Análisis de Supervivencia

Estimación no paramétrica

Sergio M. Nava Muñoz

2025-06-01

Table of contents

1 Estimación no paramétrica

1.1 Temario de la Sesión

- **Fundamentos:** ¿Qué es el análisis de supervivencia y cómo se estructuran los datos (tiempo, evento y censura)?
- **El Estimador Kaplan-Meier:** Introducción al método no paramétrico fundamental para estimar la función de supervivencia cuando hay datos censurados.
- **Cálculo e Interpretación:** Un ejemplo paso a paso para calcular e interpretar una curva de Kaplan-Meier.
- **Comparación entre Grupos:** Uso de la prueba Log-Rank para determinar si existen diferencias significativas entre las curvas de supervivencia.
- **Aplicación Práctica en R:** Implementación de estas técnicas utilizando paquetes como `survival` y `survminer`.

1.2 La función de distribución acumulada empírica (FDAE)

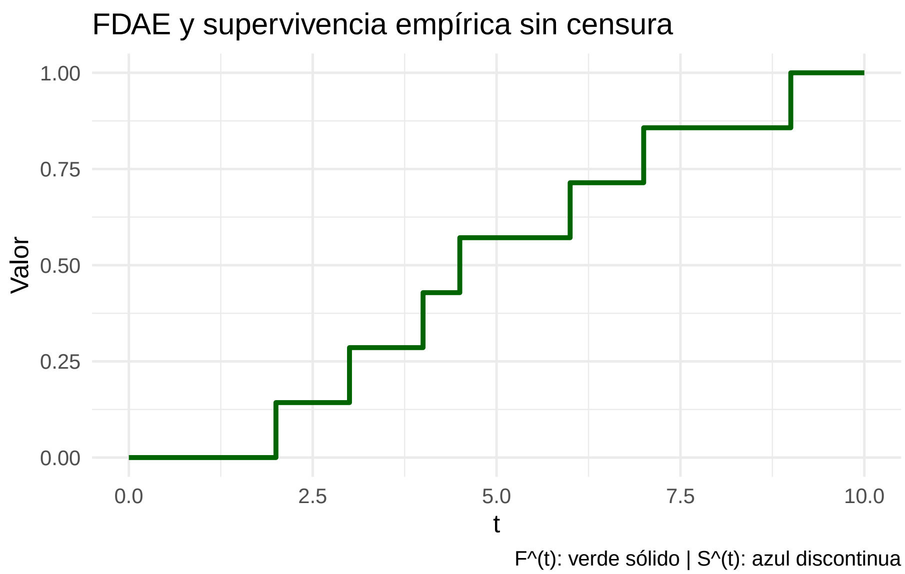
Dada una muestra de tiempos de falla sin censura:

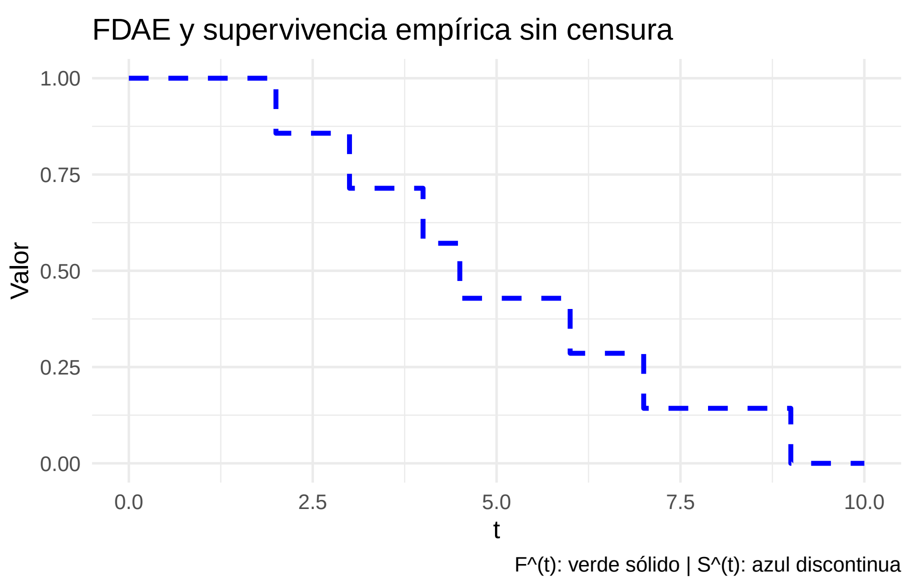
$$\hat{F}(t) = \frac{\#\{T_i \leq t\}}{n}$$

Es un estimador escalonado, que da saltos en cada observación.
La función de supervivencia empírica se define como:

$$\hat{S}(t) = 1 - \hat{F}(t)$$

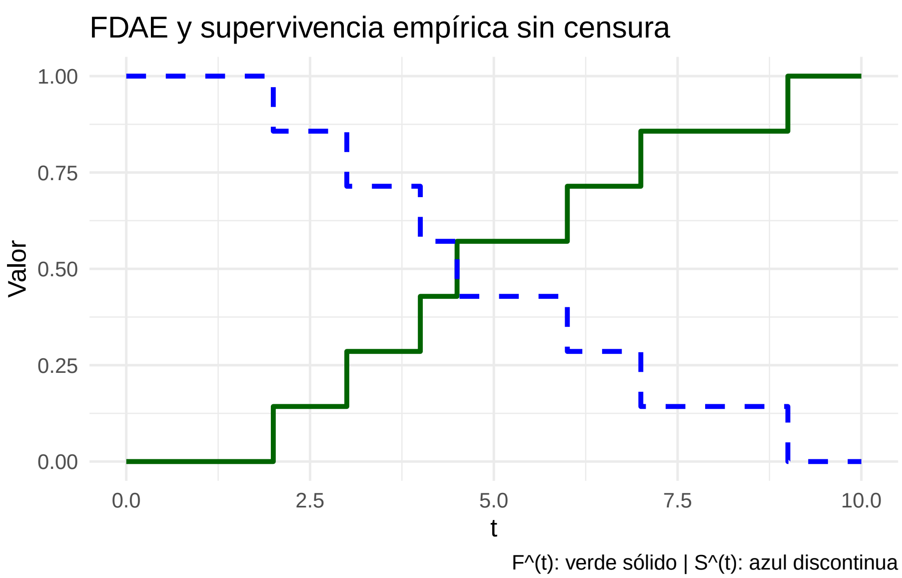
Limitación: no puede manejar adecuadamente datos censurados.





1.3 Ejemplo en R: FDAE

t	F_hat	S_hat
0.0	0.0000000	1.0000000
2.0	0.1428571	0.8571429
3.0	0.2857143	0.7142857
4.0	0.4285714	0.5714286
4.5	0.5714286	0.4285714
6.0	0.7142857	0.2857143
7.0	0.8571429	0.1428571
9.0	1.0000000	0.0000000
10.0	1.0000000	0.0000000



1.4 Estimador de Kaplan-Meier

Cuando hay censura, la FDAE no es válida. Kaplan-Meier estima la función de supervivencia como:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{m_i}{n_i}\right)$$

donde:

- m_i: número de eventos en el tiempo t_i
- n_i: número de individuos en riesgo justo antes de t_i

Es un estimador escalonado que **ajusta el denominador** cuando hay censura.

Ejemplo

Table 2: Comparación entre FDAE, Supervivencia Empírica y Kaplan-Meier

tiempo	status	FDAE	S_empirica	Kaplan_Meier
2.0	1	0.1667	0.8333	0.8750
3.0	1	0.3333	0.6667	0.7500
4.0	1	0.5000	0.5000	0.6250
4.5	0	0.5000	0.5000	0.6250
6.0	1	0.6667	0.3333	0.4688
7.0	1	0.8333	0.1667	0.3125
9.0	0	0.8333	0.1667	0.3125
10.0	1	1.0000	0.0000	0.0000

2 Cálculo e interpretación de KM

2.1 Esquema General de Datos

Table 3: Esquema General de Datos con Subíndices

No. Indiv.	t	D	X_1	X_2	...	X_p
1	t_1	D_1	X_{11}	X_{12}	$\cdots$	X_{1p}
2	t_2	D_2	X_{21}	X_{22}	$\cdots$	X_{2p}
...	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	t_n	D_n	X_{n1}	X_{n2}	$\cdots$	X_{np}

Table 4: Disposición alternativa de los datos ordenados

Tiempos de fallo ordenados $t_{(f)}$	Núm. de fallos m_f	Censurados en $[t_{(f)}, t_{(f+1)})$, q_f	Conjunto de riesgo $R(t_{(f)})$
$t_{(0)}$	m_0	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

Disposición alternativa de los datos ordenados

Una disposición alternativa de los datos se muestra a continuación.
Esta organización es la base sobre la cual se derivan las curvas de supervivencia de Kaplan-Meier.

- La primera columna de la tabla presenta los tiempos de supervivencia ordenados de menor a mayor. $t_{(f)}$
- La segunda columna muestra el conteo de fallos en cada uno de los tiempos de fallo distintos. m_f
- La tercera columna presenta los conteos de censura, denotados por q_f, correspondientes a las personas censuradas en el intervalo de tiempo que inicia en el tiempo de fallo $t_{(f)}$ y termina justo antes del siguiente tiempo de fallo, $t_{(f+1)}$. q_f
- La última columna muestra el conjunto de riesgo, que representa el grupo de individuos que han sobrevivido al menos hasta el tiempo $t_{(f)}$. $R(t_{(f)})$

Ejemplo: Tiempos de remisión (semanas) para dos grupos de pacientes con leucemia

Grupo 1 ($n = 21$) — *Tratamiento*
6, 6, 6, 7, 10,
13, 16, 22, 23,
6^+, 9^+, 10^+, 11^+,
17^+, 19^+, 20^+,
25^+, 32^+, 32^+,
34^+, 35^+
Grupo 2 ($n = 21$) — *Placebo*
1, 1, 2, 2, 3,
4, 4, 5, 5,
8, 8, 8, 8,
11, 11, 12, 13,
15, 17, 22, 23

Nota: el símbolo $^+$ denota observaciones censuradas.

Grupo	# Fallos	# Censurados	Total
Grupo 1	9	12	21
Grupo 2	21	0	21

Estadísticos descriptivos:

- $\bar{T}_1$ (ignorando censuras): 17.1
- $\bar{T}_2$: 8.6

Table 6: Grupo 1 (tratamiento): Tiempos de fallo ordenados

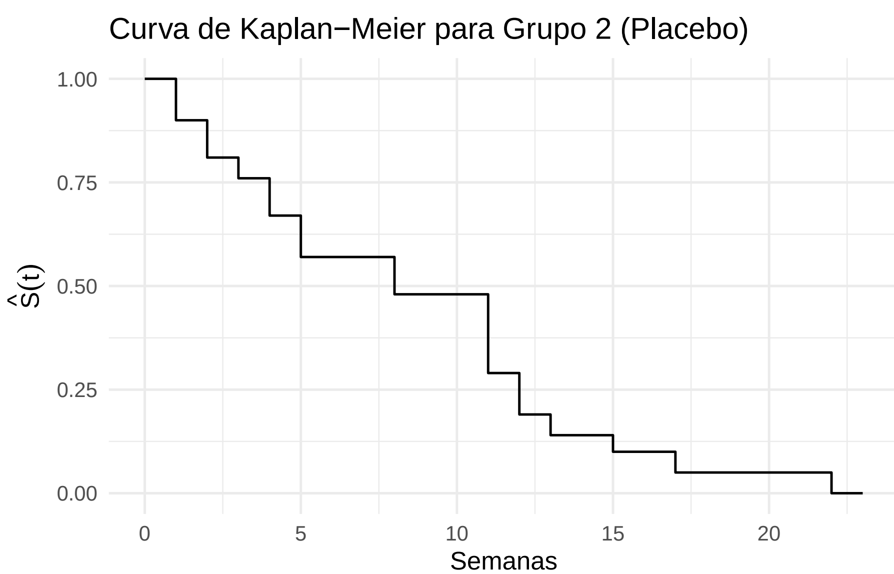
$t_{(f)}$	n_f	m_f	q_f
0	21	0	0
6	21	3	1
7	18	1	1
10	17	1	2
13	15	1	0
16	11	1	3
22	7	1	0
23	2	1	5
>23	—	—	—

Table 7: Grupo 2 (placebo): Tiempos de fallo ordenados

$t_{(f)}$	n_f	m_f	q_f
0	21	0	0
1	21	2	0
2	19	2	0
3	17	1	0
4	16	2	0
5	14	2	0
8	12	4	0
11	8	2	0
12	6	2	0
13	4	1	0
15	3	1	0
17	2	1	0
22	1	1	0
23	1	1	0

Table 8: Grupo 2 (placebo): Estimación de la función de supervivencia empírica (Kaplan-Meier)

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	1.00
1	21	2	0	0.90
2	19	2	0	0.81
3	17	1	0	0.76
4	16	2	0	0.67
5	14	2	0	0.57
8	12	4	0	0.48
11	8	2	0	0.29
12	6	2	0	0.19
13	4	1	0	0.14
15	3	1	0	0.10
17	2	1	0	0.05
22	1	1	0	0.00
23	1	1	0	0.00



Interpretación

- $\hat{S}(t_{(f)}) = \dfrac{\text{Número de sujetos sobrevivientes después de } t_{(f)}}{21}$
- No hay censura en el Grupo 2.

- Se utilizó el método de Kaplan-Meier para estimar la función de supervivencia.

Ejemplo: Cálculo de la función de supervivencia empírica

Table 9: Grupo 2 (placebo): Estimación de la función de supervivencia empírica (Kaplan-Meier)

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	1.00
1	21	2	0	0.90
2	19	2	0	0.81
3	17	1	0	0.76
4	16	2	0	0.67
5	14	2	0	0.57
8	12	4	0	0.48
11	8	2	0	0.29
12	6	2	0	0.19
13	4	1	0	0.14
15	3	1	0	0.10
17	2	1	0	0.05
22	1	1	0	0.00
23	1	1	0	0.00

Sea $\hat{S}(4)$ la probabilidad estimada de supervivencia más allá de la semana 4:

$$\hat{S}(4) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16} = \frac{14}{21} = 0.67$$

Esto equivale a:

- $\Pr(T > t_{(0)}) = \frac{21}{21} = 1$
- $\Pr(T > t_{(1)} \mid T \geq t_{(1)}) = \frac{19}{21}$
- $\Pr(T > t_{(2)} \mid T \geq t_{(2)}) = \frac{19}{19}$
- $\Pr(T > t_{(3)} \mid T \geq t_{(3)}) = \frac{16}{17}$
- $\Pr(T > t_{(4)} \mid T \geq t_{(4)}) = \frac{14}{16}$

Donde 16 es el número de individuos en riesgo en la semana 4.
Para $t = 8$:

$$\hat{S}(8) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16} \times \frac{12}{14} \times \frac{8}{12} = \frac{8}{21}$$

Fórmula KM:

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{m_j}{n_j}\right)$$

donde m_j es el número de eventos (fallos) en $t_{(j)}$ y n_j el número en riesgo.

Table 10: Grupo 1 (tratamiento): Estimación paso a paso de la función de supervivencia KM

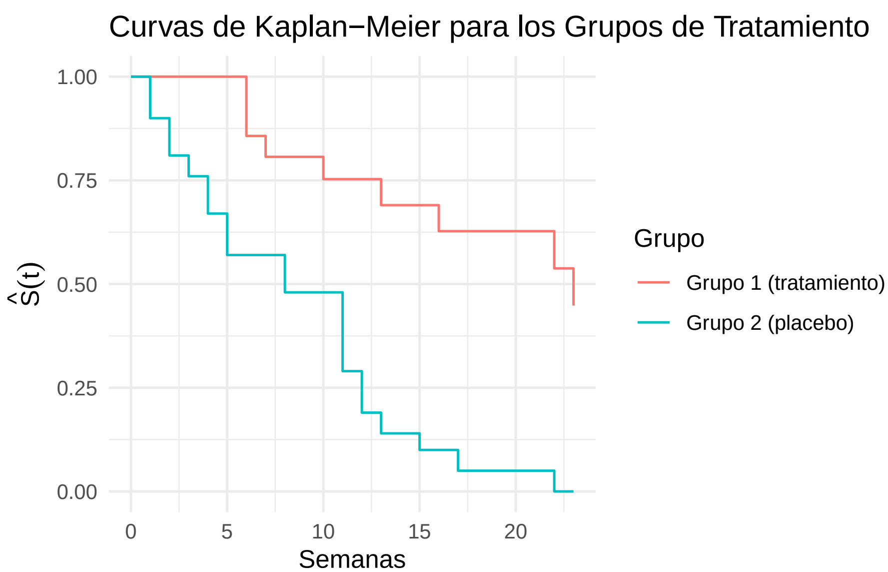
$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	1
6	21	3	1	$18/21 = 0.8571$
7	17	1	1	$0.8571 \times 16/17 = 0.8067$
10	15	1	2	$0.8067 \times 14/15 = 0.7529$
13	12	1	1	$0.7529 \times 11/12 = 0.6902$
16	11	1	2	$0.6902 \times 10/11 = 0.6275$
22	7	1	1	$0.6275 \times 6/7 = 0.5378$
23	6	1	1	$0.5378 \times 5/6 = 0.4482$

Cálculo de otras estimaciones de supervivencia

Las demás estimaciones de supervivencia se calculan multiplicando la estimación en el tiempo de fallo inmediatamente anterior por una fracción. Por ejemplo:

- La fracción es $\frac{18}{21}$ para sobrevivir más allá de la semana 6, porque 21 sujetos permanecen hasta la semana 6 y 3 de ellos no sobreviven más allá de esa semana.
- La fracción es $\frac{16}{17}$ para sobrevivir más allá de la semana 7, ya que 17 personas permanecen hasta la semana 7 y 1 de ellas no sobrevive más allá de esa semana.

Las demás fracciones se calculan de manera similar.



2.2 III. Características Generales de las Curvas de Kaplan-Meier

2.2.1 Fórmula general de KM

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times \Pr(T > t_{(f)} \mid T \geq t_{(f)})$$

2.2.2 Fórmula producto-límite (KM)

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \Pr(T > t_{(i)} \mid T \geq t_{(i)})$$

2.2.3 Ejemplo

Table 11: Grupo 1 (tratamiento): Estimación paso a paso de la función de supervivencia KM

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	1
6	21	3	1	$18/21 = 0.8571$
7	17	1	1	$0.8571 \times 16/17 = 0.8067$
10	15	1	2	$0.8067 \times 14/15 = 0.7529$
13	12	1	1	$0.7529 \times 11/12 = 0.6902$
16	11	1	2	$0.6902 \times 10/11 = 0.6275$

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
22	7	1	1	0.6275 × 6/7 = 0.5378
23	6	1	1	0.5378 × 5/6 = 0.4482

2.2.3.1 Para $t = 10$:

$$\hat{S}(10) = 0.8067 \times \frac{14}{15} = 0.7529$$

También se puede expresar como:

$$\hat{S}(10) = \frac{18}{21} \times \frac{16}{17} \times \frac{14}{15}$$

2.2.3.2 Para $t = 16$:

$$\hat{S}(16) = 0.6902 \times \frac{10}{11} = 0.6274$$

O bien:

$$\hat{S}(16) = \frac{18}{21} \times \frac{16}{17} \times \frac{14}{15} \times \frac{11}{12} \times \frac{10}{11}$$

2.3 Justificación Matemática de la Fórmula KM

Sea:

- $A = \{T \geq t_{(f)}\}$
- $B = \{T > t_{(f)}\}$

Entonces:

$$\Pr(A \cap B) = \Pr(B) = \hat{S}(t_{(f)})$$

Dado que no hay fallos en $t_{(f-1)} < T < t_{(f)}$:

$$\Pr(A) = \Pr(T \geq t_{(f-1)}) = \hat{S}(t_{(f-1)})$$

Y por la regla de la probabilidad condicional:

$$\Pr(B \mid A) = \Pr(T > t_{(f)} \mid T \geq t_{(f)})$$

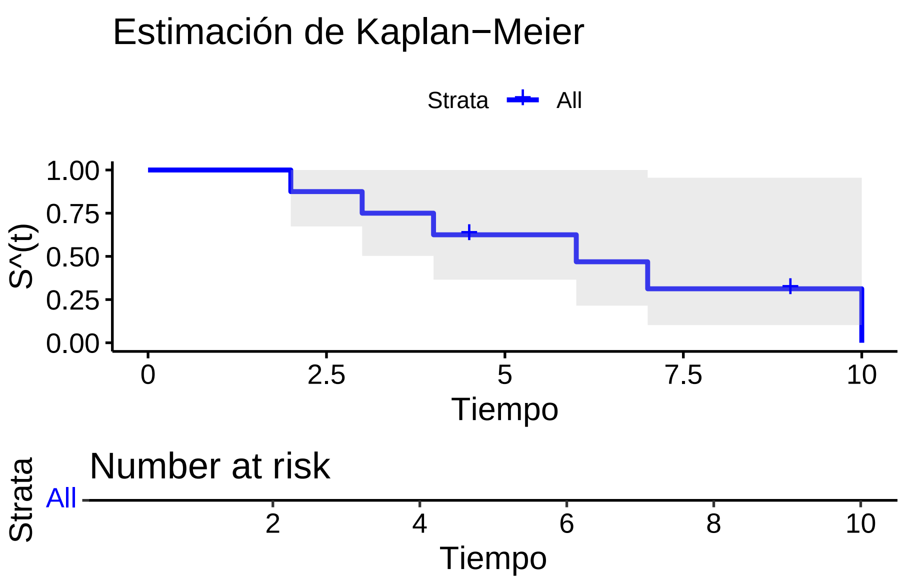
Por lo tanto, usando $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B \mid A)$:

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \cdot \Pr(T > t_{(f)} \mid T \geq t_{(f)})$$

2.4 Ejemplo en R: Kaplan-Meier

Table 12: Tabla de tiempos y estatus de censura

ID	tiempo	evento
Ind 1	2.0	1
Ind 2	3.0	1
Ind 3	4.0	1
Ind 4	4.5	0
Ind 5	6.0	1
Ind 6	7.0	1
Ind 7	9.0	0
Ind 8	10.0	1




```
Call: survfit(formula = surv_obj ~ 1, data = datos)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2      8       1    0.875   0.117        0.673        1.000
    3      7       1    0.750   0.153        0.503        1.000
    4      6       1    0.625   0.171        0.365        1.000
    6      4       1    0.469   0.187        0.215        1.000
```

```
   7       3         1     0.312    0.178           0.102           0.955
  10       1         1     0.000      NaN              NA              NA
```

3 Aplicación

3.1 Uso en R

- Librería `survival`:

```
library(survival)
Surv(tiempo, status)
```

- Este objeto puede usarse en:
  - Surv() codifica la información de tiempo y censura.
  - survfit() ajusta curvas de supervivencia (Kaplan-Meier).
  - coxph() para modelos de Cox

3.1.1 La función `Surv()` de `survival`

```
library(survival)

# Censura derecha
tiempos <- c(5, 8, 12, 3, 10)
evento <- c(1, 0, 1, 1, 0)  # 1 = evento, 0 = censurado

datos <- Surv(tiempos, evento)
datos
```

```
[1]  5   8+ 12   3  10+
```

- Crea un objeto de clase `Surv`.
- Es la base para ajustar modelos de supervivencia.

3.1.2 Visualizando `Surv()` con tipos de censura

```
# Censura izquierda
tiempos <- c(5, 8, 12, 3, 10)
evento <- c(1, 0, 1, 1, 0)
Surv(tiempos, evento, type = "left")
```

```
[1]  5   8- 12   3  10-
```

```
# Censura por intervalo
inferior <- c(2, 6, 7, 5, 1)
superior <- c(4, 6, 9, 6, 3)
```

```r
evento <- c(3, 0, 3, 0, 3)  # 3 = intervalo
Surv(inferior, superior, type = "interval2")
```

```
[1] [2, 4] 6      [7, 9] [5, 6] [1, 3]
```

3.1.3 Ajuste con `survfit()`

```r
library(survival)

# Datos con censura derecha
tiempos <- c(5, 8, 12, 3, 10)
evento <- c(1, 0, 1, 1, 0)
datos <- Surv(tiempos, evento)
print(datos)
```

```
[1]  5   8+ 12   3  10+
```

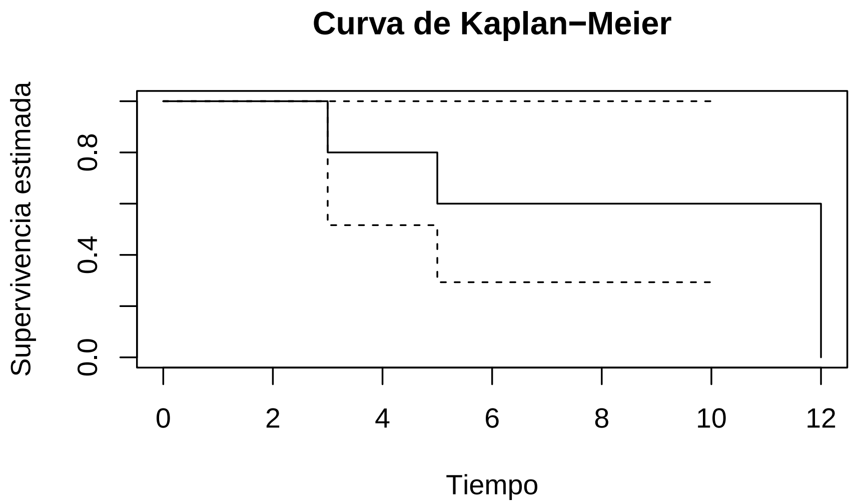
```r
modelo <- survfit(datos ~ 1)  # sin covariables
summary(modelo)
```

```
Call: survfit(formula = datos ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    3      5       1      0.8   0.179        0.516            1
    5      4       1      0.6   0.219        0.293            1
   12      1       1      0.0     NaN           NA           NA
```

- `survfit()` ajusta una curva de Kaplan-Meier.

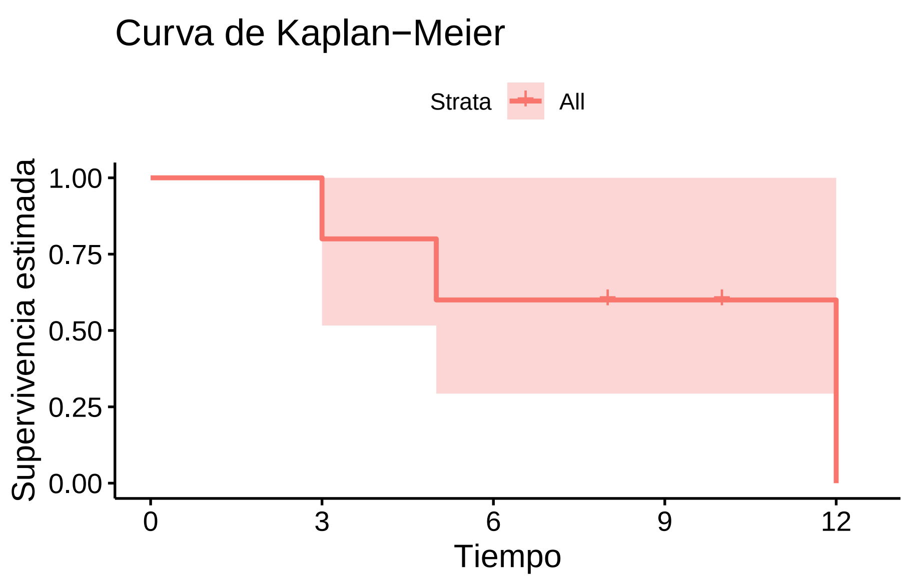
3.1.4 Graficando la curva de supervivencia

```r
plot(modelo, xlab = "Tiempo", ylab = "Supervivencia estimada",
     main = "Curva de Kaplan-Meier")
```



Puedes usar ggsurvplot() del paquete **`survminer`** para una mejor presentación visual.

```
survminer::ggsurvplot(modelo,data=datos, xlab = "Tiempo", ylab = "Supervivencia estimada",
     title = "Curva de Kaplan-Meier")
```



4 Intervalo de Confianza en Kaplan-Meier

4.1 Intervalo de Confianza en Kaplan-Meier

- La estimación de la supervivencia mediante Kaplan-Meier es una curva escalonada.
- Cada punto de la curva tiene asociado un **intervalo de confianza (IC)**.
- El IC refleja la **incertidumbre** de la estimación en cada tiempo debido a datos censurados y al tamaño muestral.

4.2 Fórmula general del IC

Para una estimación de la probabilidad de supervivencia ŜKM(t) en un tiempo 't' dado, el IC del 95% se calcula como:

$$\hat{S}_{KM}(t) \pm 1.96 \cdot \sqrt{\mathrm{Var}(\hat{S}_{KM}(t))}$$

- $\hat{S}_{KM}(t)$: estimador de Kaplan-Meier
- $\mathrm{Var}(\hat{S}_{KM}(t))$: varianza estimada con la fórmula de Greenwood, ver Klein & Moeschberger (2003).
- 1.96: cuantil de la normal estándar para un 95% de confianza

---

4.3 Fórmula de Greenwood para la Varianza

$$\mathrm{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 \cdot \sum_{j:t_j \le t} \frac{m_j}{n_j(n_j - m_j)}$$

Donde:

- m_j: número de fallas en el tiempo t_j
- n_j: número de sujetos en riesgo justo antes de t_j

4.4 Ejemplo de cálculo

Table 13: Grupo 1 (tratamiento): Estimación paso a paso de la función de supervivencia KM

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	1
6	21	3	1	18/21 = 0.8571

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
7	17	1	1	$0.8571 \times 16/17 = 0.8067$
10	15	1	2	$0.8067 \times 14/15 = 0.7529$
13	12	1	1	$0.7529 \times 11/12 = 0.6902$
16	11	1	2	$0.6902 \times 10/11 = 0.6275$
22	7	1	1	$0.6275 \times 6/7 = 0.5378$
23	6	1	1	$0.5378 \times 5/6 = 0.4482$

- A las 10 semanas,

$$\hat{S}(10) = 0.753$$

- Eventos y riesgos previos:
  - Semana 6: $m_f = 3, n_f = 16 \rightarrow \frac{3}{16 \cdot 13} = 0.0144$
  - Semana 7: $m_f = 1, n_f = 17 \rightarrow \frac{1}{17 \cdot 16} = 0.0037$
  - Semana 10: $m_f = 1, n_f = 15 \rightarrow \frac{1}{15 \cdot 14} = 0.0048$

$$\sum = 0.0144 + 0.0037 + 0.0048 = 0.0229$$

$$\text{Var}(\hat{S}(10)) = (0.753)^2 \cdot 0.0229 = 0.013$$

IC del 95%:

$$0.753 \pm 1.96 \cdot \sqrt{0.013} = (0.571, 0.935)$$

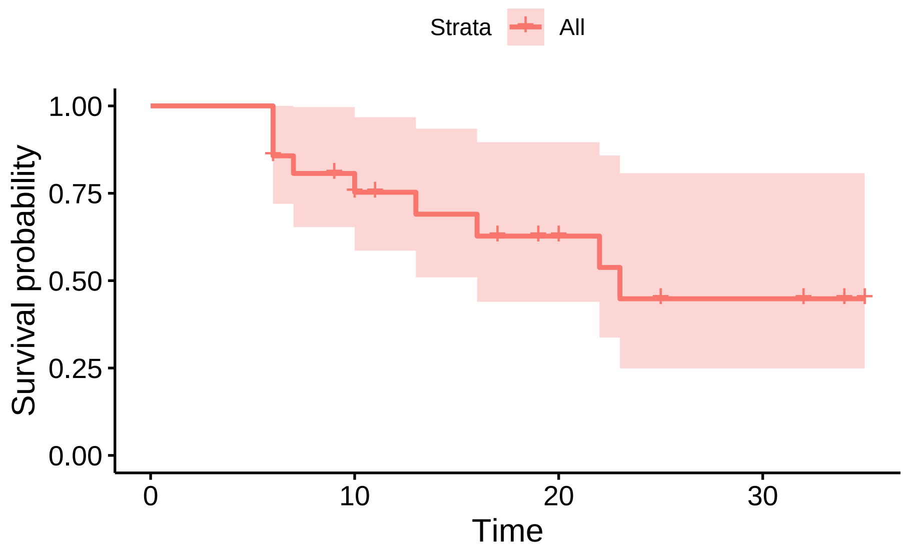
4.5 Visualización en R

```r
library(survival)

tratamiento <- data.frame(tiempo = c(6, 6, 6, 7, 10, 13, 16, 22, 23,
                                     6, 9, 10, 11, 17, 19, 20, 25, 32, 32, 34, 35),
                          status = c(rep(1,9),rep(0,12)))

ajuste <- survfit(Surv(tiempo, status) ~ 1, data = tratamiento)
summary(ajuste)
```

```
Call: survfit(formula = Surv(tiempo, status) ~ 1, data = tratamiento)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
```

```
  22        7          1      0.538  0.1282          0.337           0.858
  23        6          1      0.448  0.1346          0.249           0.807
```

```
ggsurvplot(fit=ajuste,data=tratamiento)
```



4.6 Consideraciones finales de los intervalos de confianza

- Los IC permiten visualizar la precisión de las curvas de supervivencia.
- Son especialmente útiles para comparar entre grupos.

5 Supervivencia Mediana

5.1 ¿Qué es la supervivencia mediana?

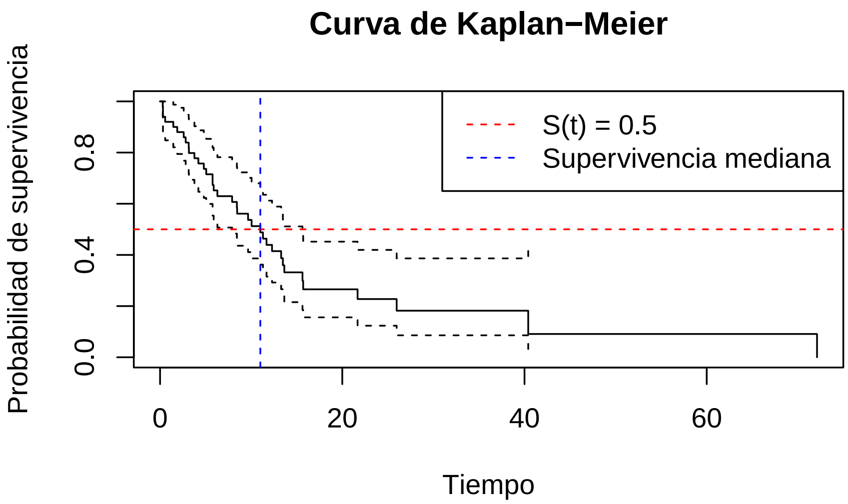
- Es el tiempo en el cual la **probabilidad de supervivencia** se reduce al **50%**.
- También se define como el tiempo `t` tal que

$$S(t) = 0.5$$

.

- Es preferida sobre la **media** cuando hay datos **censurados**, ya que la media puede no estar bien definida si la cola derecha no está completamente observada.

5.2 Ejemplo visual



5.3 ¿Por qué es útil?

- Es **robusta** frente a valores extremos.
- Resume la distribución del tiempo hasta el evento.
- Tiene una **interpretación clara**: la mitad de los individuos han experimentado el evento para ese tiempo.

5.4 Cálculo en R

```r
summary(ajuste)$table["median"]
```

```
  median 
11.00339 
```

```r
print(ajuste)
```

```
Call: survfit(formula = surv_obj ~ 1)

      n events median 0.95LCL 0.95UCL
[1,] 50     36     11    7.91    15.6
```

Este valor representa la **supervivencia mediana** estimada a partir de los datos.

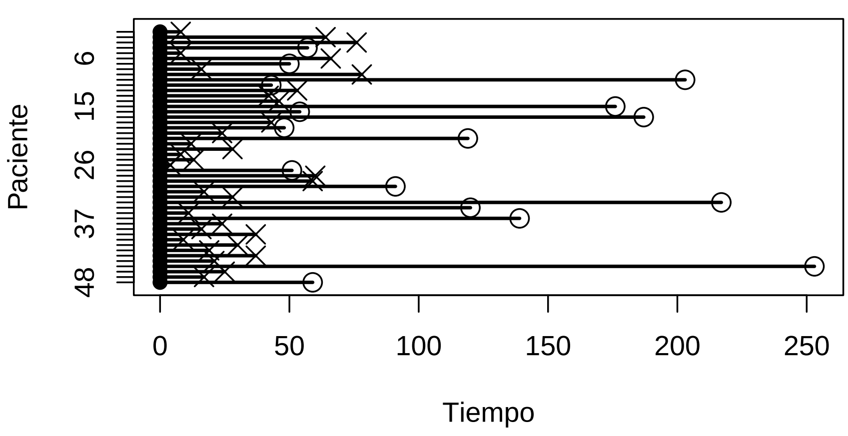
5.5 Conjunto de datos `gastricXelox` de la biblioteca `asaur`

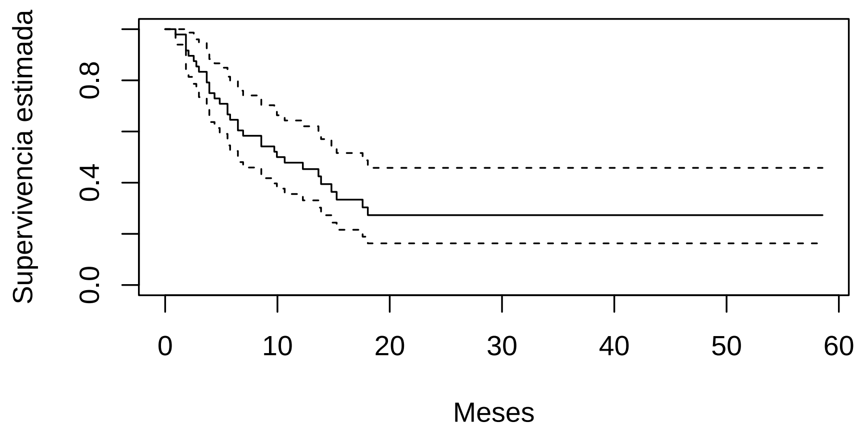
```
library(asaur)
data("gastricXelox")
```

Table 14: Ejemplo

paciente	tiempo	status
1	8	1
2	64	1
3	76	1
4	57	0
5	8	1
6	66	1

- Tiempo: semanas hasta progresión o muerte
- `delta = 1` si hubo evento, `0` si censurado
- Los datos se desordenaron para este ejemplo



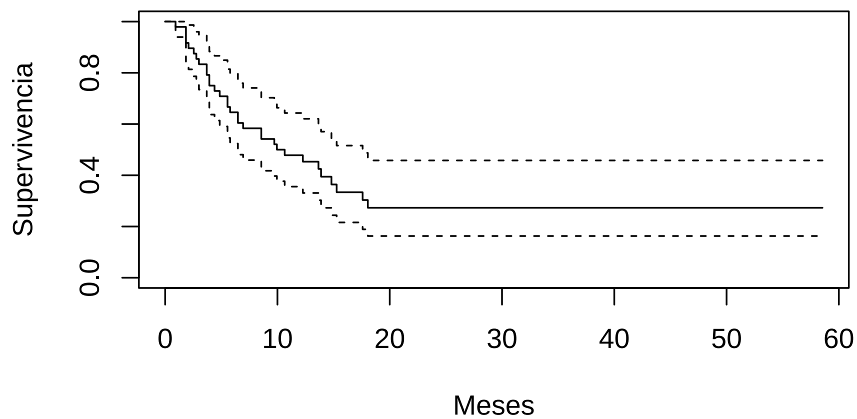


5.6 Ejercicio

- Usar R para:
  - Estimar la curva de supervivencia de `gastricXelox`
  - Obtener la mediana de supervivencia
  - Graficar con intervalo de confianza

```
Call: survfit(formula = Surv(timeMonths, delta) ~ 1, data = gastricXelox)

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.926     48       1    0.979  0.0206        0.940        1.000
 1.851     47       3    0.917  0.0399        0.842        0.998
 2.083     44       1    0.896  0.0441        0.813        0.987
 2.545     43       1    0.875  0.0477        0.786        0.974
 2.777     42       1    0.854  0.0509        0.760        0.960
 3.008     41       1    0.833  0.0538        0.734        0.946
 3.702     40       2    0.792  0.0586        0.685        0.915
 3.934     38       2    0.750  0.0625        0.637        0.883
 4.397     36       1    0.729  0.0641        0.614        0.866
 4.860     35       1    0.708  0.0656        0.591        0.849
 5.554     34       2    0.667  0.0680        0.546        0.814
 5.785     32       1    0.646  0.0690        0.524        0.796
 6.479     31       2    0.604  0.0706        0.481        0.760
 6.942     29       1    0.583  0.0712        0.459        0.741
 8.562     28       2    0.542  0.0719        0.418        0.703
```

9.719	26	1	0.521	0.0721	0.397	0.683
9.950	25	1	0.500	0.0722	0.377	0.663
10.645	23	1	0.478	0.0722	0.356	0.643
12.264	19	1	0.453	0.0727	0.331	0.620
13.653	16	1	0.425	0.0735	0.303	0.596
13.884	14	1	0.394	0.0742	0.273	0.570
14.810	13	1	0.364	0.0744	0.244	0.544
15.273	12	1	0.334	0.0742	0.216	0.516
17.587	11	1	0.303	0.0734	0.189	0.487
18.050	10	1	0.273	0.0720	0.163	0.458



6 Comparación entre grupos (Log-Rank Test)

6.1 Objetivo

- Comparar curvas de supervivencia entre **dos o más grupos**.
- Detectar diferencias globales en el riesgo de eventos a lo largo del tiempo.

6.2 Hipótesis

$$H_0 : S_1(t) = S_2(t) \quad \forall t H_A : S_1(t) \neq S_2(t) \quad \text{para al menos un valor de } t$$

- Prueba **no paramétrica**
- Se basa en la comparación entre **observados** y **esperados**

6.3 Fundamento de la prueba

En cada tiempo de fallo:

- Se registra el número de eventos observados (O_{1j}, O_{2j})
- Se calcula el número esperado bajo H_0 (E_{1j}, E_{2j})

Se acumulan a lo largo del tiempo:

$$Z = \sum_j (O_{1j} - E_{1j})$$

y la varianza:

$$\text{Var}(Z) = \sum_j V_j$$

Cálculo del número esperado bajo H_0

En la prueba de log-rank, bajo la hipótesis nula $H_0 : S_1(t) = S_2(t)$, se asume que **las tasas de fallo son iguales en ambos grupos**. Por tanto, el número esperado de fallos para cada grupo en el tiempo de fallo $t_{(f)}$ se calcula como:

- Número total de fallos en $t_{(f)}$:

$$m_f = m_{1f} + m_{2f}$$

- Número total en riesgo en $t_{(f)}$:

$$n_f = n_{1f} + n_{2f}$$

- **Esperado en el grupo 1**:

$$e_{1f} = \frac{n_{1f}}{n_f} \cdot m_f$$

- **Esperado en el grupo 2**:

$$e_{2f} = \frac{n_{2f}}{n_f} \cdot m_f$$

Este cálculo se repite en cada tiempo de fallo $t_{(f)}$ y los valores se acumulan para calcular el estadístico de prueba:

$$Z = \sum_f (m_{1f} - e_{1f}), \quad \text{Var}(Z) = \sum_f \frac{n_{1f} n_{2f} m_f (n_f - m_f)}{n_f^2 (n_f - 1)}$$

6.4 Estadístico de prueba

$$\chi^2 = \frac{(O_1 - E_1)^2}{\text{Var}(Z)} \sim \chi^2_{(1)}$$

Se compara con la distribución χ^2 con 1 grado de libertad (para dos grupos).

6.5 Tabla Expandida (Datos de Remisión)

Table 15: Tabla expandida completa: cálculo detallado para prueba de log-rank

$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	m_f	n_f	e_{1f}	e_{2f}	e_{1f}	e_{2f}	$m_{1f}-e_{1f}$	$m_{2f}-e_{2f}$
1	0	2	21	21	2	42	1.00	1.00	(21/42 × 2)	(21/42 × 2)	-1.00	1.00
2	0	2	21	19	2	40	1.05	0.95	(21/40 × 2)	(21/40 × 2)	-1.05	1.05
3	0	1	21	17	1	38	0.55	0.45	(21/38 × 1)	(21/38 × 1)	-0.55	0.55
4	0	2	21	16	2	37	1.14	0.86	(21/37 × 2)	(21/37 × 2)	-1.14	1.14
5	0	2	21	14	2	35	1.20	0.80	(21/35 × 2)	(21/35 × 2)	-1.20	1.20
6	3	0	21	12	3	33	1.91	1.09	(21/33 × 3)	(21/33 × 3)	1.09	-1.09
7	1	0	17	12	1	29	0.59	0.41	(17/29 × 1)	(17/29 × 1)	0.41	-0.41
8	0	4	16	12	4	28	2.29	1.71	(16/28 × 4)	(16/28 × 4)	-2.29	2.29
10	1	0	15	8	1	23	0.65	0.35	(15/23 × 1)	(15/23 × 1)	0.35	-0.35
11	0	2	13	8	2	21	1.24	0.76	(13/21 × 2)	(13/21 × 2)	-1.24	1.24

$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	m_f	n_f	e_{1f}	e_{2f}	e_{1f}	e_{2f}	$m_{1f} - e_{1f}$	$m_{2f} - e_{2f}$
12	0	1	12	6	1	18	0.67	0.33	(12/18 $\times$ 1)	(12/18 $\times$ 1)	-0.67	0.67
13	1	1	12	5	2	17	1.41	0.59	(12/17 $\times$ 2)	(12/17 $\times$ 2)	-0.41	0.41
15	0	1	11	4	1	15	0.73	0.27	(11/15 $\times$ 1)	(11/15 $\times$ 1)	-0.73	0.73
16	1	0	11	3	1	14	0.79	0.21	(11/14 $\times$ 1)	(11/14 $\times$ 1)	0.21	-0.21
17	0	1	10	3	1	13	0.77	0.23	(10/13 $\times$ 1)	(10/13 $\times$ 1)	-0.77	0.77
22	1	1	7	2	2	9	1.56	0.44	(7/9 $\times$ 2)	(7/9 $\times$ 2)	-0.56	0.56
23	1	1	6	1	2	7	1.71	0.29	(6/7 $\times$ 2)	(6/7 $\times$ 2)	-0.71	0.71
Totales	9	21					19.26	10.74			-10.26	10.26

- **Esperado en el grupo 1**: $e_{1f} = \frac{n_{1f}}{n_f} \cdot m_f$
- **Esperado en el grupo 2**: $e_{2f} = \frac{n_{2f}}{n_f} \cdot m_f$

Significado de las columnas de la tabla expandida

Table 16: Descripción de las columnas en la tabla expandida de log-rank

Columna	Significado
f	Índice del tiempo de fallo ordenado
$t_{(f)}$	Tiempo observado de fallo número f
m_{1f}	Número de fallos en el grupo 1 en $t_{(f)}$
m_{2f}	Número de fallos en el grupo 2 en $t_{(f)}$
n_{1f}	Número en riesgo en el grupo 1 justo antes de $t_{(f)}$
n_{2f}	Número en riesgo en el grupo 2 justo antes de $t_{(f)}$
e_{1f}	Número esperado de fallos en el grupo 1 bajo H_0

e_{2f}	Número esperado de fallos en el grupo 2 bajo H_0
$m_{1f} - e_{1f}$	Diferencia entre observados y esperados en el grupo 1
$m_{2f} - e_{2f}$	Diferencia entre observados y esperados en el grupo 2

6.6 Ejemplo (Grupo Tratamiento vs Placebo)

Ejemplo: Tiempos de remisión (semanas) para dos grupos de pacientes con leucemia

Grupo 1 ($n = 21$) — *Tratamiento*
6, 6, 6, 7, 10, 13, 16, 22, 23,
6^+, 9^+, 10^+, 11^+,
17^+, 19^+, 20^+,
25^+, 32^+, 32^+, 34^+, 35^+

Nota: el símbolo $^+$ denota observaciones censuradas.

Grupo 2 ($n = 21$) — *Placebo*
1, 1, 2, 2, 3,
4, 4, 5, 5,
8, 8, 8, 8,
11, 11, 12, 13,
15, 17, 22, 23

Grupo	# Fallos	# Censurados	Total
Grupo 1	9	12	21
Grupo 2	21	0	21

```r
library(survival)

datos <- data.frame(tiempo = c(6, 6, 6, 7, 10, 13, 16, 22, 23,
                               6, 9, 10, 11, 17, 19, 20, 25, 32, 32, 34, 35,
                               1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11,
                               11, 12, 13, 15, 17, 22, 23),
                    status = c(rep(1,9),rep(0,12), rep(1,21)),
                    grupo = factor(c(rep("Tratamiento",21),rep("Placebo",21))))

survdiff(Surv(tiempo, status) ~ grupo, data = datos)
```

```
Call:
survdiff(formula = Surv(tiempo, status) ~ grupo, data = datos)

                   N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=Placebo     21       21     10.8      9.76      16.8
```

```
grupo=Tratamiento 21        9     19.2      5.45      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

6.7 Interpretación de la salida

- Se obtiene un valor de χ^2 y un valor-p.
- Si $p < \alpha$, se **rechaza** H_0: hay evidencia de que las curvas difieren.
- Si $p \geq \alpha$, no se rechaza H_0: no hay evidencia suficiente.

6.8 Generalización de la prueba de log-rank (k grupos)

Sea k el número de grupos a comparar.

En cada tiempo de fallo $t_{(f)}$:

- m_{if}: número de fallos en el grupo i.
- n_{if}: número en riesgo en el grupo i.
- $m_f = \sum_{i=1}^{k} m_{if}$: total de fallos.
- $n_f = \sum_{i=1}^{k} n_{if}$: total en riesgo.

Valor esperado para el grupo i:

$$e_{if} = \frac{n_{if}}{n_f} \cdot m_f$$

6.9 Estadístico de prueba para k grupos

Sea $O_i = \sum_f m_{if}$ y $E_i = \sum_f e_{if}$

El estadístico log-rank generalizado es:

$$X^2 = (O - E)^T \Sigma^{-1} (O - E)$$

donde:

- $O = (O_1, \ldots, O_{k-1})$
- $E = (E_1, \ldots, E_{k-1})$
- Σ es la matriz de covarianza de O

Distribución asintótica:

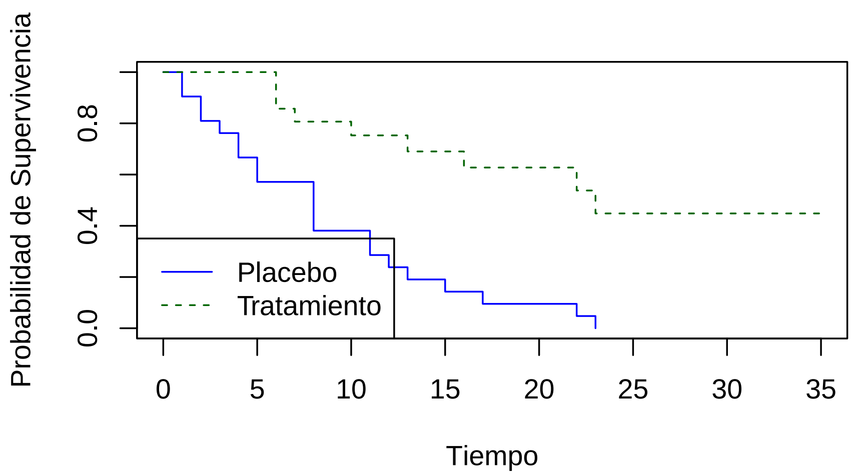
$$X^2 \sim \chi^2_{k-1}$$

Se **rechaza** H_0 si el valor-p es menor al nivel de significancia.

6.10 Consideraciones

- Sensible a diferencias en tiempos largos si hay censura temprana.
- La prueba de log-rank **asume riesgos proporcionales**.
- No considera covariables — usar modelo de Cox si se desea controlar otras variables.

6.11 Visualización



6.12 Conclusión

- La prueba de log-rank es útil para **comparar curvas de supervivencia** entre grupos.
- Es ampliamente usada por su simplicidad y poder bajo riesgos proporcionales.

6.13 Actividad práctica guiada

Datos: `lung` del paquete `survival`.

Pasos:

1. Cargar datos con `data(lung)`
2. Crear objeto `Surv(time, status)`
3. Estimar curvas por `sex`
4. Probar igualdad con log-rank

7 Referencias

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd ed.). Springer.