Una introducción al análisis de sobrevivencia usando el software R (Parte 1)

Diego I. Gallardo Mateluna

Departamento de Matemática, Facultad de Ingeniería
Universidad de Atacama

Copiapó, Chile

Minicurso para la Sociedad Chilena de Estadística (SOCHE)

diego.gallardo@uda.cl

02 de Diciembre de 2020

Resumen

- Motivación.
- Elementos de interés en el análisis de sobrevivencia.
- Censura.
- Estimadores no paramétricos de la función de sobrevivencia.
- Modelos paramétricos.
- Modelos de regresión en análisis de sobrevivencia.
- Inferencia.
- Residuos.
- Otros tópicos.

Motivación

Motivación

Asumamos que estamos interesados en el tiempo hasta que ocurra un cierto evento de interés. Usualmente este evento de interés es conocido como "**falla**".

- **Medicina**: tiempo hasta que un tratamiento surta efecto.

Motivación

Asumamos que estamos interesados en el tiempo hasta que ocurra un cierto evento de interés. Usualmente este evento de interés es conocido como "**falla**".

- **Medicina**: tiempo hasta que un tratamiento surta efecto.
- **Ingeniería**: tiempo hasta que un componente electrónico se deteriore.

Motivación

Asumamos que estamos interesados en el tiempo hasta que ocurra un cierto evento de interés. Usualmente este evento de interés es conocido como "**falla**".

- **Medicina**: tiempo hasta que un tratamiento surta efecto.
- **Ingeniería**: tiempo hasta que un componente electrónico se deteriore.
- **Educación**: tiempo hasta que un alumno adquiera cierto conocimiento.

Motivación

Asumamos que estamos interesados en el tiempo hasta que ocurra un cierto evento de interés. Usualmente este evento de interés es conocido como "**falla**".

- **Medicina**: tiempo hasta que un tratamiento surta efecto.
- **Ingeniería**: tiempo hasta que un componente electrónico se deteriore.
- **Educación**: tiempo hasta que un alumno adquiera cierto conocimiento.
- **Ecología**: tiempo hasta que un cierto animal alcance un determinado tamaño.

Motivación

Asumamos que estamos interesados en el tiempo hasta que ocurra un cierto evento de interés. Usualmente este evento de interés es conocido como "**falla**".

- **Medicina**: tiempo hasta que un tratamiento surta efecto.
- **Ingeniería**: tiempo hasta que un componente electrónico se deteriore.
- **Educación**: tiempo hasta que un alumno adquiera cierto conocimiento.
- **Ecología**: tiempo hasta que un cierto animal alcance un determinado tamaño.
- Etc.

Motivación

Es necesario definir los siguientes elementos:

- El **tiempo de inicio** del estudio debe ser definido de forma precisa, de forma que las observaciones puedan ser comparables en el origen del estudio.

Motivación

Es necesario definir los siguientes elementos:

- El **tiempo de inicio** del estudio debe ser definido de forma precisa, de forma que las observaciones puedan ser comparables en el origen del estudio.
- La **escala de medida**: usualmente el tiempo (cronológico).

Motivación

Es necesario definir los siguientes elementos:

- El **tiempo de inicio** del estudio debe ser definido de forma precisa, de forma que las observaciones puedan ser comparables en el origen del estudio.
- La **escala de medida**: usualmente el tiempo (cronológico).
- El **evento de interés**: nos enfocaremos en el caso en que sólo puede ocurrir debido a una única causa.

Motivación

Es necesario definir los siguientes elementos:

- El **tiempo de inicio** del estudio debe ser definido de forma precisa, de forma que las observaciones puedan ser comparables en el origen del estudio.
- La **escala de medida**: usualmente el tiempo (cronológico).
- El **evento de interés**: nos enfocaremos en el caso en que sólo puede ocurrir debido a una única causa.

Observación: Los estudios que involucran una respuesta temporal usualmente son de una "**larga**" duración. Sin embargo, a pesar de ser largos en duración, muchos estudios clínicos terminan antes de que todos los individuos dentro del estudio "**fallen**". De esta forma, tenemos observaciones "**incompletas**" dentro de nuestro estudio.

Motivación

Nos centraremos en la **censura a la derecha**, asociada a los estudios **prospectivos**. Es decir, aquellos estudios que se diseñan y se comienzan a realizar en el presente, pero cuyos datos se analizan transcurrido un determinado tiempo, en el futuro.

Funciones de interés en análisis de sobrevivencia

Funciones de interés

La variable aleatoria no negativa T, usualmente continua, que representa el tiempo de falla es generalmente especificada en análisis de sobrevivencia por su función de sobrevivencia o por la función de riesgo.

Funciones de interés

La variable aleatoria no negativa T, usualmente continua, que representa el tiempo de falla es generalmente especificada en análisis de sobrevivencia por su función de sobrevivencia o por la función de riesgo.

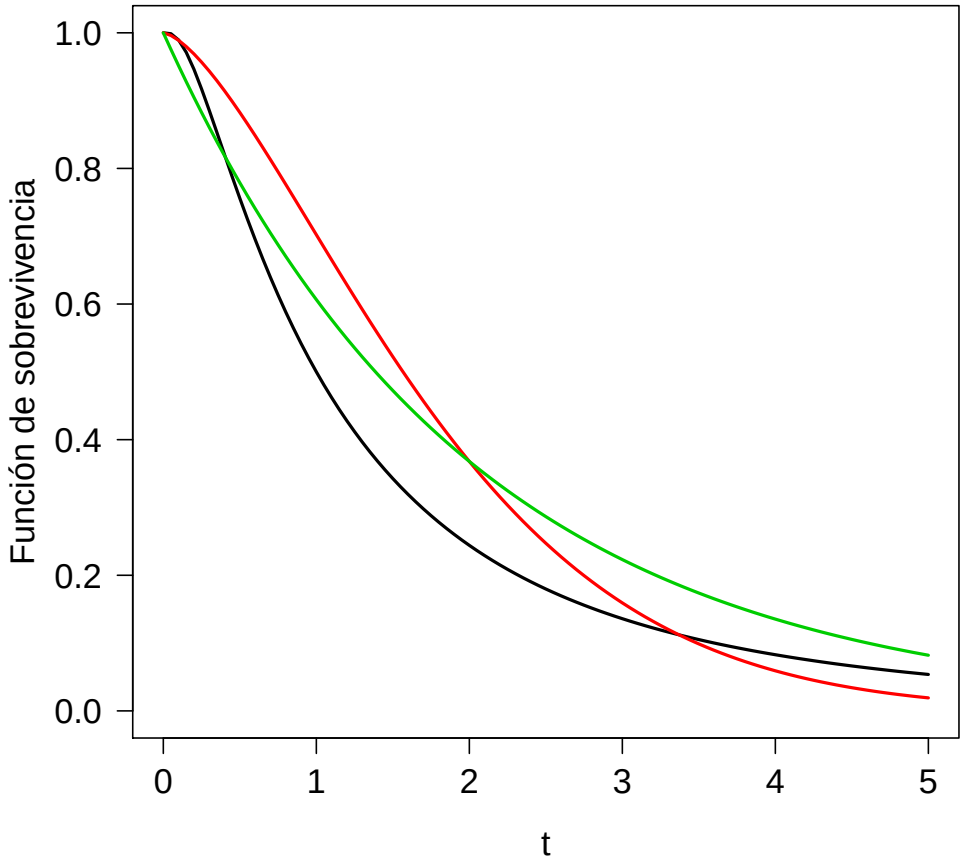
Función de sobrevivencia: Es la probabilidad de una observación no fallar hasta un cierto tiempo t, es decir,

$$S(t) = P(T > t) = 1 - F(t),$$

en que $F(\cdot)$ es la función de distribución acumulada.

Funciones de interés

Ejemplos de funciones de sobrevivencia.



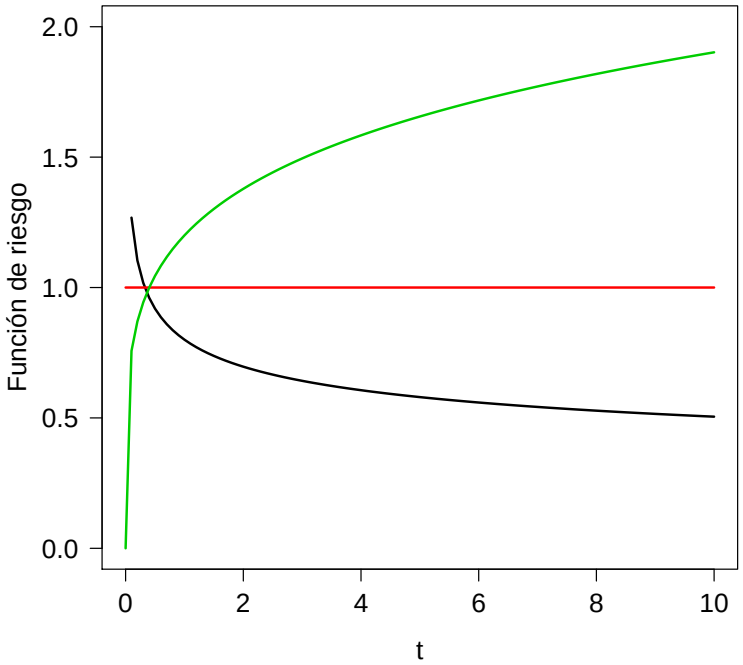
Funciones de interés

Función de riesgo o **tasa de falla**: Es la tasa de falla instantánea en el tiempo t condicional a que la observación sobrevivió hasta el tiempo t.

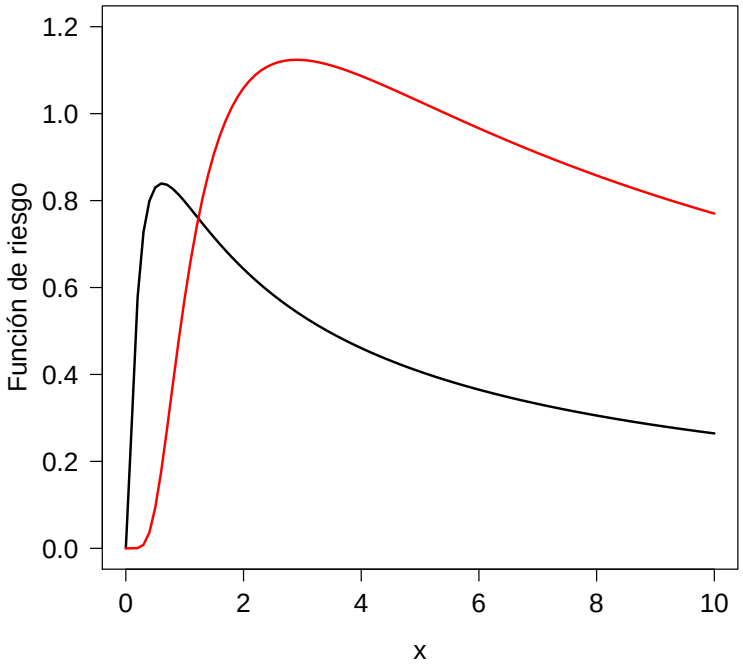
$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h \mid T \geq t)}{h} = \frac{f(t)}{S(t)}$$

Funciones de interés

Ejemplo de funciones de riesgo



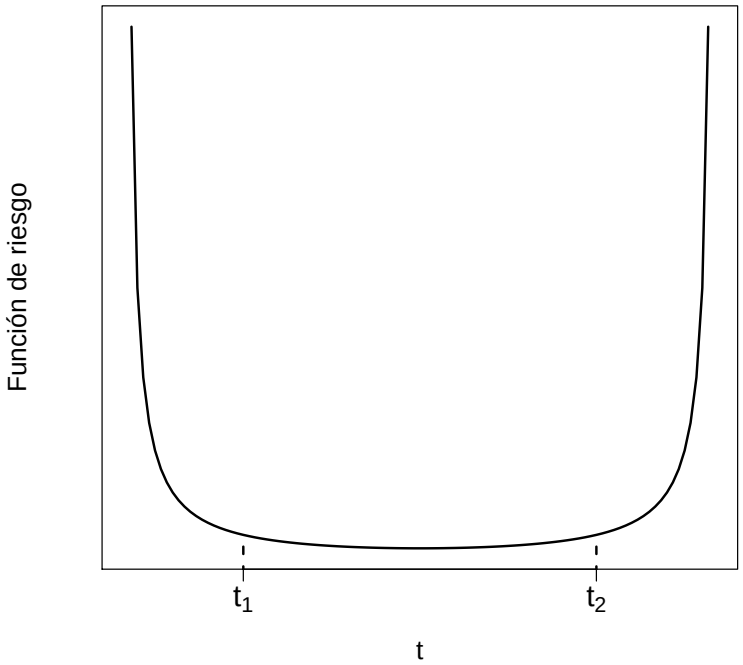
Monótona



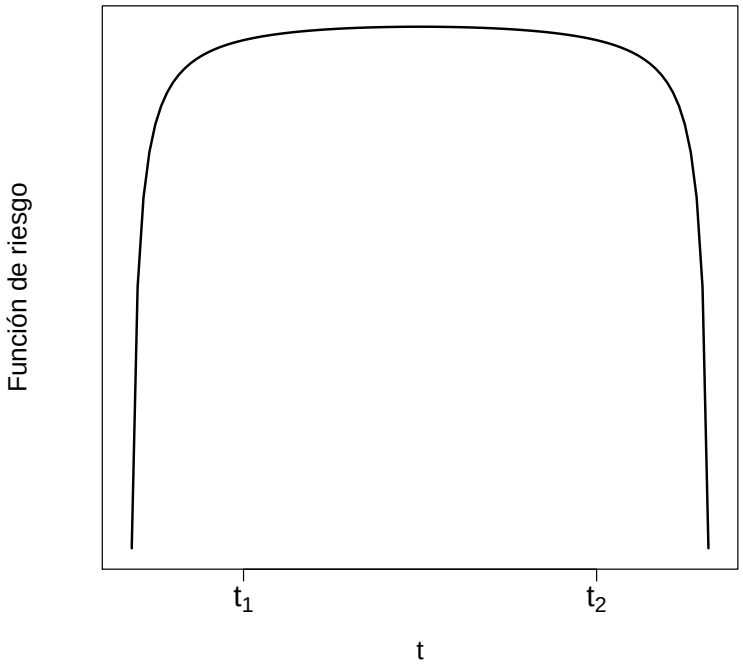
No monótona

Funciones de interés

Ejemplo de funciones de riesgo



Con forma de bañera



Bañera invertida

Funciones de interés

Función de riesgo acumulada.

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Funciones de interés

Función de riesgo acumulada.

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Observación: Hay una relación biunívoca entre $f(t)$, $S(t)$, $\lambda(t)$ y $\Lambda(t)$.

Funciones de interés

Función dada	Función requerida $f(t)$	Función requerida $S(t)$
$f(t)$	-	$\int_t^\infty f(u)du$
$S(t)$	$-\frac{dS(t)}{dt}$	-
$\lambda(t)$	$\lambda(t)\exp\left\{-\int_0^t \lambda(u)du\right\}$	$\exp\left\{-\int_0^t \lambda(u)du\right\}$
$\Lambda(t)$	$\frac{d\Lambda(t)}{dt}\exp\{-\Lambda(t)\}$	$\exp\{-\Lambda(t)\}$

Funciones de interés

	Función requerida	
Función dada	$\lambda(t)$	$\Lambda(t)$
$f(t)$	$\frac{f(t)}{\int_t^\infty f(u)du}$	$-\log\left(\int_t^\infty f(u)du\right)$
$S(t)$	$-\frac{dS(t)/dt}{S(t)}$	$-\log S(t)$
$\lambda(t)$	-	$\int_0^t \lambda(u)du$
$\Lambda(t)$	$\frac{d\Lambda(t)}{dt}$	-

Funciones de interés

Por ejemplo, para el modelo **Weibull** con parámetros $\lambda, \nu > 0$ tenemos que

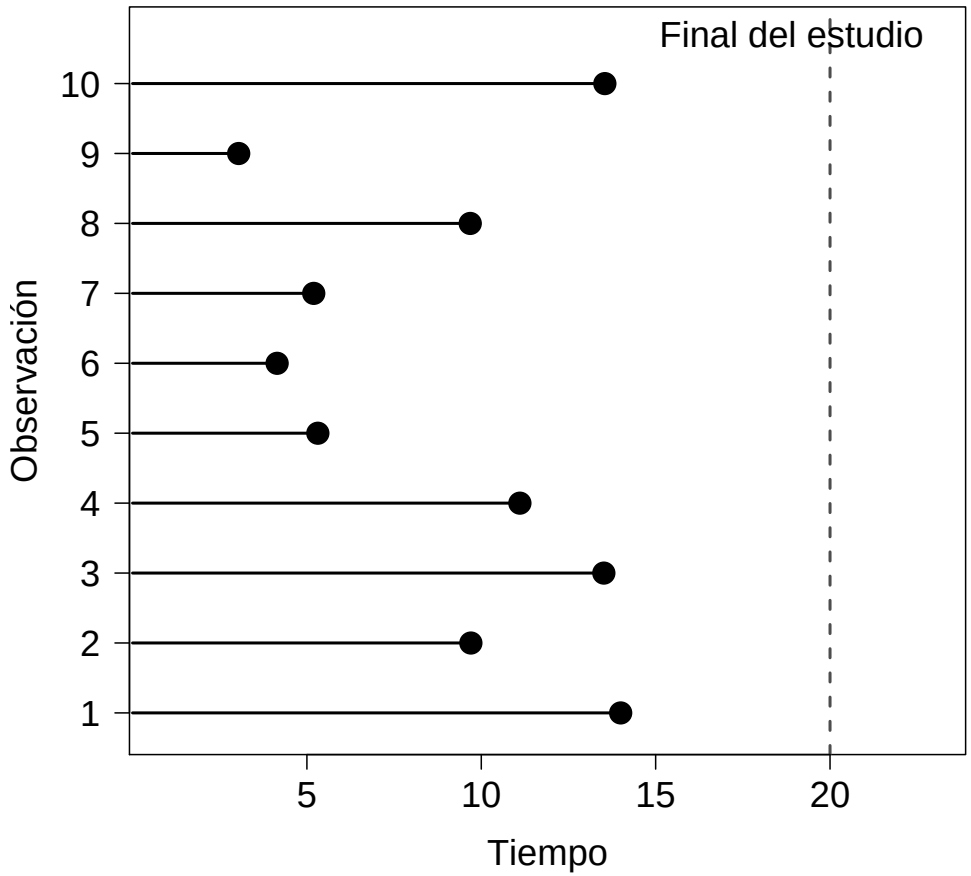
$$\begin{aligned} f(t) &= \lambda\nu t^{\nu-1} \exp\{-\lambda t^{\nu}\} \\ S(t) &= \exp\{-\lambda t^{\nu}\} \\ \lambda(t) &= \lambda\nu t^{\nu-1} \\ \Lambda(t) &= \lambda t^{\nu} \end{aligned}$$

con $t > 0$.

Censura

Tipos de censura

Caso a: **datos completos**.



Tipos de censura

Este es el caso usual que se ve en los cursos de **Inferencia Estadística**.

Podemos asumir para el tiempo de falla alguna distribución para datos positivos (**exponencial**, **Weibull**, **gamma**, **Birnbaum-Saunders**, etc.)

Tipos de censura

Este es el caso usual que se ve en los cursos de **Inferencia Estadística**.

Podemos asumir para el tiempo de falla alguna distribución para datos positivos (**exponencial**, **Weibull**, **gamma**, **Birnbaum-Saunders**, etc.)

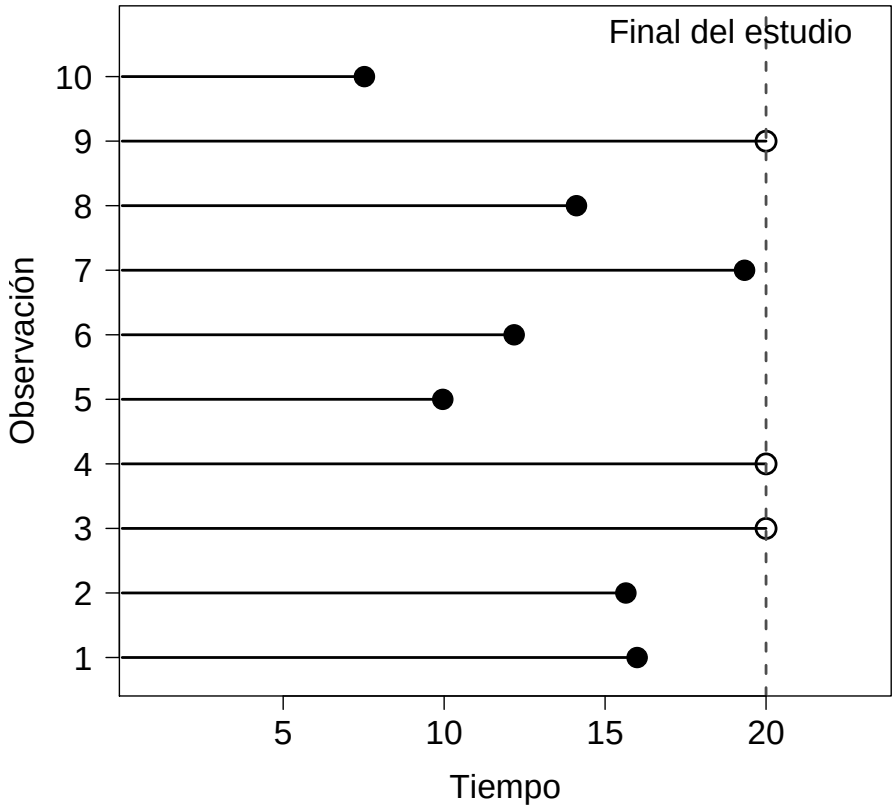
Bajo la suposición que las observaciones son independientes, la función de verosimilitud del modelo es dada por

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta}),$$

en que $\boldsymbol{\theta}$ es el vector de parámetros del respectivo modelo.

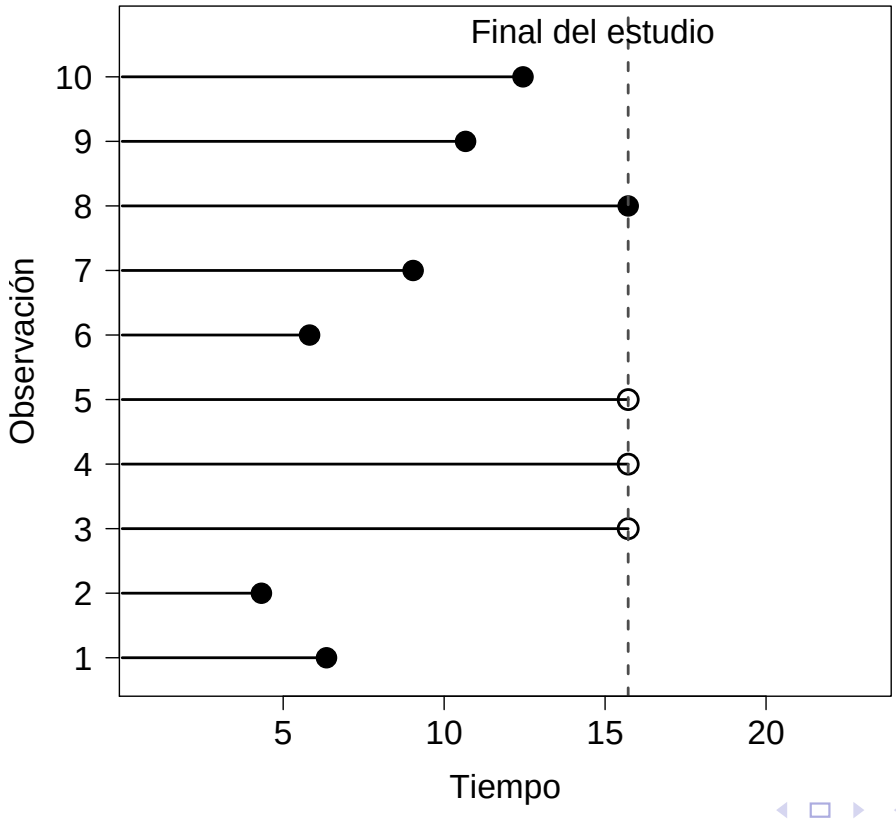
Tipos de censura

Caso b: datos con **censura tipo I**.
Cuando el estudio será terminado después de un período establecido de tiempo (usualmente prefijado antes de iniciar el estudio).



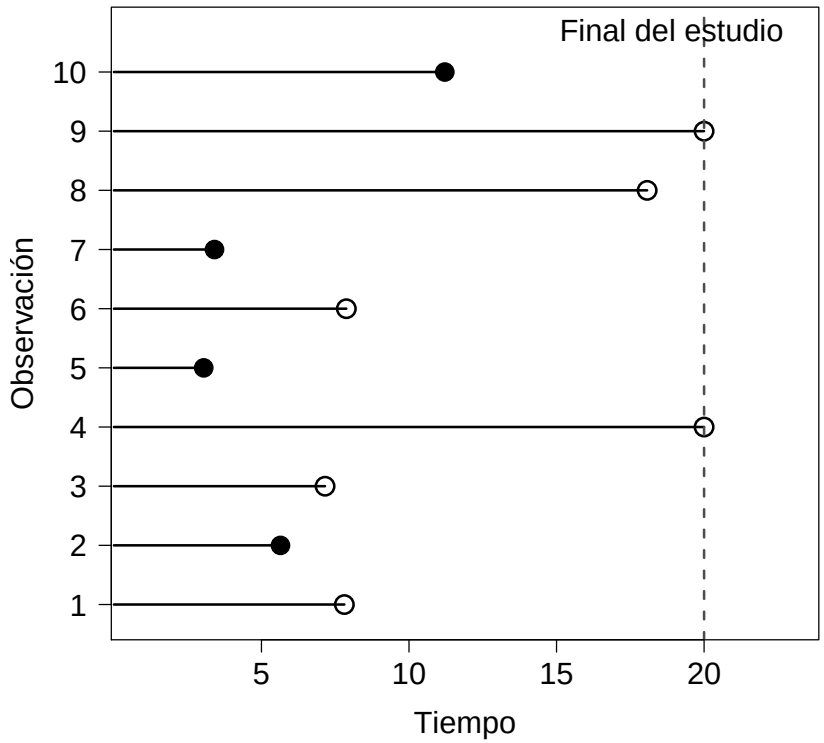
Tipos de censura

Caso c: datos con **censura tipo II**.
Cuando el estudio será terminado después de que un número establecido de observaciones "fallen" (usualmente prefijado antes de iniciar el estudio).



Tipos de censura

Caso d: datos con **censura aleatoria**.
En el contexto médico es la más común en la práctica. Ocurre cuando los pacientes son retirados del estudio sin haberse observado la "falla", o bien, porque el paciente muere por una causa diferente de la enunciada en el evento de interés.



Tipos de censura

Otros tipos de censura (en relación al tipo de estudio).

- **Censura a la izquierda** (estudios retrospectivos).
- **Censura intervalar**.

Representación de los datos de sobrevivencia

Los datos de sobrevivencia para el individuo i $(i = 1, \ldots, n)$ son representados, en forma general, como

$$(t_i, \delta_i),$$

en que t_i es el **tiempo de falla o censura** y δ_i es la **indicadora de falla**, es decir,

$$\delta_i = \begin{cases} 1 & \text{, si } t_i \text{ es un tiempo de falla} \\ 0 & \text{, si } t_i \text{ es un tiempo de censura} \end{cases}$$

Representación de los datos de sobrevivencia

Desde un punto de vista general, estamos asumiendo que para cada individuo existen dos tiempos que están "**compitiendo**":

- El tiempo de **falla**, digamos Y_i, $i = 1, \ldots, n$.
- El tiempo de **censura**, digamos C_i, $i = 1, \ldots, n$.

Representación de los datos de sobrevivencia

Desde un punto de vista general, estamos asumiendo que para cada individuo existen dos tiempos que están "**compitiendo**":

- El tiempo de **falla**, digamos Y_i, $i = 1, \ldots, n$.
- El tiempo de **censura**, digamos C_i, $i = 1, \ldots, n$.

Nuestras variables aleatorias son $T_i = \min(Y_i, C_i)$ y $\Delta_i = I(Y_i \leq C_i)$, cuyas realizaciones son $(t_1, \delta_1), \ldots, (t_n, \delta_n)$.

Representación de los datos de sobrevivencia

Ejemplo:

Los tiempos de **remisión** (en meses) de pacientes con cierto tipo de cáncer son dados por

$$(15.6, 1), (13.2, 1), (20.3, 0), \ldots, (6.4, 0)$$

Representación de los datos de sobrevivencia

Ejemplo:

Los tiempos de **remisión** (en meses) de pacientes con cierto tipo de cáncer son dados por

$$(15.6, 1), (13.2, 1), (20.3, 0), \ldots, (6.4, 0)$$

Otra forma común de presentar la información es usar el signo $+$ para denotar los tiempos censurados. Para los tiempos mencionados anteriormente, tendríamos

$$15.6, 13.2, 20.3^{+}, \ldots, 6.4^{+}$$

Estimadores no paramétricos de la función de sobrevivencia

Estimadores NP de la función de sobrevivencia

Ante la ausencia de censuras, el estimador natural de la función de sobrevivencia es basado en el estimador **empírico** de la **función de distribución acumulada**.

$$\widehat{F}_{emp}(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i \leq t)$$

Estimadores NP de la función de sobrevivencia

Ante la ausencia de censuras, el estimador natural de la función de sobrevivencia es basado en el estimador **empírico** de la **función de distribución acumulada**.

$$\widehat{F}_{emp}(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i \leq t)$$

y

$$\widehat{S}_{emp}(t) = 1 - \widehat{F}_{emp}(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i > t).$$

Estimadores NP de la función de sobrevivencia

Estimador de **Kaplan-Meier** (K-M, Kaplan and Meier, 1958) es dado por

$$\widehat{S}_{KM}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right),$$

en que

- $t_1 < t_2 < \ldots < t_k$, son los k **tiempos distintos y ordenados de falla**;
- d_j es el **número de fallas** en t_j, $j = 1, 2, \ldots, k$;
- n_j es el **número de individuos en riesgo** en t_j. Es decir, los individuos que no han fallado y que no han sido censurados hasta el instante inmediatamente anterior a t_j.

Estimadores NP de la función de sobrevivencia

Observaciones sobre el estimador de K-M

- Es una función tipo "escalera".

Estimadores NP de la función de sobrevivencia

Observaciones sobre el estimador de K-M

- Es una función tipo "escalera".
- En ausencia de censuras, corresponde al estimador empírico de la función de sobrevivencia.

Estimadores NP de la función de sobrevivencia

Observaciones sobre el estimador de K-M

- Es una función tipo "escalera".
- En ausencia de censuras, corresponde al estimador empírico de la función de sobrevivencia.
- Fácil de calcular.

Estimadores NP de la función de sobrevivencia

Observaciones sobre el estimador de K-M

- Es una función tipo "escalera".
- En ausencia de censuras, corresponde al estimador empírico de la función de sobrevivencia.
- Fácil de calcular.
- Es **insesgado** para $S(t)$ cuando $n \to +\infty$.

Estimadores NP de la función de sobrevivencia

Observaciones sobre el estimador de K-M

- Es una función tipo "escalera".
- En ausencia de censuras, corresponde al estimador empírico de la función de sobrevivencia.
- Fácil de calcular.
- Es **insesgado** para $S(t)$ cuando $n \to +\infty$.
- Si la observación más grande de la muestra es censurada, entonces

$$\widehat{S}_{KM}(t) > 0, \quad \forall t > t_k.$$

Estimadores NP de la función de sobrevivencia

Observaciones sobre el estimador de K-M

- Es una función tipo "escalera".
- En ausencia de censuras, corresponde al estimador empírico de la función de sobrevivencia.
- Fácil de calcular.
- Es **insesgado** para $S(t)$ cuando $n \to +\infty$.
- Si la observación más grande de la muestra es censurada, entonces

$$\widehat{S}_{KM}(t) > 0, \quad \forall t > t_k.$$

- No permite la incorporación de **covariables**.

Estimadores NP de la función de sobrevivencia

Estimadores de la varianza del estimador de K-M:

- **Fórmula de Greenwood**.

$$\widehat{Var}\left(\widehat{S}_{KM}(t)\right) = \left[\widehat{S}_{KM}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(d_j - n_j)}.$$

Estimadores NP de la función de sobrevivencia

Estimadores de la varianza del estimador de K-M:

- **Fórmula de Greenwood**.

$$\widehat{Var}\left(\widehat{S}_{KM}(t)\right) = \left[\widehat{S}_{KM}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(d_j - n_j)}.$$

Para un t fijo, un **intervalo de confianza** aproximado de $100(1-\alpha)\%$ de confianza es dado por

$$\widehat{S}_{KM}(t) \mp z_{\alpha/2}\sqrt{\widehat{Var}\left(\widehat{S}_{KM}(t)\right)},$$

en que $z_{\alpha/2}$ es el percentil $\alpha/2$ de la distribución normal estándar.

Estimadores NP de la función de sobrevivencia

- **Transformación log-log** (Kalbfleish and Prentice, 1980). Defina $\widehat{U}(t) = \log(-\log(\widehat{S}_{KM}(t)))$

$$\widehat{Var}\left(\widehat{U}(t)\right) = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j(d_j-n_j)}}{\left[\log \widehat{S}_{KM}(t)\right]^2}.$$

Para un t fijo, un **intervalo de confianza** aproximado de $100(1-\alpha)\%$ de confianza es dado por

$$\widehat{S}_{KM}(t)^{\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}\left(\widehat{U}(t)\right)}\right\}}$$

Estimadores NP de la función de sobrevivencia

Estimador de **Nelson-Aalen** (N-A, Nelson, 1972; Aalen, 1978) es dado por

$$\widehat{\Lambda}_{NA}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j}\right),$$

En la literatura también es conocido como el estimador de **Fleming-Harrington**.

Estimadores NP de la función de sobrevivencia

Estimador de **Nelson-Aalen** (N-A, Nelson, 1972; Aalen, 1978) es dado por

$$\widehat{\Lambda}_{NA}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j}\right),$$

En la literatura también es conocido como el estimador de **Fleming-Harrington**.

El estimador de N-A para la función de sobrevivencia es dado por

$$\widehat{S}_{NA}(t) = \exp\left\{-\widehat{\Lambda}_{NA}(t)\right\}.$$

Estimadores NP de la función de sobrevivencia

Estimadores de la varianza del estimador de N-A:

- Aalen (1978).

$$\widehat{Var}\left(\widehat{\Lambda}_{NA}(t)\right) = \sum_{j:t_j<t}\left(\frac{d_j}{n_j^2}\right).$$

Y por lo tanto,

$$\widehat{Var}\left(\widehat{S}_{NA}(t)\right) = \left[\widehat{S}_{NA}(t)\right]^2 \sum_{j:t_j<t}\left(\frac{d_j}{n_j^2}\right).$$

Estimadores NP de la función de sobrevivencia

Datos de melanoma (paquete *timereg*). Son 205 pacientes diagnosticados con cáncer de melanoma recolectados por el Hospital universitario de Odense.

- **status**: código indicando el estado final del paciente (1:muerte debido al melanoma; 2: vivo; 3: muerte por otra causa).
- **days**: tiempo de sobrevida (en días).
- **sex**: 0: femenino; 1: masculino.
- **ulc**: presencia de úlceras. (1: presente; 0: ausente).
- **thick**: tamaño del tumor (en 1/100 mm).

Llamaremos **t1**, **d1** y **sex1** al tiempo (en años), indicadora de falla y sexo, respectivamente.

Estimadores NP de la función de sobrevivencia

Datos de cáncer de pulmón (paquete *survival*)

- **status**: código indicando el estado final del paciente (1:censura; 2: muerte).
- **time**: tiempo de sobrevida (en días).
- **sex**: 1: masculino; 2: femenino.
- **ph.karno**: score de Karnofsky, indicado por un médico. (mal:0 - bien:100).

NOTA: para la observación 206 de la variable **ph.karno** hay un missing. Llamaremos **t2**, **d2** y **sex2** al tiempo (en años), indicadora de falla y sexo, respectivamente.

Estimadores NP de la función de sobrevivencia

Por ejemplo, para definir datos de sobrevivencia en **R** podemos usar la función `Surv` del paquete `survival`

```
> require(survival); require(timereg)
> data(melanoma); attach(melanoma)
> t1=days/365.25
> d1=ifelse(status==1, 1, 0)
> Surv(t1, d1)
  [1]  0.02737851+  0.08213552+  0.09582478+  0.27104723+
  [5]  0.50650240   0.55852156   0.57494867   0.63518138
  [9]  0.63518138+  0.76386037   0.80766598   0.97193703+
 [13]  1.05681040   1.16632444   1.28405202   1.34976044+
 [17]  1.44832307   1.70020534   1.72210815   1.80424367
 ...
```

Estimadores NP de la función de sobrevivencia

La función **survfit** permite calcular los estimadores de **K-M** y de **N-A**, además de sus respectivas varianzas e intervalos de confianza. La forma general de usar esta función es

```
survfit(formula, data, weights, subset, na.action,
        newdata, individual=F, conf.int=.95, se.fit=T,
        type=c("kaplan-meier","fleming-harrington", "fh2"),
        error=c("greenwood","tsiatis"),
        conf.type=c("log","log-log","plain","none"),
        conf.lower=c("usual", "peto", "modified"))
```

Estimadores NP de la función de sobrevivencia

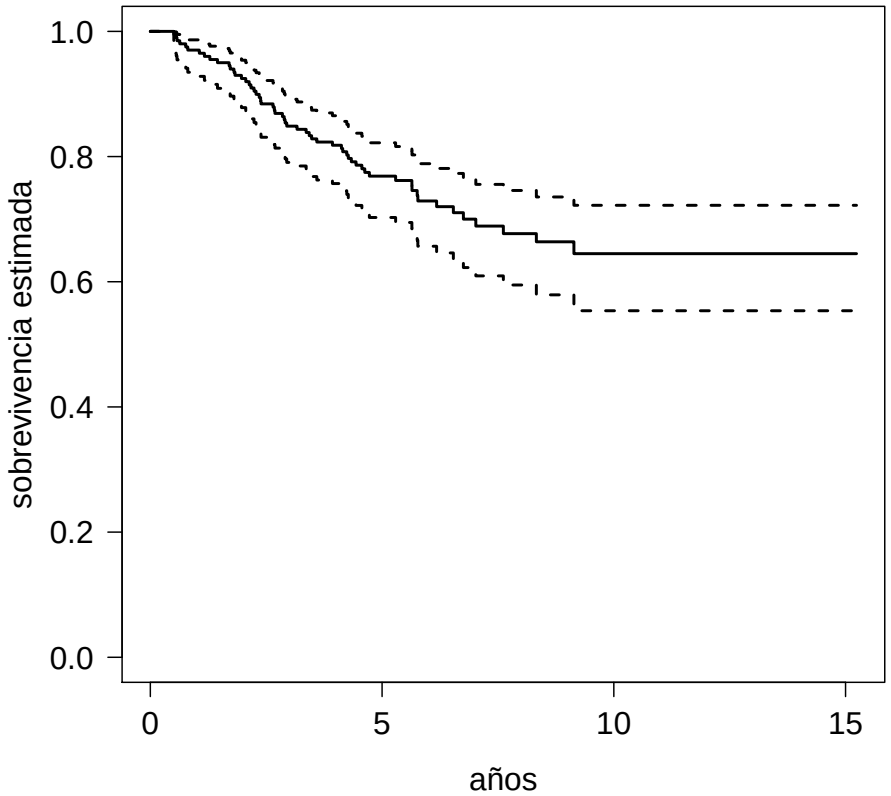
La función **survfit** permite calcular los estimadores de **K-M** y de **N-A**, además de sus respectivas varianzas e intervalos de confianza. La forma general de usar esta función es

```
survfit(formula, data, weights, subset, na.action,
        newdata, individual=F, conf.int=.95, se.fit=T,
         type=c("kaplan-meier","fleming-harrington", "fh2"),
         error=c("greenwood","tsiatis"),
         conf.type=c("log","log-log","plain","none"),
        conf.lower=c("usual", "peto", "modified"))
```

Estimadores NP de la función de sobrevivencia

Por ejemplo, para calcular el estimador de K-M usamos

```
> KM.0=survfit(Surv(t1, d1) ~ 1, conf.type="log-log")
> plot(KM.0, ...)
```



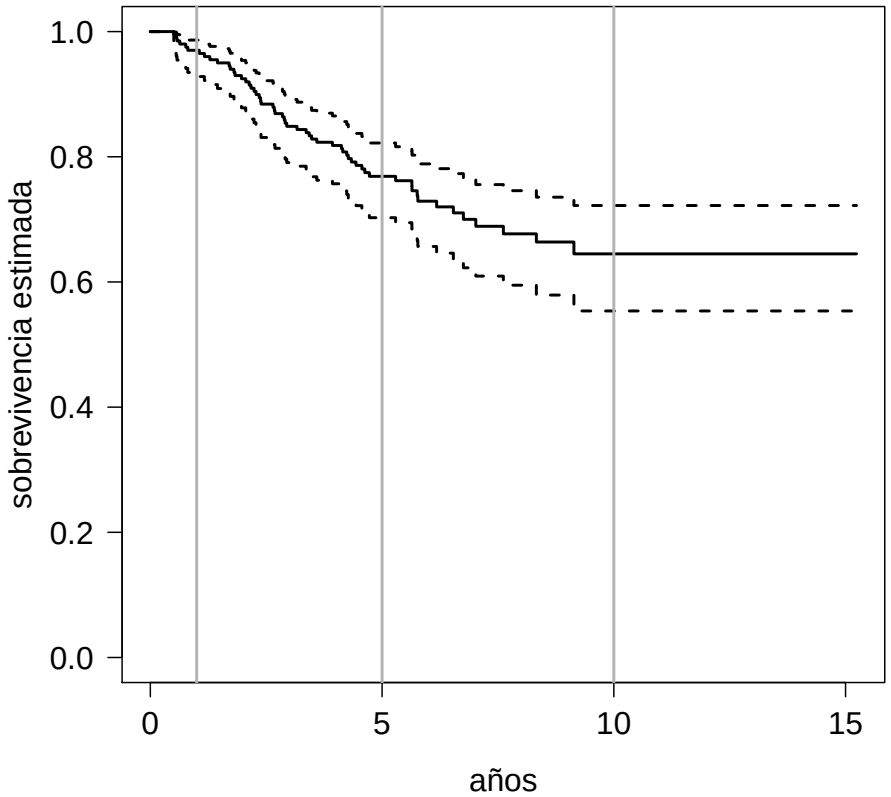
NOTA: para calcular el estimador de N-A, basta adicionar **type="fleming-harrington"**.

Estimadores NP de la función de sobrevivencia

```
> summary(KM.0)

Call: survfit(formula = Surv(t1, d1) ~ 1, conf.type = "log-log")

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.507    201       1    0.995 0.00496        0.965        0.999
 0.559    200       1    0.990 0.00700        0.961        0.998
 0.575    199       1    0.985 0.00855        0.954        0.995
 0.635    198       1    0.980 0.00985        0.948        0.992
 0.764    196       1    0.975 0.01100        0.941        0.990
 0.808    195       1    0.970 0.01202        0.935        0.986
 1.057    193       1    0.965 0.01297        0.928        0.983
 ...
 6.177     80       1    0.720 0.03438        0.646        0.781
 6.538     75       1    0.710 0.03523        0.635        0.773
 6.754     69       1    0.700 0.03619        0.623        0.765
 7.023     63       1    0.689 0.03729        0.609        0.756
 7.617     57       1    0.677 0.03854        0.595        0.746
 8.329     52       1    0.664 0.03994        0.579        0.735
 9.139     35       1    0.645 0.04307        0.554        0.722
```

Estimadores no paramétricos de la función de sobrevivencia

Por ejemplo. Si no consideramos covariables, ¿Cuál es la probabilidad de sobrevivir 1, 5 o 10 años para los pacientes con cáncer de melanoma? Es decir, ¿$S(1)$, $S(5)$ y $S(10)$?



Estimadores NP de la función de sobrevivencia

```
> summary(KM.0)
Call: survfit(formula = Surv(t1, d1) ~ 1, conf.type = "log-log")

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.507    201       1    0.995 0.00496        0.965        0.999
 ...
 0.808    195       1    0.970 0.01202        0.935        0.986
 1.057    193       1    0.965 0.01297        0.928        0.983
 1.166    192       1    0.960 0.01384        0.922        0.980
 ...
 4.726    131       1    0.769 0.03033        0.703        0.822
 5.292    110       1    0.762 0.03085        0.695        0.816
 5.643     95       1    0.754 0.03155        0.685        0.809
 ...
 7.617     57       1    0.677 0.03854        0.595        0.746
 8.329     52       1    0.664 0.03994        0.579        0.735
 9.139     35       1    0.645 0.04307        0.554        0.722
```

Estimadores NP de la función de sobrevivencia

```
> summary(KM.0)
Call: survfit(formula = Surv(t1, d1) ~ 1, conf.type = "log-log")

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.507    201       1    0.995 0.00496        0.965        0.999
 ...
 0.808    195       1    0.970 0.01202        0.935        0.986
 1.057    193       1    0.965 0.01297        0.928        0.983
 1.166    192       1    0.960 0.01384        0.922        0.980
 ...
 4.726    131       1    0.769 0.03033        0.703        0.822
 5.292    110       1    0.762 0.03085        0.695        0.816
 5.643     95       1    0.754 0.03155        0.685        0.809
 ...
 7.617     57       1    0.677 0.03854        0.595        0.746
 8.329     52       1    0.664 0.03994        0.579        0.735
 9.139     35       1    0.645 0.04307        0.554        0.722
```

Estimadores no paramétricos de la función de sobrevivencia

¿Cuál es la probabilidad de sobrevivir 1, 5 o 10 años para los pacientes con cáncer de melanoma?

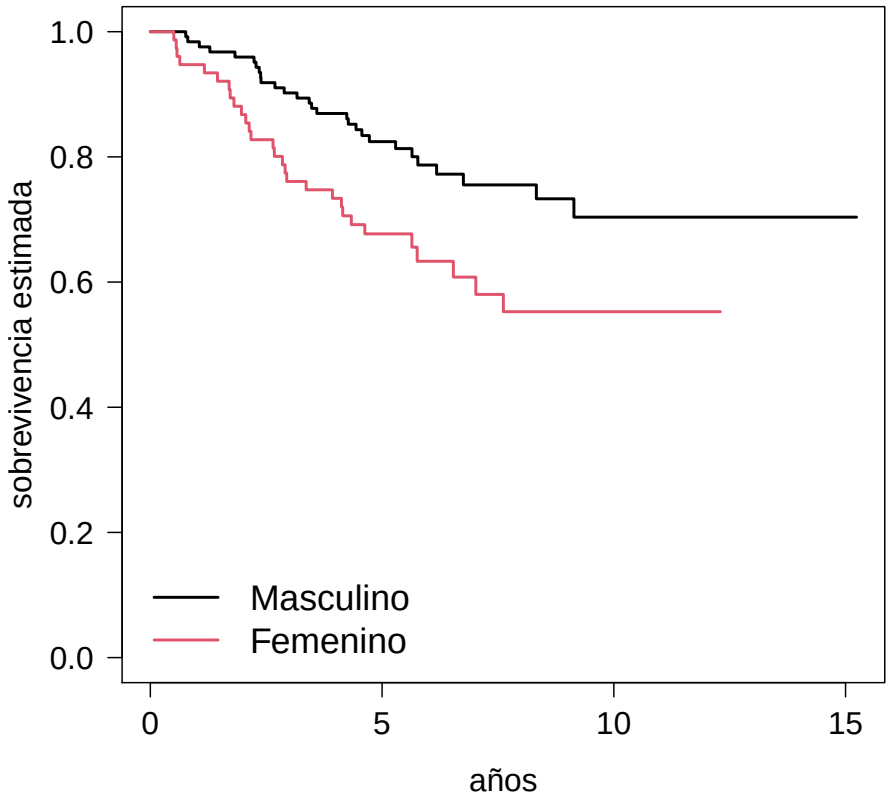
- $\widehat{S}_{KM}(1) = 0.970$, (0.935 - 0.986)
- $\widehat{S}_{KM}(5) = 0.769$, (0.703 - 0.822)
- $\widehat{S}_{KM}(10) = 0.645$, (0.554 - 0.722)

Se presenta la estimativa puntual y el respectivo Intervalo al 95% de confianza.

Estimadores NP de la función de sobrevivencia

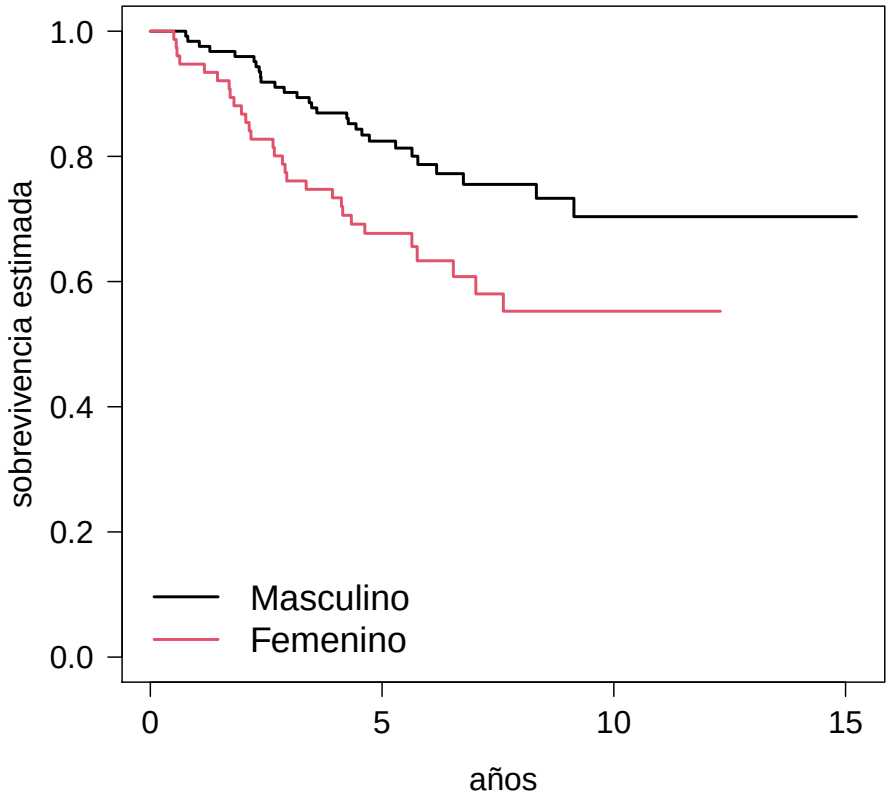
Si queremos calcular el estimador de K-M para los diferentes niveles de una variable usamos

```
> KM.sex=survfit(Surv(t1, d1) ~ sex1, conf.type="log-log")
> plot(KM.sex, ...)
```



Estimadores no paramétricos de la función de sobrevivencia

¿Cuál es la probabilidad de sobrevivir 1 año para las mujeres con cáncer de melanoma? ¿Y para los hombres?



Estimadores NP de la función de sobrevivencia

```
> summary(KM.sex)

Call: survfit(formula = Surv(t1, d1) ~ sex1, conf.type = "log-log")

                sex1=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
0.764    124       1    0.992 0.00803        0.944        0.999
0.808    123       1    0.984 0.01131        0.937        0.996
1.057    121       1    0.976 0.01384        0.927        0.992
1.284    120       1    0.968 0.01593        0.916        0.988
...

                sex1=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
0.507     76       1    0.987  0.0131        0.910        0.998
0.559     75       1    0.974  0.0184        0.899        0.993
0.575     74       1    0.961  0.0223        0.883        0.987
0.635     73       1    0.947  0.0256        0.866        0.980
1.166     72       1    0.934  0.0284        0.849        0.972
1.448     70       1    0.921  0.0310        0.832        0.964
...
```

Estimadores no paramétricos de la función de sobrevivencia

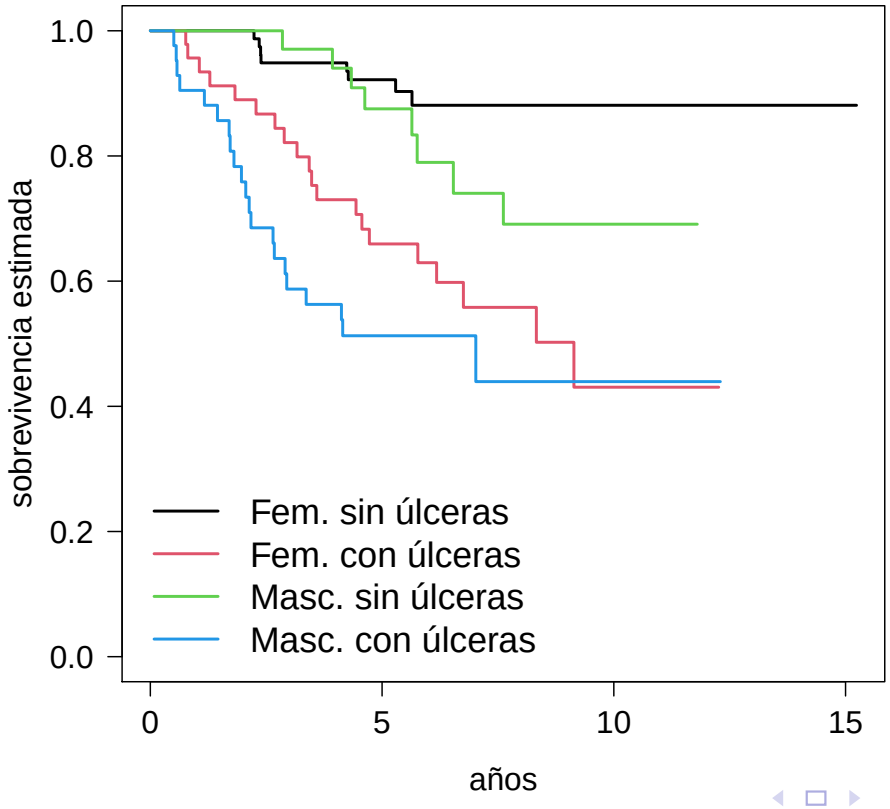
¿Cuál es la probabilidad de sobrevivir 1 años para las mujeres con cáncer de melanoma? ¿Y para los hombres?

- Para las mujeres
  - $\widehat{S}_{KM}(1) = 0.984, (0.937 - 0.996)$.
- Para los hombres
  - $\widehat{S}_{KM}(1) = 0.947, (0.866 - 0.980)$.

Estimadores NP de la función de sobrevivencia

También es posible construir el estimador de K-M para más de una covariable de forma conjunta

```
KM.su=survfit(Surv(t1, d1) ~ sex1+ulc,
   conf.type="log-log")
plot(KM.su, ...)
```



Estimadores NP de la función de sobrevivencia

Se puede construir el estimador de K-M (o de N-A) para una covariable continua?
No, pero se puede categorizar dicha variable.

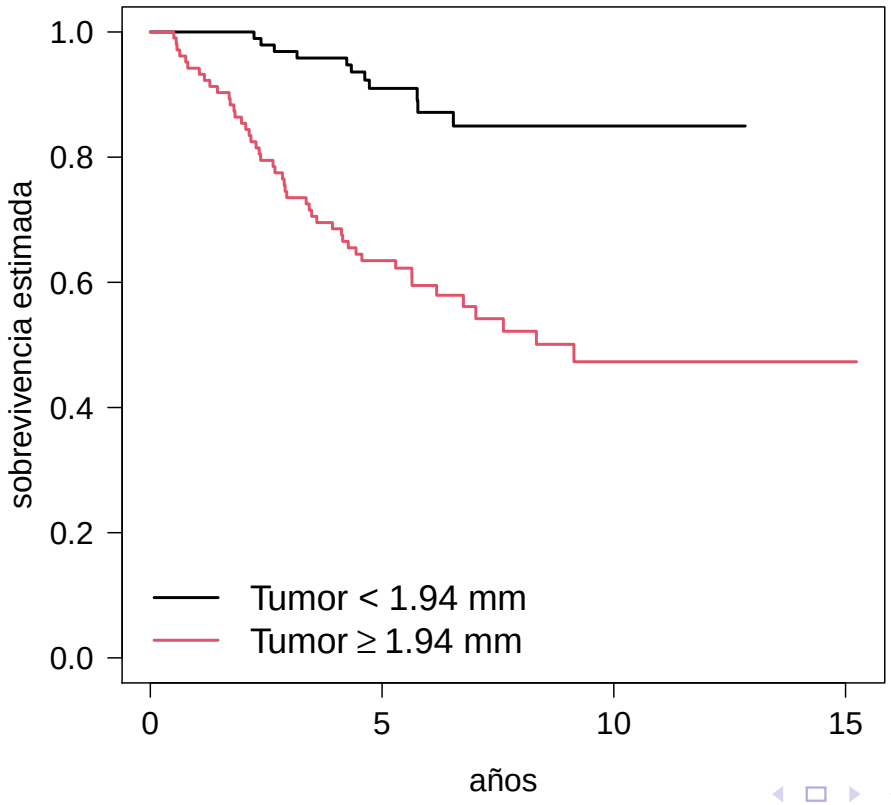
Estimadores NP de la función de sobrevivencia

Se puede construir el estimador de K-M (o de N-A) para una covariable continua?
No, pero se puede categorizar dicha variable.
En el ejemplo, consideremos la variable *thick*, pero multiplicada por 100 para llevarla a mm (en vez de 1/100 mm que es la escala original).

```
> summary(thick)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.10    0.97    1.94    2.92    3.56   17.42
```

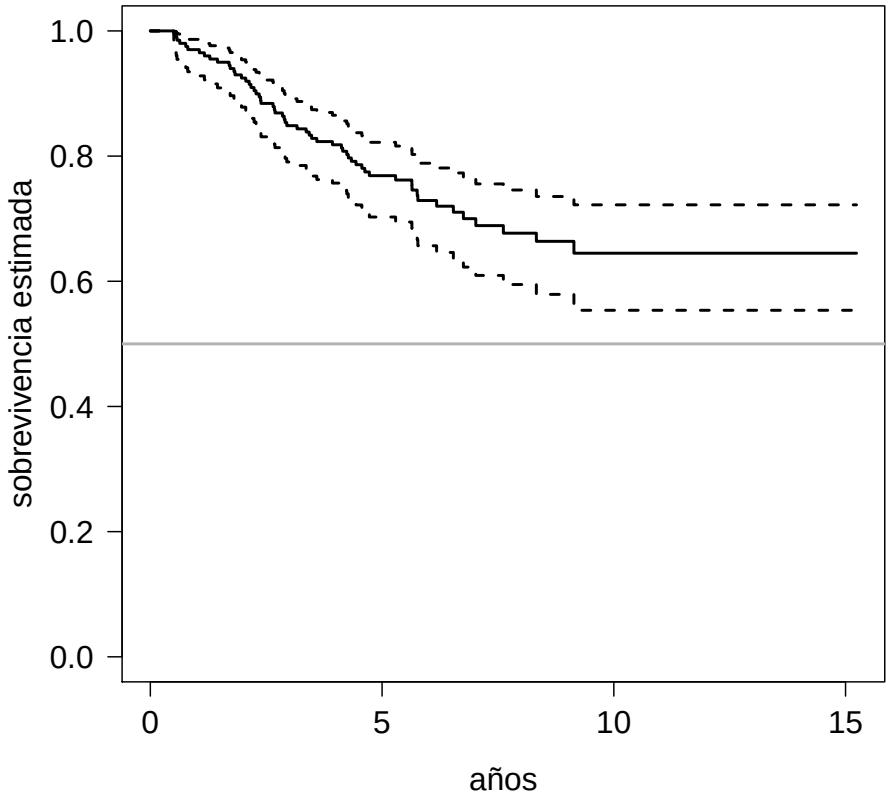
Estimadores NP de la función de sobrevivencia

```
> thick.cat=ifelse(thick<1.94,0,1)
> KM.thick=survfit(Surv(t1, d1) ~ thick.cat,
> conf.type="log-log")
> plot(KM.thick, ...)
```



Estimadores NP de la función de sobrevivencia

¿Cómo estimar la mediana (u otro percentil) usando el estimador de Kaplan-Meier?



Para este conjunto de datos no se puede estimar la mediana basándonos en el estimador de K-M.

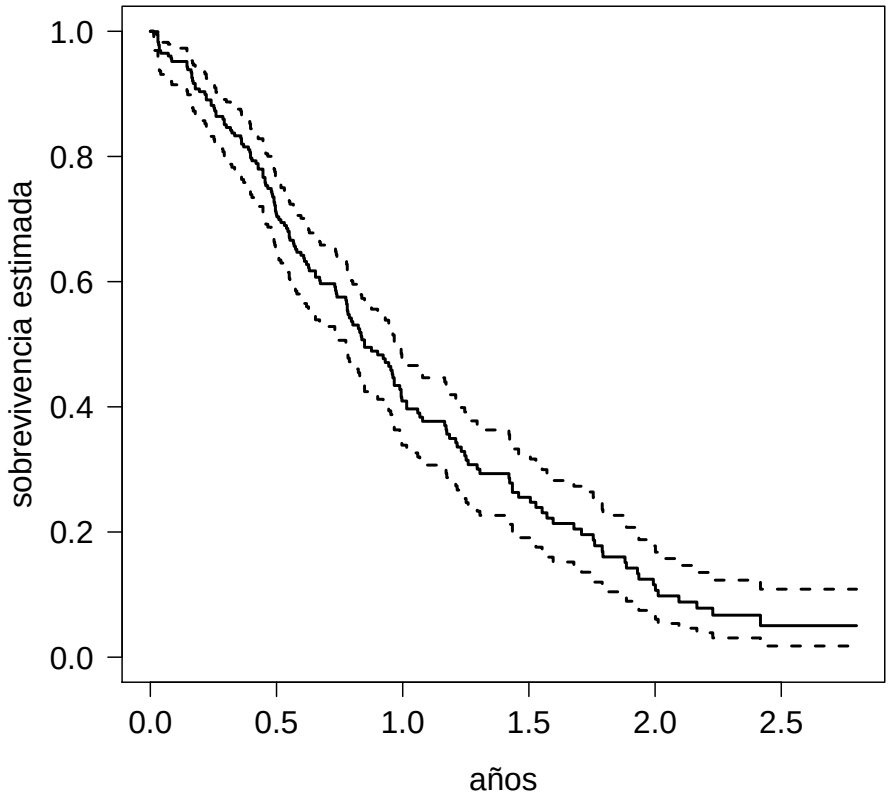
Estimadores NP de la función de sobrevivencia

```
> KM.0=survfit(Surv(t1, d1) ~ 1, conf.type="log-log")
> KM.0

Call: survfit(formula = Surv(t1, d1) ~ 1,
      conf.type = "log-log")

      n  events  median 0.95LCL 0.95UCL
    205      57      NA      NA      NA
```

Como $\widehat{S}_{KM}(t) > 0.5, \forall t > 0$, entonces no se puede obtener una estimativa para la mediana basado en este estimador.

Estimadores NP de la función de sobrevivencia

¿Cómo estimar la mediana (u otro percentil) usando el estimador de Kaplan-Meier?



(Conjunto de datos de cáncer de pulmón)

Estimadores NP de la función de sobrevivencia

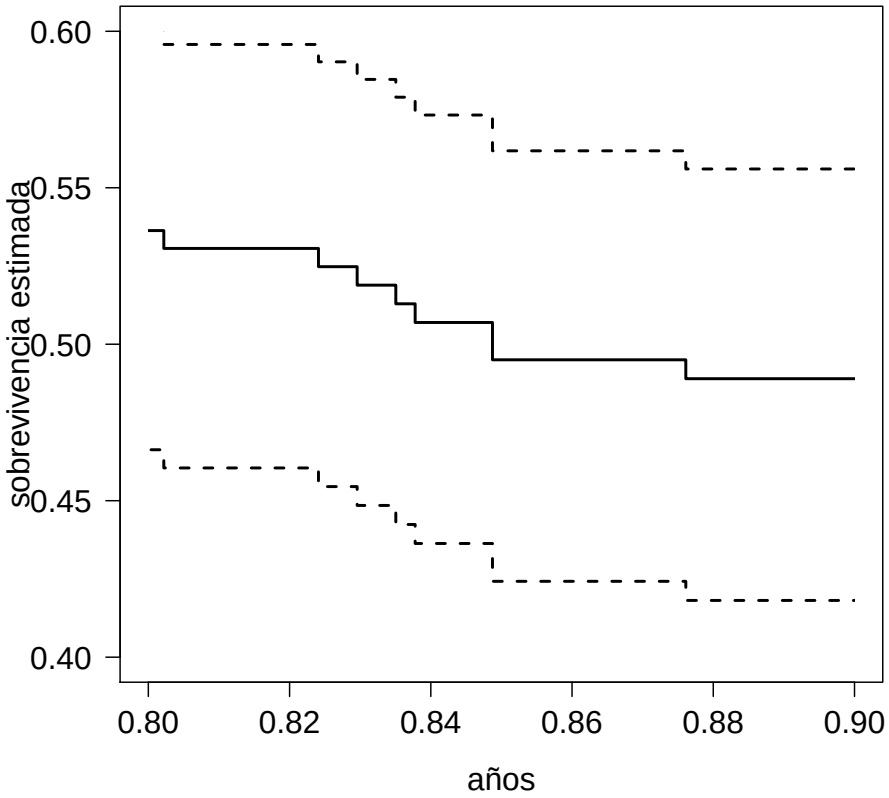
```
> KM.00=survfit(Surv(t2, d2) ~ 1,
          conf.type="log-log")
> KM.00

Call: survfit(formula = Surv(t2, d2) ~ 1, conf.type = "log-log")

    n  events  median 0.95LCL 0.95UCL
  227 164.000   0.849   0.778   0.988
```

Estimadores NP de la función de sobrevivencia

```
> summary(KM.00)

Call: survfit(formula = Surv(t2, d2) ~ 1, conf.type = "log-log")

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.0137    227       1   0.9956 0.00440       0.9691        0.999
 0.0301    226       3   0.9824 0.00873       0.9537        0.993
 0.0329    223       1   0.9780 0.00974       0.9479        0.991
 ...
 0.8350     87       1   0.5152 0.03504       0.4445        0.581
 0.8378     86       1   0.5092 0.03514       0.4384        0.576
 0.8487     85       2   0.4972 0.03532       0.4262        0.564
 0.8761     82       1   0.4912 0.03541       0.4201        0.558
 0.9008     81       1   0.4851 0.03548       0.4140        0.552
 ...
```

Estimadores NP de la función de sobrevivencia

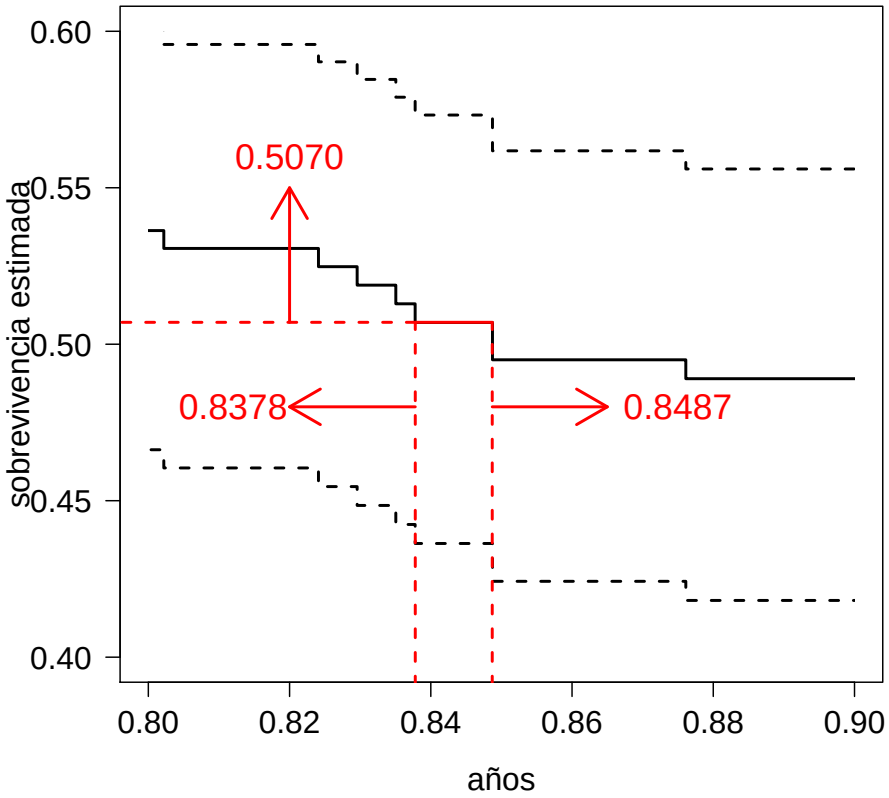
¿Cómo estimar la mediana (u otro percentil) usando el estimador de Kaplan-Meier?



(Conjunto de datos de cáncer de pulmón)

Estimadores NP de la función de sobrevivencia

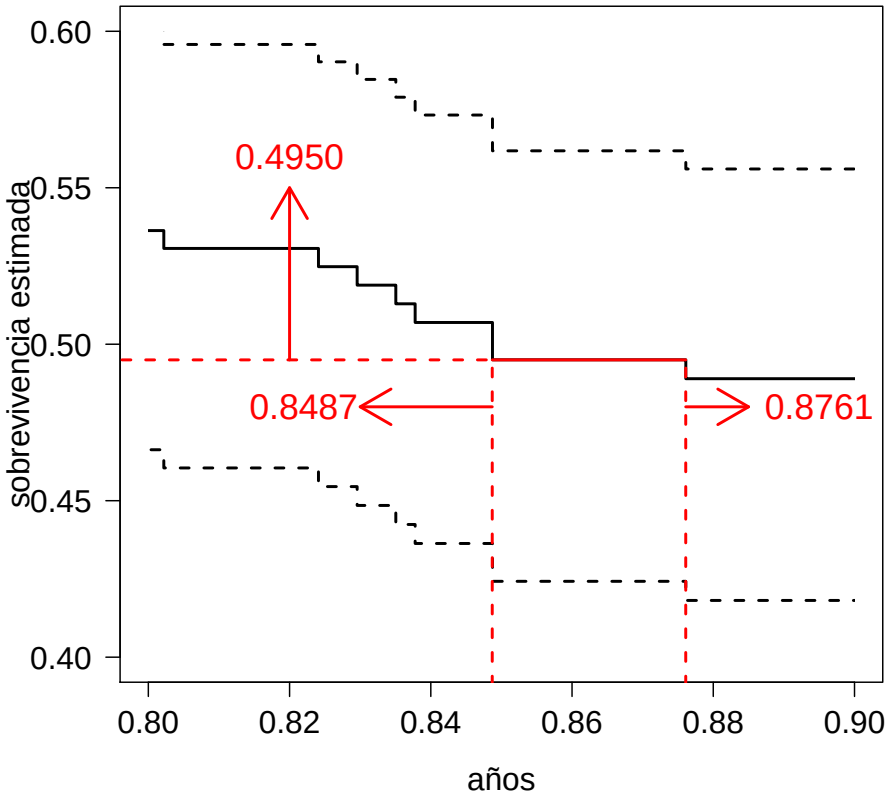
¿Cómo estimar la mediana (u otro percentil) usando el estimador de Kaplan-Meier?



(Conjunto de datos de cáncer de pulmón)

Estimadores NP de la función de sobrevivencia

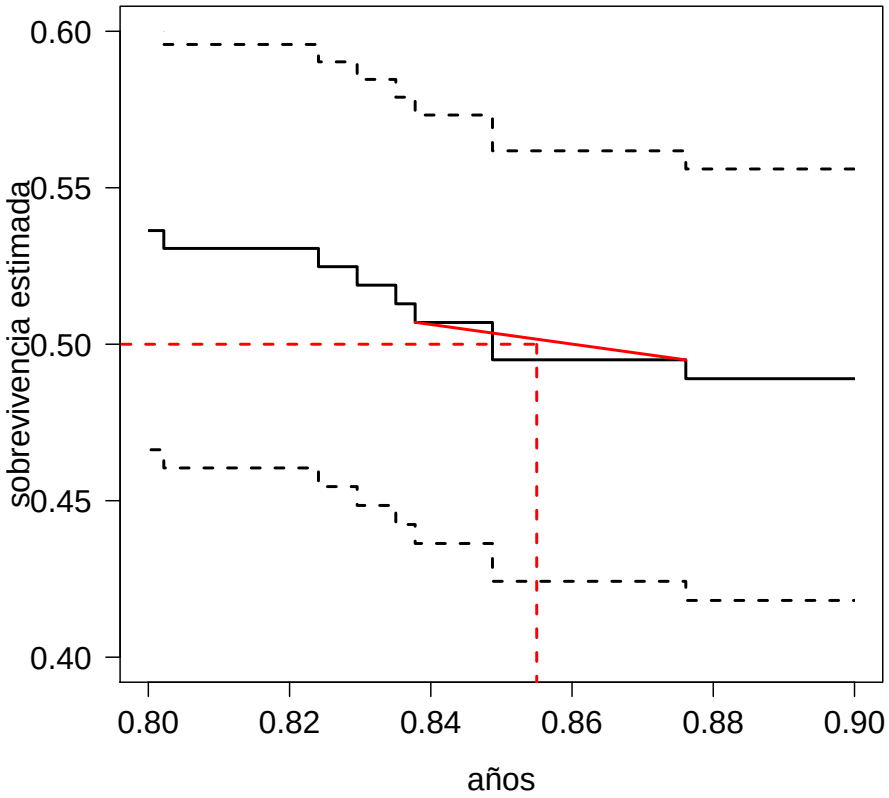
¿Cómo estimar la mediana (u otro percentil) usando el estimador de Kaplan-Meier?



(Conjunto de datos de cáncer de pulmón)

Estimadores NP de la función de sobrevivencia

¿Cómo estimar la mediana (u otro percentil) usando el estimador de Kaplan-Meier?



Interpolación lineal entre los puntos (0.8378, 0.5070) y (0.8761, 0.4950).

Estimadores NP de la función de sobrevivencia

¿Cómo estimar la media usando el estimador de Kaplan-Meier?

Estimadores NP de la función de sobrevivencia

¿Cómo estimar la media usando el estimador de Kaplan-Meier?
Se puede chequear que, para cualquier distribución positiva en que la media exista, se tiene que

$$\begin{aligned}\mathbb{E}(T) &= \int_0^{+\infty} t \times f(t)dt \\ &= \int_0^{+\infty} S(t)dt.\end{aligned}$$

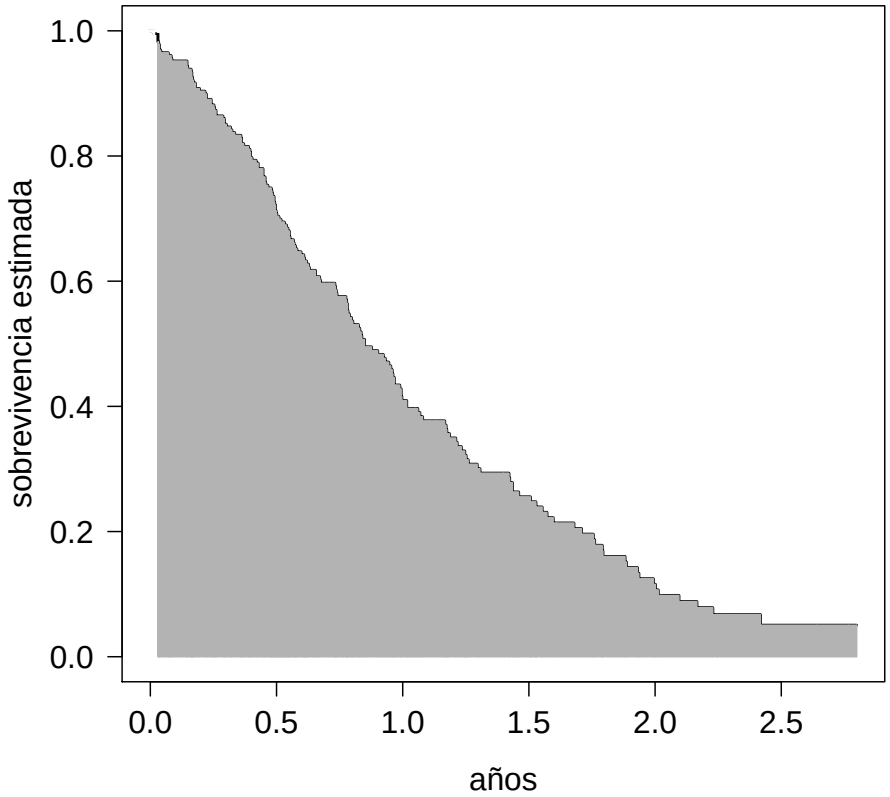
Estimadores NP de la función de sobrevivencia

¿Cómo estimar la media usando el estimador de Kaplan-Meier?
Se puede chequear que, para cualquier distribución positiva en que la media exista, se tiene que

$$\begin{aligned}\mathbb{E}(T) &= \int_0^{+\infty} t \times f(t)dt \\ &= \int_0^{+\infty} S(t)dt.\end{aligned}$$

En otras palabras, la media de la distribución es el área bajo la curva de sobrevivencia de la variable aleatoria T. (siempre y cuando ésta exista).

Estimadores NP de la función de sobrevivencia



Ejemplo del área bajo la curva para el conjunto de datos de cáncer de pulmón.

Nota: Recuerde que si la observación más grande de la muestra es censurada, entonces $\widehat{S}_{KM}(t) > 0$, $\forall t > t_k$.

Estimadores NP de la función de sobrevivencia

De forma general, se puede chequear que, para cualquier distribución positiva se tiene que

$$\mathbb{E}(T^k) = \int_0^{+\infty} t^{k-1} S(t) dt.$$

Es decir, $\mathbb{E}(T^k)$ es el área bajo la curva de la función $t^{k-1}S(t)$, siempre y cuando esa esperanza exista. De esta forma, una fórmula para la varianza puede ser deducida.

Estimadores NP de la función de sobrevivencia

Tanto la media como la varianza estimada a través del estimador de Kaplan-Meier pueden ser obtenidas usando la función `enparCensored` del paquete `EnvStats`

```
enparCensored(x, censored, censoring.side = "left",
    correct.se = FALSE, left.censored.min = "DL",
    right.censored.max = "DL", ci = FALSE,
    ci.method = "normal.approx", ci.type = "two-sided",
    conf.level = 0.95, pivot.statistic = "z",
    ci.sample.size = NULL, n.bootstraps = 1000)
```

Estimadores NP de la función de sobrevivencia

Tanto la media como la varianza estimada a través del estimador de Kaplan-Meier pueden ser obtenidas usando la función `enparCensored` del paquete `EnvStats`

```
enparCensored( x, censored, censoring.side = "left",
    correct.se = FALSE, left.censored.min = "DL",
     right.censored.max = "DL", ci = FALSE,
    ci.method = "normal.approx", ci.type = "two-sided",
    conf.level = 0.95, pivot.statistic = "z",
    ci.sample.size = NULL, n.bootstraps = 1000)
```

x: tiempos observados. **censored**: indicador de censura. **censoring.side**: censura a la izquiera (left) o a la derecha (right). **right.censored.max**: Valor de censura máximo DL (detection limit, detecta automáticamente la censura más alta); puede especificarse un valor.

Estimadores NP de la función de sobrevivencia

```
> enparCensored(t2, 1-d2, censoring.side = "right")
$distribution
[1] "None"
$sample.size
[1] 227
...
$percent.censored
[1] 27.7533
$parameters
      mean         sd    se.mean
1.03347528 0.72608416 0.05409535
$n.param.est
[1] 2
$method
[1] "Kaplan-Meier"
...
```

Estimadores NP de la función de sobrevivencia

De esta forma, para el conjunto de datos de cáncer de pulmón tenemos que $\widehat{\mathbb{E}(T)} = 1.0335$, $\widehat{\sqrt{Var(T)}} = 0.7261$ y s.e.$\left(\widehat{\mathbb{E}(T)}\right) = 0.0541$.

Estimadores NP de la función de sobrevivencia

De esta forma, para el conjunto de datos de cáncer de pulmón tenemos que $\widehat{\mathbb{E}(T)} = 1.0335$, $\sqrt{\widehat{Var(T)}} = 0.7261$ y s.e.$\left(\widehat{\mathbb{E}(T)}\right) = 0.0541$.

Note que la media y desviación estándar usual de los tiempos (ignorando que son datos censurados a la derecha) son 0.8357 y 0.5767, respectivamente.

Estimadores NP de la función de sobrevivencia

De esta forma, para el conjunto de datos de cáncer de pulmón tenemos que $\widehat{\mathbb{E}(T)} = 1.0335$, $\sqrt{\widehat{Var(T)}} = 0.7261$ y s.e.$\left(\widehat{\mathbb{E}(T)}\right) = 0.0541$.

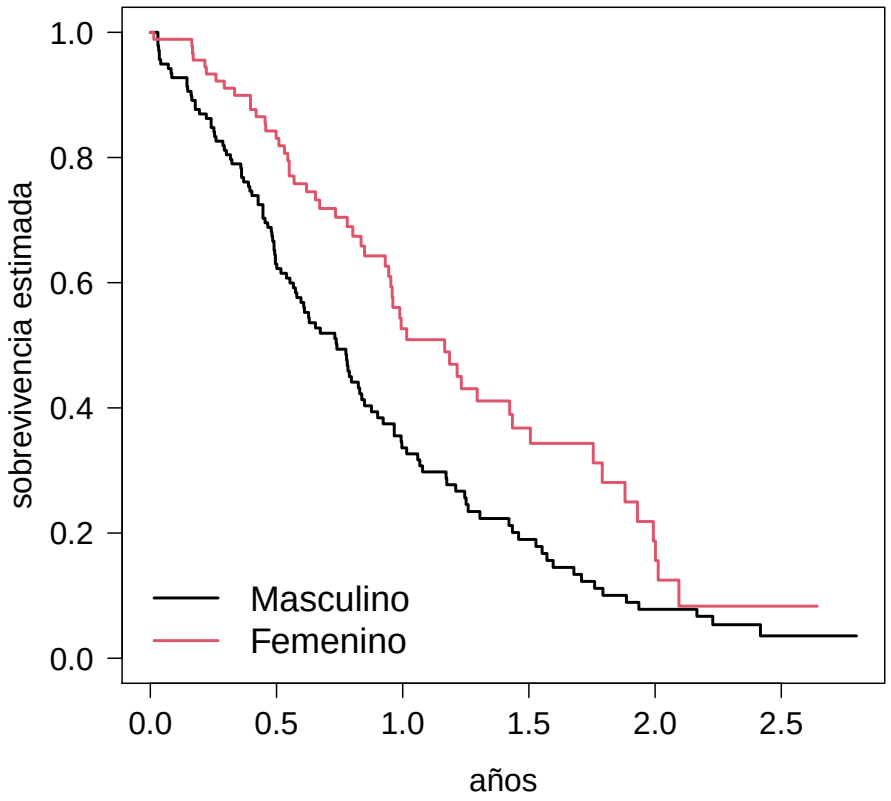
Note que la media y desviación estándar usual de los tiempos (ignorando que son datos censurados a la derecha) son 0.8357 y 0.5767, respectivamente.

NOTA: En presencia de covariables, el mismo procedimiento podría ser aplicado para cada grupo, de forma de obtener estimaciones para la media y desviación estándar en cada factor de una covariable, o bien, en combinaciones de factores de covariables.

Estimadores NP de la función de sobrevivencia

En el ejemplo de cáncer de pulmón.

```
KM.s.2=survfit(Surv(t2, d2) ~ sex2, conf.type="log-log")
plot(KM.s.2, ...)
```



Estimadores NP de la función de sobrevivencia

```
> KM.s.2
Call: survfit(formula = Surv(t2, d2) ~ sex2, conf.type = "log-log")

        n events median 0.95LCL 0.95UCL
sex2=1 137    111  0.739   0.580   0.849
sex2=2  90     53  1.166   0.945   1.435
```

Estimadores NP de la función de sobrevivencia

```
> hombres<-which(sex2==1)
> enparCensored(t2[hombres], 1-d2[hombres], censoring.side = "right")
$distribution
[1] "None"
$sample.size
[1] 137
...
$percent.censored
[1] 18.9781
$parameters
      mean         sd    se.mean
0.89718785 0.68364094 0.06303053
$n.param.est
[1] 2
$method
[1] "Kaplan-Meier"
...
```

Estimadores NP de la función de sobrevivencia

```
> mujeres<-which(sex2==2)
> enparCensored(t2[mujeres], 1-d2[mujeres], censoring.side = "right")
$distribution
[1] "None"
$sample.size
[1] 90
...
$percent.censored
[1] 41.11111
$parameters
      mean         sd    se.mean
1.24819736 0.72382830 0.09012297
$n.param.est
[1] 2
$method
[1] "Kaplan-Meier"
...
```

Estimadores NP de la función de sobrevivencia

En otras palabras,

	Hombres	Mujeres
$\widehat{\mathbb{E}(T)}$	0.8972	1.2482
$\widehat{\sqrt{Var(T)}}$	0.6836	0.7238
s.e.$\left(\widehat{\mathbb{E}(T)}\right)$	0.0630	0.0901
$\widehat{Me(T)}$	0.7390	1.1660
$\bar{t}$	0.7790	0.9280
S_t	0.5839	0.5571
me(t)	0.6160	0.8008

Comparación de curvas de sobrevivencia

Comparación de curvas de sobrevivencia

Considere que queremos comparar las curvas de sobrevivencia en dos grupos, digamos 1 y 2. En otras palabras, queremos testear las hipótesis

$$H_0 : S_1(t) = S_2(t), \quad \forall t > 0 \qquad \text{versus} \qquad H_1 : \text{Lo contrario.}$$

Asumamos que $t_1 < t_2 < \ldots < t_k$ son los diferentes tiempos de falla observados al combinar todos los tiempos observados en ambos grupos.

Comparación de curvas de sobrevivencia

En cada tiempo de falla t_j, $j = 1, \ldots, k$, podemos generar una tabla de contingencia como la siguiente

	Grupo		
	1	2	
Fallas hasta t_j	d_{1j}	d_{2j}	d_j
No fallas hasta t_j	$n_{1j} - d_{ij}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

Comparación de curvas de sobrevivencia

En este contexto, el primer test propuesto fue el de *logrank* (Mantel, 1966), cuya estadística es dada por

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2},$$

en que $w_{2j} = n_{2j}d_j/n_j$ y $(V_j)_2 = n_{2j}n_{1j}d_j(n_j - d_j)/(n_j^2(n_j - 1)^2)$.
Sobre H_0, $T \sim \chi^2_{(1)}$.

Comparación de curvas de sobrevivencia

En este contexto, el primer test propuesto fue el de *logrank* (Mantel, 1966), cuya estadística es dada por

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2},$$

en que $w_{2j} = n_{2j}d_j/n_j$ y $(V_j)_2 = n_{2j}n_{1j}d_j(n_j - d_j)/(n_j^2(n_j - 1)^2)$.
Sobre H_0, $T \sim \chi^2_{(1)}$.

NOTA: es recomendable usar este test si se satisface que el cuociente de las funciones de riesgo de ambos grupos es aproximadamente constante (i.e., propiedad de riesgos proporcionales).

Comparación de curvas de sobrevivencia

Otras propuestas en la litetura:

$$T = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2},$$

en que u_j, $j = 1, \ldots, k$ son pesos asociados a las tablas de contingencia de cada uno de los tiempos de falla. Casos particulares:

Comparación de curvas de sobrevivencia

Otras propuestas en la litetura:

$$T = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2},$$

en que u_j, $j = 1, \ldots, k$ son pesos asociados a las tablas de contingencia de cada uno de los tiempos de falla. Casos particulares:

- $u_j = 1$: test de logrank.

Comparación de curvas de sobrevivencia

Otras propuestas en la litetura:

$$T = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2},$$

en que u_j, $j = 1, \ldots, k$ son pesos asociados a las tablas de contingencia de cada uno de los tiempos de falla. Casos particulares:

- $u_j = 1$: test de logrank.
- $u_j = n_j$: test de Gehan-Breslow (Gehan, 1965; Breslow, 1970).

Comparación de curvas de sobrevivencia

Otras propuestas en la litetura:

$$T = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2},$$

en que u_j, $j = 1, \ldots, k$ son pesos asociados a las tablas de contingencia de cada uno de los tiempos de falla. Casos particulares:

- $u_j = 1$: test de logrank.
- $u_j = n_j$: test de Gehan-Breslow (Gehan, 1965; Breslow, 1970).
- $u_j = \sqrt{n_j}$: test de Tarone and Ware (1977).

Comparación de curvas de sobrevivencia

Otras propuestas en la litetura:

$$T = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2},$$

en que u_j, $j = 1, \ldots, k$ son pesos asociados a las tablas de contingencia de cada uno de los tiempos de falla. Casos particulares:

- $u_j = 1$: test de logrank.
- $u_j = n_j$: test de Gehan-Breslow (Gehan, 1965; Breslow, 1970).
- $u_j = \sqrt{n_j}$: test de Tarone and Ware (1977).
- $u_j = \widetilde{S}(t_{j-1})n_j/(n_j + 1)$: test de Peto-Peto (1972); Prentice and Marek (1979), con $\widetilde{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j+1-d_j}{n_j+1}\right)$.

Comparación de curvas de sobrevivencia

Otras propuestas en la litetura:

$$T = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2},$$

en que u_j, $j = 1, \ldots, k$ son pesos asociados a las tablas de contingencia de cada uno de los tiempos de falla. Casos particulares:

- $u_j = 1$: test de logrank.
- $u_j = n_j$: test de Gehan-Breslow (Gehan, 1965; Breslow, 1970).
- $u_j = \sqrt{n_j}$: test de Tarone and Ware (1977).
- $u_j = \widetilde{S}(t_{j-1}) n_j/(n_j + 1)$: test de Peto-Peto (1972); Prentice and Marek (1979), con $\widetilde{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j+1-d_j}{n_j+1}\right)$.
- $u_j = [\widehat{S}_{KM}(t_{j-1})]^{\rho}$: test de Harrington-Fleming (1982).

Comparación de curvas de sobrevivencia

La función `logrank_test` del paquete `coin` permite implementar estos tests

```
logrank_test(formula, ties.method = c("mid-ranks", "Hothorn-Lausen",
                                      "average-scores"),
      type = c("logrank", "Gehan-Breslow", "Tarone-Ware", "Prentice",
               "Prentice-Marek", "Andersen-Borgan-Gill-Keiding",
               "Fleming-Harrington", "Gaugler-Kim-Liao", "Self"),
      rho = NULL, gamma = NULL, ...)
```

Comparación de curvas de sobrevivencia

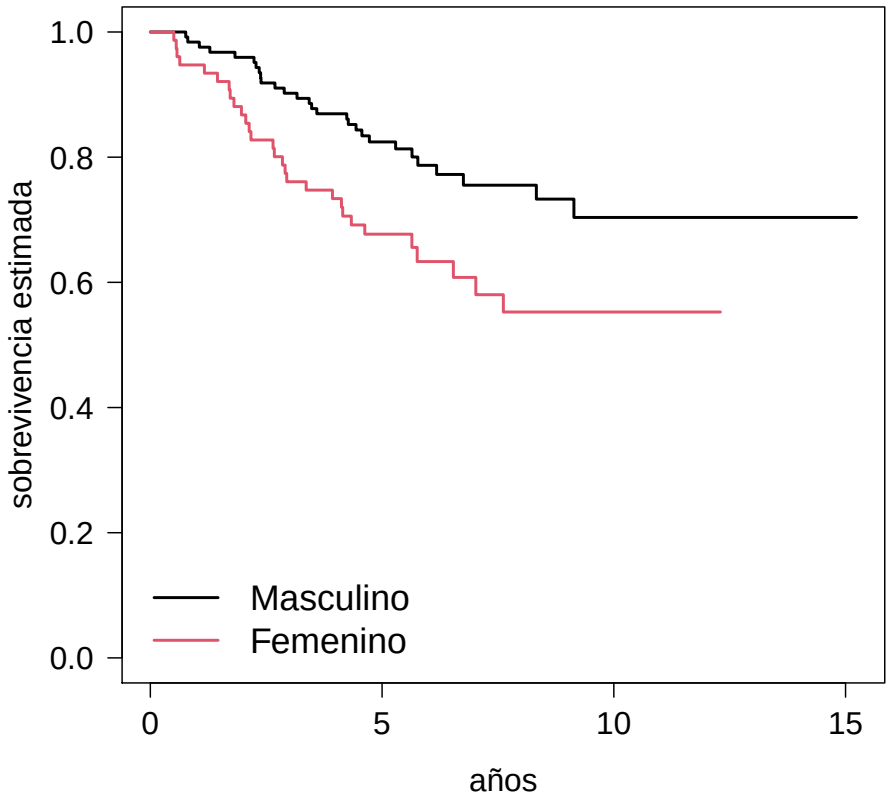
La función `logrank_test` del paquete `coin` permite implementar estos tests

```
logrank_test(formula, ties.method = c("mid-ranks", "Hothorn-Lausen",
                                      "average-scores"),
      type = c("logrank", "Gehan-Breslow", "Tarone-Ware", "Prentice",
               "Prentice-Marek", "Andersen-Borgan-Gill-Keiding",
               "Fleming-Harrington", "Gaugler-Kim-Liao", "Self"),
      rho = NULL, gamma = NULL, ...)
```

Estimadores NP de la función de sobrevivencia

Volvamos a nuestro ejemplo del cáncer de melanoma...

```
> KM.sex=survfit(Surv(t1, d1) ~ sex1)
> plot(KM.sex, ...)
```



Estimadores NP de la función de sobrevivencia

Por ejemplo, para hacer el test de log-rank para comparar las curvas de sobrevivencia por sexo usamos

```
> logrank_test(Surv(t1, d1) ~ as.factor(sex1))

        Asymptotic Two-Sample Logrank Test

data:  Surv(t1, d1) by as.factor(sex1) (0, 1)
Z = 2.4917, p-value = 0.01271
alternative hypothesis: true theta is not equal to 1
```

Estimadores NP de la función de sobrevivencia

Análogamente, para hacer el test de Tarone and Ware para comparar las curvas de sobrevivencia por sexo usamos

```
> logrank_test(Surv(t1, d1) ~ as.factor(sex1),
                type="Tarone-Ware")

        Asymptotic Two-Sample Tarone-Ware Test

data:  Surv(t1, d1) by as.factor(sex1) (0, 1)
Z = 2.6169, p-value = 0.008873
alternative hypothesis: true theta is not equal to 1
```

Estimadores NP de la función de sobrevivencia

Análogamente, para hacer el test de Fleming and Harrington (con $\rho = 0.5$) para comparar las curvas de sobrevivencia por sexo usamos

```
> logrank_test(Surv(t1, d1) ~ as.factor(sex1),
                type=""Fleming-Harrington", rho=0.5)

        Asymptotic Two-Sample Fleming-Harrington Test

data:  Surv(t1, d1) by as.factor(sex1) (0, 1)
Z = 2.5559, p-value = 0.01059
alternative hypothesis: true theta is not equal to 1
```

Estimadores NP de la función de sobrevivencia

Un resumen de los 5 tests se muestra a continuación

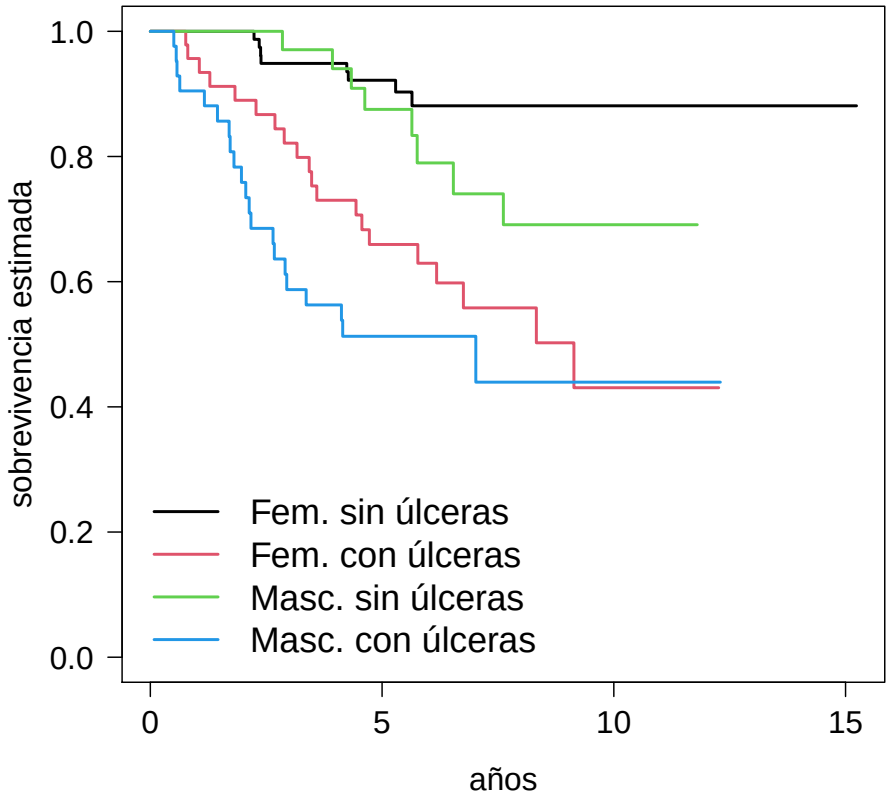
Test	Z	p-value
log-rank	2.4917	0.012710
Gehan-Breslow	2.6743	0.007489
Tarone-Ware	2.6169	0.008873
Prentice-Marek	2.6144	0.008939
Fleming-Harrington	2.5559	0.010590

Note que todos los tests rechazan la hipótesis de igualdad usando un 5% de significación, mientras que si se usa un 1% de significación los tests de log-rank y de Fleming-Harrington no rechazan la hipótesis nula.

Estimadores NP de la función de sobrevivencia

Observación: el test también permite la comparación de 3 o más grupos.

```
KM.su=survfit(Surv(t1, d1) ~ sex1 + ulc,
conf.type="log-log")
plot(KM.su, ...)
```



Estimadores NP de la función de sobrevivencia

Podemos crear una nueva variable indicando cada uno de los grupos

```
> grupos=rep(NA, length=length(sex1))
> for(i in 1:length(sex1))
+ {
+ if(sex1[i]==0 & ulc[i]==0) grupos[i]=1
+ if(sex1[i]==0 & ulc[i]==1) grupos[i]=2
+ if(sex1[i]==1 & ulc[i]==0) grupos[i]=3
+ if(sex1[i]==1 & ulc[i]==1) grupos[i]=4
+ }
> table(grupos)

grupos
 1  2  3  4
79 47 36 43
```

Estimadores NP de la función de sobrevivencia

Así, para hacer el test de log-rank para comparar las curvas de sobrevivencia por sexo y úlceras usamos

```
> logrank_test(Surv(t1, d1) ~ as.factor(grupos))

        Asymptotic K-Sample Logrank Test

data:  Surv(t1, d1) by as.factor(grupos) (1, 2, 3, 4)
chi-squared = 30.63, df = 3, p-value = 1.017e-06
```

Estimadores NP de la función de sobrevivencia

Un resumen de los 5 tests para este caso se muestra a continuación

Test	χ^2	p-value
log-rank	30.630	1.017e-06
Gehan-Breslow	33.971	2.009e-07
Tarone-Ware	32.617	3.879e-07
Prentice-Marek	32.720	3.689e-07
Fleming-Harrington	31.710	6.025e-07

Note que todos los tests rechazan la hipótesis de igualdad usando cualquier nivel de significación usual.

Modelos paramétricos

Modelos paramétricos

Distribución Weibull

Modelos paramétricos

- Distribución Weibull

$$f(t;\boldsymbol{\theta}) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\},$$

$$S(t;\boldsymbol{\theta}) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\},$$

$$\Lambda(t;\boldsymbol{\theta}) = \left(\frac{t}{\alpha}\right)^\gamma, \qquad t, \alpha, \gamma > 0,$$

en que $\boldsymbol{\theta} = (\gamma, \alpha)$. Denotaremos como WEI$(\gamma, \alpha)$

Modelos paramétricos

- Distribución Weibull

$$f(t;\boldsymbol{\theta}) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\},$$

$$S(t;\boldsymbol{\theta}) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\},$$

$$\Lambda(t;\boldsymbol{\theta}) = \left(\frac{t}{\alpha}\right)^\gamma, \qquad t, \alpha, \gamma > 0,$$

  en que $\boldsymbol{\theta} = (\gamma, \alpha)$. Denotaremos como WEI$(\gamma, \alpha)$

- $\gamma = 1$: modelo exponencial.
- α es un parámetro de escala y γ es un parámetro de forma.
- Función de riesgo es monótona: puede ser constante ($\gamma = 1$), decreciente ($\gamma < 1$) o creciente ($\gamma > 1$).

Modelos paramétricos

Si $T \sim$WEI(γ, α)

- $\mathbb{E}(T) = \alpha\Gamma(1 + 1/\gamma)$;
- $\text{Var}(T) = \alpha^2 \left[\Gamma(1 + 2/\gamma) - \{\Gamma(1 + 1/\gamma)\}^2\right]$;
- $t_q = \alpha[-\log(1 - q)]^{1/\gamma}$ es el percentil $100 \times q$ de la distribución, con $q \in (0, 1)$,

siendo $\Gamma(\cdot)$ la función gamma.

Modelos paramétricos

Además si $T \sim$WEI(γ, α), entonces $Y = \log(T) \sim VE(\mu, \sigma)$ (distribución valor extremo)

$$f(y; \mu, \sigma) = \frac{1}{\sigma} \exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left\{\frac{y-\mu}{\sigma}\right\}\right\},$$

$$S(y; \mu, \sigma) = \exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}, \quad y, \mu \in \mathbb{R}, \sigma > 0,$$

en que $\mu = \log(\alpha)$ y $\sigma = 1/\gamma$ son parámetros de localización y escala, respectivamente.

Modelos paramétricos

En R, para el modelo Weibull las siguientes funciones estás implementadas en el paquete base de R

```
> dweibull(x, shape, scale=1, log=FALSE)
> pweibull(x, shape, scale=1, lower.tail = TRUE, log.p = FALSE)
> qweibull(x, shape, scale=1, lower.tail = TRUE, log.p = FALSE)
> rweibull(n, shape, scale=1)
```

que corresponden a la función de densidad (`dweibull`), distribución acumulada (`pweibull`), cuantil (`qweibull`) y un generador de números aleatorios (`rweibull`), respectivamente.

Modelos paramétricos

La función de distribución acumulada de $T \sim$WEI$(\gamma = 1.3, \alpha = 10)$, evaluada en los tiempos 1, 3 y 5, puede ser calculada como

```
> pweibull(c(1,3,5), shape=1.3, scale=10)
[1] 0.0488835 0.1886482 0.3337739
```

Modelos paramétricos

La función de distribución acumulada de $T \sim$WEI$(\gamma = 1.3, \alpha = 10)$, evaluada en los tiempos 1, 3 y 5, puede ser calculada como

```
> pweibull(c(1,3,5), shape=1.3, scale=10)
[1] 0.0488835 0.1886482 0.3337739
```

Mientras que la función de sobrevivencia de la misma variable aleatoria evaluada en los mismos puntos se calcula como

```
> pweibull(c(1,3,5), shape=1.3, scale=10, lower.tail=FALSE)
[1] 0.9511165 0.8113518 0.6662261
```

Modelos paramétricos

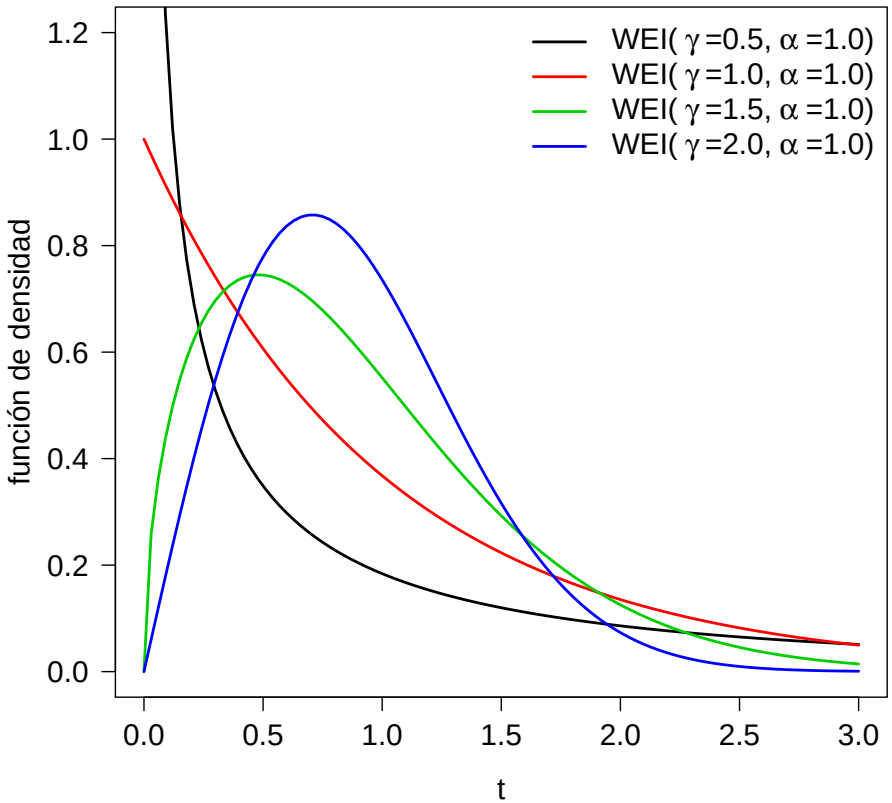
La función de riesgo y de riesgo acumulada no están implementadas en R, pero en base a las funciones anteriores de pueden implementar fácilmente

```
> hweibull<-function(x, shape, scale=1)
   {
       exp(dweibull(x, shape, scale=1, log=TRUE) -
          pweibull(x, shape, scale=1, log=TRUE))
   }

> Hweibull<-function(x, shape, scale=1)
  {-pweibull(x, shape, scale, lower.tail=FALSE, log=TRUE)}
```

Modelos paramétricos

Para graficar la función de densidad del modelo WEI, basta usar

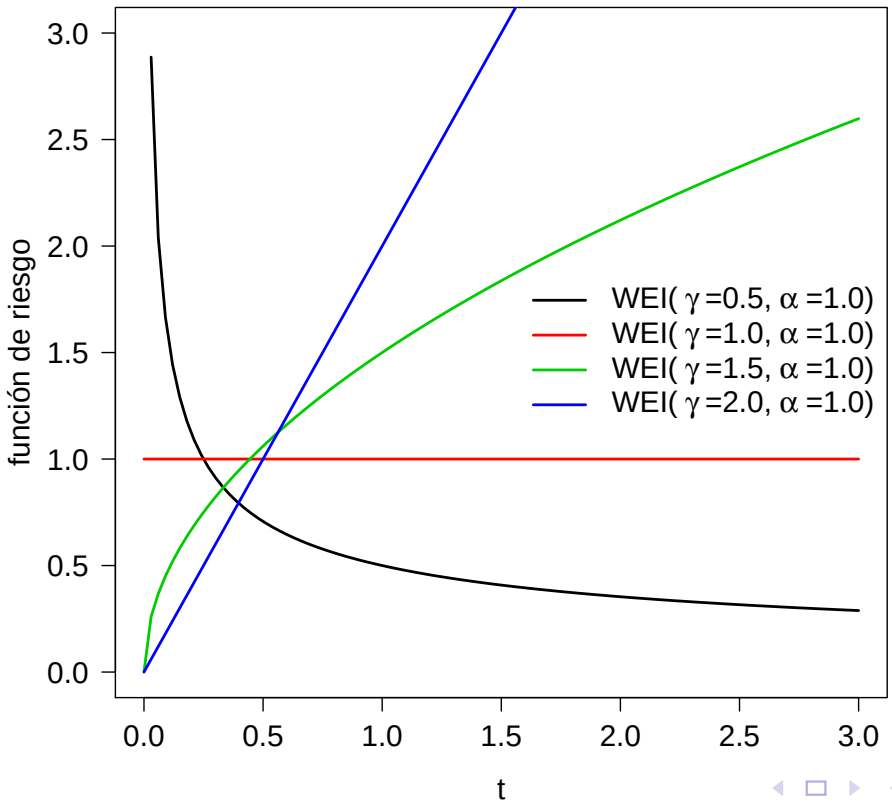
```
> curve(dweibull(x, shape=0.5, scale=1.0),...)
```



Modelos paramétricos

Análogamente, para graficar la función de riesgo del modelo WEI, basta usar

```
curve(hweibull(x, shape=0.5, scale=1.0),...)
```



Modelos paramétricos

Distribución log-normal

Modelos paramétricos

- Distribución log-normal

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

$$S(t;\boldsymbol{\theta}) = 1-\Phi\left(\frac{\log(t)-\mu}{\sigma}\right),$$

$$\lambda(t;\boldsymbol{\theta}) = \frac{f(t;\boldsymbol{\theta})}{S(t;\boldsymbol{\theta})}, \qquad t,\sigma>0, \mu\in\mathbb{R}$$

  en que $\boldsymbol{\theta} = (\mu, \sigma)$.

- σ es un parámetro de escala y μ es un parámetro de localización.
- Función de riesgo es no monótona: es creciente hasta alcanzar un valor máximo y luego es decreciente.
- $Y = \log(T) \sim N(\mu, \sigma^2)$.

Modelos paramétricos

Si $T \sim \text{LN}(\mu, \sigma)$

- $\mathbb{E}(T) = \exp(\mu + \sigma^2/2)$;
- $\text{Var}(T) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$;
- $t_q = \exp(\mu + z_q\sigma)$ es el percentil $100 \times q$ de la distribución, con $q \in (0, 1)$.

Modelos paramétricos

En R, para el modelo LN las siguientes funciones estás implementadas en el paquete base de R

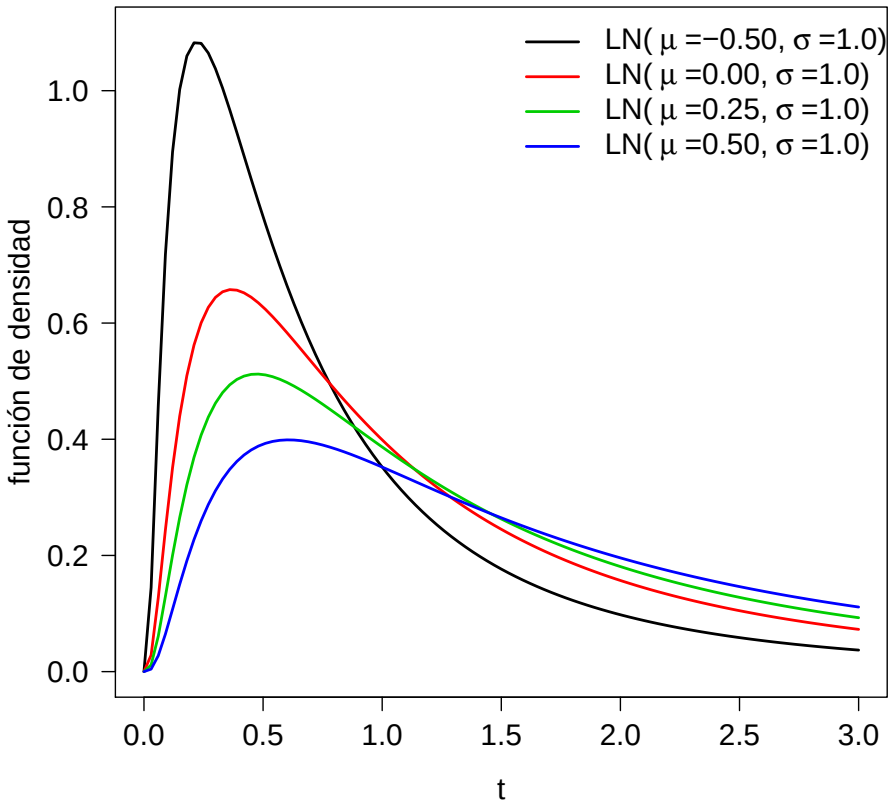
```
> dlnorm(x, meanlog = 0, sdlog = 1, log = FALSE)
> plnorm(q, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)
> qlnorm(p, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)
> rlnorm(n, meanlog = 0, sdlog = 1)
```

que corresponden a la función de densidad (`dlnorm`), distribución acumulada (`plnorm`), cuantil (`qlnorm`) y un generador de números aleatorios (`rlnorm`), respectivamente.

Modelos paramétricos

Para graficar la función de densidad del modelo LN, basta usar

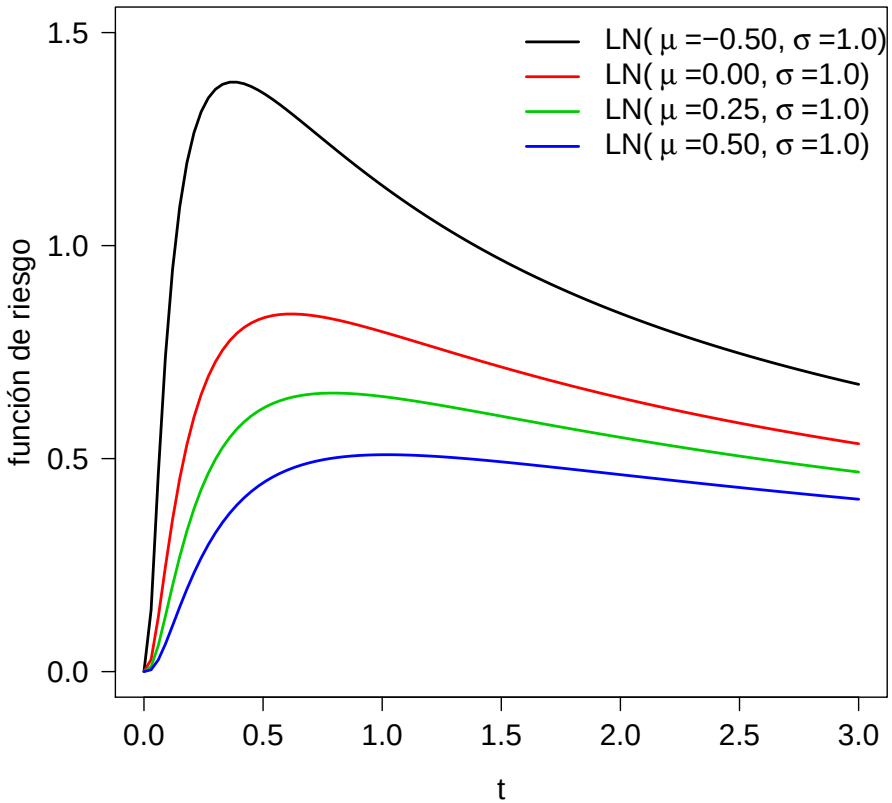
```
> curve(dlnorm(x, meanlog=-0.5, sdlog=1.0),...)
```



Modelos paramétricos

Para graficar la función de densidad del modelo LN, basta usar

```
> curve(hlnorm(x, meanlog=-0.5, sdlog=1.0),...)
```



Distribución log-logística

Modelos paramétricos

- Distribución log-logística

$$f(t;\boldsymbol{\theta}) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left(1 + \left(\frac{t}{\alpha}\right)^\gamma\right)^{-2},$$

$$S(t;\boldsymbol{\theta}) = \frac{1}{1 + (t/\alpha)^\gamma},$$

$$\lambda(t;\boldsymbol{\theta}) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha\left[1 + (t/\alpha)^\gamma\right]}, \qquad t, \gamma, \alpha > 0.$$

  en que $\boldsymbol{\theta} = (\alpha, \gamma)$. Denotaremos como LL$(\gamma, \alpha)$.

- α es un parámetro de escala y γ es un parámetro de forma.
- Función de riesgo es no monótona para $\gamma > 1$: es creciente hasta alcanzar un valor máximo y luego es decreciente; monótona decreciente para $\gamma \leq 1$.

Modelos paramétricos

Si $T \sim \mathsf{LL}(\gamma, \alpha)$

- $\mathbb{E}(T) = \dfrac{\pi\alpha}{\gamma \sin(\pi/\gamma)}$;
- $\mathrm{Var}(T) = \dfrac{2\pi\alpha^2}{\gamma \sin(2\pi/\gamma)} - \mathbb{E}^2(T)$;
- $t_q = \alpha \left[\frac{q}{1-q}\right]^{1/\gamma}$ es el percentil $100 \times q$ de la distribución, con $q \in (0, 1)$.

Modelos paramétricos

Si $T \sim \text{LL}(\gamma, \alpha)$, entonces $Y = \log(T) \sim L(\mu, \sigma)$ (distribución logística)

$$f(y; \mu, \sigma) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\right\} \left(1 + \exp\left\{\frac{y-\mu}{\sigma}\right\}\right)^{-2},$$

$$S(y; \mu, \sigma) = \left(1 + \exp\left\{\frac{y-\mu}{\sigma}\right\}\right)^{-1}, \quad y, \mu \in \mathbb{R}, \sigma > 0,$$

en que $\mu = \log(\alpha)$ y $\sigma = 1/\gamma$ son parámetros de localización y escala, respectivamente.

Modelos paramétricos

En R, para el modelo LL las siguientes funciones estás implementadas en el paquete `actuar` de R

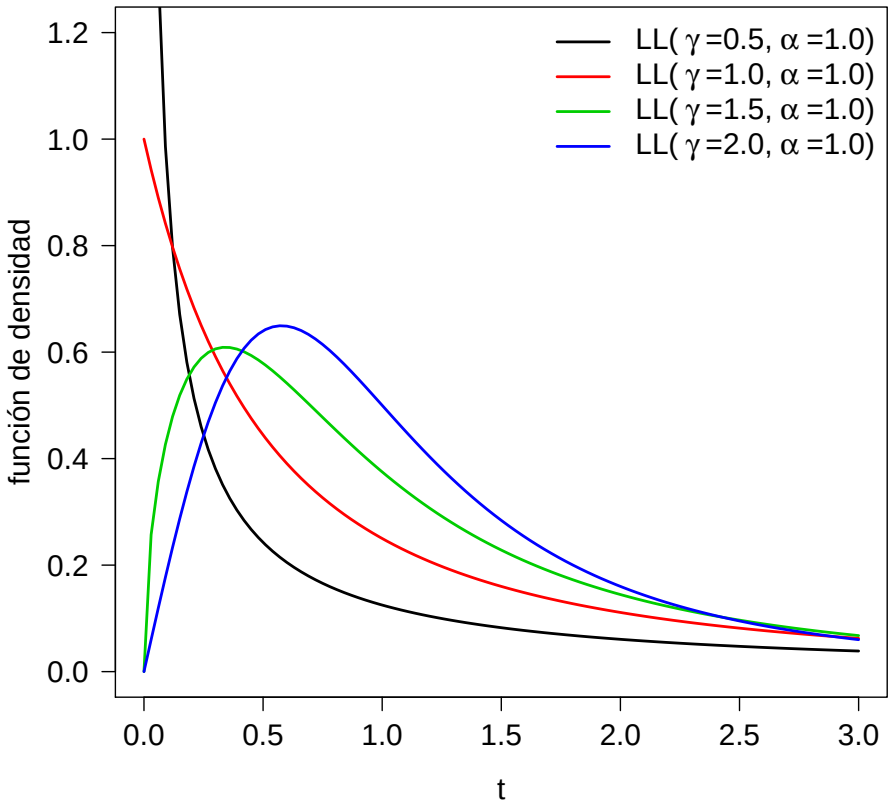
```
> dllogis(x, shape, rate = 1, scale = 1/rate, log = FALSE)
> pllogis(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)
> qllogis(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)
> rllogis(n, shape, rate = 1, scale = 1/rate)
```

que corresponden a la función de densidad (`dllogis`), distribución acumulada (`pllogis`), cuantil (`qllogis`) y un generador de números aleatorios (`rllogis`), respectivamente.

Modelos paramétricos

Para graficar la función de densidad del modelo LN, basta usar

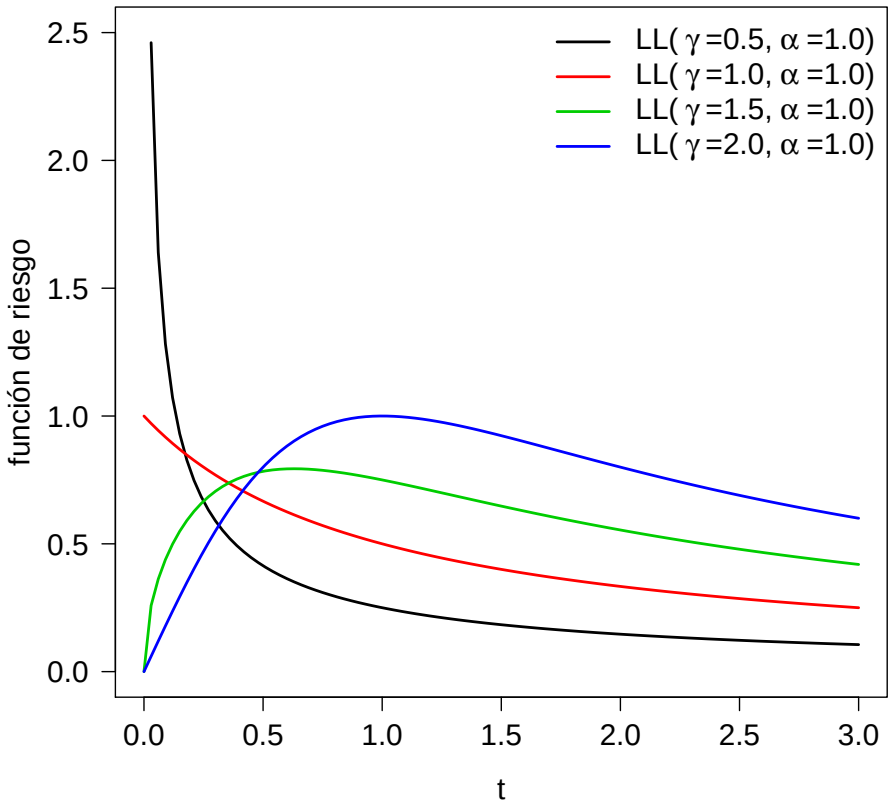
```
> curve(dllogis(x, shape=0.5, scale=1.0), ...)
```



Modelos paramétricos

Para graficar la función de densidad del modelo LN, basta usar

```
> curve(hllogis(x, shape=0.5, scale=1.0),...)
```



Modelos paramétricos

¿Cómo decidir qué modelo usar? (Antes de ajustarlo)

Modelos paramétricos

Note que para el modelo WEI(γ, α), tenemos que

$$S(t;\boldsymbol{\theta}) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}.$$

Es decir,

$$\begin{aligned}\log\left(-\log S(t;\boldsymbol{\theta})\right) &= \gamma\left(\log t - \log\alpha\right)\\ &= b_0^* + b_1^* \log t,\end{aligned}$$

en que $b_0^* = -\gamma\log\alpha$ y $b_1^* = \gamma$.

Modelos paramétricos

Note que para el modelo WEI(γ, α), tenemos que

$$S(t;\boldsymbol{\theta}) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}.$$

Es decir,

$$\begin{aligned}\log\left(-\log S(t;\boldsymbol{\theta})\right) &= \gamma\left(\log t - \log\alpha\right)\\ &= b_0^* + b_1^* \log t,\end{aligned}$$

en que $b_0^* = -\gamma\log\alpha$ y $b_1^* = \gamma$. Por lo tanto, en poblaciones homogéneas, se sugiere construir el gráfico de $\log t$ versus $\log\left(-\log\left(\widehat{S}_{KM}(t)\right)\right)$. Si los puntos de ese gráfico se aproximan de una recta, entonces el modelo Weibull es apropiado para esos datos.

Modelos paramétricos

Para el modelo LN(μ, σ), tenemos que

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

Es decir,

$$\begin{aligned}\Phi^{-1}\left(S(t;\boldsymbol{\theta})\right) &= -\frac{(\log(t)-\mu)}{\sigma} \\ &= b_0^* + b_1^* \log t,\end{aligned}$$

en que $b_0^* = \mu/\sigma$ y $b_1^* = -1/\sigma$.

Modelos paramétricos

Para el modelo LN(μ, σ), tenemos que

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

Es decir,

$$\begin{aligned}\Phi^{-1}\left(S(t;\boldsymbol{\theta})\right) &= -\frac{(\log(t)-\mu)}{\sigma} \\ &= b_0^* + b_1^* \log t,\end{aligned}$$

en que $b_0^* = \mu/\sigma$ y $b_1^* = -1/\sigma$. Por lo tanto, en poblaciones homogéneas, se sugiere construir el gráfico de $\log t$ versus $\Phi^{-1}\left(\widehat{S}_{KM}(t)\right)$. Si los puntos de ese gráfico se aproximan de una recta, entonces el modelo log-normal es apropiado para esos datos.

Modelos paramétricos

Finalmente, para el modelo LL(γ, α), tenemos que

$$S(t;\boldsymbol{\theta}) = \frac{1}{1+(t/\alpha)^{\gamma}}.$$

Es decir,

$$\begin{aligned}\log\left(\frac{1}{S(t;\boldsymbol{\theta})}-1\right) &= \gamma\left(\log t - \log\alpha\right)\\ &= b_0^* + b_1^*\log t,\end{aligned}$$

en que $b_0^* = -\gamma\log(\alpha)$ y $b_1^* = \gamma$.

Modelos paramétricos

Finalmente, para el modelo LL(γ, α), tenemos que

$$S(t;\boldsymbol{\theta}) = \frac{1}{1+(t/\alpha)^{\gamma}}.$$

Es decir,

$$\begin{aligned}\log\left(\frac{1}{S(t;\boldsymbol{\theta})}-1\right) &= \gamma\left(\log t - \log\alpha\right)\\ &= b_0^* + b_1^*\log t,\end{aligned}$$

en que $b_0^* = -\gamma\log(\alpha)$ y $b_1^* = \gamma$. Por lo tanto, en poblaciones homogéneas, se sugiere construir el gráfico de $\log t$ versus $\log\left(\frac{1}{\widehat{S}_{KM}(t)}-1\right)$. Si los puntos de ese gráfico se aproximan de una recta, entonces el modelo log-logístico es apropiado para esos datos.

Modelos paramétricos

Para el ejemplo de cáncer de pulmón tenemos que

```
##Calcula el estimador de KM
> KM.0=survfit(Surv(t2, d2) ~ 1, conf.type="log-log")
```

Modelos paramétricos

Para el ejemplo de cáncer de pulmón tenemos que

```
##Calcula el estimador de KM
> KM.0=survfit(Surv(t2, d2) ~ 1, conf.type="log-log")

##Captura los valores de los tiempos de falla
> t0=KM.0$time[which(KM.0$surv>0 & KM.0$surv<1)]
```

Modelos paramétricos

Para el ejemplo de cáncer de pulmón tenemos que

```
##Calcula el estimador de KM
> KM.0=survfit(Surv(t2, d2) ~ 1, conf.type="log-log")

##Captura los valores de los tiempos de falla
> t0=KM.0$time[which(KM.0$surv>0 & KM.0$surv<1)]

##Captura la sobrev. evaluada en los tiempos de falla
> S0=KM.0$surv[which(KM.0$surv>0 & KM.0$surv<1)]
```

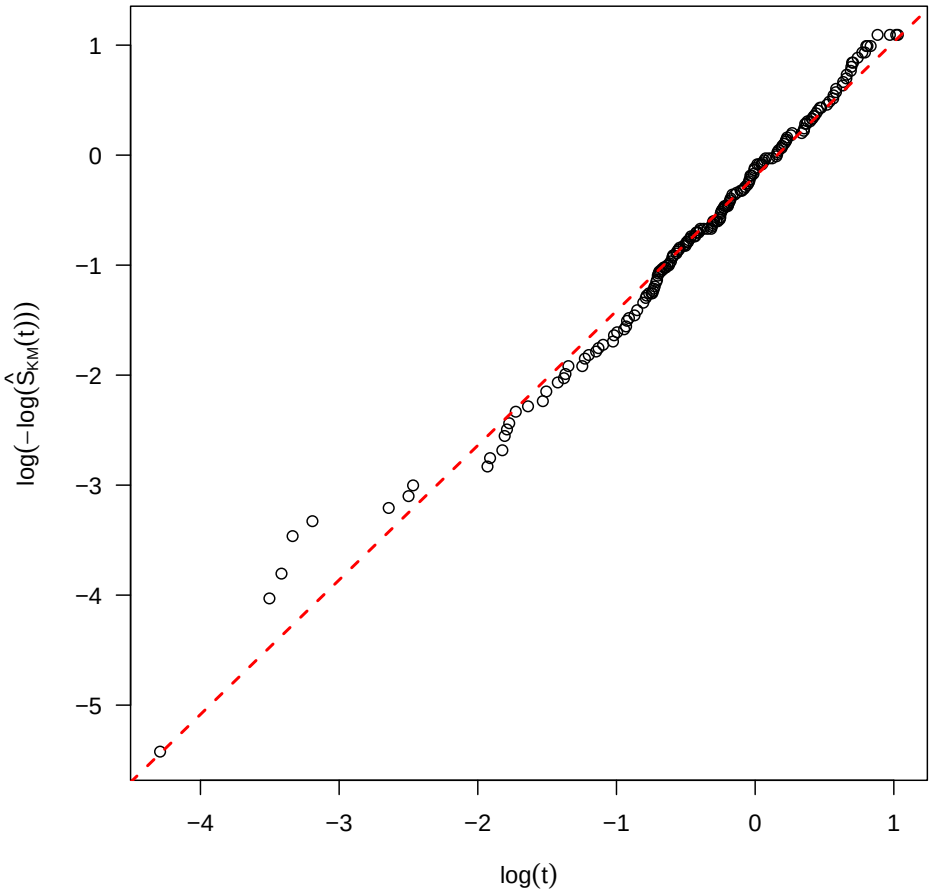
Modelos paramétricos

Para el ejemplo de cáncer de pulmón tenemos que

```
##Calcula el estimador de KM
> KM.0=survfit(Surv(t2, d2) ~ 1, conf.type="log-log")

##Captura los valores de los tiempos de falla
> t0=KM.0$time[which(KM.0$surv>0 & KM.0$surv<1)]

##Captura la sobrev. evaluada en los tiempos de falla
> S0=KM.0$surv[which(KM.0$surv>0 & KM.0$surv<1)]

##Hacemos un modelo de regresión entre log(-log(S0)) y log(t0)
> m.w<-lm(log(-log(S0))~log(t0))
```

Modelos paramétricos

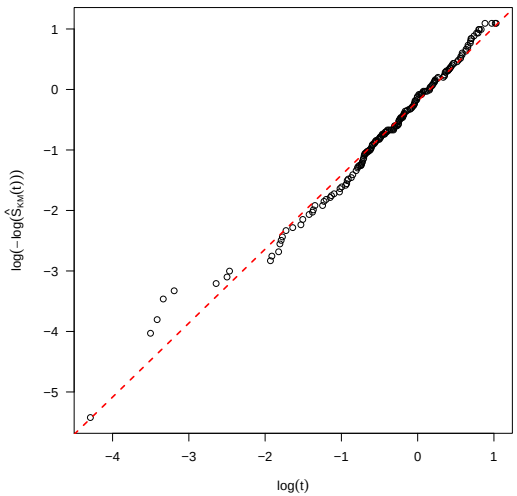
```
> plot(log(t0), log(-log(S0)), ...)
> abline(coef(m.w)[1],coef(m.w)[2], ...)
```



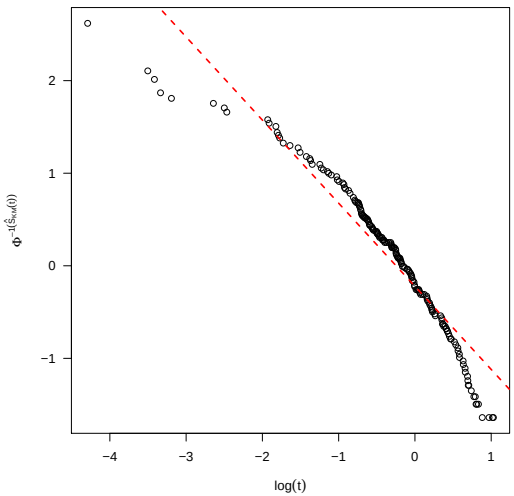
$R_a^2 = 0.984$

Modelos paramétricos

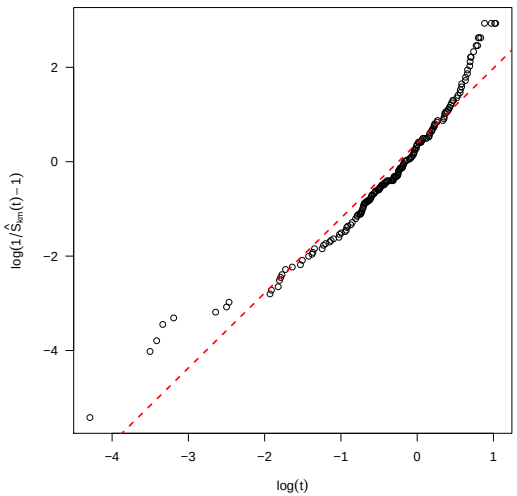
Para el conjunto de datos cáncer de pulmón (sin considerar covariables)



WEI ($R_a^2 = 0.984$)



LN ($R_a^2 = 0.913$)



LL ($R_a^2 = 0.934$)

Modelos paramétricos

¿Y si tengo covariables?

Modelos paramétricos

¿Y si tengo covariables?

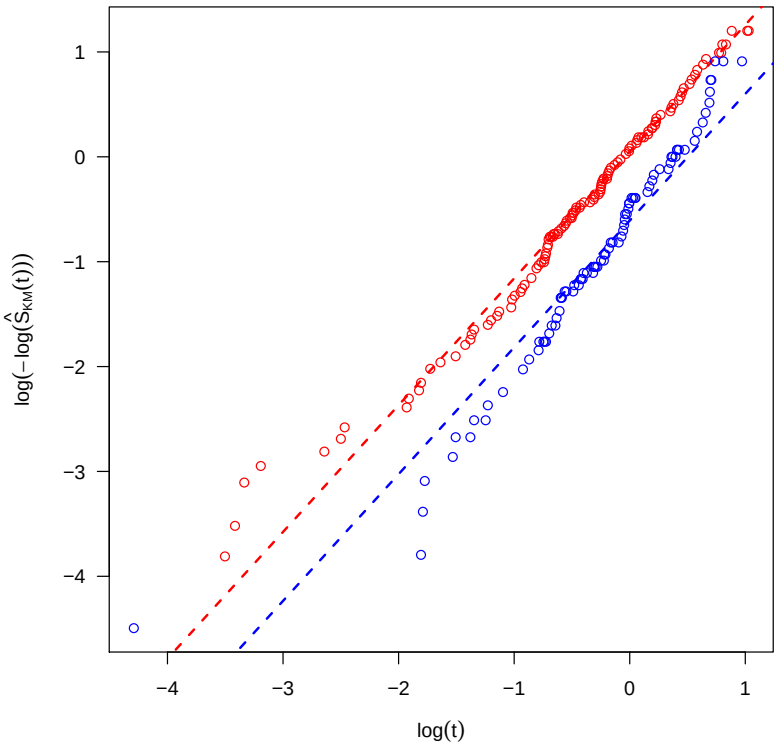
Repita el proceso anterior para cada factor de la(s) covariable(s).

Modelos paramétricos

```
> hombres<-which(sex2==0); mujeres<-which(sex2==1)
> KM.h=survfit(Surv(t2[hombres], d2[hombres]) ~ sex2[hombres])
> KM.m=survfit(Surv(t2[mujeres], d2[mujeres]) ~ sex2[mujeres])

> th<-summary(KM.h)$time; Sh<-summary(KM.h)$surv
> tm<-summary(KM.m)$time; Sm<-summary(KM.m)$surv

> t0<-c(th,tm); S0<-c(Sh,Sm); sex.f<-c(rep(0,length(th)),rep(1,length(tm)))
> m.w<-lm(log(-log(S0))~log(t0)+sex.f)
```

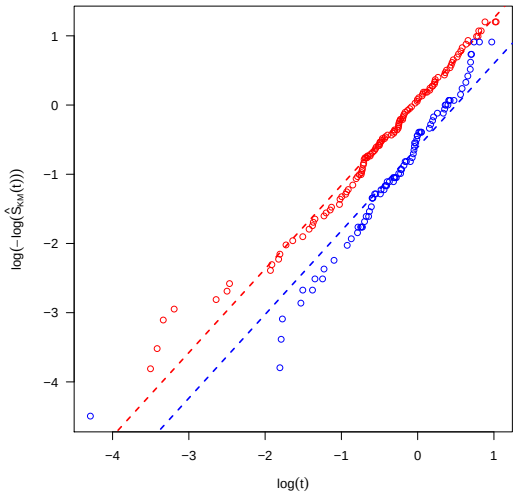
Modelos paramétricos

```
> plot(log(t0), log(-log(S0)), ...)
> points(log(th), log(-log(Sh)), ...)
> abline(coef(m.w)[1],coef(m.w)[2], ...)
> points(log(tm), log(-log(Sm)), ...)
> abline(coef(m.w)[1]+coef(m.w)[3], coef(m.w)[2], ...)
```

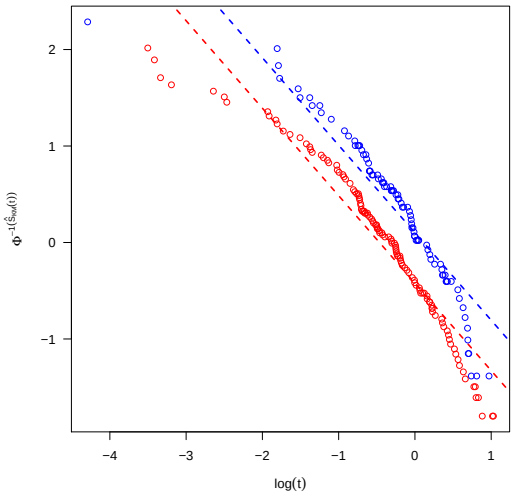


Modelos paramétricos

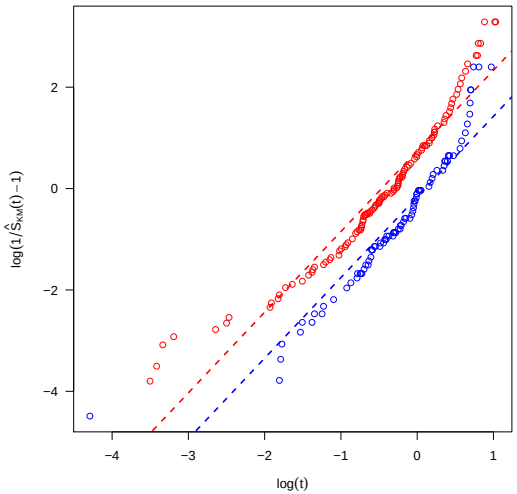
Para el conjunto de datos de cáncer de pulmón (considerando la covariable sexo)



WEI ($R_a^2 = 0.958$)



LN ($R_a^2 = 0.899$)



LL ($R_a^2 = 0.915$)

Referencias Bibliográficas

Referencias

Referencias Bibliográficas

- Aalen, O.O. (1978). Nonparametric Inference for a Family of Counting Processes. Annals of Statistics, **6**, 701-726.
- Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika, **57**, 579-594.
- Dutang, C., Goulet, V., Pigeon, M. (2008). actuar: An R Package for Actuarial Science. Journal of Statistical Software, **25**, 1-37.
- Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika, **52**, 203-224.
- Harrington, D.P., Fleming, T.R. (1982). A class of rank test procedures for censored survival data. Biometrika, **69**, 553-66.
- Hosmer, D.W., Lemeshow, S. (1999). Applied Survival Analysis. John Wiley and Sons, New York.

Referencias Bibliográficas

- Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). Implementing a class of permutation tests: The coin package. Journal of Statistical Software, **28**, 1-23.
- Kalbfleish, J.D, Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York.
- Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc., **53**, 457-481.
- Klein, J.P., Moeschberger, M.L. (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag, New York.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports. **50**, 163-70.

Referencias Bibliográficas

- Millard SP (2013). EnvStats: An R Package for Environmental Statistics. Springer, New York. ISBN 978-1-4614-8455-4. URL: https://www.springer.com.
- Nelson, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. Technometrics, **14**, 945-965.
- Peto R., Peto J. (1972) Asymptotically efficient rank invariant test procedures. J R Stat Soc A, **135**, 185-198.
- Prentice, R.L., Marek, P. (1979). A qualitative discrepancy between censored data rank tests. Biometrics **35**, 861-867.
- Tarone, R.E., Ware, J. (1977) On Distribution-Free Tests for Equality of Survival Distributions. Biometrika, **64**, 156-160.
- Therneau T (2020). A Package for Survival Analysis in R. R package version 3.1-12. URL: https://CRAN.R-project.org/package=survival.