

Capítulo 2

Procedimientos No Paramétricos

Un paso inicial en el análisis de datos de supervivencia es presentar resúmenes gráficos o numéricos de los tiempos de supervivencia para unidades experimentales en cierto grupo. Tales resúmenes pueden dar pauta a un análisis más detallado de los datos. Los datos de supervivencia pueden ser resumidos convenientemente a través de estimadores de la función de supervivencia y de la fuerza de mortalidad. Los métodos para encontrar estos estimadores son llamados *no paramétricos* o *libres de distribución*, ya que no necesitan que se hagan suposiciones específicas sobre la distribución de los tiempos de supervivencia.

2.1. Estimación de la Función de supervivencia

Supóngase que se tiene una muestra de tiempos de supervivencia donde ninguna de las observaciones tiene censura. La función de supervivencia es la probabilidad de que un individuo sobreviva por un tiempo mayor o igual a t . Esta función se

puede estimar por la *función de supervivencia empírica*:

$$\tilde{S}(t) = \frac{\text{Número de individuos con tiempos de supervivencia } \geq t}{\text{Número total de individuos en los datos}}.$$

En forma equivalente, $\tilde{S}(t) = 1 - \tilde{F}(t)$, donde $\tilde{F}(t)$ es la función de distribución empírica., que es el cociente del número total de individuos vivos en el tiempo t entre el número total de individuos en el estudio. Nótese que $\tilde{S}(t) = 1$ para $t < t_{(1)}$, donde $t_{(1)}$ representa la observación mas chica; además, $\tilde{S}(t) = 0$ para $t \geq t_{(n)}$, donde $t_{(n)}$ es la observación mas grande.

El método para estimar la función de supervivencia usando el cociente no se puede usar cuando hay observaciones con censura. A continuación se describen algunos métodos no paramétricos que permiten estimar $S(t)$ en presencia de datos censurados.

2.1.1. El Estimador Actuarial

El *estimador actuarial*, o la *tabla de vida*, de una función de supervivencia se obtiene al dividir el periodo de observación en una serie de intervalos de tiempo. Estos intervalos no deben de ser de igual magnitud aunque, en general, lo son. Supóngase que el j -ésimo intervalo de un total de m intervalos, $j = 1, 2, \dots, m$, abarca de t'_j a t'_{j+1} , y sean d_j y c_j el número de muertes y el número de tiempos de supervivencia censurados respectivamente, en este intervalo. Sea n_j el número de individuos que estan vivos, y por lo tanto en riesgo de morir, al principio del j -ésimo intervalo. Partiendo de la suposición de que el proceso de censura es tal que los tiempos de supervivencia ocurren uniformemente durante el j -ésimo intervalo, de manera tal que el número promedio de individuos que estan en riesgo durante este intervalo es

$$n'_j = n_j - c_j/2,$$

el cual representa el número ajustado de individuos en riesgo.

En el j -ésimo intervalo la probabilidad de muerte se puede estimar con d_j/n'_j , y entonces la probabilidad de supervivencia correspondiente es $(n'_j - d_j)/n'_j$. Ahora, la probabilidad de que un individuo sobreviva después del tiempo t'_k es el producto de las probabilidades de que un individuo sobreviva los $k - 1$ intervalos anteriores, y entonces el estimador actuarial es

$$S^*(t) = \prod_{j=1}^k \left(\frac{n'_j - d_j}{n'_j} \right), \quad t \in [t'_k, t'_{k+1}),$$

para $k = 1, \dots, m$. Aquí se puede comprobar que $S^*(t) = 1$ para $t < t_{(1)}$, y que $S^*(t) = 0$ para $t \geq t_{(n)}$.

El estimador actuarial es sensible a la elección de los intervalos usados, exactamente en la misma forma en que la gráfica de un histograma depende de la elección de los intervalos de clase. El estimador actuarial es adecuado en situaciones en las que los tiempos de las muestras son desconocidos, y la única información disponible es el número de muertes y el número de observaciones censuradas que ocurren en una serie de intervalos de tiempo.

Cuando los tiempos de supervivencia son conocidos, el estimador actuarial se puede usar pero la agrupación de los intervalos conlleva a la pérdida de información.

Ejemplo 2.1. Los datos en la Tabla 2.1, obtenidos de Krall, Uthoff y Harley (1975), corresponden a 48 pacientes con edades de 50 a 80 años de edad. Varios de estos pacientes no habían muerto al final del estudio, por lo que las observaciones correspondientes son tiempos censurados por la derecha. Por el momento solo será necesario concentrarse en las columnas **time**, los tiempos de supervivencia (en meses), y **status**, el estatus de censura (1=sin censura, 0=con censura).

Los tiempos de supervivencia se agrupan para obtener el número de pacientes que mueren d_j , y el número que está censurado c_j , en cada uno de los cinco años del estudio, y después en los tres restantes. El número de individuos en riesgo al

Tabla 2.1: Tiempos de supervivencia (en meses) de pacientes en un estudio sobre mieloma múltiple.

patient	time	status	age	sex	bun	ca	hb	pccells	protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14	9	0
20	56	0	66	1	18	11	12.5	90	0
21	88	1	63	1	21	9	14	42	1
22	24	1	67	1	10	10	12.4	44	0
23	51	1	60	2	10	10	10.1	45	1
24	4	1	74	1	48	9	6.5	54	0
25	40	0	72	1	57	9	12.8	28	1
26	8	1	55	1	53	12	8.2	55	0
27	18	1	51	1	12	15	14.4	100	0
28	5	1	70	2	130	8	10.2	23	0
29	16	1	53	1	17	9	10	28	0
30	50	1	74	1	37	13	7.7	11	1
31	40	1	70	2	14	9	5	22	0
32	1	1	67	1	165	10	9.4	90	0
33	36	1	63	1	40	9	11	16	1
34	5	1	77	1	23	8	9	29	0
35	10	1	61	1	13	10	14	19	0
36	91	1	58	2	27	11	11	26	1
37	18	0	69	2	21	10	10.8	33	0
38	1	1	57	1	20	9	5.1	100	1
39	18	0	59	2	21	10	13	100	0
40	6	1	61	2	11	10	5.1	100	0
41	1	1	75	1	56	12	11.3	18	0
42	23	1	56	2	20	9	14.6	3	0
43	15	1	62	2	21	10	8.8	5	0
44	18	1	60	2	18	9	7.5	85	1
45	12	0	71	2	46	9	4.9	62	0
46	12	1	60	2	6	10	5.5	25	0
47	17	1	65	2	28	8	7.5	8	0
48	3	0	59	1	90	10	10.2	6	1

Tabla 2.2: Estimador Actuarial de la función de supervivencia para los datos de mieloma múltiple.

Intervalo	Periodo	d_j	c_j	n_j	n'_j	$(n'_j - d_j)/n'_j$	$S^*(t)$
1	[0, 12)	16	4	48	46.0	0.6521	0.6522
2	[12, 24)	10	4	28	26.0	0.6154	0.4013
3	[24, 36)	1	0	14	14.0	0.9286	0.3727
4	[36, 48)	3	1	13	12.5	0.7600	0.2832
5	[48, 60)	2	2	12	11.0	0.8181	0.2317
6	[60, ∞)	4	1	6	5.5	0.2727	0.0632

principio de cada intervalo n_j se calcula junto con el número ajustado de individuos en riesgo n'_j . Finalmente, se estima la probabilidad de muerte en cada intervalo, y estas cantidades se usan para obtener el estimador de $S(t)$. Los cálculos se muestran en la Tabla 2.2.

Cuando se usa el lenguaje de programación R para obtener la Tabla... es posible usar la función `survfit`. El procedimiento se muestra a continuación:

```
# Invoca la libreria de supervivencia
library(survival)

# Pon en un objeto los datos
mye <- read.table("C:/Mis documentos/multmye.dat",header=T)

# Pon en un objeto cantidades sobre los datos
mye.surv <- survfit(Surv(time,status),data=mye)

# los diferentes tiempos de supervivencia sin repeticion
tiempos <- mye.surv$time

# el numero de individuos en riesgo en cada tiempo:
n.riesgo <- mye.surv$n.risk

# el numero de eventos (muertes) en cada tiempo de supervivencia
n.eventos <- mye.surv$n.event

# Encuentra las d_j
```

```

limites <- c(12, 24, 36, 48, 60, Inf)
M <- length(limites)
sum.n.vent <- 1:M
  <- 1:M
for(i in 1:M){
sum.n.vent[i] <- sum(n.eventos[tiempos < limites[i]])}

lag.sum <- 1:M
lag.sum[1] <- 0
lag.sum[2:M] <- sum.n.vent[1:(M-1)]

d.j <- sum.n.vent - lag.sum

# Calcula n_j
n.j <- 1:M
n.j[1] <- n.riesgo[1]
for(i in 1:(M-1)){
n.j[1+i] <- rev(n.riesgo[tiempos <= limites[i]])[1]
}

# Calcula c_j

cum.c.j <- 1:M
for (i in 1:M){
cum.c.j[i] <- sum(1 - mye$status[mye$time < limites[i]])
}

lag.cum.c.j <- 1:M
lag.cum.c.j[1] <- 0
lag.cum.c.j[2:M] <- cum.c.j[1:(M-1)]

c.j <- cum.c.j - lag.cum.c.j

n.j.prim <- n.j - c.j/2

p.j <- (n.j.prim - d.j)/n.j.prim

S.j <- cumprod(p.j)

```

El Estimador Kaplan-Meier

La determinación del estimador Kaplan-Meier de la función de supervivencia de una muestra que incluye tiempos censurados es parecida a la del estimador actuarial. Sin embargo, cada uno de los intervalos se forma de manera tal que cada tiempo de muerte determina los límites.

Supóngase que hay n individuos y sus tiempos de supervivencia correspondientes son t_1, t_2, \dots, t_n . Algunas de estas observaciones pueden estar censuradas por la derecha, y es posible que existan observaciones con el mismo tiempo de supervivencia observado. Supóngase que hay r tiempos de con muertes observadas (observaciones sin censura), con $r \leq n$. Después de ordenar las observaciones en orden ascendente, el j -ésimo tiempo observado se denota $t_{(j)}$, para $j = 1, \dots, r$; de esta forma, los r tiempos de muerte ordenados son $t_{(1)} < t_{(2)} < \dots < t_{(r)}$. El número de individuos que se encuentran con vida antes del tiempo $t_{(j)}$, incluyendo los que están por morir en este tiempo, se denotan con n_j , y d_j denota el número que muere en este tiempo. El intervalo que va de $t_{(j)} - \delta$ a $t_{(j)}$, donde δ es un tiempo infinitesimal, incluye un tiempo de una muerte. Como hay n_j individuos que sobreviven poco antes de $t_{(j)}$, y hay d_j muertes en $t_{(j)}$, la probabilidad de que un individuo muera durante el intervalo que va de $t_{(j)} - \delta$ a $t_{(j)}$ se estima con d_j/n_j . La probabilidad de supervivencia estimada en el intervalo es entonces $(n_j - d_j)/n_j$.

Es posible que varios tiempos de supervivencia censurados ocurran al mismo tiempo que uno o más muertes, de manera tal que el tiempo de muerte y los tiempos censurados ocurren simultáneamente. En este caso, cuando se calcula n_j , los tiempos de supervivencia censurados se toman como si ocurrieran inmediatamente después de del tiempo de la muerte.

En la forma en que se construyen los intervalos de tiempo, el intervalo que va de $t_{(j)}$ a $t_{(j+1)} - \delta$, que es el tiempo inmediatamente anterior al siguiente tiempo

de muerte, no contiene ninguna muerte. Por lo tanto, la probabilidad de sobrevivir de $t_{(j)}$ a $t_{(j+1)} - \delta$ es uno, y la probabilidad conjunta de sobrevivir de $t_{(j)} - \delta$ a $t_{(j)}$ y de $t_{(j)}$ a $t_{(j+1)} - \delta$ se puede estimar con $(n_j - d_j)/n_j$. Tomando el límite $\delta \rightarrow 0$, $(n_j - d_j)/n_j$ se convierte en un estimador de la probabilidad de supervivencia de $t_{(j)}$ a $t_{(j+1)}$.

Suponiendo que la ocurrencia de los eventos en la muestra es independiente de individuo a individuo, el estimado de la función de supervivencia en cualquier intervalo que va de $t_{(k)}$ a $t_{(k+1)}$, $k = 1, \dots, r$, donde $t_{(r+1)} = \infty$, se calcula como la probabilidad de supervivencia después de $t_{(k)}$. Esta es la probabilidad de sobrevivir en el intervalo $t_{(k)}$ a $t_{(k+1)}$ y todos los intervalos anteriores. Este es el *estimador Kaplan-Meier* de la función de supervivencia, el cual está dado por

$$\hat{S}(t) = \prod_{j=1}^k \left(\frac{n_j - d_j}{n_j} \right), \quad t \in [t_{(k)}, t_{(k+1)})$$

para $k = 1, \dots, r$, con $\hat{S}(t) = 1$ para $t < t_{(1)}$. Si la observación más grande es un tiempo censurado t^* , entonces $\hat{S}(t)$ no está definido para $t > t^*$. Por otra parte, si la observación más grande es un tiempo sin censura, i.e. $t_{(r)}$, entonces $n_r = d_r$, y por lo tanto $\hat{S}(t)$ es cero para $t \geq t_{(r)}$.

La gráfica del estimador Kaplan-Meier de la función de supervivencia es una función escalonada, en la cual las probabilidades de supervivencia son constantes entre los tiempos de muerte adyacentes y decrece en cada tiempo de muerte. El estimador Kaplan-Meier es también conocido como el *estimador de límite producto* (en inglés, *product limit estimator*).

Nótese que si no hay tiempos con censura en los datos, entonces $n_j - d_j = n_{j+1}$, $j = 1, 2, \dots, k$, y entonces el estimador Kaplan-Meier se expresa como

$$\hat{S}(t) = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times \dots \times \frac{n_{k+1}}{n_k}.$$

Esto se reduce a $\hat{S}(t) = n_{k+1}/n_1$, para $k = 1, 2, \dots, r - 1$, con $\hat{S}(t) = 1$ para $t < t_{(1)}$

Tabla 2.3: Tiempo en semanas para la discontinuidad del DIU.

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	107*

Tabla 2.4: Estimador Kaplan-Meier para los datos DIU

Intervalo	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0-	18	0	1.0000	1.0000
10-	18	1	0.9444	0.9444
19-	15	1	0.9333	0.8815
30-	13	1	0.9231	0.8137
36-	12	1	0.9167	0.7459
59-	8	1	0.8750	0.6526
75-	7	1	0.8571	0.5594
93-	6	1	0.8333	0.4662
97-	5	1	0.8000	0.3729
107	3	1	0.6667	0.2486

y $\hat{S}(t) = 0$ para $t \geq t_{(r)}$. Como n_1 es el número de individuos en riesgo antes de la primera muerte, i.e. $n_1 = n$, y $n_{k+1} = 0$ es el número de individuos con tiempos de supervivencia mayores o iguales a t_{k+1} , entonces $\hat{S}(t)$ es simplemente la función de supervivencia empírica.

Ejemplo 2.2. Los datos en la Tabla 2.3 se refieren al número de semanas desde el comienzo de un dispositivo intrauterino (DIU) hasta su discontinuidad (WHO, 1987). Los datos están dados para 18 mujeres quienes tienen edades de 18 a 35 años y han tenido dos embarazos. Los tiempos de discontinuidad que son censuradas tienen un asterisco. El estimador Kaplan-Meier se muestra en la Tabla 2.4. La gráfica de $\hat{S}(t)$ se observa en la Figura 2.1. Nótese que como el tiempo mas largo tiene censura, $\hat{S}(t)$ no está definido después de $t = 107$.

El language R incluye en su librería `survival` funciones listas para tabular y graficar el estimador Kaplan-Meier. El procedimiento se describe a continuación:

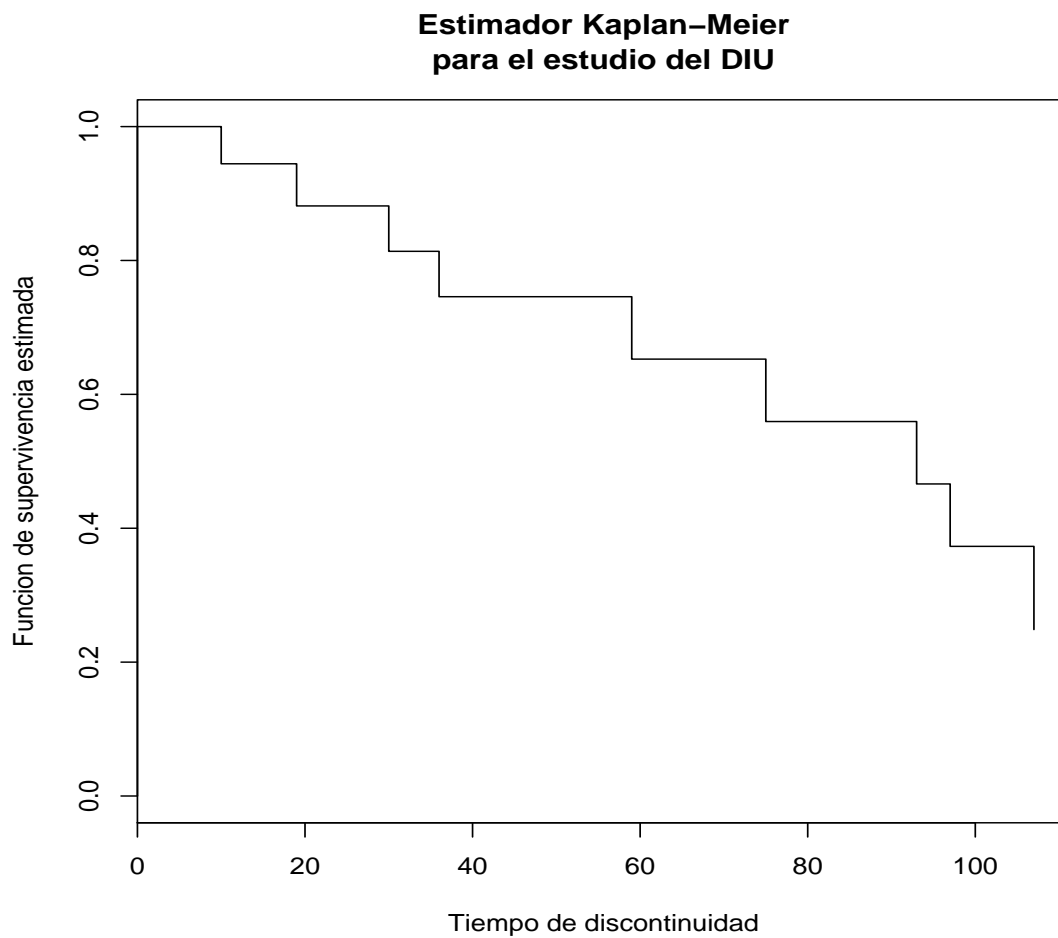


Figura 2.1: Estimador Kaplan-Meier de la función de supervivencia para los datos del DIU.

```
# Pon los datos en el formato data.frame
diu <- data.frame(tiempo=c(10, 13, 18, 19, 23, 30, 36, 38, 54,
                          56, 59, 75, 93, 97, 104, 107, 107, 107),
                 estatus=c(1, 0, 0, 1, 0, 1, 1, 0, 0,
                          0, 1, 1, 1, 1, 0, 1, 0, 0))

library(survival)
```

```
# Encuentra el estimador Kaplan-Meier
diu.surv <- survfit(Surv(tiempo, estatus), data=diu)

# Muestra el estimador y otras cantidades importantes
summary(diu.surv)

# Obten una grafica del estimador
plot(diu.surv, conf.int=FALSE, mark.time=FALSE)
title(main="Estimador Kaplan-Meier\npara el estudio del DIU",
      xlab="Tiempo de discontinuidad",
      ylab="Funcion de supervivencia estimada")
```