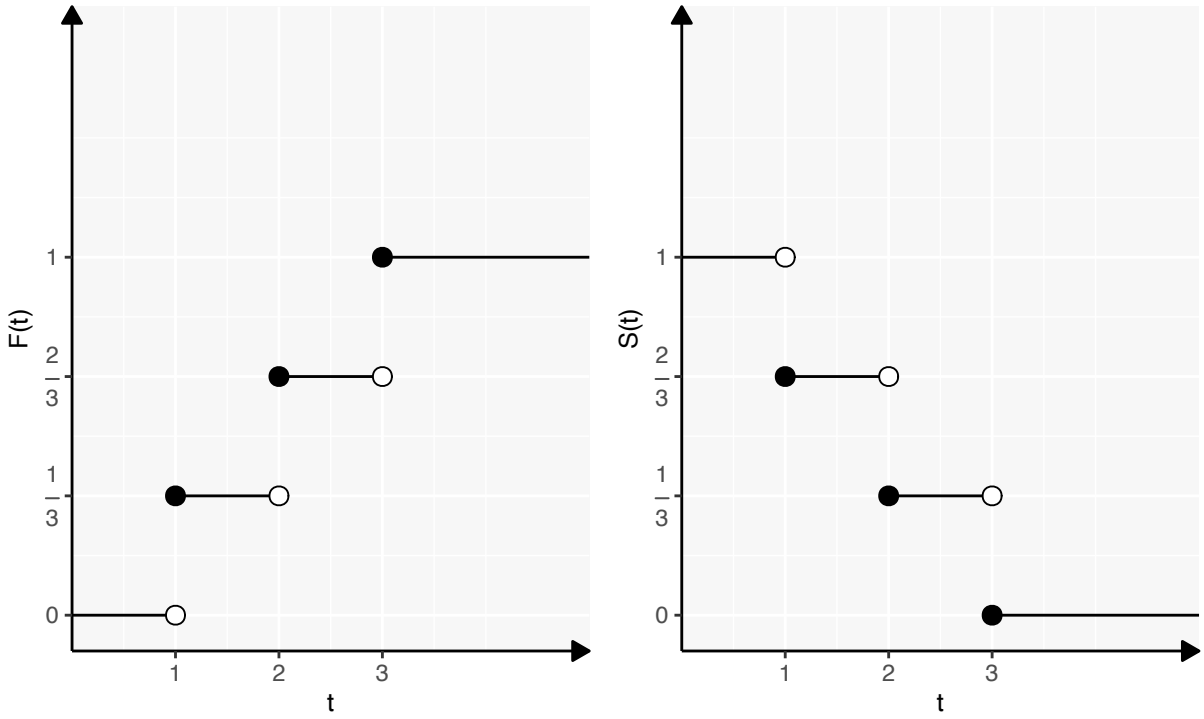
Ejemplo

Suponga que se tiene la siguiente distribución para la variable aleatoria T: $\mathbb{P}(T = j) = \frac{1}{3}$; $\quad j = 1, 2, 3$

Para este caso se tienen los siguientes resultados junto con las siguientes representaciones gráficas:

$$S(t) = \begin{cases} 1 & -\infty < t < 1 \\ \frac{2}{3} & 1 \leq t < 2 \\ \frac{1}{3} & 2 \leq t < 3 \\ 0 & t \geq 3 \end{cases} \qquad h(t) = \begin{cases} \frac{f(1)}{S(0)} = \frac{1}{3} & t = 1 \\ \frac{f(2)}{S(1)} = \frac{1}{2} & t = 2 \\ \frac{f(3)}{S(2)} = 1 & t = 3 \end{cases}$$

[3] $\mathbb{P}(T \geq u_k) = \mathbb{P}(T > u_{k-1}) = S(u_{k-1})$ por el hecho de estar tratando con una variable discreta, ya que el siguiente valor de u_{k-1} en la lista $\{u_1, \ldots, u_{k-1}, u_k\}$ es u_k. Además, véase que el uso de $f_T(u_k)$ en este caso discreto es simplemente notación y no se debe confundir con la función de densidad de la v.a T.



3.3. Función de Riesgo Acumulado

Caso continuo

Esta función, denotada por $H(t)$, es importante en la medición de la frecuencia con que ocurren los fallos en el tiempo y en el análisis de residuos para el ajuste de algunos modelos. $H(t)$ se define como sigue:

$$H(t) = \int_0^t h(u)du$$

Hemos visto anteriormente que:

$$S(t) = exp(-H(t))$$

entonces

$$H(t) = -log(S(t))$$

Podemos calcular a $f(t)$ en términos de $h(t)$ y $H(t)$:

$$f(t) = h(t)S(t) = h(t)exp(-H(t))$$

Caso discreto

$H(t)$ se define como:

$$H(t) = \sum_{k:u_k \leq t} h(u_k)$$

Una definición alternativa es:

$$H(t) = -\sum_{k:u_k \leq t} log(1 - h(u_k))$$

Capítulo 4

Parámetros Poblacionales

Si conocemos la distribución de T (cuando sea posible) es importante extraer cierta información que nos ayude a sacar conclusiones sobre los tiempos de supervivencia; a saber, información sobre su(s) parámetro(s). Aunque es habitual pensar a T como una variable aleatoria continua, a continuación presentaremos parámetros poblacionales en términos de la función de superviviencia para el caso continuo y discreto.

4.1. Media

Caso continuo

Para el caso continuo se tiene:

$$\mu = \mathbb{E}[T] = \int_0^\infty \mathbb{P}(T > t) = \int_0^\infty S(t)dt$$

La demostración se queda como ejercicio al alumno.

Caso discreto

μ en este caso es de la forma:

$$\mu = \mathbb{E}[T] = \sum_{k=1}^{\infty} u_k f(u_k) = \sum_{k=1}^{\infty} S(u_k)$$

La demostración se queda como ejercicio al alumno.

4.2. Varianza

Caso continuo

Para este caso la varianza se ve como:

$$\sigma^2 = Var[T] = \mathbb{E}[T^2] - (\mathbb{E}[T])^2 = 2\int_0^\infty tS(t)dt - \left(\int_0^\infty S(t)dt\right)^2$$

La demostración de la igualdad anterior se obtiene si se observa que:

$$\mathbb{E}[T^2] = \int_0^\infty t^2 f(t)dt = -t^2 S(t)|_0^\infty + 2\int_0^\infty tS(t)dt = 2\int_0^\infty tS(t)dt$$

Y ocupando que:

$$\mathbb{E}[T] = \int_0^\infty S(t)dt$$

Caso discreto

De manera similar:

$$\sigma^2 = Var[T] = \mathbb{E}[T^2] - (\mathbb{E}[T])^2 = 2\sum_{k=1}^{\infty} u_k S(u_k) - \left(\sum_{k=1}^{\infty} S(u_k)\right)^2$$

La demostración se queda de ejercicio al alumno.

4.3. Función de Media Residual

Para individuos de edad x este parámetro, denotado por $mr(x)$, mide la esperanza de vida residual; esto es, "la esperanza de vida que les queda a partir de la edad x".

Caso continuo

Se define como:

$$mr(x) = \mathbb{E}[T - x|T > x] = \frac{\mathbb{E}[T-x]}{\mathbb{P}(T>x)} = \frac{\int_x^\infty (t-x)f(t)dt}{S(x)} = \frac{\int_x^\infty S(t)dt}{S(x)}$$

Caso discreto

Para este caso es:

$$mr(u_x) = \frac{\sum_{k=x}^{\infty} S(u_k)}{S(u_x)}$$

4.4. Cuantiles de Orden p

El cálculo de diversos cuantiles nos permitirá hacer comparaciones entre diversos grupos de sujetos, además de obtener información sobre los tiempos de falla, por ejemplo, el tiempo mediano de falla (a veces es mejor el cálculo de la mediana que la media).

Caso continuo

El cuantil o percentil, t_p, de orden p de la variable aleatoria continua T será aquel que:

$$S(t_p) = 1 - p$$

Si queremos el tiempo mediano de los tiempos de supervivencia entonces debemos calcular $t_{0.5}$ talque:

$$S(t_{0.5}) = 0.5$$

Caso discreto

En este caso t_p es:

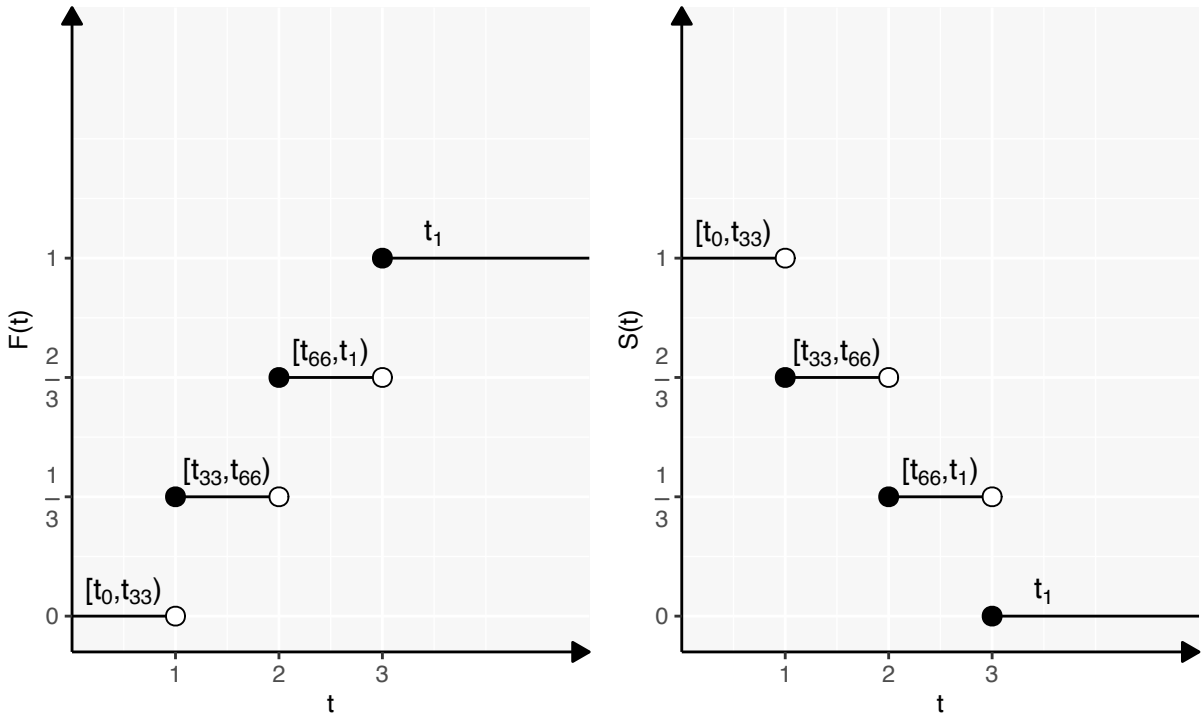
$$t_p = inf\{t : S(t) \leq 1 - p\}$$

Ejemplo 1

Siguiendo el ejemplo presentado en la sección 3.2, se tiene lo siguiente:

- Función de supervivencia de T: $S(t) = \mathbb{P}(T > t) = \begin{cases} 1 & \text{si } t = 1 \\ \frac{2}{3} & \text{si } t = 2 \\ \frac{1}{3} & \text{si } t = 3 \\ 0 & \text{si } t > 3 \end{cases}$
- Función de riesgo: $h(t) = \begin{cases} \frac{1}{3} & \text{si } t = 1 \\ \frac{1}{2} & \text{si } t = 2 \\ 1 & \text{si } t = 3 \end{cases}$
- $\mu = \mathbb{E}(T) = \sum_{t=1}^{3} S(t) = \sum_{t=1}^{3} t \cdot f(t) = 2.$
- $mrl(2) = \frac{\sum_{k=2}^{\infty} S(u_k)}{S(2)} = \frac{2/3+1/3+0}{2/3} = \frac{3}{2} = 1.5.$
- $t_{0.75} = inf\{t : S(t) \leq 0.15\} = 3$

Con la finalidad de aclarar la distribución de los cuantiles en una distribución discreta, se dejan las siguientes gráficas correspondientes a este ejercicio.



Ejemplo 2

Sea T una v.a con distribución uniforme continua en $(0, 100)$ unidades días.

1. Encontrar la función de supervivencia y evaluar la supervivencia para 30 y 35 años.
2. Encontrar la función de riesgo y evaluar el riesgo para 60 días.

3. Encontrar la esperanza de riesgo residual a los 75 días.

Soluciones:

1.
$$f_T(t) = \begin{cases} \frac{1}{100} & \forall \ t \in [0,100] \\ 0 & e.o.c \end{cases} \implies S(t) = 1 - \int_0^t \frac{1}{100} dv = 1 - \frac{t}{100}$$
$$\implies S(30) = \mathbb{P}(T > 30) = 1 - \frac{30}{100}; \ \ S(35) = 1 - \frac{35}{100}$$

2.
$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{100}}{1 - \frac{1}{100}} = \frac{1}{100 - t} \implies h(60) = \frac{1}{100 - 60} = \frac{1}{40}$$

3.
$$mrl(x) = \mathbb{E}(T - 75 | T > 75) = \frac{\mathbb{E}(T - 75)}{\mathbb{P}(T > 75)} = \frac{\int_x^\infty (t - 75) f(t) dt}{S(75)} = \frac{\int_x^{100} (t - 75) \frac{1}{100} dt}{1 - \frac{75}{100}}$$
$$\implies mrl(75) = \frac{1}{25} \int_{75}^{100} (t - 75) dt = \frac{1}{25} \left(\frac{100^2}{2} - \frac{75^2}{2} \right) - \frac{75(100 - 75)}{25} = \frac{25}{2} = 12.5$$

Otra solución

Demostrandose que $\int_x^\infty (t - x) f(t) dt = \int_x^\infty S(t) dt$

$$\implies mrl(75) = \frac{\int_{75}^{100} (t - 75) \frac{1}{100} dt}{S(75)} = \frac{\int_{75}^{100} S(75) dt}{S(75)} = 12.5$$

Capítulo 5

Modelos Paramétricos

Existen varios modelos paramétricos que se emplean en el análisis de supervivencia, esto se debe a que pueden representar de manera adecuada el comportamiento de ciertos fenómenos. La motivación para usar un modelo en particular es, por lo general, empírica; o bien, con base en la información que proporcione algún modelo **No paramétrico**. Las familias paramétricas más importantes son: *Exponencial, Weibull, Log-Normal, Log-logística* y *Gamma.*

5.1. Modelo Exponencial

El modelo exponencial es el más importante debido a su amplia aplicación, por ejemplo, puede emplearse en estudios para determinar el tiempo de vida útil de algunos artículos manufacturados. Este modelo juega un papel fundamental análogo a la *distribución normal* en la inferencia estadística tradicional.

Función de Densidad

$$f(t) = h(t)S(t) = \lambda exp(-\lambda t) = \lambda e^{-\lambda t}$$

Para esta distribución λ es un parámetro que modifica la escala de la distribución. Este es comúnmente llamado ***tasa*** definido por $1/s$ donde s es el verdadero parámetro de escala[1] de la distribución. Es muy común utilizar este parámetro en lugar del parámetro de escala ya que simplifica la expresión matemática.

Función de Supervivencia

$$S(t) = exp\left(-\int_0^t h(u)du\right) = exp\left(-\int_0^t \lambda du\right) = exp(-\lambda t) = e^{-\lambda t}$$

Si asumimos **una tasa de riesgo invariante en el tiempo**, $h(t) = \lambda$ con $\lambda > 0$, generamos el modelo exponencial:

Función de Riesgo

Tomamos $\lambda > 0$ y hacemos:

$$h(t) = \lambda$$

Aunque el supuesto de una función de riesgo constante resulta ser una restricción considerable, el modelo exponencial no deja de ser útil e importante en variedad de aplicaciones. Cabe destacar, que este modelo cumple con la propiedad de *pérdida de memoria*; a saber, se cumple que:

[1] Un parámetro de escala modifica la escala o *dispersión* de la distribución ya sea que entre mayor sea este parámetro mayor dispersión se tendrá o se tenga el caso contrario como en una *log-logistica*. En el siguiente enlace se puede encontrar mayor información sobre esto así como animaciones para un mejor entendimiento.

$$\mathbb{P}(T > t + x | T > t) = \mathbb{P}(T > x)$$

Parámetros

Si $T \sim Exp(\lambda)$ entonces:

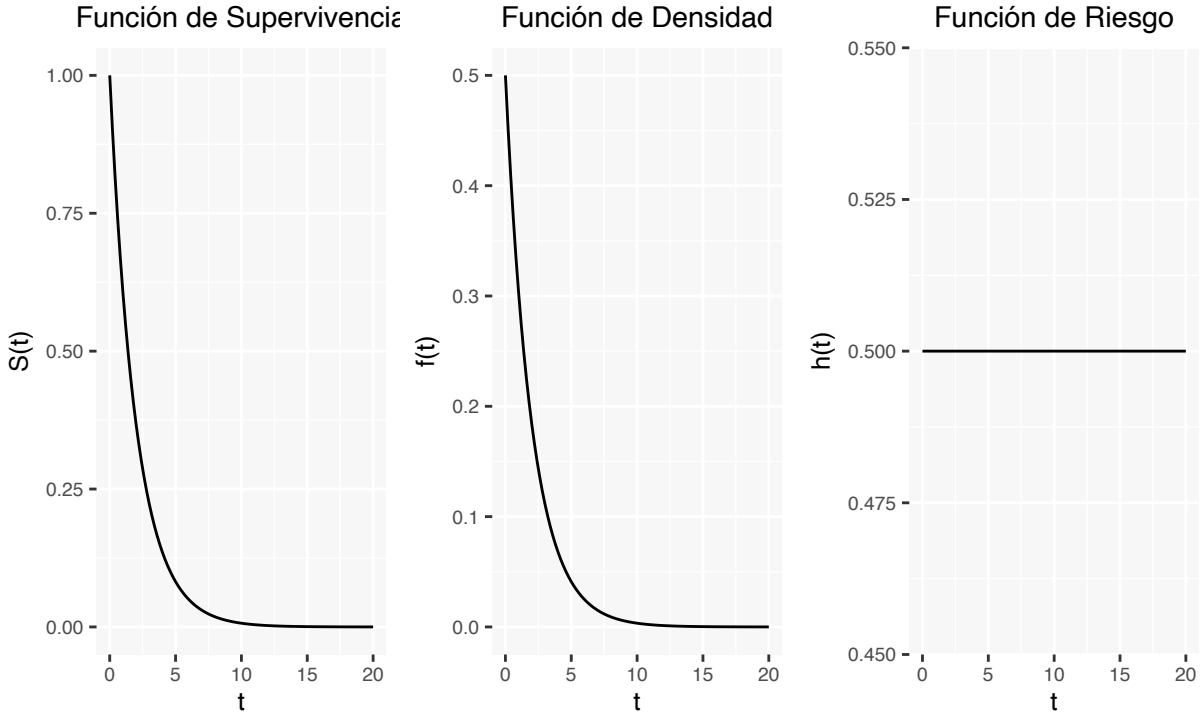
$$\mathbb{E}[T] = \int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}$$

Y

$$Var[T] = \mathbb{E}[T^2] - \mathbb{E}[T]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Gráficas

A continuación se muestran gráficas de una *Exponencial* con $\lambda = 0.5$



R

En el lenguaje de programación `R` se pueden obtener resultados de esta distribución como una muestra pseudo aleatoria y aproximaciones numéricas para la distribución y densidad de esta así como la obtención de los cuantiles.

En general, el sufijo para esta distribución es `*exp`, de tal manera que se tienen las siguientes funciones mencionando la interpretación de sus resultados:

- `rexp(n, rate = 1)`: Muestra pseudo aleatoria de tamaño n.
- `dexp(x, rate = 1, log = FALSE)`: Valores de $f(x)$.
- `pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)`: Valores de $F(x)$.
- `qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)`: Cuantil t_p.

Algo que hay que recordar es que al utilizar la función, por ejemplo, `rexp(n, r)` se respeta el orden de los parámetros por lo que `r` será el valor correspondiente a la tasa y no parámetro de escala es decir el valor $r = \lambda$. Se menciona esta por que puede ser común que se confunda `r` con $1/\lambda$, por lo que se recomienda tener precaución.

5.2. Modelo Weibull

El modelo Weibull es una generalización del modelo exponencial, se agrega un parámetro de forma[2] γ y se mantiene el parámetro de escala λ. Este modelo es uno de los más utilizados para tiempos de falla: tiene utilidad en la vida de algunos artículos manufacturados, así como en los tiempos de aparición de tumores en medicina.

Función de Densidad

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}} = h(t)S(t)$$

Obsérvese que si $\gamma = 1$ entonces el modelo Weibull se reduce al modelo exponencial.

Función de Supervivencia

$$S(t) = exp\left(-\int_0^t h(u)du\right) = exp\left(-\int_0^t \lambda\gamma(\lambda u)^{\gamma-1}du\right) = exp(-(\lambda t)^{\gamma}) = e^{-(\lambda t)^{\gamma}}$$

Si asumimos, en general, una **función de riesgo monótona: creciente o decreciente**, se puede obtener el modelo Weibull.

Función de Riesgo

Para este modelo $h(t)$ está dada por:

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}; \quad \gamma > 0, \lambda > 0 \text{ y } t > 0$$

Parámetros

Si $T \sim Weibull(\gamma, \lambda)$ entonces:

$$\mathbb{E}[T] = \frac{1}{\lambda}\Gamma\left(\frac{1}{\gamma}+1\right)$$

Y además

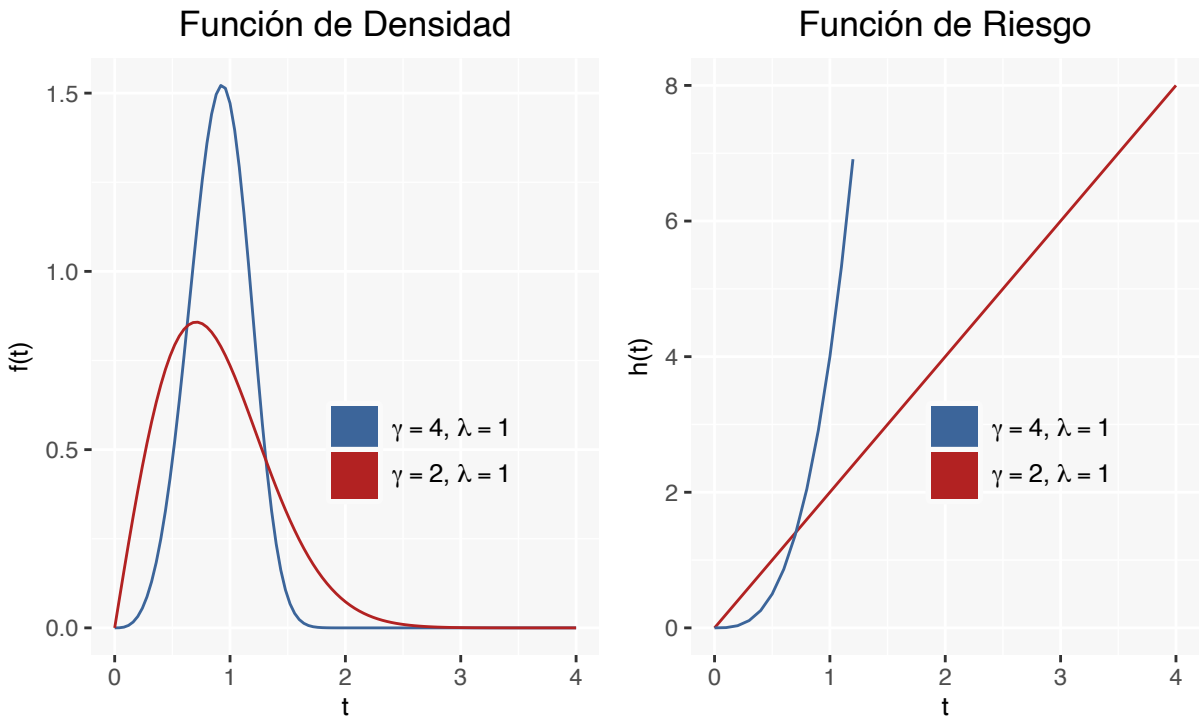
$$Var[T] = \frac{1}{\lambda^2}\Gamma\left(\frac{2}{\gamma}+1\right) - \left(\frac{1}{\lambda}\Gamma\left(\frac{1}{\gamma}+1\right)\right)^2 = \frac{1}{\lambda^2}\left[\Gamma\left(\frac{2}{\gamma}+1\right) - \left(\Gamma\left(\frac{1}{\gamma}+1\right)\right)^2\right]$$

Debe resultar sencillo para el alumno demostrar las igualdades de la esperanza y varianza.

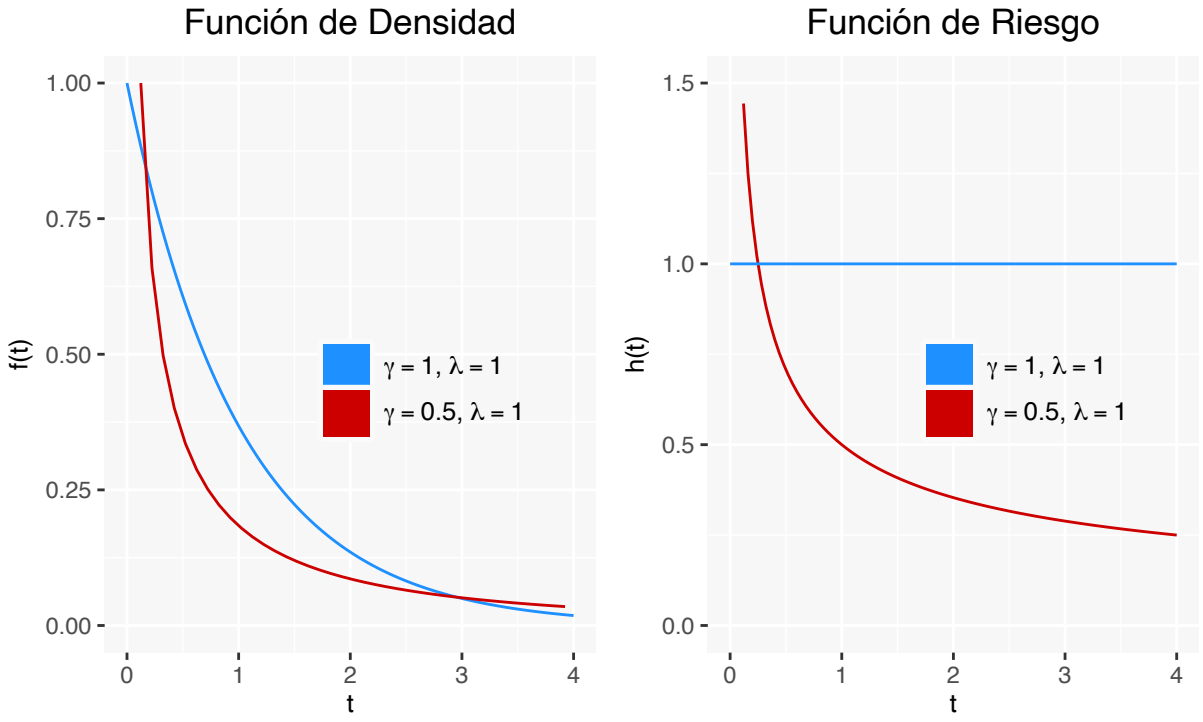
Gráficas

Las dos gráficas siguientes muestran curvas de $f(t)$ y $h(t)$ para el modelo Weibull con $\lambda = 1$ y $\gamma = 2, 4$:

[2] Un parámetro de forma es aquel que no es de escala ni de localización o una función de estos, ya que su único propósito es modificar la forma de la distribución de manera distinta a estos anteriores, es decir que no sólo traslada o afecta la variabilidad de la misma.



Y para el modelo Weibull con $\lambda = 1$ y $\gamma = 0.5, 1$ se tienen las siguientes gráficas:



Como bien hemos dicho, si suponemos una función de riesgo monótona creciente ó decreciente la distribución Weibull puede ser generada. Concretamente, si $\gamma > 1$ entonces $h(t)$ es estrictamente creciente, si $\gamma < 1$ entonces $h(t)$ es estrictamente decreciente. Cuando $\gamma = 1$ se tiene la distribución exponencial.

R

Para este caso el sufijo referente a esta distribución es `*weibull`, de tal manera que se tienen las siguientes funciones mencionando la interpretación de sus resultados:

- `rweibull(n, shape, scale = 1)`: Muestra pseudo aleatoria de tamaño n.
- `dweibull(x, shape, scale = 1, log = FALSE)`: Valores de $f(x)$.
- `pweibull(q, shape, scale = 1, lower.tail = TRUE, log.p = FALSE)`: Valores de $F(x)$.

- `qweibull(p, shape, scale = 1, lower.tail = TRUE, log.p = FALSE)`: Cuantil t_p.

Así como en el caso de la exponencial y en todos los casos posteriores, se respeta el orden de los parámetros por lo que para aplicar los parámetros tal como se vieron en esta sección se puede utilizar esta función de la siguiente manera: `dweibull`(γ, $1/\lambda$) o sin importar el orden indicando de manera explicita que `shape` $= \gamma$ y `scale` $= 1/\lambda$. Esto último debido a que la función de densidad para este modelo en R esta considerada como $f(x) = \frac{\gamma}{\lambda}(\frac{x}{\lambda})^{\gamma-1}e^{-(\frac{x}{\lambda})^{\gamma}}$.

5.3. Modelo Log-Normal

El modelo Log-Normal tiene estrecha relación con la distribución *Normal*. De hecho, el tiempo de supervivencia T se dice que sigue una distribución Log-Normal, si $Y = ln(T)$ se distribuye $N(\mu, \sigma^2)$.

La distribución Log-Normal se ha utilizado como modelo en el tiempo de falla de aislantes eléctricos y en el tiempo de aparición de cáncer pulmonar. También se utiliza en poblaciones que son una mezcla de tiempos de vida cortos y largos. A pesar de esto, este modelo es criticado por ser decreciente para valores grandes de t, lo cual parece inadecuado en algunas situaciones.

Función de Densidad

Para este modelo, $f(t)$ está dada por:

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

En este caso, μ y σ pasan a ser los parámetros de escala y de forma de la distribución, los cuales hacen diferencia con los de la distribución normal, ya que en tal caso μ sería un parámetro de localización y σ seguiría siendo el parámetro de escala. Para agregar un parámetro de localización a la distribución $log - normal$ bastaría camiar x por $x - \theta$ donde θ sería tal parámetro.

Función de Supervivencia

$$S(t) = \int_t^{\infty} f(u)du = 1 - \int_0^t f(u)du = 1 - \int_0^t \frac{1}{\sqrt{2\pi}\sigma u} exp\left(-\frac{1}{2}\left(\frac{ln(u)-\mu}{\sigma}\right)^2\right) du$$

Tomando la notación de la función de distribución de una *Normal estándar*, $\Phi()$, se tiene:

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

Función de Riesgo

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\sqrt{2\pi}\sigma t} exp\left(-\frac{1}{2}\left(\frac{ln(t)-\mu}{\sigma}\right)^2\right)}{1 - \Phi\left(\frac{ln(t)-\mu}{\sigma}\right)}$$

Si bien la expresión de $h(t)$ parece ser complicada, su gráfica resulta ser más interesante. Ésta toma el valor de cero en $t = 0$, crece hasta un valor máximo y luego tiende a cero cuando $t \to \infty$. Véase sección de gráficas.

Parámetros

Si $T \sim LogNormal(\mu, \sigma^2)$ entonces:

$$\mathbb{E}[T] = exp\left(\mu + \frac{\sigma^2}{2}\right)$$

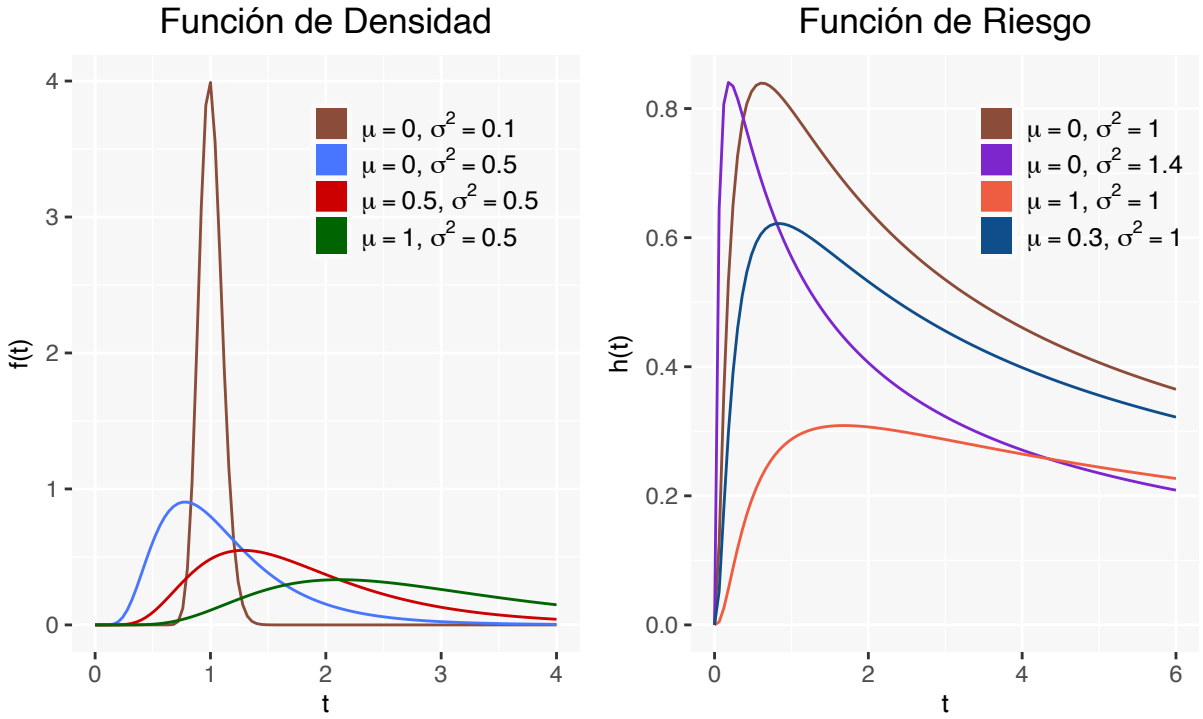
Y además

$$Var[T] = exp(2\mu + 2\sigma^2) - exp(2\mu + \sigma^2)$$

La demostración de las igualdades de media y varianza se dejan al alumno.

Gráficas

A continuación algunos ejemplos de las gráficas de $f(t)$ y $h(t)$ para la distribución Log-Normal:



R

Para esta distribución el sufijo respectivo será `*lnorm`, de tal manera que se tienen las siguientes funciones mencionando la interpretación de sus resultados:

- `rlnorm(n, meanlog = 0, sdlog = 1)`: Muestra pseudo aleatoria de tamaño n.
- `dlnorm(x, meanlog = 0, sdlog = 1, log = FALSE)`: Valores de $f(x)$.
- `plnorm(q, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)`: Valores de $F(x)$.
- `qlnorm(p, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)`: Cuantil t_p.

En este caso, es evidente que `meanlog` correspondería al parámetro de escala μ y `sdlog` al parámetro de forma σ. Finalmente se menciona que estos parámetros no se deben confundir explícitamente con los de una distribución normal.

Ejemplo

El tiempo de muerte en días después de un trasplante de médula sigue una distribución $log - normal$ con $\mu = 3.177, \sigma = 2.084$. Calcular lo siguiente

1. La media y la mediano tiempo de muerte
2. La probabilidad de que un individuo sobreviva 200 días después de un trasplante.

Soluciones

1.

- Mediana: $t_{0.5} = e^{\mu}$ cuando $T \sim log - normal \implies$ Mediana de tiempo de muerte: $t_{0.5} = e^{3.177} = 23.97$.

- Media: $\mathbb{E}(T) = e^{\mu+\sigma^2/2}$ cuando $T \sim log - normal \implies$ Media de tiempo de muerte: $\mathbb{E}(T) = e^{3.177+\frac{(2.084)^2}{2}} = 210.29$ días.

Entonces si sobrevivieron 23.97, queda una gran cantidad de días donde, en promedio son 210.29.

2. $S(200) = 1 - \Phi\left(\frac{\ln(200)-3.177}{2.084}\right) = 1 - 0.8438 = 0.15436$.

5.4. Modelo Log-Logístico

El modelo Log-Logístico es derivado de la distribución *Logística*. Se dice que T tiene distribución Log-Logística si $Y = ln(T)$ sigue una distribución Logística con parámetros μ y σ^2.

Función de Densidad

$$f(t) = \frac{\alpha\lambda(\lambda t)^{\alpha-1}}{(1+(\lambda t)^{\alpha})^2}$$

con $\alpha = \frac{1}{\sigma}$ y $\lambda = exp(-\frac{\mu}{\sigma}) > 0$. α es el parámetro de forma y λ es el parámetro de escala.

Función de Supervivencia

$$S(t) = \frac{1}{1+(\lambda t)^{\alpha}}$$

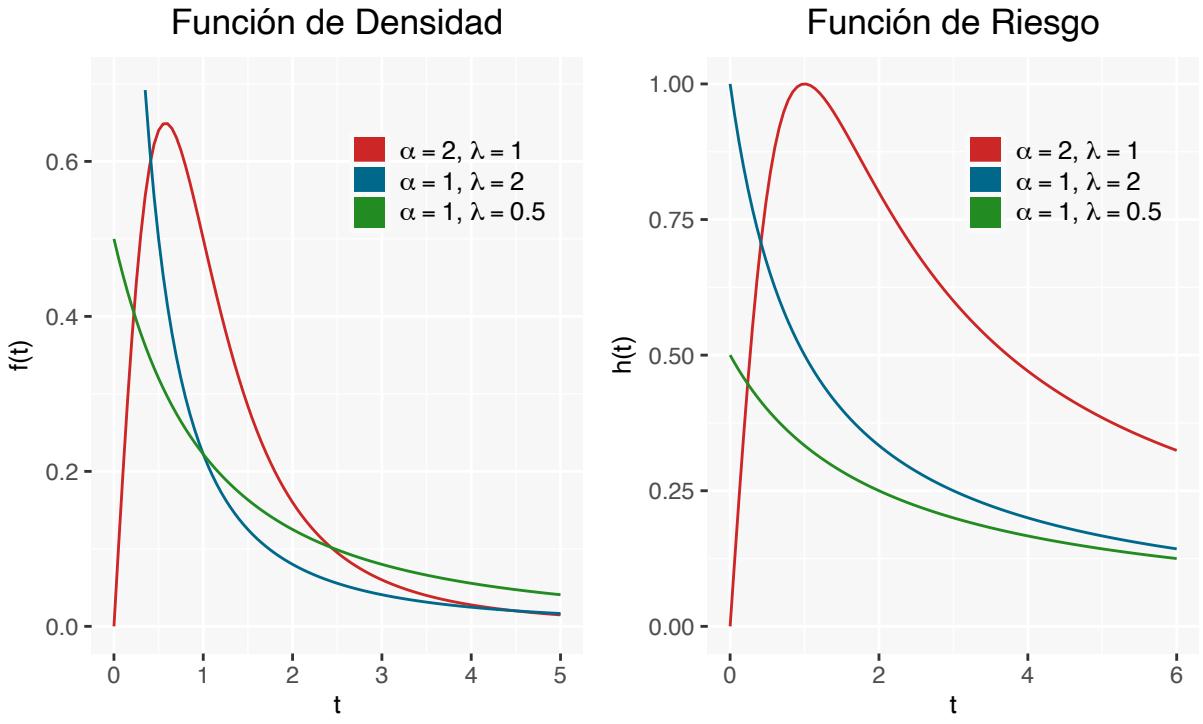
Función de Riesgo

$$h(t) = \frac{f(t)}{S(t)} = \frac{\alpha\lambda(\lambda t)^{\alpha-1}}{1+(\lambda t)^{\alpha}}$$

La función de riesgo es monótona decreciente para $\alpha \leq 1$, y para $\alpha > 1$ la función de riesgo crece hasta alcanzar un máximo en $t = (\frac{\alpha-1}{\lambda})^{\frac{1}{\alpha}}$ y luego decrece a cero cuando $t \to \infty$.

Gráficas

Algunos ejemplos de las gráficas de $f(t)$ y $h(t)$ para la distribución Log-Logística:



R

Para esta distribución se recomienda descargar y utilizar el paquete `actuar`, en la cual vienen un conjunto de funciones similares a las anteriores, de hecho el sufijo respectivo será `*llogis` y, de manera análoga a las anteriores funciones, se tiene lo siguiente:

- `rllogis(n, shape, rate = 1, scale = 1/rate)`: Muestra pseudo aleatoria de tamaño n.
- `dllogis(x, shape, rate = 1, scale = 1/rate, log = FALSE)`: Valores de $f(x)$.
- `pllogis(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)`: Valores de $F(x)$.
- `qllogis(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)`: Cuantil t_p.

En este caso, como en la exponencial y la weibull, `rate` correspondería al parámetro λ y `shape` al parámetro de forma α.

5.5. Modelo Gamma

La distribución Gamma, que incluye a las distribuciones *exponencial* y *ji-cuadrada*, ha sido utilizada como un modelo para problemas de confiabilidad industrial, hepatogramas en adultos normales y en pacientes con cirrosis, en supervivencia de plaquetas, entre otros. Este modelo tiene dos parámetros: β es el parámetro de forma y λ es el parámetro que modifica de escala; estrictamente hablando λ es la tasa: $\lambda = \frac{1}{s}$ donde s sería el verdadero parámetro de escala.

Función de Densidad

$f(t)$ está dada por:

$$f(t) = \frac{\lambda^\beta}{\Gamma(\beta)} t^{\beta-1} exp(-\lambda t); \quad \lambda, \beta > 0$$

Función de Supervivencia

$$S(t) = 1 - Ig(\lambda t, \beta)$$

donde[3]:

$$Ig(t, \beta) = \frac{1}{\Gamma(\beta)} \int_0^t u^{\beta-1} e^{-u} du$$

Al igual que un modelo $Weibull(\lambda, alpha = 1)$, la distribución exponencial es un caso particular del modelo $Gamma(\lambda, \beta = 1)$. Cuando $\beta \to \infty$, modelo gamma se aproxima a una distribución normal.

Función de Riesgo

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{\lambda^\beta}{\Gamma(\beta)} t^{\beta-1} exp(-\lambda t)}{1 - Ig(\lambda t, \beta)}$$

Esta función de riesgo tiene distintos comportamientos:

- Es **monótona creciente** para $\beta > 1$. En este caso sucede que $h(0) = 0$ y $\underset{t \to \infty}{h(t)} \to \lambda$. Además, la moda de la distribución es $t = \frac{\beta - 1}{\lambda}$

[3]Esta función se llama *función gamma incompleta* y en algunas fuentes se puede encontrar esta en particular como ***lower incomplete gamma function***. Cuando se tiene $\int_x^\infty u^{\beta-1} e^{-u} du = \Gamma(x, \beta)$ se le conoce como ***upper incomplete gamma function***, la cual es una generalización de la función gamma: $\Gamma(x, 0) = \Gamma(x)$. Originalmente esta función no tiene el cociente que se esta utilizando en este caso, esto es un efecto de normalizar la función como se puede ver en el siguiente enlace. Algunas veces esta función es conocida directamente como función gamma incompleta como en (Klein and Moeschberger, 2006)

- Es **monótona decreciente** con $\beta < 1$. En tal caso $h(0) = \infty$ y $\underset{t\to\infty}{h(t)} \to \lambda$.

Parámetros

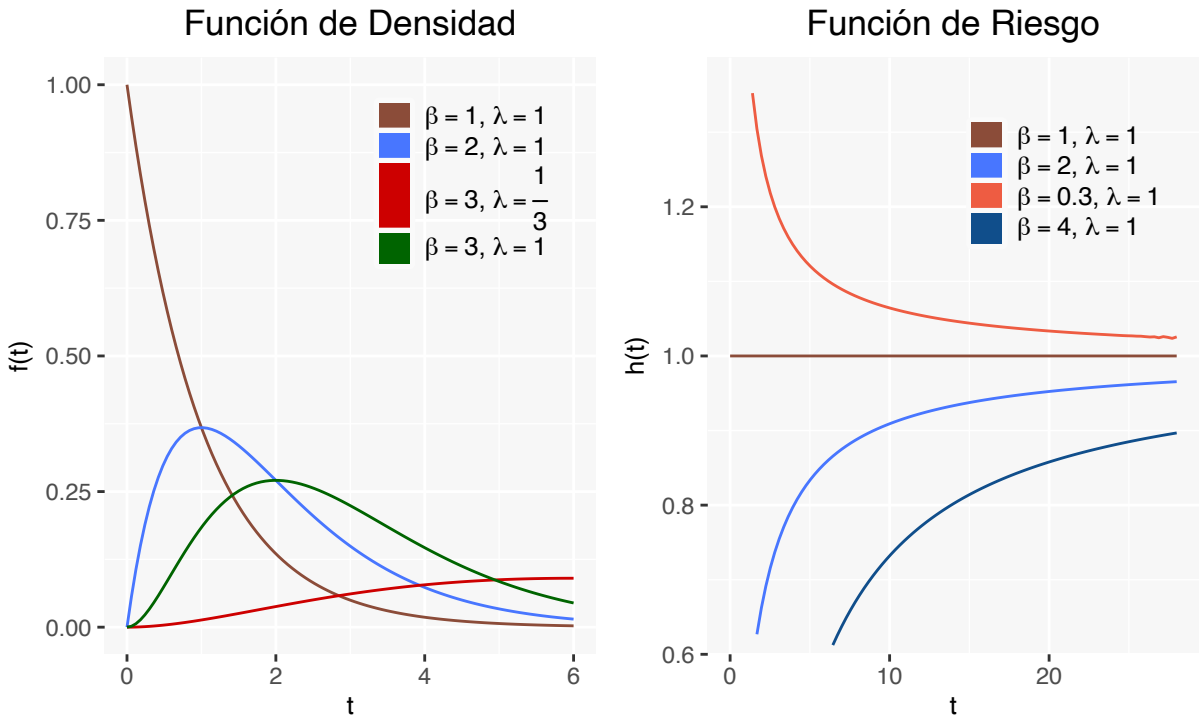
Si $T \sim Gamma(\beta, \lambda)$ entonces:

$$\mathbb{E}[T] = \frac{\beta}{\lambda} \quad \text{y} \quad Var[T] = \frac{\beta}{\lambda^2}$$

Las demostraciones de las igualdades de la esperanza y varianza se quedan de ejercicio al alumno.

Gráficas

Se muestran ejemplos de las gráficas de $f(t)$ y $h(t)$ para el modelo Gamma:



R

Para esta distribución el sufijo respectivo será `*gamma`, de tal manera que se tienen las siguientes funciones mencionando la interpretación de sus resultados:

- `rgamma(n, shape, rate = 1, scale = 1/rate)`: Muestra pseudo aleatoria de tamaño n.
- `dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)`: Valores de $f(x)$.
- `pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)`: Valores de $F(x)$.
- `qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)`: Cuantil t_p.

Como en el caso de la exponencial, se recomienda tener precaución ya que el parámetro `rate` es el que corresponde al parámetro λ en este trabajo.

5.6. Modelo Gamma Generalizada

Finalmente, se presenta el modelo Gamma Generalizada con dos parámetros de forma y un parámetro de que modifica la escala α, β y λ respectivamente.

Función de Densidad

$f(t)$ está dada por:

$$f(t) = \frac{\alpha\lambda^{\beta\alpha}}{\Gamma(\beta)} t^{\alpha\beta-1} \exp(-\lambda t)^{\alpha}; \quad \lambda, \alpha, \beta > 0$$

Función de Supervivencia

$$S(t) = 1 - Ig((\lambda t)^{\alpha}, \beta)$$

Esta distribución se reduce a las siguientes

- *Exponencial* cuando $\beta = \alpha = 1$.
- *Weibull* cuando $\beta = 1$.
- *Gamma* cuando $\alpha = 1$.
- Tiende a la log-normal cuando $\beta \to \infty$.

R

Para esta distribución se recomienda descargar y utilizar el paquete `ggamma`, en la cual vienen un conjunto de funciones similares a las anteriores. Aquí el respectivo sufijo será `*ggamma` y, de manera análoga a las anteriores funciones, se tiene lo siguiente:

- `rggamma(n, a, b, k)`: Muestra pseudo aleatoria de tamaño n.
- `dggamma(x, a, b, k, log = F)`: Valores de $f(x)$.
- `pggamma(p, a, b, k, lower.tail = TRUE, log.p = FALSE)`: Valores de $F(x)$.
- `qggamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)`: Cuantil t_p.

Para este caso, se tiene una parametrización de uso común en distintas fuentes. `a` $= \frac{1}{\lambda}$, `b` $= \alpha$ y `k` $= \beta$.

Ejemplo

El tiempo de vida en meses de cierta especie de ratón sigue una distribución $gamma(\beta = 3, \lambda = 0.2)$.

1. Calcular la probabilidad de que un ratón sobreviva más de 18 meses.
2. ¿Cuál es la probabilidad de que un ratón muera en el primer año de vida?
3. ¿Cuál es la esperanza de vida media de esta especia?

Soluciones

1. $S(18) = 1 - F(18) = 1 - 0.6972532 = 0.3027468$
2. $F(12) = 0.4302913$
3. $\mathbb{E}(T) = \frac{\beta}{\lambda} = \frac{3}{0.2} = 15$ meses.

Cabe señalar que hasta este punto **no hemos introducido datos censurados a los modelos**.

Capítulo 6

La Función de Verosimilitud con Censura y Truncamiento

Cuando se conoce la distribución de T, hacer estimaciones (usualmente por el método de máxima verosimilitud) sobre parámetros poblacionales resulta de suma importancia, pues a partir de ellos no sólo podemos conocer *estimadores* de la media, varianza o cuantiles; sino que también podemos conocer $\hat{S}(t)$, $\hat{h}(t)$, etc. En la sección anterior revisamos los modelos paramétricos más usados en análisis de supervivencia, sin embargo, no contemplamos datos censurados o truncados en dichos modelos. En este apartado veremos la función de verosimilitud considerando censura y truncamiento, y con base en esta función haremos estimaciones para algunas distribuciones.

6.1. Caso General

De acuerdo al tipo de observación, se tienen las siguientes contribuciones a la función de verosimilitud:

Exactas T_i	$f(t_i)$
Censurada por la derecha $T_i > C_i$	$\mathbb{P}(T_i > C_i) = S(C_i)$
Censurada por la izquierda $T_i < C_i$	$\mathbb{P}(T_i < C_i) = 1 - S(C_i)$
Censurada por la intervalo $L_i < T_i \leq R_i$	$\mathbb{P}(L_i < T_i \leq R_i) = S(L_i) - S(R_i)$
Truncado por la izquierda $T_i \vert T_i > u_i$	$\mathbb{P}(T_i \vert T_i > u_i) = \frac{f(t_i)}{S(u_i)}$
Truncado por la derecha $T_i \vert T_i \leq v_i$	$\mathbb{P}(T_i \vert T_i \leq v_i) = \frac{f(t_i)}{1-S(v_i)}$

Entonces la función de verosimilitud es:

$$\mathscr{L} = \prod_{i \in D} f(t_i) \prod_{i \in R} S(C_i) \prod_{i \in L} (1 - S(C_i)) \prod_{i \in I} [S(L_i) - S(R_i)]$$

donde

- D: Conjunto de tiempos de fallo.
- R: Conjunto de observaciones censuradas por la derecha.
- L: Conjunto de observaciones censuradas por la izquierda.
- I: Conjunto de observaciones censuradas por intervalo.

Cuando hay datos truncados, se sustituye $f(t_i)$ por $\frac{f(t_i)}{S(u_i)}$ y $S(C_i)$ por $\frac{S(C_i)}{S(u_i)}$.

6.2. Censura por la Derecha Tipo I

Suponga que se tiene una muestra aleatoria de n individuos con tiempos de vida $T_1, T_2, ..., T_n$(v.a.i.i.d) y que está asociado a cada individuo un tiempo fijo de censura $C_i > 0$. Se observa a T_i solamente si

$T_i \leq C_i$, por lo que los datos son parejas: (t_i, δ_i), $i = 1, 2, ..., n$ donde $t_i = min(T_i, C_i)$ y:

$$\delta_i = \begin{cases} 0 & \text{si } t_i = C_i \\ 1 & \text{si } t_i = T_i \end{cases}$$

Entonces la función de verosimilitud para datos con este tipo de censura es de la forma:

$$\mathscr{L} = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

donde $\sum_{i=1}^{n} \delta_i$ representa el total de los tiempos de vida observados.

Considerando que $f(t) = h(t)S(t)$, entonces:

$$\begin{aligned} \mathscr{L} &= \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \\ &= \prod_{i=1}^{n} h(t_i)^{\delta_i} exp\{-H(t_i)\} \end{aligned}$$

6.3. Censura por la Derecha Tipo II

Supongamos que se observan los r tiempos de falla $T_{(1)} < T_{(2)} < ... < T_{(r)}$ más pequeños, dejando a $n - r$ tiempos censurados por la derecha, de una muestra de tamaño n. De modo que, los datos serán los r tiempos de fallo más pequeños de $T_1, T_2, ..., T_n$.

T tiene f.d.p $f(t)$ y función de supervivencia $S(t)$. Entonces la f.d.p. conjunta de $T_{(1)}, T_{(2)}, ..., T_{(r)}$ es:

$$\mathscr{L} = \frac{n!}{(n-r)!} \left\{ \prod_{i=1}^{r} f(t_i) \right\} \left[S\left(t_{(r)}\right) \right]^{n-r}$$

donde $\prod_{i=1}^{r} f(t_i)$ es la parte correspondiente a las r fallas observadas y $[S(t_{(r)})]^{n-r}$ constituye la aportación de las observaciones censuradas después de la $r - esima$ falla observada.

Observemos que si quitamos el término constante $\frac{n!}{(n-r)!}$ y definiendo a la función indicadora:

$$\delta_i = \begin{cases} 0 & \text{si } T_i > T_{(r)} \\ 1 & \text{si } T_i \leq T_{(r)} \end{cases}$$

Entonces la función de verosimilitud para este tipo de censura se puede reesctibir:

$$\mathscr{L} \propto \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

que es la función de verosimilitud con datos censurados por la derecha tipo 1 revisada en la sección anterior.

6.4. Censura Aleatoria

Supongamos que para cada individuo se tienen T_i y C_i variables aleatorias independientes, con funciones de densidad $f_T(t)$ y $f_C(t)$, y con funciones de superviviencia $S_T(t)$ y $S_C(t)$ respectivamente. Sean (T_i, C_i), $i = 1, 2, ..., n$ parejas de observaciones independientes, la variable que observamos es $t_i = min(T_i, C_i)$ y se define:

$$\delta_i = \begin{cases} 0 & \text{si } T_i > C_i \\ 1 & \text{si } T_i \leq C_i \end{cases}$$

por lo que, las observaciones son las parejas (t_i, δ_i) con $i = 1, 2, ..., n$.

La función de verosimilitud es entonces:

$$\mathscr{L} = \prod_{i=1}^{n} [S_C(t_i)]^{1-\delta_i} [f_C(t_i)]^{\delta_i} [f_T(t_i)]^{\delta_i} [S_T(t_i)]^{1-\delta_i}$$

Similar al caso se censura por la derecha tipo 2, se puede demostrar que la expresión anterior es proporcional a $\prod_{i=1}^{n} f_T(t_i)^{\delta_i} S_T(t_i)^{1-\delta_i}$.

6.5. Truncamiento por la Izquierda

Consideremos los tiempos de fallo truncados por la izquierda, es decir, T_i tal que $T_i \geq u_i$ para ser observado (u_i es el valor de truncamiento).

En este caso, las observaciones serán (u_i, t_i, δ_i) con $t_i \geq u_i$ tiempo de fallo y δ_i indicador de censura por la derecha. Por lo que:

$$\mathscr{L} = \prod_{i=1}^{n} \left\{ \frac{f(t_i)}{S(u_i)} \right\}^{\delta_i} \left\{ \frac{S(t_i)}{S(u_i)} \right\}^{1-\delta_i} = \prod_{i=1}^{n} \{h(t_i)\}^{\delta_i} \left\{ \frac{S(t_i)}{S(u_i)} \right\}$$

Obsérvese que el hecho de dividir la función de verosimilitud entre $S(u_i)$ se limita a que los datos sean truncados por la izquierda[1].

6.6. Truncamiento por la Derecha

Consideremos tiempos de fallo T_i tal que $T_i \leq v_i$ para que sea observado. Entonces las observaciones serán (t_i, v_i) para todo $i = 1, ..., n$.

Por lo que:

$$\mathscr{L} = \prod_{i=1}^{n} \left\{ \frac{f(t_i)}{1 - S(v_i)} \right\} = \prod_{i=1}^{n} \mathbb{P}(T_i | T_i < v_i)$$

6.7. Estimaciones para Algunos Modelos

Modelo Exponencial

Sean los tiempos de fallo T_i independientes y provenientes de una distribución exponencial:

[1] De primera vista, pareciera que la segunda igualdad de la expresión anterior es incorrecta y que debería ser $\prod_{i=1}^{n} \{h(t_i)\}^{\delta_i} \left\{ \frac{S(t_i)}{S(u_i)} \right\}^{1-\delta_i}$ lo cual no es correcto. Se recomienda desarrollar la primera igualdad y relacionar lo necesario para obtener $h(t_i)$.

- $f(t) = \lambda e^{-\lambda t}$
- $S(t) = e^{-\lambda t}$
- $h(t) = \lambda$
- $H(t) = \lambda t$

Entonces, la función de verosimilitud con censura tipo I es:

$$\mathscr{L} = \prod_{i}^{n} \lambda^{\delta_i} e^{-\lambda t_i} = \lambda^{\sum_{i=1}^{n} \delta_i} e^{-\lambda \sum_{i=1}^{n} t_i}$$

Sea $r = \sum_{i=1}^{n} \delta_i$ el número de observaciones exactas (no censuradas). Entonces:

$$\mathscr{L} = \lambda^{r} e^{-\lambda \sum_{i=1}^{n} t_i}$$

Si queremos un estimador para λ, hacemos:

$$\frac{\partial ln(\mathscr{L})}{\partial \lambda} = \frac{\partial [r ln(\lambda) - \lambda \sum_{i=1}^{n} t_i]}{\partial \lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

$$\Longrightarrow \hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i}$$

Observe que si *no hay datos censurados* entonces $\sum_{i=1}^{n} \delta_i = n$, por lo que $\hat{\lambda} = \frac{1}{\bar{t}}$, y este es el estimador que conocemos desde el curso de inferencia.

Ejemplo

El siguiente ejercicio fue basado en el ejercicio 3.6 del libro (Klein and Moeschberger, 2006). Los siguientes datos consisten en los tiempos de recaída y los tiempos de muerte después de la recaída de 10 pacientes con trasplante de médula ósea. Suponga que el tiempo hasta la recaída tiene una distribución exponencial con la tasa de riesgo λ.

Paciente	Tiempo de recaida en meses
1	5
2	8
3	12
4	24
5	32
6	17
7	16+
8	17+
9	19+
10	30+

Los datos censurados están representados con un '+'

- a) Calcule la tasa de recaída.
- b) Calcule la probabilidad de no recaer en 16 meses.

Sabemos que los estimadores máximo-verosímiles cumplen la propiedad de invarianza, de modo que:

a) $\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i} = \frac{6}{\sum_{i=1}^{10} t_i} = \frac{6}{180} = 3.3333333\,\%$

b) $\hat{S}(16) = e^{-\hat{\lambda} t} = e^{-0.033(16)} = 0.5866463$

Modelo Weibull

Supongamos que se tienen los tiempos de fallo T_i con censura, provenientes de una distribución Weibull:

- $f(t) = \lambda\gamma(\lambda t)^{\gamma-1}e^{-(\lambda t)^\gamma}$
- $S(t) = e^{-(\lambda t)^\gamma}$
- $h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$

Entonces, su respectiva función de verosimilitud es de la forma:

$$\mathscr{L} = \prod_i^n \left[\lambda\gamma(\lambda t_i)^{\gamma-1}e^{-(\lambda t_i)^\gamma}\right]^{\delta_i}\left[e^{-(\lambda t_i)^\gamma}\right]^{1-\delta_i} = \prod_i^n \left[\lambda\gamma(\lambda t_i)^{\gamma-1}\right]^{\delta_i}\left[e^{-(\lambda t_i)^\gamma}\right]$$

Por lo que:

$$\ln(\mathscr{L}) = \sum_{i=1}^n \delta_i \ln(\lambda\gamma(\lambda t_i)^{\gamma-1}) - (\lambda t_i)^\gamma$$

Si suponemos que $r = \sum_{i=1}^n \delta_i$ y desarrollamos la expresión anterior, obtenemos:

$$\ln(\mathscr{L}) = r\ln(\lambda\gamma) + (\gamma-1)r\ln(\lambda) + (\gamma-1)\sum_{i=1}^n \delta_i \ln(t_i) - \lambda^\gamma \sum_{i=1}^n t_i^\gamma$$

Los estimadores máximo-verosímiles de λ y γ se obtienen derivando $\ln(\mathscr{L})$ con respecto a λ y γ, igualando a cero y evaluando en $\hat{\lambda}$ y $\hat{\gamma}$:

$$\frac{\partial \ln(\mathscr{L})}{\partial\lambda} = r\frac{\gamma}{\lambda\gamma} + \frac{(\gamma-1)r}{\lambda} - \gamma\lambda^{\gamma-1}\sum_{i=1}^n t_i^\gamma = 0$$

$$\frac{\partial \ln(\mathscr{L})}{\partial\gamma} = r\frac{\lambda}{\lambda\gamma} + r\ln(\lambda) + \sum_{i=1}^n \delta_i\ln(t_i) - \left[\lambda^\gamma\ln(\lambda)\sum_{i=1}^n t_i^\gamma + \lambda^\gamma\sum_{i=1}^n t_i^\gamma\ln(t_i)\right] = 0$$

Simplificando ambas ecuaciones tenemos:

$$r\gamma - \gamma\lambda^\gamma\sum_{i=1}^n t_i^\gamma = 0$$

$$\frac{r}{\gamma} + r\ln(\lambda) + \sum_{i=1}^n \delta_i\ln(t_i) - \lambda^\gamma\left[\ln(\lambda)\sum_{i=1}^n t_i^\gamma - \sum_{i=1}^n t_i^\gamma\ln(t_i)\right] = 0$$

Despejamos a λ de la primera ecuación:

$$\hat{\lambda} = \left(\frac{r}{\sum_{i=1}^n t_i^{\hat{\gamma}}}\right)^{\frac{1}{\hat{\gamma}}}$$

Y sustituyendo en la segunda ecuación:

$$\begin{aligned}&\tfrac{r}{\hat{\gamma}} + r\ln\left(\left(\tfrac{r}{\sum_{i=1}^n t_i^{\hat{\gamma}}}\right)^{\frac{1}{\hat{\gamma}}}\right) + \textstyle\sum_{i=1}^n \delta_i\ln(t_i)\\ &-\left(\tfrac{r}{\sum_{i=1}^n t_i^{\hat{\gamma}}}\right)\left[\ln\left(\left(\tfrac{r}{\sum_{i=1}^n t_i^{\hat{\gamma}}}\right)^{\frac{1}{\hat{\gamma}}}\right)\textstyle\sum_{i=1}^n t_i^{\hat{\gamma}} - \sum_{i=1}^n t_i^{\hat{\gamma}}\ln(t_i)\right] \quad = 0\end{aligned}$$

La expresión anterior es una ecuación *no lineal* de $\hat{\gamma}$, cuya solución es únicamente mediante un método numérico. Una vez que se ha obtenido el valor $\hat{\gamma}$, éste se sustituye en $\left(\frac{r}{\sum_{i=1}^{n} t_i^{\hat{\gamma}}}\right)^{\frac{1}{\hat{\gamma}}}$ para así obtener, finalmente, el valor de $\hat{\lambda}$.

Ecuaciones *no lineales*, como hemos visto en el modelo *Weibull*, aparecen en la estimación de parámetros de las distribuciones *Log-Normal*, *Log-Logística* y *Gamma*; resulta que, al considerar datos con censura y truncamiento, las estimaciones se complican en los modelos paramétricos. No obstante, tenemos ayuda de las computadoras y software que nos permitirán realizar estimaciones adecuadas.

Parte III

Estudio no paramétrico

Capítulo 7

Modelos No Paramétricos para la Función de Supervivencia

Los métodos estadísticos más utilizados en el análisis de supervivencia son los *No Paramétricos*. Debido a que sólo cuando se conoce la distribución que siguen los tiempos de falla, las estimaciones con métodos paramétricos será adecuada. La eficiencia de los métodos no paramétricos radica en que los datos no sigan una distribución teórica.

Puesto que nuestro interés es estimar $S(t)$, en los métodos no paramétricos las curvas de supervivencia, por lo general, se producen usando uno de dos métodos: el **análisis actuarial** o el **método límite-producto de Kaplan-Meier**.

Es importante mencionar que, cuando **no** hay datos censurados podemos estimar $S(t)$ de manera sencilla mediante la función empírica:

$$\hat{S}(t) = \hat{\mathbb{P}}(T > t) = \frac{\#t_i > t}{n}$$

donde la muestra aleatoria es: $t_{(1)} \leq t_{(2)} \leq ... \leq t_{(n)}$.

En este caso, la función será escalonada con decrementos $\frac{1}{n}$ si todas las observaciones son distintas, o con decrementos $\frac{d}{n}$ si hay d tiempos de falla iguales a t.

Los métodos que se presentan a continuación incluyen la percepción del censuramiento en los datos; quizá el alumno está familiarizado con alguno de ellos, pues se estudian en cursos previos.

7.1. Método Actuarial (Tabla de Vida)

El análisis actuarial divide el tiempo en intervalos y calcula la supervivencia en cada intervalo. La longitud del intervalo depende de la frecuencia con que ocurre el suceso de interés.

Se asume que:

- Todos los abandonos durante un intervalo dado ocurren aleatoriamente durante dicho intervalo.
- Las personas que se retiran del estudio tienen un destino parecido a las que se quedan.
- El periodo de tiempo durante el cual una persona entra en el estudio no tiene efecto en la respuesta.

Dividimos el eje del tiempo en $k+1$ intervalos $l_j = (a_{j-1}, a_j]$. Entonces para cada elemento de una muestra aleatoria de tamaño n, se observa un tiempo de fallo T o un valor censurado por la derecha C.

Definimos:

- n_j: número de individuos en riesgo (vivos o no censurados) al tiempo a_{j-1}.
- d_j: número de fallas en el intervalo l_j.
- c_j: número de individuos que se censuran en el intervalo l_j.

El número de individuos sin falla al inicio de l_j es n_j.

Suponga que la función de supervivencia para los tiempos de falla es $S(t) = \mathbb{P}(T > t)$. Entonces:

$$S(a_j) = \mathbb{P}(T > a_j) = \mathbb{P}(T > a_0)\mathbb{P}(T > a_1|T > a_0)...\mathbb{P}(T > a_j|T > a_{j-1})$$

Sea:

- $S_j = S(a_j)$
- $p_j = \mathbb{P}(T > a_j|T > a_{j-1}) = \frac{S_j}{S_{j-1}}$ (Sobrevivencia hasta a_j después de haber sobrevivido hasta a_{j-1})
- $q_j = 1 - p_j = \mathbb{P}(T \leq a_j|T > a_{j-1}) = \frac{S_{j-1}-S_j}{S_{j-1}}$ (No sobrevivir hasta a_j después de haber sobrevivido hasta a_{j-1})

Donde $S_0 = 1$, $S_{k+1} = 0$, $q_{k+1} = 1$. Por lo tanto:

$$S_j = p_1 p_2 ... p_j$$

El objetivo es estimar S_j con base en la estimación de $p_j = 1 - q_j$.

Si en l_j **no** hay observaciones censuradas, entonces estimamos S_j por medio de:

$$\hat{q_j} = \frac{d_j}{n_j}$$

Por otro lado, si en l_j **hay** observaciones censuradas, y suponiendo que las censuras se distribuyen uniformemente, entonces S_j se puede obtener mediante:

$$\hat{q_j} = \frac{d_j}{n_j - \frac{c_j}{2}}$$

De modo que, para este caso, se tiene:

$$\hat{S}_j = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i - \frac{c_i}{2}}\right)$$

Ejemplo

El siguiente ejemplo es tomado del libro (Collett, 2015), ejemplo 1.3 :

Supervivencia de pacientes con myeloma múltiple.

Myeloma múltiple es una enfermedad caracterizada por la acumulación múltiple de células plasmáticas anormales, un tipo de células blancas de la sangre, en la médula ósea. La proliferación de las células plasmáticas anormales dentro de los huesos causa dolor y la destrucción del tejido óseo. El objetivo de un estudio realizado en el Centro Médico de la Universidad del Oeste de Virginia, USA, fue examinar la asociación entre los valores de ciertas variables explicativas(covariables) y el tiempo de supervivencia de los pacientes. En el estudio, el tiempo de supervivencia fue medido en meses, desde el diagnóstico hasta la muerte por myeloma múltiple.

La siguiente tabla muestra un **fragmento** de los resultados obtenidos en el estudio. En ésta se relaciona a un total de 48 pacientes, todos ellos estaban entre los 50 y 80 años. Algunos de estos pacientes no habían muerto durante el tiempo que el estudio fue completado, por lo que estos individuos contribuyeron con tiempos censurados por la derecha. La codificación del estatus de supervivencia de un individuo en la tabla es codificado con un 0 si la observación es censurada y 1 si fue muerte por myeloma.

Patient number	Survival time	Status	Age	Sex	Bun	Ca	Hb	Pcells	Protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12.0	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14.0	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13.0	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14.0	9	0
20	56	0	66	1	18	11	12.5	90	0

Ahora bien, se busca estimar $S(t)$ mediante la construcción de la tabla de vida. La información registrada en otras variables explicativas será ignorada.

Se consideran los intervalos de tiempo, para cada uno se calcula el número de pacientes que fallecieron d_j,el número de datos censurados c_j, el número en riesgo de muerte al inicio de cada uno de estos intervalos n_j, y el número ajustado en riesgo $n_j^* = n_j - \frac{c_j}{2}$ (dado que hay datos censurados). Finalmente, la probabilidad de supervivencia en cada intervalo es estimada(multiplicando cada p_j).

Los cálculos son presentados a continuación:

Interval	Time period	d_j	c_j	n_j	n_j^*	p_j	$S(t)$
1	0-	16	4	48	46.0	0.6521739	0.6521739
2	12-	10	4	28	26.0	0.6153846	0.4013378
3	24-	1	0	14	14.0	0.9285714	0.3726708
4	36-	3	1	13	12.5	0.7600000	0.2832298
5	48-	2	2	9	8.0	0.7500000	0.2124224
6	60-	4	1	5	4.5	0.1111111	0.0236025

Y la curva de supervivencia es la dada en la figura 7.1:

7.2. Estimador Producto-Límite (Kaplan-Meier)

El estimador producto-límite fue propuesto por Kaplan y Meier en 1958 como el estimador máximo-verosímil de la función de supervivencia.

El método de Kaplan-Meier calcula la supervivencia cada vez que un paciente muere. Da proporciones exactas de supervivencia debido a que utiliza tiempos de supervivencia precisos.

La característica distintiva del análisis con este método, es que la proporción acumulada que sobrevive se calcula para el tiempo de supervivencia individual de cada paciente, en contraste con la agrupación de los tiempos de supervivencia en intervalos hechos en la tabla de vida. Por esta razón es especialmente útil para estudios que utilizan un número pequeño de pacientes.

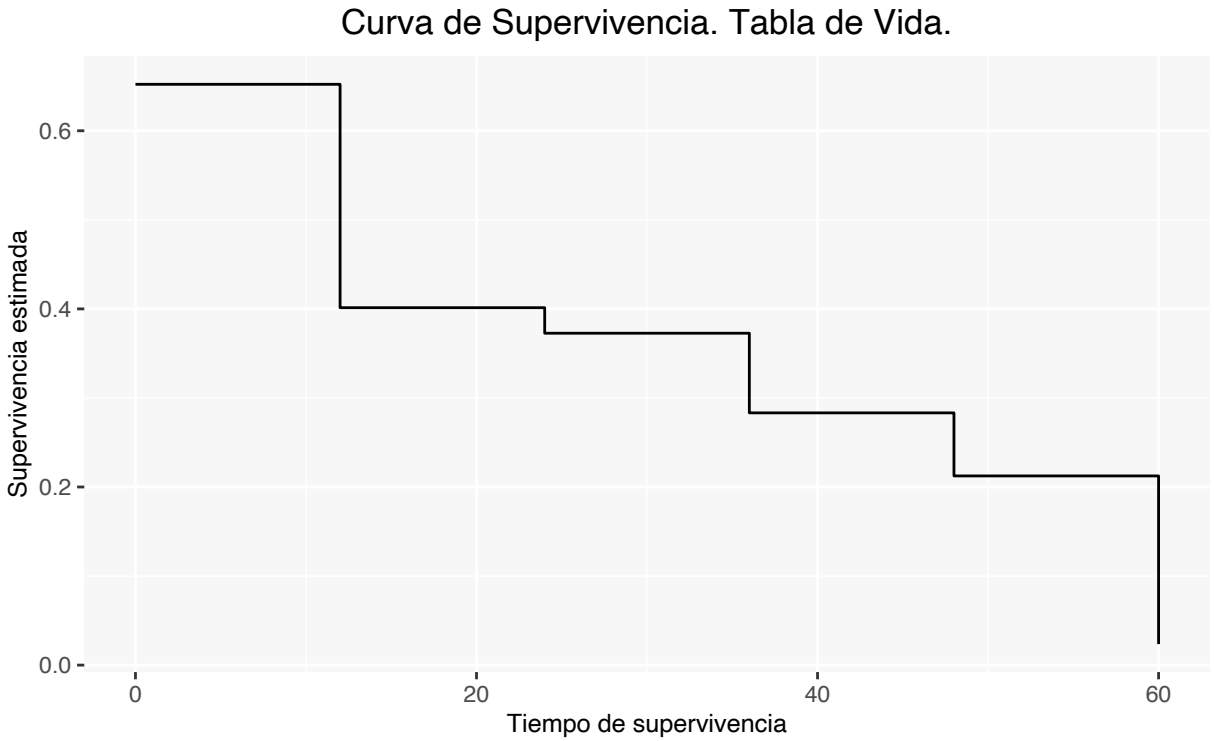


Figura 7.1: Supervivencia estimada para los pacientes con myeloma múltiple con el método actuarial

Este método asume que:

- Las personas que se retiran del estudio tienen un destino parecido a las que se quedan.
- El periodo de tiempo durante el cual una persona entra en el estudio no tiene efecto independiente en la respuesta.

7.2.1. Construcción del Estimador K-M

Es natural pensar a T como una variable aleatoria continua, y por tanto, teóricamente no es posible tener observaciones iguales. No obstante, en la práctica los tiempos de supervivencia son medidos en escalas como: días, meses, años, etcétera; por lo que, hay posibilidad de tener observaciones repetidas. Por esta razón conviene modelar a T como una variable aleatoria discreta. La idea del estimador *K-M* es la siguiente:

Sea $T_1, T_2, ..., T_n$ una m.a. de una población discreta con soporte en $\{u_1, u_2, ...\}$.

La muestra observada de T se puede representar como (t_i, δ_i) para $i = 1, 2, ..., n$ donde:

$$\delta_i = \begin{cases} 0 & \text{si } t_i \textit{ es censurado} \\ 1 & \text{si } t_i \textit{ no es censurado} \end{cases}$$

Entonces la función de verosimilitud será:

$$\mathscr{L} = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Desarrollando la expresión anterior tenemos

$$\mathscr{L} = \prod_{i=1}^{n} (h(u_k) S(u_{k-1})^{\mathbb{1}_{t_i=u_k}})^{\delta_i} (S(u_k)^{\mathbb{1}_{k=max\{j:u_j \leq t_i\}}})^{1-\delta_i}$$

Sea

$$d_k = \sum_{i=1}^{n} \mathbb{1}_{(t_i=u_k,\delta_i=1)} \text{ (numero de tiempos de fallo iguales a } u_k)$$
$$n_k = \sum_{i=1}^{n} \mathbb{1}_{(t_i\geq u_k)} \text{ (numero de individuos en riesgo al tiempo } u_k)$$

$$\Longrightarrow \mathscr{L} = \prod_k (h(u_k))^{d_k}(1-h(u_k))^{n_k-d_k}$$

Ahora maximizamos la función de verosimilitud para $h(u_k)$:

$$\ln(\mathscr{L}) = \sum_k \{d_k \ln(h(u_k)) + (n_k - d_k)\ln(1-h(u_k))\}$$

$$\Longrightarrow \frac{\partial \ln(\mathscr{L})}{\partial h(u_k)} = \frac{d_k}{h(u_k)} - \frac{(n_k-d_k)}{(1-h(u_k))} = 0$$

despejando $h(u_k)$ se tiene

$$\therefore \hat{h}(u_k) = \frac{d_k}{n_k}$$

Dado que los estimadores máximo-verosímiles cumplen con el principio de invarianza, y ocupando que $S(t) = \prod_{k:u_k\leq t}(1-h(u_k))$ (visto anteriormente) tenemos:

$$\hat{S}(t) = \prod_{k:u_k\leq t}\left(1-\frac{d_k}{n_k}\right)$$

Y es así es como se deriva el estimador *K-M.*

Proposición: $\mathbb{E}[\hat{h}(u_k)] = h(u_k)$ (Insesgamiento).

Ejemplo

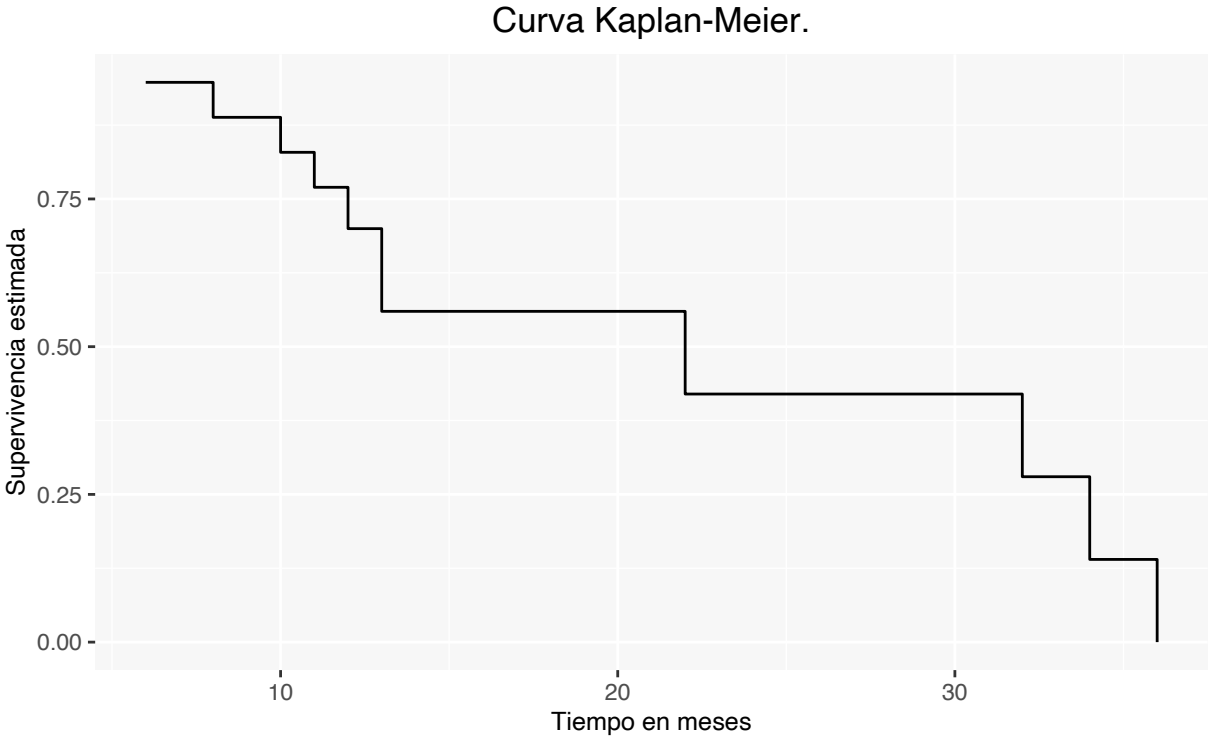
Se obtuvieron los tiempos de remisión de 20 pacientes con osteosarcoma, a los que se trataba con 3 meses de quimioterapia después de amputación.

- 11 pacientes recayeron a los 6, 8, 10, 11, 12, 13, 13, 22, 32, 34 y 36 meses.
- 8 pacientes se retiraron vivos al final del estudio contribuyendo 3, 7, 7, 11, 14, 16, 20 y 20 meses de observación, sin haber sufrido recaídas.
- Un paciente rehusó continuar la terapia a los 11 meses y se retiró del estudio libre de enfermedad.

La siguiente tabla muestra la forma de estimar $S(t)$ por el método *K-M*:

time	d_j	n_k	c_k	$1 - \frac{d_k}{n_k}$	$S(t)$
6	1	19	1	0.9473684	0.9473684
8	1	16	2	0.9375000	0.8881579
10	1	15	0	0.9333333	0.8289474
11	1	14	2	0.9285714	0.7697368
12	1	11	0	0.9090909	0.6997608
13	2	10	0	0.8000000	0.5598086
22	1	4	4	0.7500000	0.4198565
32	1	3	0	0.6666667	0.2799043
34	1	2	0	0.5000000	0.1399522
36	1	1	0	0.0000000	0.0000000

Y la gráfica de $\hat{S}(t)$ es:



Ejercicio

Suponga que disponemos de los datos de supervivencia de 10 pacientes que han sido aleatoriamente asignados a los tratamientos A y B.

- A: 3, 5, 7, 9+, 18
- B: 12, 19, 20, 20+, 33+

Construya la función de supervivencia para cada tratamiento y grafíquelas. ¿Qué se puede decir de los tratamientos a partir de las gráficas?

Capítulo 8

Algunas Estimaciones sobre Modelos No Paramétricos

Cuando se emplean los modelos no paramétricos, es necesario hacer inferencia más allá de la estimación puntual. Para tal propósito, conocer la varianza de algunos estimadores de interés, resulta de gran ayuda.

8.1. Estimación de la Varianza para el Estimador de $S(t)$

8.1.1. Tabla de Vida

Para la tabla de vida se tienen los siguientes resultados:

$$\hat{q_j} = \frac{d_j}{n_j - \frac{c_j}{2}} \quad \hat{S}_j = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i - \frac{c_i}{2}}\right)$$

Entonces un estimador de la varianza para $\hat{S}_j$ es:

$$\hat{Var}(\hat{S}_j) = \hat{S}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{\hat{p}_i(n_i - \frac{c_i}{2})}$$

Si deseamos obtener intervalos de confianza puntuales para $S(t)$, podemos partir de la distribución asintótica de $\hat{S}_j$:

$$\hat{S}_j \sim N(S_j, \hat{Var}(\hat{S}_j))$$

De donde, un intervalo de confianza para $S(t_j)$ al $(1-\alpha) * 100\,\%$ es:

$$\hat{S}_j \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{S}_j)}$$

8.1.2. Kaplan-Meier

Partiendo de que

$$\hat{S}(t) = \prod_{k:u_k \leq t} (1 - h_k)$$

Si aplicamos ln tenemos:

$$\ln(\hat{S(t)}) = \sum_{k:u_k \leq t} \ln(1 - h_k)$$

$$\Longrightarrow Var(\ln(\hat{S}(t))) = \sum_{k:u_k \leq t} Var(\ln(1 - h_k))$$

Desarrollando con series de Taylor (método Delta) se obtiene:

$$Var(\hat{S}(t)) = \hat{S}^2(t) \sum_{k:u_k \leq t} \frac{Var(\hat{h}_k)}{(1 - \hat{h}_k)^2}$$

Uno de los **estimadores** más comunes para $Var(\hat{S}(t))$ es el **estimador de Greenwood**; éste supone que el número de individuos que sobreviven a lo largo del intervalo que empieza en t_j tiene una distribución *Binomial* con parámetros n_j y p_j (p_j es la verdadera probabilidad de supervivencia a lo largo del intervalo). De manera que, el estimador Greenwood es:

$$\hat{Var}(\hat{S}(t)) \approx \hat{S}^2(t) \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

- El estimador del error estándar será: $\sqrt{\hat{Var}(\hat{S}(t))}$.
- Cuando $n \to \infty$, $\hat{S}(t)$ tiene una distribución normal: $\hat{S}(t) \sim N(S(t), Var(\hat{S}(t)))$.
- El intervalo puntual al $(1 - \alpha) * 100\,\%$ de confianza para $S(t_0)$ será: $\hat{S}(t_0) \pm Z_{1-\alpha/2}\sqrt{\hat{Var}(\hat{S}(t_0))}$[1].

Ejercicio

Un estudio consistió en medir el tiempo(en meses) en que los pacientes desarrollaron un cierto tipo de tumor. Los resultados que se obtuvieron fueron:

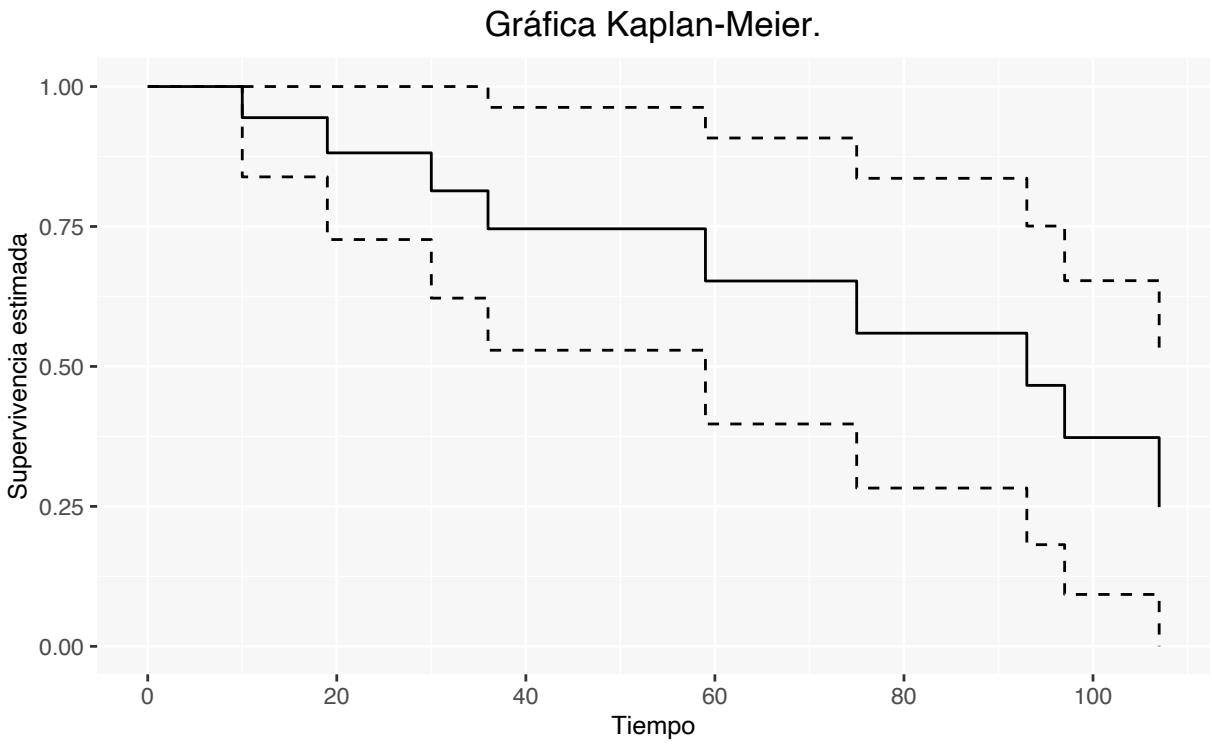
$$10, 13^+, 18^+, 19, 23^+, 30, 36, 38^+, 54^+, 56^+, 59, 75, 93, 97, 104^+, 107, 107^+, 107^+$$

La siguiente tabla muestra $\hat{S}(t)$ por el método de *K-M*, además de los intervalos de confianza puntuales utilizando el estimador de *Greenwood*:

time	d_k	n_k	c_k	$S(t)$	std error	Lower_95 %	Upper_95 %
10	1	18	0	0.9444444	0.0539903	0.8386254	1.0000000
19	1	15	2	0.8814815	0.0789892	0.7266655	1.0000000
30	1	13	1	0.8136752	0.0977770	0.6220358	1.0000000
36	1	12	0	0.7458689	0.1106702	0.5289593	0.9627785
59	1	8	3	0.6526353	0.1303197	0.3972133	0.9080573
75	1	7	0	0.5594017	0.1411673	0.2827189	0.8360845
93	1	6	0	0.4661681	0.1451991	0.1815830	0.7507531
97	1	5	0	0.3729345	0.1429930	0.0926734	0.6531955
107	1	3	3	0.2486230	0.1392472	0.0000000	0.5215425

La gráfica correspondiente es:

[1] Posiblemente al realizar las cuentas, como en el ejemplo que se encuentra en esta sub sección, se obtengan valores negativos en alguno de los límites inferiores del intervalo de confianza. Esto es porque la expresión matemática del intervalo de confianza permite este tipo de valores, pero al tratarse de una función de supervivencia $S(t)$ su rango serán valores en $[0, 1]$, por lo que, por convención, se coloca el valor 0 cuando se tiene un número negativo.



8.2. Estimadores de la Función de Riesgo Acumulada

Recordemos que:

$$H(t) = -log(S(t))$$

Entonces podemos estimar $H(t)$ de una manera sencilla ($\hat{S}(t)$ es obtenido por *K-M*):

$$\hat{H}_1(t) = -log(\hat{S}(t))$$

No obstante, una alternativa para conocer $H(t)$ es utilizar el **estimador Nelson-Aalen**. Este estimador resulta útil primordialmente en el análisis de datos para la selección entre modelos paramétricos para el tiempo de fallo T, además proporciona un estimador para $h(t)$. El estimador *Nelson-Aalen* asume que $H(t)$ es la suma de riesgos, esto es:

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

Entonces el estimador de $H(t)$ será:

$$\hat{H}_2(t) = \sum_{j:u_j \leq t} \hat{h}_j = \sum_{j:u_j \leq t} \frac{d_j}{n_j}$$

Observe que si tomamos $\hat{H}_2(t)$, podemos estimar de otra manera a $S(t)$. Es decir, $S(t)$, usando **Nelson-Aalen**, es:

$$\hat{S}_2(t) = exp\{-\hat{H}_2(t)\} = exp\left\{-\sum_{j:u_j \leq t} \frac{d_j}{n_j}\right\}$$

Un estimador de la varianza de $\hat{H}_2(t)$ es:

$$\hat{Var}(\hat{H}_2(t)) = \sum_{j:t_j \leq t} \frac{d_j}{n_j^2}$$

Y el intervalo de confianza puntual al $(1-\alpha) * 100\,\%$ para $H_2(t_0)$ es:

$$\hat{H}_2(t_0) \pm Z_{1-\alpha/2} \frac{\sqrt{\hat{Var}\{\hat{H}_2(t)\}}}{\hat{H}_2(t_0)}$$

Ejemplo

Siguiendo con el ejemplo anterior, que mide el tiempo (en meses) en que los pacientes desarrollaron un cierto tipo de tumor.

La tabla muestra $\hat{S}(t)$ por el método de *K-M*, entonces podemos calcular $\hat{H}_1(t) = -log(\hat{S}(t))$. Y por otro lado usar el estimador Nelson-Aalen para obtener $\hat{H}_2(t)$.

time	d_k	n_k	c_k	$S(t)$	$H_1(t)$	$H_2(t)$
10	1	18	0	0.9444444	0.0571584	0.0555556
19	1	15	2	0.8814815	0.1261513	0.1222222
30	1	13	1	0.8136752	0.2061940	0.1991453
36	1	12	0	0.7458689	0.2932054	0.2824786
59	1	8	3	0.6526353	0.4267368	0.4074786
75	1	7	0	0.5594017	0.5808874	0.5503358
93	1	6	0	0.4661681	0.7632090	0.7170024
97	1	5	0	0.3729345	0.9863526	0.9170024
107	1	3	3	0.2486230	1.3918177	1.2503358

8.3. Estimación Puntual de la Media

Puede ser de interés estudiar algunos parámetros poblacionales que dependen de la función de supervivencia. Por ejemplo: la media, la mediana y cualquier cuantil o percentil.

Hemos visto en la sección 4.1 que:

$$\mu = \int_0^\infty S(t)dt$$

Entonces podemos reemplazar $S(t)$ con el estimador de $\hat{S}(t)$(obtenido por *K-M*), por lo que:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt$$

Observación: El estimador será apropiado si la observación más grande del conjunto de datos es un tiempo de falla y **NO** una observación censurada.

Si la última observación(la más grande) es una observación censurada, entonces $\hat{\mu}$ se calcula como:

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt$$

Donde τ es el valor que determina el tiempo más grande al que una persona puede sobrevivir.

Un estimador de la varianza de $\hat{\mu}_\tau$ con $t_1, t_2, ..., t_k$ tiempos de fallo observados, es:

$$\hat{Var}(\hat{\mu_\tau}) = \sum_{i=1}^{k} \left\{ \int_{t_i}^{\tau} \hat{S}(t)dt \right\}^2 \cdot \frac{d_i}{n_i(n_i - d_i)}$$

El intervalo al $(1-\alpha)100\,\%$ de confianza para μ_τ es:

$$\hat{\mu}_\tau \pm Z_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\mu}_\tau)}$$

NOTA: Es importante que, cuando se esté calculando la media de supervivencia $\hat{\mu}_\tau$, se verifique que el último valor no sea censurado y se observe en qué intervalo se está calculando la media.

8.4. Estimación de Cuantiles

Recordemos que los cuantiles de orden p, t_p, es el mínimo valor t tal que $S(t) \leq 1-p$. Usando el estimador $\hat{S}(t)$ obtenido por *K-M*, tenemos:

$$\hat{t}_p = inf\{t : \hat{S}(t) \leq 1-p\}$$

Por lo que, si deseamos calcular la **mediana estimada**:

$$\hat{t}_{0.5} = inf\{t : \hat{S}(t) \leq 0.5\}$$

8.5. Bandas de Confianza para la Función de Supervivencia

Deseamos encontrar dos v.a. $L(t)$ y $U(t)$ tales que:

$$\mathbb{P}(L(t) \leq S(t) \leq U(t)) = 1-\alpha \quad \forall t_L \leq t \leq t_U$$

Entonces $[L(t), U(t)]$ serán las bandas al $(1-\alpha) * 100\,\%$ de confianza de $S(t)$. Puede usted decir ¿cuál es la diferencia entre *bandas de confianza* e *intervalos de confianza puntuales*?

Hay dos métodos de aproximación para las bandas de confianza. La primera aproximación fue propuesta por *Nair* (Nair, 1984) y esencialmente proporciona límites de confianza que son proporcionales a los intervalos de confianza puntuales, estas bandas son llamadas *bandas de probabilidad iguales* o *bandas EP*[2]. Caso contrario, la segunda aproximación, propuesta por *Hall* y *Wellner*[3] (Hall and Wellner, 1980), establece bandas no proporcionales a los intervalos de confianza puntuales.

Nuestro interés es obtener las bandas de confianza para $S(t)$; afortunadamente hoy en día se cuenta con software que nos ayuda en dicha tarea, específicamente en **R** es sencillo obtenerlo.

[2]Tanto para estas bandas como las de Hall & Wellner se consideran los D eventos distintos del tiempo $t_1 < t_2 < \cdots < t_D$; $t_L < t_U$ el rango de tiempo para las bandas de confianza, por lo que t_U es menor o igual al evento de tiempo más grande. t_L, en este caso es más grande o igual al tiempo de evento más pequeño y para ambas bandas se tiene que $\alpha_L = \frac{n\hat{\sigma}^2_S(t_L)}{1+n\hat{\sigma}^2_S(t_L)}$ y $\alpha_U = \frac{n\hat{\sigma}^2_S(t_U)}{1+n\hat{\sigma}^2_S(t_U)}$. Las bandas EP al $100(1-\alpha)\,\%$ están dadas por

$$\hat{S}(t) - e_\alpha(\alpha_L, \alpha_U)\hat{S}(t)\sigma_S(t) \leq S(t) \leq \hat{S}(t) + e_\alpha(\alpha_L, \alpha_U)\hat{S}(t)\sigma_S(t)$$

para todos los $t_L \leq t \leq t_U$, donde $e_\alpha(\alpha_L, \alpha_U)$ es un cuantil tal que $\alpha = \mathbb{P}\left(\sup_{\alpha_L \leq u \leq \alpha_U} \frac{|W^0(u)|}{[u(1-u)]^{1/2}} > e_\alpha(\alpha_L, \alpha_U) \right)$ donde $W^0(u)$ es un puente browninano con $0 \leq u \leq 1$ y $\hat{S}(t)\sigma_S(t)$ se derivada del estimador de Greenwood.

[3]Las bandas Hall and Wellner al $100(1-\alpha)\,\%$ están dadas por $\hat{S}(t) - h_\alpha(\alpha_L, \alpha_U)\sqrt{n}\left[1 + n\hat{\sigma}^2_S(t)\right]\hat{S}(t) \leq S(t) \leq \hat{S}(t) + h_\alpha(\alpha_L, \alpha_U)\sqrt{n}\left[1 + n\hat{\sigma}^2_S(t)\right]\hat{S}(t)$ para todos los $t_L \leq t \leq t_U$ donde los valores críticos $h_\alpha(\alpha_L, \alpha_U)$ esta dado por $\alpha = \mathbb{P}\left(\sup_{\alpha_L \leq u \leq \alpha_U} |W^0(u)| > h_\alpha(\alpha_L, \alpha_U) \right)$. En este caso, t_L puede ser cero. Se puede consultar más sobre estas expresiones en el siguiente enlace.

8.6. Diagnóstico para el Uso de Modelos Paramétricos

Si existe la necesidad de encontrar un modelo paramétrico que ajuste "bien" a los datos, se pueden emplear algunos métodos gráficos que nos den algunos *indicios* sobre la posible *distribución* de los datos. Tales indicios pueden ser proporcionados por los métodos no paramétricos.

8.6.1. Gráficas de las Funciones de Supervivencia

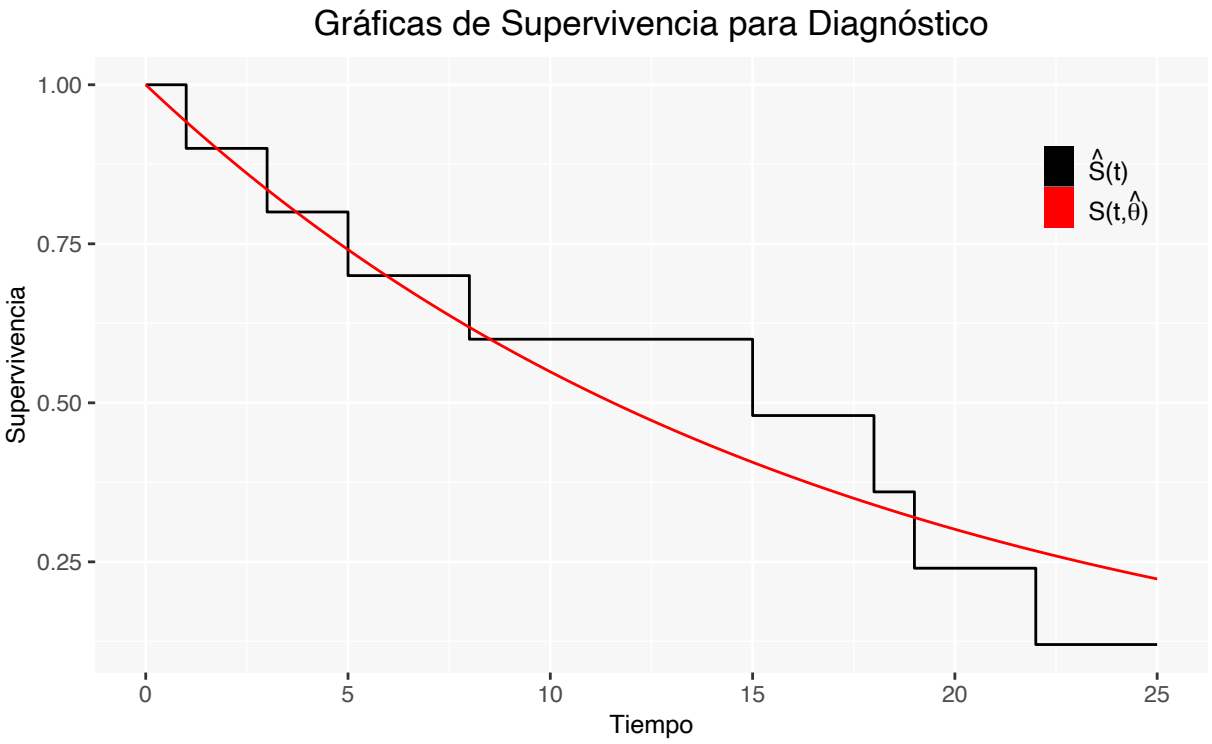
Se conoce la función de supervivencia paramétrica $S(t;\theta)$ y se tiene un estimador $\hat{\theta}$. Si el modelo paramétrico es *adecuado* entonces $S(t;\hat{\theta})$ y $\hat{S}(t)$(estimada por K-M) deben ser *similares*.

Ejemplo

Se tienen los siguientes tiempos de supervivencia:

$$8, 5, 10^{+}, 1, 3, 18, 22, 15, 25^{+}, 19$$

Se desea saber si estos datos ajustan a un modelo *exponencial* con $\lambda = 0.06$. Realizamos las gráficas de supervivencia:



De acuerdo a la gráfica anterior, ¿Existe algún indicio de que los datos siguen una distribución exponencial con $\lambda = 0.06$?

8.6.2. Gráfica $P - P$

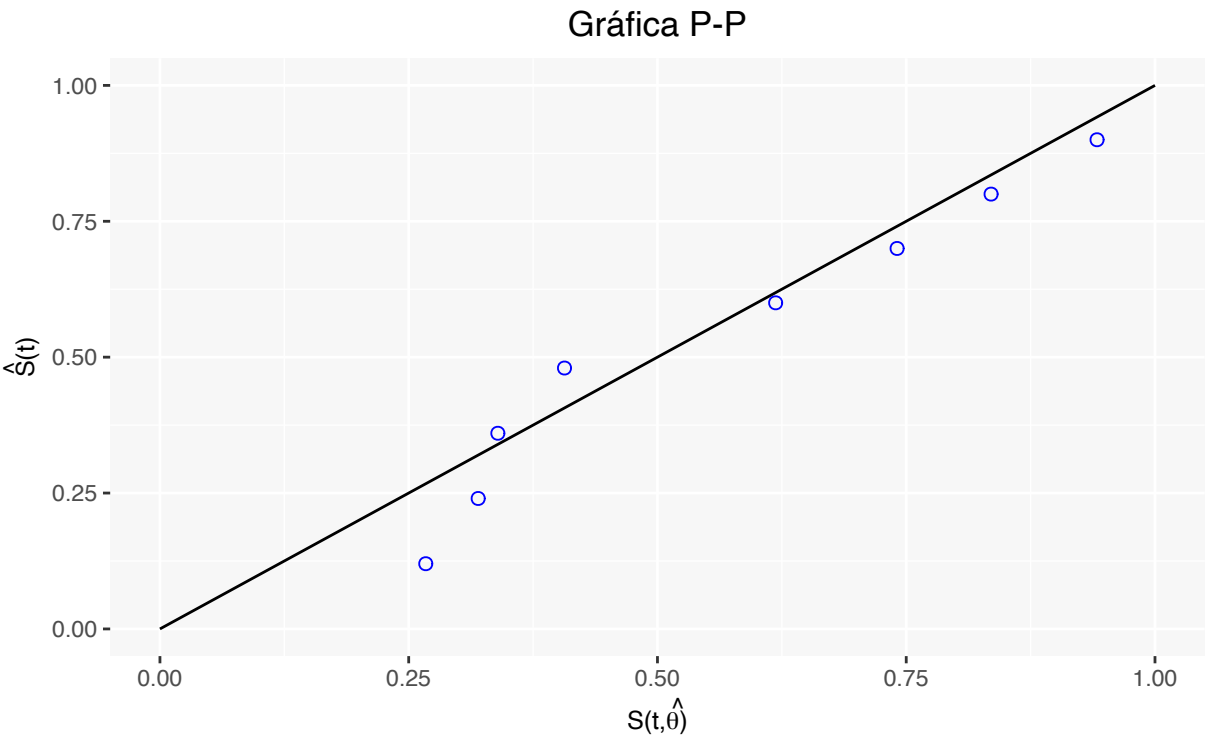
Si el modelo paramétrico es adecuado entonces los puntos:

$$\left(S(t_j;\hat{\theta}), \hat{S}(t_j)\right)$$

Deberán caer en la recta identidad ($\hat{S}(t_j)$ es estimada por K-M). Es llamada *gráfica* $P-P$ pues se gráfica *Probabilidad vs Probabilidad.*

Ejemplo

Tomando los datos del ejemplo anterior, queremos ver si los datos se distribuyen posiblemente exponencial con $\lambda = 0.06$. Si hacemos la gráfica P-P se tiene:

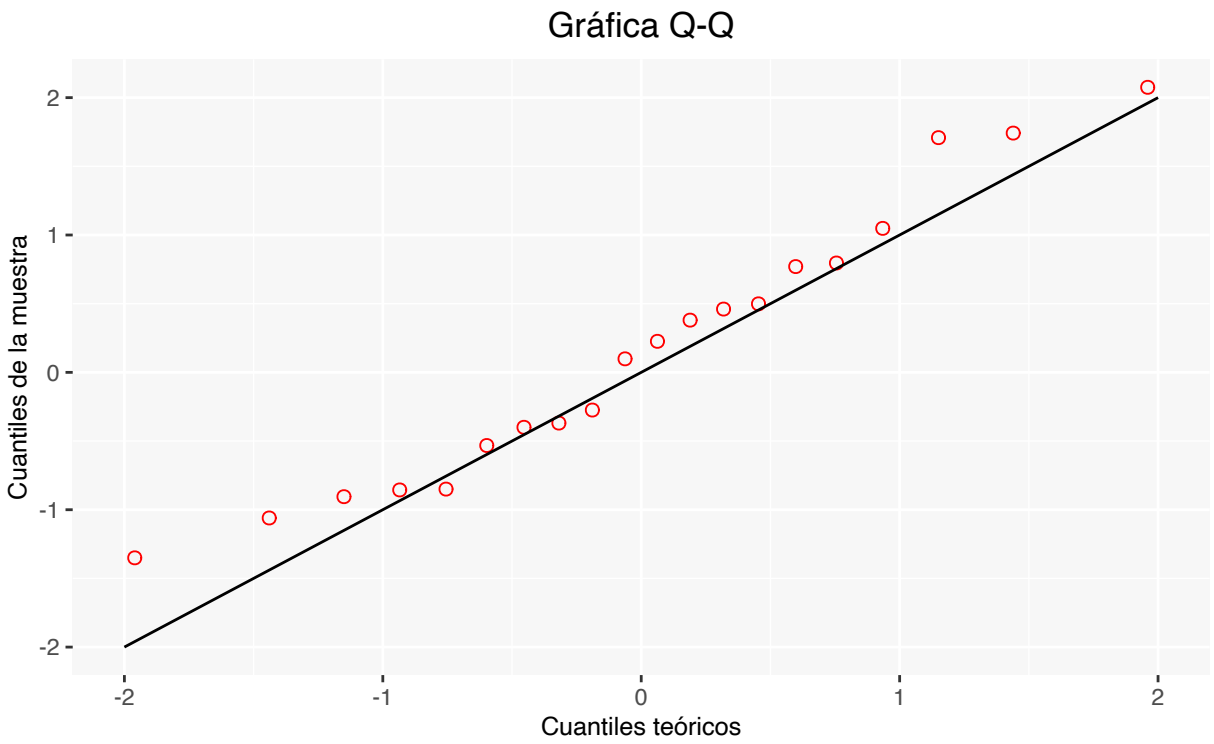


De acuerdo a la gráfica $P-P$, ¿Es el modelo exponencial con $\lambda = 0.06$ adecuado a los datos?

8.6.3. Gráfica $Q-Q$

La *gráfica* $Q-Q$ consiste en graficar diversos cuantiles de los **datos** *vs* diversos cuantiles del modelo paramétrico propuesto. Si el modelo teórico propuesto es "adecuado" a los datos entonces la gráfica será cercana a la función identidad.

A continuación se muestra un ejemplo de la gráfica $Q-Q$:



8.6.4. Linearización de la Función de Supervivencia

Cuando tenemos un modelo paramétrico $S(t;\theta)$ pero no conocemos $\hat{\theta}$ podemos hacer *linearización de la función de supervivencia*. Es decir, si existen g_1 y g_2 tales que $g_1\{S(t;\theta)\}$ es una función *lineal* de $g_2(t)$ entonces podemos graficar $g_1(\hat{S}(t))$ *vs* $g_2(t)$. Y si la familia paramétrica es adecuada, la gráfica se aproxima a una linea recta. Para este método gráfico se utiliza el estimador **Nelson-Aalen** para $\hat{S}(t)$.

Modelo Exponencial

Tenemos que:

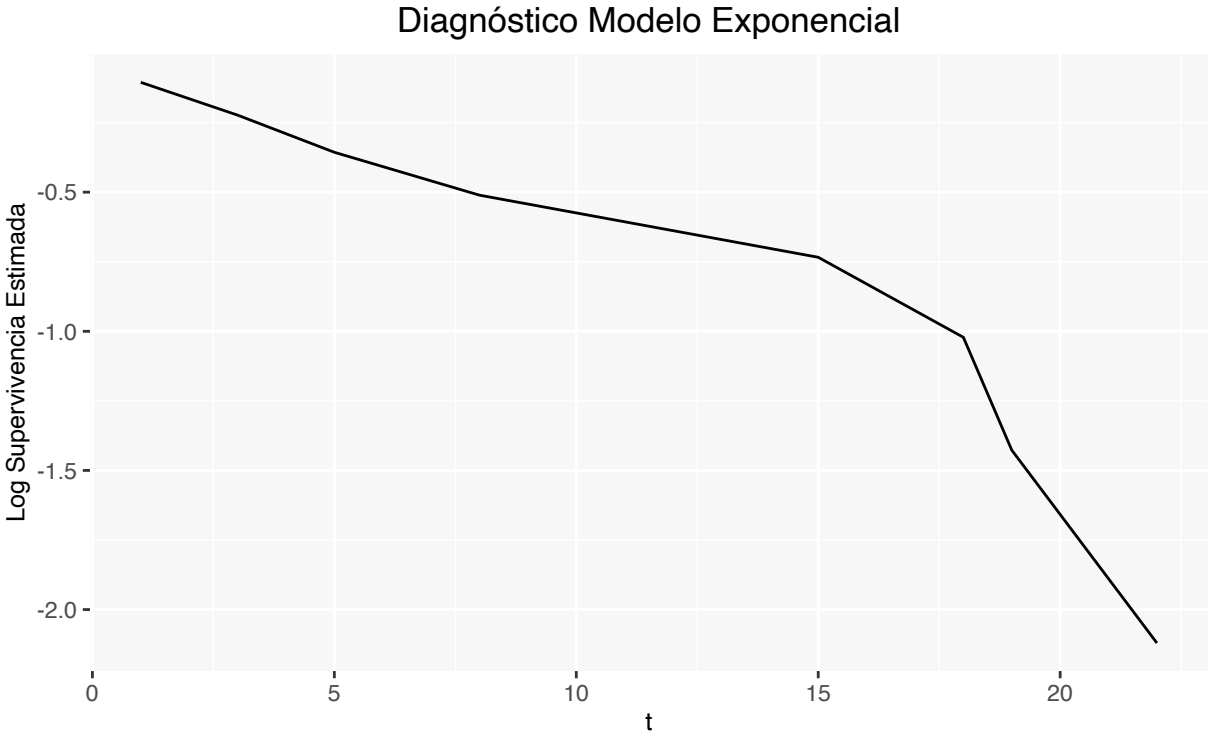
$$S(t) = e^{-\lambda t}$$

$$\Longrightarrow log\{S(t)\} = -\lambda t$$

De manera que $g_1 = log\{S(t)\}$ y $g_2 = t$. Entonces debemos graficar $log(\hat{S}(t))$ *vs* t, y si los datos se ajustan a un modelo exponencial, dicha gráfica debe ser aproximadamente una linea recta con pendiente $-\lambda$.

Ejemplo

Consideremos los datos: $8, 5, 10^{+}, 1, 3, 18, 22, 15, 25^{+}, 19$, veamos si se pueden ajustar posiblemente a un modelo exponencial a través de la gráfica de linearización. Entonces tenemos:



¿Qué puede decir a partir de la gráfica anterior?, ¿Puede asumirse que los datos se ajustan a un modelo exponencial?

Modelo Weibull

Para el modelo Weibull tenemos que:

$$S(t) = e^{-(\lambda t)^{\gamma}}$$

$$\Longrightarrow -log\{S(t)\} = (\lambda t)^{\gamma}$$

$$\Longrightarrow log\{-log\{S(t)\}\} = \gamma(log(\lambda) + log(t))$$

Entonces $g_1 = log\{-log\{S(t)\}\}$ y $g_2 = log(t)$. Por lo que debemos graficar $log\{-log\{\hat{S}(t)\}\}$ *vs* $log(t)$, y si los datos se ajustan a un modelo Weibull, la gráfica debe ser aproximadamente una linea recta con intercepto $\gamma log(\lambda)$ y pendiente γ.

Modelo Log-Logístico

La función de supervivencia es:

$$S(t) = (1 + \lambda t^{\alpha})^{-1}$$

$$\Longrightarrow -log\{S(t)\} = log(1 + \lambda t^{\alpha})$$

$$\Longrightarrow log\{exp\{-log\{S(t)\}\} - 1\} = log(\lambda) + \alpha log(t)$$

La gráfica de diagnóstico es: $log\{exp\{-log\{\hat{S}(t)\}\} - 1\}$ *vs* $log(t)$. Si el modelo Log-Logístico es adecuado a los datos, entonces la gráfica es aproximadamente una linea recta con intercepto $log(\lambda)$ y pendiente α.

Modelo Log-Normal

Para este modelo, la gráfica que debemos emplear para el diagnóstico es $\Phi^{-1}[1 - exp(-log(\hat{S}(t)))]$ *vs* $log(t)$, donde Φ^{-1} son los cuantiles de una *Normal estándar*. Una linea recta(aproximadamente) indica una posible distribución Log-Normal en los datos.

Los métodos gráficos que hemos visto son una herramienta que proporcionan indicios sobre la posible distribución de los datos(si nuestro interés es encontrar un modelo paramétrico). Evidentemente una gráfica no **demuestra nada** por lo que tendríamos que realizar pruebas formales de **bondad de ajuste** (pruebas que consideren la censura y/o truncamiento).

Capítulo 9

Pruebas de Hipótesis

Las pruebas de hipótesis juegan un papel importante en la inferencia estadística. En el contexto del análisis de supervivencia, resulta primordial comparar poblaciones en cuanto a sus funciones de supervivencia, pues de ello podemos saber, por ejemplo, la *eficacia* de un tratamiento respecto a otro, los tiempos de aparición de un tumor en dos grupos, entre otras cosas.

El objetivo de la comparación de poblaciones en el análisis de supervivencia es similar a aquellos procedimientos diseñados para comparar estadísticos provenientes de muestras independientes, como la *prueba t*, la *prueba de los signos*, la *prueba U de Mann-Whitney*(1947), la *prueba de Kruskal-Wallis*(1952), etcétera. Todas estas pruebas de comparación se utilizan para evaluar diferencias entre estadísticos que han sido estimados basados en la información que se obtiene de subgrupos poblacionales independientes entre si. No obstante, dichas pruebas **no consideran la censura** en los datos, y por esta razón se imposibilita su aplicación "directa" en datos de supervivencia.

Las pruebas más utilizadas para comparar funciones de supervivencia, las cuales consideran la censura en los datos, son: la *prueba Log-Rank* propuesta por Mantel-Haenszel(1959), la *prueba generalizada de Wilcoxon* propuesta por Gehan(1965), la *prueba de Peto-Peto*(1972), la *prueba de Tarone-Ware*(1977), la *prueba de Harrington-Fleming*(1982) que generaliza parte de las pruebas anteriores y una versión más general propuesta por Fleming et al. (1987). En esta sección veremos la prueba de *Log-Rank* y la prueba generalizada de *Wilcoxon*, que esencialmente son la misma salvo una ponderación.

9.1. Comparación de 1 población

Suponiendo que la tasa de riesgo real en la población estudiada es $h(t)$, el objetivo en esta prueba es determinar, estadisticamente, si la tasa de riesgo $h_0(t)$, la cual está completamente especificado en el intervalo $(0, \tau)$, es adecuada para las observaciones a tratar para un tiempo fijo τ, es decir:

$$H_0 : h(t) = h_0(t) \ \forall \ t \leq \tau \quad vs \quad H_a : h(t) \neq h_0(t); \quad \text{p.a} \ \ t \leq \tau$$

Recordando que el estimador de la función de riesgo acumulado es $\hat{H}_2(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$, entonces $\frac{d_i}{n_i}$ es un estimador de la tasa de riesgo en t_i.

Cuando la hipótesis nula es cierta, el valor esperado de la tasa de riesgo en t_i es $h_0(ti)$. Ahora, sea $\omega(t)$ la función de pesos tal que $\omega(t) = 0$ si $n_t = 0$, es decir que no se otorga algún peso cuando no hay elementos en dicho tiempo.

Considerando el clásico estadístico de prueba $Z(\tau) = O(\tau) - \mathbb{E}(\tau)$ (elementos observados menos esperados) [1] y considerando que D es el número de tiempos en la muestra, se tiene lo siguiente:

$$Z(\tau) = O(\tau) - \mathbb{E}(\tau) = \sum_{i=1}^{D} \omega(ti) \frac{d_i}{n_i} - \int_0^{\tau} \omega(s) h_0(s) ds$$

[1] Ver el parecido con el estadístico utilizado en pruebas de independencia.

Sólo para confirmar, $O(\tau)$ representa el número de eventos al tiempo τ. Cuando H_0 es cierta

$$Var(Z(\tau)) = \int_0^\tau \omega^2(s)\frac{h_0(s)}{n_s}ds$$

Entonces, para muestras grandes el siguiente estadístico tiene una distribución $\chi^2_{(1)}$[2].

$$T = \frac{[Z(\tau)]^2}{Var(Z(\tau))} \sim \chi^2_{(1)}$$

Cuando τ es igual al tiempo mayor en el estudio:

$$\mathbb{E}[\tau] = Var(Z(\tau)) = \sum_{j=1}^{n}(H_0(T_j) - H_0(L_j))$$

donde $H_0(t)$ es la función de riesgo acumulado bajo la hipótesis nula, L_j es la edad de entrada y T_j la edad de salida.

Ejemplo

Se tiene una muestra de 26 pacientes con cierta enfermedad. Se desea probar que la tasa de riesgo es similar a la tasa de mortalidad de la población de Iowa en 1960. Realizando los cálculos correspondientes, se tiene lo siguiente

$$Z(\tau) = O(26) - \mathbb{E}(26) = 15 - 4.4740 \implies T = \frac{(15 - 4.4740)^2}{4.4740} = 24.76457$$

El $p - value$ de $T = 24.76457$ (6.4777279e-07) es cercano a cero $\implies$ se rechaza la hipótesis nula. $\therefore$ La tasa de riesgo NO es similar a la tasa de mortalidad de Iowa en 1960.

9.2. Prueba Log-Rank

Mantel-Haenszel(1959) propusieron un estadístico que permite relacionar las pruebas de asociación de las tablas de contingencia con los contrastes de igualdad de funciones de supervivencia entre subgrupos poblacionales.

Suponga que se quiere contrastar las funciones de supervivencia de dos grupos poblacionales, digamos Grupo 1 y Grupo 2:

$$H_0 : S_1(t) = S_2(t) \quad \forall t > 0 \quad vs \quad H_a : S_1(t) \neq S_2(t) \quad \text{p.a } t > 0$$

Suponga además que hay k tiempos diferentes de **ocurrencia** del evento(fallas) en el grupo *combinado*, digamos: $t_{(1)}, t_{(2)}, ..., t_{(k)}$ y que en el momento t_i ocurren d_{1i} eventos en el primer grupo y d_{2i} eventos en el segundo, para todo $i = 1, 2, ..., k$. En cada momento t_i, hay n_{1i} individuos en riesgo en el primer grupo y n_{2i} individuos en el segundo.

En consecuencia, en el momento t_i habrá $d_i = d_{1i} + d_{2i}$ (fallas totales) y $n_i = n_{1i} + n_{2i}$ (total de individuos en riesgo). La siguiente tabla muestra el número de ocurrencias del evento en el momento t_i, para el Grupo 1 y Grupo 2:

[2] ¿Qué distribución de conteo tiene la misma varianza que su media?

Grupo	Fallos t_i	Sobrevivientes t_i	Individuos en riesgo t_i
I	d_{1i}	$n_{i1} - d_{1i}$	n_{1i}
II	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Totales	d_i	$n_i - d_i$	n_i

Si se considera que en el momento i-ésimo se tiene una población formada por dos grupos, Grupo 1 y Grupo 2, y se define la variable aleatoria d_{1i} como el número de eventos que ocurren en el Grupo 1 en el momento t_i. En ese momento se tiene una población de tamaño n_i definida por el total de individuos en riesgo, clasificada en dos subpoblaciones de tamaños n_{1i} (Grupo 1) y n_{2i} (Grupo 2).

Si se asume que el número de fallas d_i para los dos grupos combinados es una muestra aleatoria(sin reemplazo) de la población anterior, entonces la v.a. d_{1i} sigue una distribución *hipergeométrica*(n_i, d_i, n_{1i}) cuya media y varianza son:

$$e_{1i} = \mathbb{E}(d_{1i}) = n_{1i}\frac{d_i}{n_i}$$

$$V_{1i} = Var(d_{1i}) = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{n_i^2(n_i - 1)}$$

La hipótesis nula H_0 que se desea probar es que **no** hay diferencia entre las funciones de supervivencia de ambos grupos, lo que se logra evaluando la diferencia entre el número de fallas *observadas* y el número de fallas *esperadas* en cada uno de los momentos de ocurrencia, bajo los supuestos de H_0. Esto es equivalente a comparar el número de fallas ocurridas en cualquiera de los grupos con respecto al número de fallas esperadas en el grupo combinado.

De manera que, la prueba **Log-Rank** se basa en el estadístico:

$$U_L = \sum_{i=1}^{k}(d_{1i} - e_{1i})$$

Entonces bajo H_0 (las supervivencias en las dos poblaciones son iguales) tenemos que : $\mathbb{E}(U_L) = 0$ y $Var(U_L) = \sum_{i=1}^{k} V_{1i}$. En consecuencia:

$$L = \frac{U_L - \mathbb{E}(U_L)}{\sqrt{Var(U_L)}} = \frac{\sum_{i=1}^{k}(d_{1i} - e_{1i})}{\sqrt{\sum_{i=1}^{k} V_{1i}}} \sim N(0,1)$$

$$\Longrightarrow L^2 = \frac{(\sum_{i=1}^{k}(d_{1i} - e_{1i}))^2}{\sum_{i=1}^{k} V_{1i}} \sim \chi^2_{(1)}$$

Finalmente, la **estadística** que ocuparemos para la prueba **Log-Rank** es:

$$L^2 = \frac{(\sum_{i=1}^{k}(d_{1i} - e_{1i}))^2}{\sum_{i=1}^{k} V_{1i}}$$

La **Regla de Decisión** es rechazar H_0 al nivel de significancia α si:

$$L^2 > J_{1-\alpha}$$

Donde $J_{1-\alpha}$ es el cuantil $1-\alpha$ de una $\chi^2_{(1)}$. Recordemos que la regla de decisión puede obtenerse, también, a través del $p - value$.

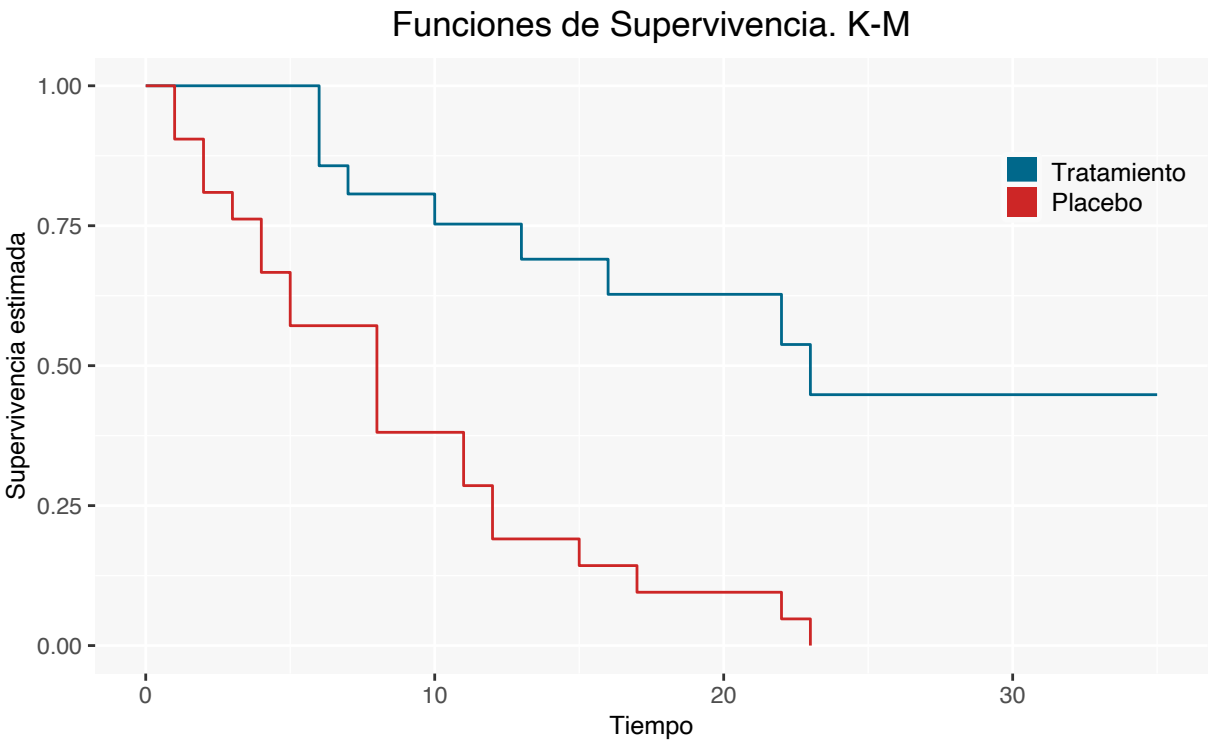
La prueba Log-Rank es muy potente para detectar diferencias cuando los logaritmos de las funciones de supervivencia son *proporcionales*, no obstante, la potencia de la prueba disminuye cuando las funciones de supervivencia se **cruzan**.

Ejemplo

Se tienen los tiempos de remisión (en semanas) para dos grupos de pacientes con leucemia. Cada grupo se conforma por 21 pacientes.

Grupo 1 (Tratamiento)	Grupo 2 (Placebo)
6, 6, 6	1, 1, 2
7, 10, 13	2, 3, 4
16, 22, 23	4, 5, 5
6+, 9+, 10+	8, 8, 8
11+, 17+, 19+	8, 11, 11
20+, 25+, 32+	12, 12, 15
32+, 34+, 35+	17, 22, 23

La forma de la función de supervivencia para cada grupo es:



Las funciones de supervivencia de los dos grupos "parecen" tener diferencia, de acuerdo a la gráfica anterior. No obstante, debemos comprobar si es una diferencia *significativa*. Enseguida corroboraremos tal afirmación mediante la prueba Log-Rank.

La siguiente tabla muestra los individuos que fallan y los individuos en riesgo de cada grupo para cada tiempo t_i $(i = 1, 2, ..., k)$, donde k=17. Además, se muestran los cálculos de las fallas esperadas para la construcción del estadístico Log-Rank:

i	t_i	d_{1i}	d_{2i}	n_{1i}	n_{2i}	e_{1i}	e_{2i}	$d_{1i}-e_{1i}$	$d_{2i}-e_{2i}$	V_{1i}
1	1	0	2	21	21	1	1	-1	1	0.49
2	2	0	2	21	19	1.05	0.95	-1.05	1.05	0.49
3	3	0	1	21	17	0.55	0.45	-0.55	0.55	0.25
4	4	0	2	21	16	1.14	0.86	-1.14	1.14	0.48
5	5	0	2	21	14	1.2	0.8	-1.2	1.2	0.47
6	6	3	0	21	12	1.91	1.09	1.09	-1.09	0.65
7	7	1	0	17	12	0.59	0.41	0.41	-0.41	0.24
8	8	0	4	16	12	2.29	1.71	-2.29	2.29	0.87
9	10	1	0	15	8	0.65	0.35	0.35	-0.35	0.23
10	11	0	2	13	8	1.24	0.76	-1.24	1.24	0.45
11	12	0	2	12	6	1.33	0.67	-1.33	1.33	0.42
12	13	1	0	12	4	0.75	0.25	0.25	-0.25	0.19
13	15	0	1	11	4	0.73	0.27	-0.73	0.73	0.2
14	16	1	0	11	3	0.79	0.21	0.21	-0.21	0.17
15	17	0	1	10	3	0.77	0.23	-0.77	0.77	0.18
16	22	1	1	7	2	1.56	0.44	-0.56	0.56	0.3
17	23	1	1	6	1	1.71	0.29	-0.71	0.71	0.2
Total		9	21			19.25	10.75	-10.25	10.25	6.26

Entonces el estadístico **Log-Rank** es:

$$L^2 = \frac{(\sum_{i=1}^{k}(d_{1i}-e_{1i}))^2}{\sum_{i=1}^{k} V_{1i}} = \frac{(-10.25)^2}{6.26} = 16.78$$

Para $\alpha = 0.05$ tenemos que $J_{0.95} = 3.8414$ por lo que:

$$L^2 = 16.78 > J_{0.95} = 3.8414$$

Por lo tanto, **se rechaza** H_0 y se concluye que **existe diferencia significativa** entre las funciones de supervivencia para el Grupo 1 (Tratamiento) y el Grupo 2 (Placebo).

9.3. Prueba Generalizada Wilcoxon

Se considera la hipótesis a probar:

$$H_0 : S_1(t) = S_2(t) \quad \forall t > 0 \quad vs \quad H_a : S_1(t) \neq S_2(t) \quad \text{p.a } t > 0$$

La prueba *generalizada de Wilcoxon* es una generalización de la prueba *Log-Rank*; Wilcoxon añade una **ponderación** a las fallas *observadas* menos las fallas *esperadas*. El estadístico en el que se basa la prueba de Wilcoxon es:

$$U_w = \sum_{i=1}^{k} n_i(d_{1i}-e_{1i})$$

Bajo la hipótesis nula H_0 tenemos que $\mathbb{E}(U_w) = 0$ y $V_w = Var(U_w) = \sum_{i=1}^{k} n_i^2 V_{1i}$. En consecuencia:

$$W = \frac{U_w - \mathbb{E}(U_w)}{\sqrt{V_w}} = \frac{\sum_{i=1}^{k} n_i(d_{1i}-e_{1i})}{\sqrt{\sum_{i=1}^{k} n_i^2 V_{1i}}} \sim N(0,1)$$

$$\Longrightarrow W^2 = \frac{(\sum_{i=1}^{k} n_i(d_{1i} - e_{1i}))^2}{\sum_{i=1}^{k} n_i^2 V_{1i}} \sim \chi^2_{(1)}$$

Entonces, la **estadística** que consideraremos para la prueba generalizada de **Wilcoxon** será:

$$W^2 = \frac{(\sum_{i=1}^{k} n_i(d_{1i} - e_{1i}))^2}{\sum_{i=1}^{k} n_i^2 V_{1i}}$$

La **Regla de Decisión** es rechazar H_0 al nivel de significancia α si:

$$W^2 > J_{1-\alpha}$$

Donde $J_{1-\alpha}$ es el cuantil $1-\alpha$ de una $\chi^2_{(1)}$.

Ejercicio

Realice la prueba generalizada de Wilcoxon de acuerdo a los datos del ejempo de la sección anterior.

9.4. Comparación de m Poblaciones

Si deseamos comparar las funciones de supervivencia de m poblaciones planteamos la siguiente hipótesis:

$$H_0 : S_1(t) = S_2(t) = ... = S_m(t) \forall t \text{ vs } H_a : S_r(t) \neq S_s(t) \text{ p.a } r \neq s; r, s = 1, ..., m$$

Para realizar la prueba, denotemos d_j el vector de fallas al tiempo t_j, $j = 1, ..., m$ con vector de medias $\mathbb{E}(d_j)$ y matriz de varianzas y covarianzas:

$$Var(d_{ij}) = \frac{n_{ij} \cdot n_{2j} \cdot d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

$$Cov(d_{ji}, ..., d_{jk}) = \frac{-n_{ji} \cdot n_{jk} \cdot d_j(n_j - d_j)}{n_j^2(n_j - 1)} \text{ , } i \neq k$$

Donde i denota el Grupo y j es el tiempo. Ahora bien, sumando sobre los tiempos de falla t_j:

$$\underline{D} = \sum_{j=1}^{k} \{\underline{d}_j - \mathbb{E}(d_j)\}$$

$$\underline{V} = \sum_{j=1}^{k} Var(\underline{d}_j)$$

Mantel-Haenszel propone probar la hipótesis de m supervivencias iguales usando la forma cuadrática $O = D^t V^{-1} D$, donde V^{-1} es la inversa generalizada de V que bajo H_0 se distribuye $\chi^2_{(m-1)}$.

Parte IV

Lleno de riesgos

Capítulo 10

Modelo de Riesgos Proporcionales

Cuando se desea comparar 2 o más grupos de tiempos-evento, si los grupos son "similares" entonces se les pueden aplicar los métodos no paramétricos vistos. Usualmente los individuos en los grupos tienen características adicionales que pueden afectar el resultado, por ejemplo, variables como: edad, género, nivel socioeconómico, consumo del alcohol, ritmo cardiaco, nivel de colesterol, etcétera.

Dichas variables pueden usarse como **covariables** (variables explicativas, factores de riesgo, variables independientes) en un modelo que explique la variable respuesta. Así, después de ajustar las covariables, la comparación de tiempos de supervivencia entre grupos deberá tener menos sesgo y ser más precisa que la simple comparación de tiempos de supervivencia.

Otro problema a resolver será predecir la distribución del tiempo de ocurrencia de cierto evento a partir de un conjunto de covariables.

Los datos estarán dados de la siguiente forma: (t_i, δ_i, x_i) con $i = 1, ...n$, donde:

- i: individuos.
- t_i: tiempo de falla o censura.
- δ_i: indicador de falla o censura.
- x_i: conjunto de covariables.

Nota:

- $X_i = (X_{i1}, X_{i2}, ..., X_{ip})$; donde p= número de covariables.
- X_i puede depender del tiempo, es decir, X_{ik} es una v.a que cambia en el tiempo (peso, colesterol, etc.).

Sea $h_i(t|X_i)$ la función de riesgo al tiempo t del individuo i dadas las covariables X_i (vector de riesgo). El **modelo** propuesto por Cox (1972) es:

$$h_i(t) = \varphi(X_i; \theta) \cdot h_0(t)$$

Donde:

- $\theta = (\theta_1, ..., \theta_p)$ es el vector de p parámetros asociados a las covariables (coeficientes de regresión).
- $\varphi(\cdot, \cdot)$ es la *función liga* de las covariables con el tiempo t.
- $h_0(t)$ es la función de riesgo *base*.

La función $\varphi(\cdot, \cdot)$ debe satisfacer que $\varphi(0, \theta) = 1$, esto para que, en ausencia de covariables, se tenga que $h_i(t) = h_0(t)$.

La forma más común de $\varphi(X_i, \theta)$ es $\varphi(X_i, \theta) = e^{X_i'\theta}$ (Supone que X_i no tiene intercepto), entonces:

$$h_i(t) = e^{X_i'\theta} \cdot h_0(t)$$

Si aplicamos *logaritmo* tenemos que:

$$\ln(h_i(t)) = X_i^{'}\theta + \ln(h_0(t))$$

$$\implies X_i^{'} \cdot \theta = ln\left(\frac{h_i(t)}{h_0(t)}\right) \tag{10.1}$$

Es decir, el cociente de la función de riesgo del individuo i con respecto al riesgo base será igual a una forma lineal de las covariables.

El nombre de riesgos proporcionales se deriva del cociente de las funciones de riesgo de dos individuos:

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{X_i^{'}\theta} \cdot h_0(t)}{e^{X_j^{'}\theta} \cdot h_0(t)} = e^{(X_i - X_j)'\theta}$$

A la expresión anterior se le conoce como **riesgo relativo** y es constante en el tiempo, cuyo valor depende simplemente de la diferencia entre valores de las covariables de los dos individuos[1].

Si $X_{1i} = 1$ y $X_{1j} = 0$, representan tratamiento y placebo respectivamente, y si todas las demás covariables se mantienen constantes, entonces e^{θ_1} es el riesgo de que se presente la falla con el tratamiento relativo a que se presente la falla con el placebo.

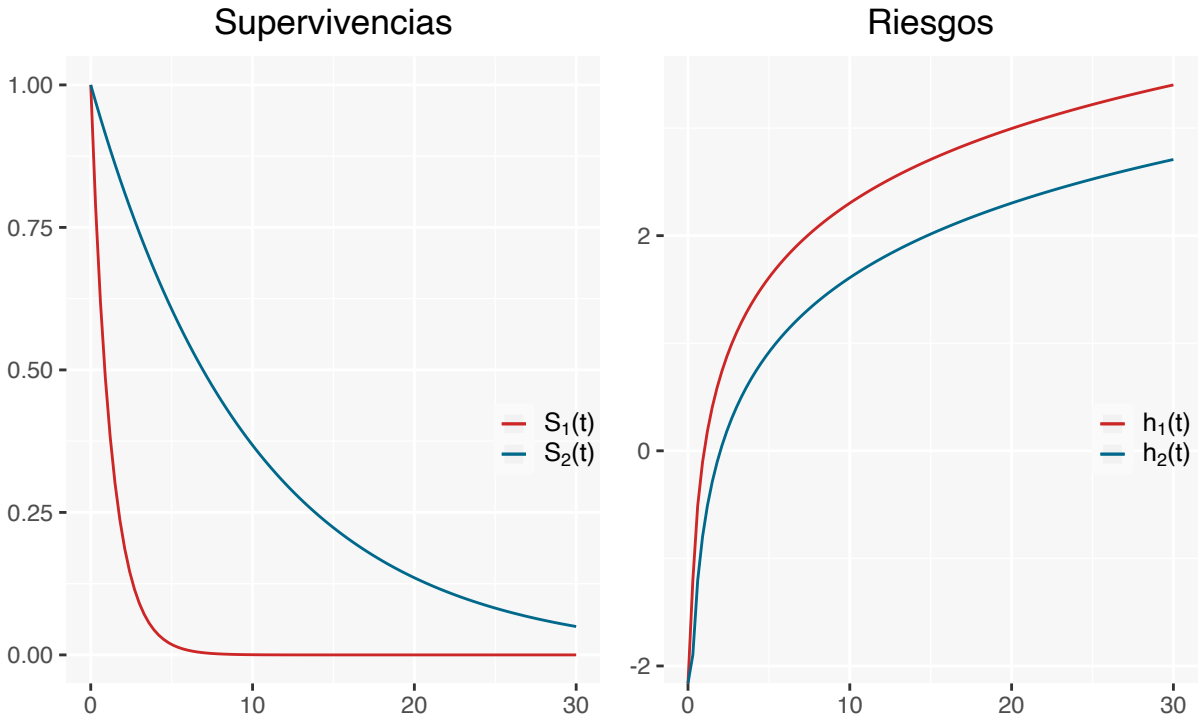
$$\frac{h_i(t)}{h_j(t)} = e^{\theta_1}$$

Bajo el modelo de riesgos proporcionales, las funciones de supervivencia y densidad del individuo i son:

$$S_i(t) = \{S_0(t)\}^{exp(X_i'\theta)}$$

$$f_i(t) = e^{(X_i'\theta)} h_0(t) \{S_0(t)\}^{exp(X_i'\theta)}$$

donde $S_0(t) = exp\{-H_0(t)\}$ es la función de supervivencia base y $H_0(t)$ la función de riesgo acumulado base. Una consecuencia del supuesto de proporcionalidad entre los riesgos de dos individuos i, j con covariables X_i, X_j es que las funciones de riesgo y supervivencia **no se intersectan**.



[1] No hay dependencia con el tiempo.

NOTA: Si $S_0(t)$ es miembro de una familia paramétrica, por lo general, $S_i(t)$ no es miembro de la misma familia. Véase los siguientes ejemplos

- Riesgo base Weibull: $h_0(t) = \lambda\alpha t^{\alpha-1}$

$$\implies h_i(t) = \lambda\alpha t^{\alpha-1} e^{X_i'\theta} = \lambda e^{X_i'\theta}\alpha t^{\alpha-1} \quad \therefore h_i \sim Weibull(\alpha, \lambda e^{X_i'\theta})$$

- Riesgo base log-logistico: $h_0(t) = \frac{\alpha\lambda t^{\alpha-1}}{(1+\lambda t^{\alpha})}$

$$\implies h_i(t) = \frac{e^{Xi'\theta}\alpha t^{\alpha-1}}{(1+\lambda t^{\alpha})}$$

- Riesgo base Gamma: $S_0(t) = 1 - lg(\lambda t, \beta)$

$$S_i(t) = \{S_0(t)\}^{e^{X_i'\theta}} = (1 - lg(\lambda t, \beta))^{e^{X_i'\theta}}$$

10.1. Inferencia sobre θ

La inferencia para los modelos de riesgos proporcionales paramétricos se hacen por máxima verosimiliutd.

Sea (t_i, δ_i, X_i)

- i : individuos.
- t_i: tiempo de fallo o censura.
- δ_i: Indicador de fallo o censura.
- X_i: Covariables.

Sean $h_0(t|\alpha, \lambda)$ y $S_0(t|\alpha, \lambda)$ funciones de riesgo y supervivencia base.

Entonces la función de verosimilitud para $(\theta, \alpha, \lambda)$ será:

$$\begin{aligned}
\mathscr{L}(\theta, \alpha, \lambda) &= \prod_{i=1}^{n} \{f_i(t)\}^{\delta_i} \{S_i(t)\}^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left\{ e^{X_i'\theta} h_0(t) S_0(t)^{e^{X_i'\theta}} \right\}^{\delta_i} \left\{ S_0(t)^{e^{X_i'\theta}} \right\}^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left\{ e^{X_i'\theta} h_0(t) \right\}^{\delta_i} \{S_0(t)^{e^{X_i'\theta}}\}
\end{aligned}$$

- Los estimadores máximo verosímil se obtienen numéricamente.
- La forma explicita de $\mathscr{L}(\theta, \alpha, \lambda)$ dependerá de la elección de h_0.
- Inferencia para los parámetros más allá de la estimación puntual se basa en los resultados asintóticos.

10.2. Estimación Semiparamétrica (Verosimilitud parcial).

El modelo de riesgos proporcionales semiparamétrico surge cuando la función de riesgo base $h_0(t)$ se considera como una parámetro desconocido, y en este caso es necesario hacer inferencia para θ y $h_0(t)$. El parámetro de interés más importante del modelo es θ y entonces $h_0(t)$ es considerado un parámetro de ruido.

Supongamos que los datos consisten en el vector de observaciones $T = (T_1, ..., T_n)$ de la densidad $f(t|\theta, \eta)$ donde θ es el vector de parámetros de interés y η es parámetro de ruido.

Sean $t_{(1)} < t_{(2)} < ... < t_{(D)}$ los tiempos de fallo observados de manera exacta.

Sea $X_{(j)}$ la variable asociada al individuo con tiempo de fallo $t_{(j)}$.

Definimos $R(t_{(j)})$ como el conjunto de todos los individuos en riesgo justo antes de $t_{(j)}$.

Entonces la **verosimilitud parcial** para θ es

$$ {}_p\mathscr{L}(\theta) = \prod_{j=1}^{D} \frac{h_j(t_{(j)})}{\sum\limits_{i \in R(t_{(j)})} h_i(t_{(j)})} = \prod_{j=1}^{D} \frac{\exp\left(X_{(j)}^{'}\theta\right)}{\sum\limits_{i \in R(t_{(j)})} \exp\left(X_i^{'}\theta\right)} $$

OBS :

- ${}_p\mathscr{L}(\theta)$ no depende de $h_0(t)$
- El numerador depende sólo de la información del individuo que falla.
- El denominador usa información de todos los individuos que aún no han experimentado fallo incluyendo censurados.
- La verosimilitud parcial se trata como cualquier función de verosimilitud (aplica logaritmo, derivada igual a cero,...)
- θ es un vector de dimensión $p \implies$ se obtendrán p derivadas parciales. La mayoría de los paquetes usan algoritmos de Newton-Raphson para resolver el sistema de ecuaciones simultaneas.
- Pruebas de hipótesis e intervalos de confianza para θ se pueden obtener con distribución asintótica normal con media θ y matriz de varianzas y covarianzas.

10.3. Estimador de Breslow ($H_0(t)$ y $S_0(t)$)

Si la funciones base son también de interés, se puede utilizar el estimador propuesto por Berslow (1974) que es una generalización del estimador de *Nelson-Aalen.*

$$ \hat{H}_0(t) = \sum_{i:t_i \leq t} \left\{ \frac{\delta_i}{\sum_{j=1}^{n} Y_j(t_i) e^{X_j^{'}\hat{\theta}}} \right\} = \frac{\text{Fallecidos}}{\text{Individuos en riesgo}} $$

donde $Y_i(t) = \mathbb{1}_{\{t_i \geq t\}}$ indicadora si $\hat{\theta} = 0$ y $\hat{H}_0(t)$ es el estimador de Nelson-Aalen visto previamente.

$$ \therefore \hat{S_0}(t) = exp\{-\hat{H}_0(t)\} = e^{\sum_{i:t_i \leq t} \left\{ \frac{\delta_i}{\sum_{j=1}^{n} Y_j(t_i) e^{X_j^{'}\hat{\theta}}} \right\}} $$

Notas

1. En el caso de que se presenten **empates** (múltiples individuos con el mismo tiempo de fallo) la ${}_p\mathscr{L}(\theta)$ debe ajustarse para que se considere la naturaleza discreta de las observaciones.
2. El modelo de riesgos proporcionales permite la incorporación de covariables dependientes en el tiempo.

Ejemplo 1

Un estudio sobre la supervivencia clasifica según la raza en: blanco, negro e hispano. Entonces las variable X_1 toma los siguientes valores:

$$ \begin{gathered} X_1 = 1 \text{ si es blanco} \\ X_1 = 2 \text{ si es negro} \\ X_1 = 3 \text{ si es hispano} \end{gathered} $$

Sin embrago, también podría plantearse como dos variables X_1 y X_2

$$ \begin{gathered} X_1 = 1 \text{ si es blanco, 0 en otro caso.} \\ X_2 = 1 \text{ si es negro, 0 en otro caso} \end{gathered} $$

Entonces, el modelo de riesgos proporcionales será: $h(t|X) = h_0 e^{\beta_1 X_1 + \beta_2 X_2}$

- Si $h(t|X_1 = 1, X_2 = 0) = h_0(t)e^{\beta_1} \longrightarrow$ Riesgo de blanco
- Si $h(t|X_1 = 0, X_2 = 1) = h_0(t)e^{\beta_2} \longrightarrow$ Riesgo de negro
- Si $h(t|X_1 = 0, X_2 = 0) = h_0(t) \longrightarrow$ Riesgo de hispano (riesgo base)

Riesgos relativos:

- Entre negro e hispano: $\frac{h(t|X_1=0,X_2=1)}{h(t|X_1=0,X_2=0)} = \frac{h_0(t)e^{\beta_2}}{h_0(t)} = e^{\beta_2}$

Es decir, e^{β_2} son las veces que el riesgo que tienen los negros en comparación con los hispanos.

- Se busca que el riesgo relativo sea $\neq 1$ para poder decir que la categoría segmenta datos.

Ejemplo 2

Un estudio con 863 pacientes con trasplante de hígado. Dos de las variable que se recabaron de los pacientes fueron, género y raza. Entonces los pacientes en el estudio se dividieron en las siguientes categorías

432 hombres blancos
92 hombres negros
286 mujeres blancas
59 mujeres negras

Para ajustar un modelo de riesgos proporcionales a estos datos, una opción es definir 3 covariables:

$Z_1 = 1$ hombre negro, 0.e.o.c
$Z_2 = 1$ hombre blanco, 0.e.o.c
$Z_3 = 1$ mujer negra, 0.e.o.c

Y el modelo para la función de riesgo será:

$$h(t|Z) = h_0(t)exp\{\theta_1 Z_1 + \theta_2 Z_2 + \theta_3 Z_3\}$$

Los estimadores máximo verosímil son $\hat{\theta_1} = 0.160$, $\hat{\theta_2} = 0.298$, $\hat{\theta_3} = 0.657$.

Riesgo relativo de hombre negro con mujer blanca :

$$h(t|Z) = \frac{h_0(t)exp\{0.160\}}{h_0(t)exp\{0\}} = e^{0.160} = 1.17$$

$\therefore$ Los hombres negros son más propensos a morir por trasplante de hígado que las mujeres blancas; es decir, por cada mujer blanca, un hombre negro (1.17) muere en el trasplante.

En el caso de un hombre negro y un hombre blanco

$$h(t|Z) = \frac{h_0(t)exp\{0.160\}}{h_0(t)exp\{0.298\}} = e^{-0.088} = 0.9157609$$

Otra opción de plantear el modelo es con 2 covariables y una interacción, y quedaría de la siguiente forma:

$Z_1 = 1$ Si es mujer, 0 e.o.c
$Z_2 = 1$ Si es negro(a), 0 e.o.c
$Z_3 = Z_1 \cdot Z_2$ Esta variable tomará el valor 1 si es mujer negra, y 0 e.o.c

Y el modelo para la función de riesgo será: $h(t|Z) = h_0 exp\{\theta_1 Z_1 + \theta_2 Z_2 + \theta_3 Z_1 Z_2\}$.

Los estimadores máximo verosímil son $\hat{\theta_1} = -.2484, \hat{\theta_2} = -.0888, \hat{\theta_3} = .7435$. Hay que notar que la interpretación de las $\theta' s$ será diferente y en este caso el parámetro de interés será el de la interacción (θ_3)

Riesgos relativos :

$$\frac{\text{Hombre negro}}{\text{Mujer blanca}} = \frac{h_0(t)exp\{-0.0888\}}{h_0 exp\{-0.2484\}} = e^{-(0.0888+0.2484)} = e^{-.1596} = 1.17$$

Véase que es el mismo resultado que se obtuvo en el modelo anterior, por lo que se sigue conservando el mismo riesgo independiente del modelo.

Ejercicios

1) Calcular el riesgo mujer negra relativo a mujer blanca.
2) Calcular el riesgo mujer negra relativo a hombre negro
3) Calcular el riesgo hombre blanco relativo a mujer blanca.

RECORDATORIO: Codificación de variables categóricas.

Una variable categórica con k clases se transforma en $k-1$ variables binarias.

Ejemplo 1

X: Color de cabello {negro, cafe, rojo, blanco}. Tenemos una variable con 4 categorías. Por lo que la transformaremos en 3 variables binarias

$X_1 = 1$ si es negro, 0 e.o.c
$X_2 = 1$ si es café. 0 e.o.c
$X_3 = 1$ si es rojo, 0.e.o.c

Con estas tres variables se cubren todas las opciones de color de cabello.

10.4. Significancia de los parámetros (Prueba de Wald)

¿Son significativos $\hat{\theta_1}, \hat{\theta_2}, \hat{\theta_3}$ en ejemplo numero 2?

Similar al caso de regresión lineal, la significancia de los parámetros radica en la importancia del riesgo proporcionado por las covariables asociadas a parámetros. Y en caso de no ser significativos los parámetros, las covariables se podrían eliminar del modelo.

Prueba de hipótesis: Nos interesa hacer pruebas sobre θ, de manera general:

$$H_0 : \theta_1 = \theta_{H_0} \text{ vs } H_1 : \theta_1 \neq \theta_{H_0}$$

donde $\theta = (\theta_1^t, \theta_2^t) = (\theta_1, \theta_2, ..., \theta_p)$

- θ_1 : Es el vector de $q * 1$ (q parámetros de interés)
- θ_2 : Los parámetros restantes $(p - q)$

Partimos la matriz de información : $I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$

Donde I_{11}, I_{22} segundas derivadas de la función de verosimilitud.

La **Prueba de Wald** se define como:

$$X_w^2 = (\hat{\theta_1} - \hat{\theta_{H_0}})' [I^*(\theta)]^{-1} (\hat{\theta_1} - \hat{\theta_{H_0}})$$

Donde $I^*(\theta)$ es la matriz $q * q$ superior de I.

- Para muestras grandes, la estadística de prueba: $X_w^2 \sim \chi^2_{(q)}$
- Un criterio de información es el proporcionado por el criterio de Aikaike (ACI):

$$AIC = -2log\mathscr{L} + kp$$

- $\mathscr{L}$: Función de verosimilitud.
- k: Cte.(usualmente igual a 2).
- p: # de parámetros en el modelo.

Para la construcción del modelo usando el modelo de riesgos proporcionales de Cox, se puede usar el estadístico de Wald para seleccionar covariables significativas[2] y considerando el valor del AIC para evaluar la mejora en el modelo[3].

10.5. Estimación de $S(t)$ después de obtener las estimaciones de los parámetros del modelo de Cox

Hasta el momento se han dado estimadores para θ y pruebas sobre dicho parámetro; es decir, teniendo el modelo de riesgos proporcionales $h(t|X) = h_0(t)e^{X'\theta}$ y ajustando dicho modelo a nuestros datos obtenemos $\hat{\theta}$. Ahora lo que se desea es la función de supervivencia con el conjunto de covariables X_i.Para ello sea $t_1 < t_2, \ldots < t_D$ todos los tiempos de falla y d_i el número de fallas del tiempo t_i. Entonces

$$\omega(t_i, \hat{\theta}) = \sum_{j \in R(t_i)} \exp\left\{\sum_{h=1}^{D} \hat{\theta}_h X_{jh}\right\}$$

Y el estimador de la función de riesgo acumulado base será la siguiente, la cual es una función escalonada a cada tiempo de falla.

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{w(t_i, \hat{\theta})}$$

Por lo que un estimador de la función de supervivencia base será el siguiente, el cual corresponde a los individuos cuyas covariables X son ceros.

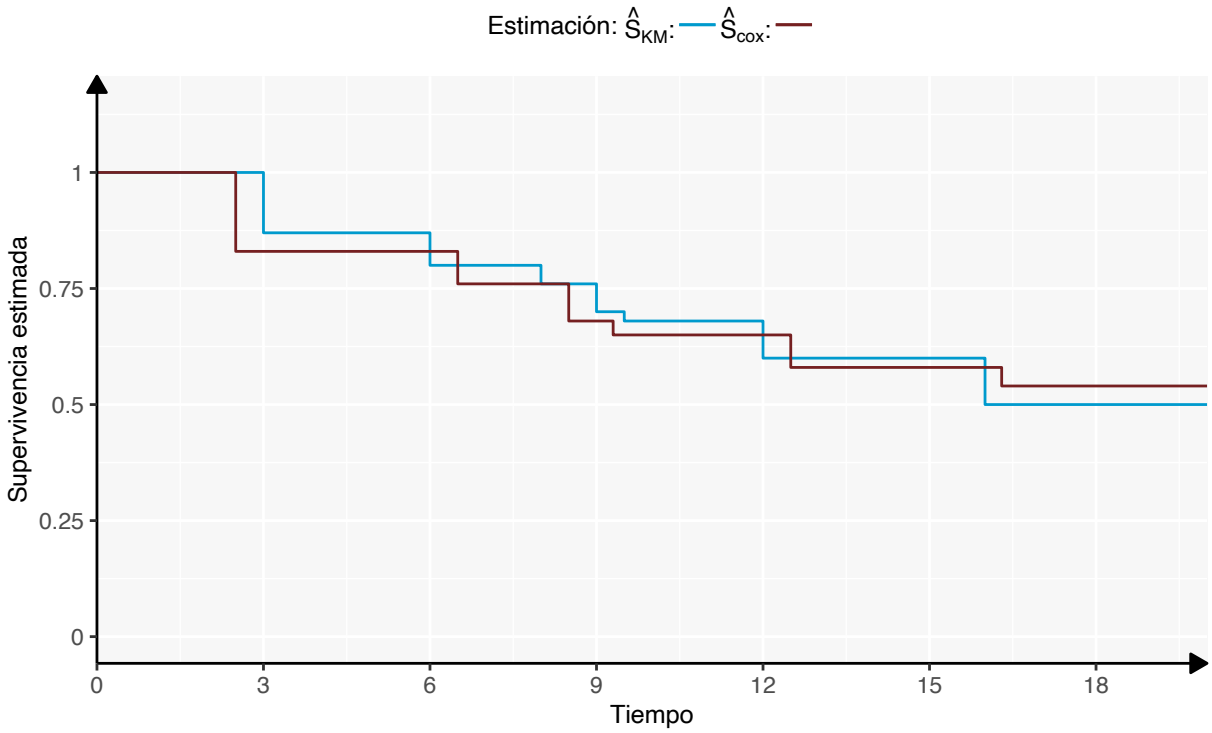
$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}$$

Para estimar la función de supervivencia para un individuo con covariables X^* se utilizará el siguiente estimador

$$\hat{S}(t|X^*) = \hat{S}_0(t) \exp\left\{\hat{\theta}' X^*\right\}$$

La siguiente gráfica es un ejemplo comparativo de dos estimaciones de funciones de supervivencia, una con el método de Kaplan-Meier y Cox.

[2]Recordar que por la ecuación (10.1), el modelo esta asumiendo una forma lineal de las covariables.
[3]Para un mayor número de variables el $AIC \uparrow$ pero para variables más significativas $AIC \downarrow$.



10.6. Verificación de ajuste de Modelo

Mediante gráficos de las funciones de supervivencia buscamos las funciones de supervivencia para los distintos valores de X's donde **no** se crucen[4].

En caso de que el modelo de Cox no cumpla con el supuesto de proporcionalidad se puede corregir el modelo mediante:

1) Agregar más covariables
2) Considerar interacciones entre las covariables
3) Introducir términos no lineales
4) Permitir que las covariables dependan del tiempo

10.7. Extensión del modelo de Cox a covariables dependientes del tiempo

Sea $X(t) = [X_1(t), ..., X_p(t)]$ es el conjunto de covariables o factores de riesgo al tiempo t que podrían afectar la distribución de la variable de supervivencia.

Entonces $X_k(t)$´s son covariables dependientes del tiempo cuyos valores cambian o permanecen constantes (como el caso anterior). Supondremos que los valores de estas covariables son predecibles (el valor es conocido).

Ejemplos de este tipo de variables son presión arterial, colesterol, tamaño del tumor, etc.

Sustituyendo en el modelo de Cox tenemos:

[4]Por el hecho de que se busca **proporcionalidad** y, por lo tanto, que la segregación separe completamente a la población.

$$\begin{aligned} h(t|X(t)) =& h_0(t) \cdot exp\{\beta^{`} \cdot Z(t)\} \\ =& h_0(t) \cdot exp\left\{\sum_{k=1}^{p} \beta_k \cdot Z(t)\right\} \end{aligned}$$

Entonces para probar el supuesto de riesgos proporcionales se crea una variable artificial:

$$X_2(t) = X_1 \cdot g(t)$$

donde X_1 es una variable fija en el tiempo $g(t)$ es una función que depende del tiempo usualmente

$$g(t) = ln(t)$$

y se ajusta un modelo de Cox para las covariables X_1 y $X_2(t)$, tenemos

$$\begin{aligned} h(t|X(t)) &= h_0 \cdot exp\{\beta_1 X_1 + \beta_2 X_2(t)\} \\ &= h_0(t) \cdot exp\{\beta_1 X_1 + \beta_2 X_1 \cdot g(t)\} \end{aligned}$$

y se realiza la prueba de hipótesis para $\beta_2 = 0$. Si se desea evaluar riesgos proporcionales para 2 individuos con diferentes valores de X_1

$$\frac{h(t|X_1)}{h(t|X_1^*)} = exp\{\beta_1(X_1 - X_1^*) + \beta_2 \cdot g(t)(X_1 - X_1^*)\}$$

la cuál dependerá del tiempo si $\beta_2 \neq 0$.

Bibliografía

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Hall, W. J. and Wellner, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika*, 67(1):133–143.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, second edition.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: a comparative study. *Technometrics*, 26(3):265–275.