

### 2.1.2. Error estándar para el estimador Kaplan-Meier

Como Kaplan-Meier es el estimador de la función de supervivencia más importante y ampliamente usado, es importante derivar el error estándar de  $\hat{S}(t)$  se deriva en esta sección.

El estimador Kaplan-Meier de la función de supervivencia evaluada en  $t$ , donde  $t$  se encuentra en intervalo que va de  $t_{(k)}$  a  $t_{(k+1)}$ , se puede expresar como

$$\hat{S}(t) = \prod_{j=1}^k \hat{p}_j, \quad k = 1, \dots, r,$$

donde  $\hat{p}_j = (n_j - d_j)/n_j$  es la probabilidad estimada de que un individuo sobreviva el intervalo de tiempo que comienza en  $t_{(j)}$ ,  $j = 1, \dots, r$ . Si se toma el logaritmo natural, se tiene que

$$\log \hat{S}(t) = \prod_{j=1}^k \log \hat{p}_j,$$

y, como los elementos en la muestra son independientes, la varianza de  $\log \hat{S}(t)$  está dada por

$$\text{var} \left\{ \log \hat{S}(t) \right\} = \sum_{j=1}^k \text{var} \{ \log \hat{p}_j \}.$$

El número de individuos que sobreviven al intervalo que comienza en  $t_{(j)}$  puede tomarse como una variable aleatoria binomial con parámetros  $n_j$  y  $p_j$ ; i.e.  $n_j$  es el número de realizaciones Bernoulli, y  $p_j$  es la probabilidad de éxito (de sobrevivir). De esta forma, la varianza del número de individuos que sobrevive  $S_j = n_j - d_j$ ,  $S_j \sim \text{BIN}(n_j, p_j)$ , está dada por

$$\text{Var}[S_j] = \text{Var}[n_j - d_j] = n_j p_j (1 - p_j).$$

Como  $\hat{p}_j = (n_j - d_j)/n_j$ , la varianza de  $\hat{p}_j$  es  $\text{Var}[S_j]/n_j^2$ , que es  $p_j(1 - p_j)/n_j$ . Por lo tanto, la varianza de  $\hat{p}_j$  se puede estimar con

$$\hat{p}_j(1 - \hat{p}_j)/n_j.$$

Para obtener la varianza de  $\log \hat{p}_j$ , se hace uso del resultado general para aproximar la varianza de una función de una variable aleatoria. De acuerdo a este resultado, la varianza de la función  $g(X)$  de la variable aleatoria  $X$  está dada por

$$\text{Var}\{g(X)\} \approx \left\{ \frac{dg(X)}{dX} \right\}^2 \text{Var}(X).$$

Este resultado es conocido como la *aproximación de la serie de Taylor* de la varianza de una función de una variable aleatoria. Usando este resultado, la varianza aproximada de  $\log \hat{p}_j$  es  $\text{Var}[\hat{p}_j]/\hat{p}_j^2$ ; por lo que un estimador aproximado de la varianza de  $\log \hat{p}_j$  es  $(1 - \hat{p}_j)/(n_j \hat{p}_j)$ , el cual en sustitución de  $\hat{p}_j$  se reduce a

$$\frac{d_j}{n_j(n_j - d_j)}.$$

Se tiene entonces que

$$\text{Var} \left[ \log \hat{S}(t) \right] \approx \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)},$$

y usando de nueva cuenta la aproximación de la serie de Taylor, se tiene que

$$\text{Var} \left[ \log \hat{S}(t) \right] \approx \frac{1}{\left[ \hat{S}(t) \right]^2} \text{Var} \left[ \hat{S}(t) \right],$$

y de esta forma

$$\text{Var} \left[ \hat{S}(t) \right] \approx \left[ \hat{S}(t) \right]^2 \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}.$$

Finalmente, el error estándar del estimador Kaplan Meier de la función de supervivencia, el cual se define como la raíz cuadrada de la varianza estimada del estimador, está dado por

$$\text{s.e.} \left[ \hat{S}(t) \right] \approx \hat{S}(t) \sqrt{\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}}, \quad \text{para } t_{(k)} \leq t < t_{(k+1)}.$$

A este resultado se le conoce como la *fórmula de Greenwood*.

Si no hay tiempos de supervivencia con censura, entonces  $n_j - d_j = n_{j+1}$  y el estimador de  $\text{Var}[\log \hat{p}_j]$  es  $(n_j - n_{j+1}) / (n_j n_{j+1})$ . El valor aproximado de  $\text{Var}[\log \hat{S}(t)]$  es

$$\sum_{j=1}^k \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j=1}^k \left( \frac{1}{n_{j+1}} - \frac{1}{n_j} \right) = \frac{n_1 - n_{k+1}}{n_1 n_{k+1}},$$

el cual se puede expresar como

$$\frac{1 - \hat{S}(t)}{n_1 \hat{S}(t)},$$

ya que  $\hat{S}(t) = n_{k+1}/n_1$  para  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r - 1$ , cuando no hay censura. Usando la aproximación de la serie de Taylor, el estimador de la varianza de  $\hat{S}(t)$  es  $\hat{S}(t)[1 - \hat{S}(t)]/n_1$ .

### 2.1.3. Intervalos de confianza para valores de la función de supervivencia

Una vez que se ha calculado el error estándar de  $\hat{S}(t)$  es posible encontrar un *intervalo de confianza* para el valor correspondiente a  $\hat{S}(t)$ . Un intervalo de confianza es un estimador por intervalos de la función de supervivencia, y el intervalo se forma de manera tal que el valor verdadero de la función de supervivencia está dentro de sus límites dada una probabilidad de que la contenga, la cual es determinada con anterioridad.

Un intervalo de confianza para el valor verdadero de la función de supervivencia en el tiempo  $t$  se puede obtener al suponer que el valor estimado de la función de supervivencia en  $t$  se distribuye normal con media  $S(t)$  y varianza estimada dada por el cuadrado de la fórmula de Greenwood. El intervalo se calcula con los puntos porcentuales de una distribución normal estándar. De esta forma, si  $z_{1-\alpha/2}$  denota el percentil de una distribución normal estándar al nivel  $1 - \alpha/2$ , i.e.  $\text{Pr}\{Z < z_{1-\alpha/2}\} = 1 - \alpha/2$  con  $Z \sim N(0, 1)$ , entonces un intervalo de confianza al coeficiente

de confianza  $1 - \alpha$  para  $S(t)$  es el intervalo con límites

$$\hat{S}(t) \pm z_{1-\alpha/2} \text{s.e.}[\hat{S}(t)].$$

Un inconveniente de este procedimiento es que los intervalos de confianza son simétricos. Cuando se obtienen límites para valores de  $\hat{S}(t)$  cercanos a cero o uno, los intervalos simétricos no son adecuados pues sus valores se encuentran fuera del intervalo  $[0, 1]$ . Una forma pragmática de solucionar el problema es reemplazar los límites que exceden uno por 1, los que son inferiores a cero por 0.

Una procedimiento alternativo es transformar  $\hat{S}(t)$  para encontrar un valor que se encuentre en el rango  $(-\infty, \infty)$ , y entonces encontrar un intervalo de confianza para el valor transformado. El intervalo de confianza resultante es entonces transformado de nuevo para obtener el intervalo de confianza para  $S(t)$ . Transformaciones posibles la logística, dada por  $\log\{S(t)/[1 - S(t)]\}$ , y la log-log, dada por  $\log\{-\log \hat{S}(t)\}$ . Para la transformación log-log, se puede usar el resultado general

$$\text{Var}[\log(-X)] \approx \frac{1}{X^2} \text{Var}[X],$$

y al hacer la sustitución  $X = \log \hat{S}(t)$  se obtiene que

$$\text{Var} \left[ \log\{-\log \hat{S}(t)\} \right] \approx \frac{1}{[\log \hat{S}(t)]^2} \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}.$$

El error estándar (s.e.) de  $\log\{-\log \hat{S}(t)\}$  es la raíz cuadrada de esta cantidad. La aproximación

$$\frac{\log\{-\log \hat{S}(t)\} - \log\{-\log S(t)\}}{\text{s.e.} \left[ \log\{-\log \hat{S}(t)\} \right]} \sim N(0, 1)$$

permite determinar límites de  $100(1 - \alpha)\%$ , que son

$$\hat{S}(t) \exp \left\{ \pm z_{1-\alpha/2} \text{s.e.}[\log\{-\log \hat{S}(t)\}] \right\}.$$

**Ejemplo 2.3.** El error estándar y los límites de confianza de correspondiente a los datos del DIU en la Tabla 2.3 se muestran en la Tabla 2.5. En estas tabla, los

Tabla 2.5: Errores estándares e intervalos de confianza de  $\hat{S}(t)$  para los datos del DIU.

Intervalo	$\hat{S}(t)$	s.e. $[\hat{S}(t)]$	I. de C. al 95 %
0-	1.0000	0.0000	
10-	0.9444	0.0540	(0.666, 0.992)
19-	0.8815	0.0790	(0.602, 0.969)
30-	0.8137	0.0978	(0.524, 0.936)
36-	0.7459	0.1107	(0.454, 0.897)
59-	0.6526	0.1303	(0.344, 0.843)
75-	0.5594	0.1412	(0.256, 0.780)
93-	0.4662	0.1452	(0.183, 0.710)
97-	0.3729	0.1430	(0.121, 0.631)
107	0.2486	0.1392	(0.047, 0.531)

errores estándares han sido calculados con la fórmula de Greenwood y los intervalos de confianza se han calculado usando la transformación log-log.

En esta tabla se puede observar que en general el error estándar del estimado de la función de supervivencia se incrementa con el tiempo. La razón de esto es que los estimadores en los últimos tiempos se basan en menos observaciones. Una gráfica de la función de supervivencia junto con los límites de confianza se muestra en la Figura 2.2.

El procedimiento para encontrar la Tabla 2.5 y la Figura 2.2 en el lenguaje R se da a continuación:

```
# 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))

# Invoca la libreria de supervivencia
library(survival)

# Obten el estimador Kaplan-Meier como un objeto de tipo survfit
```

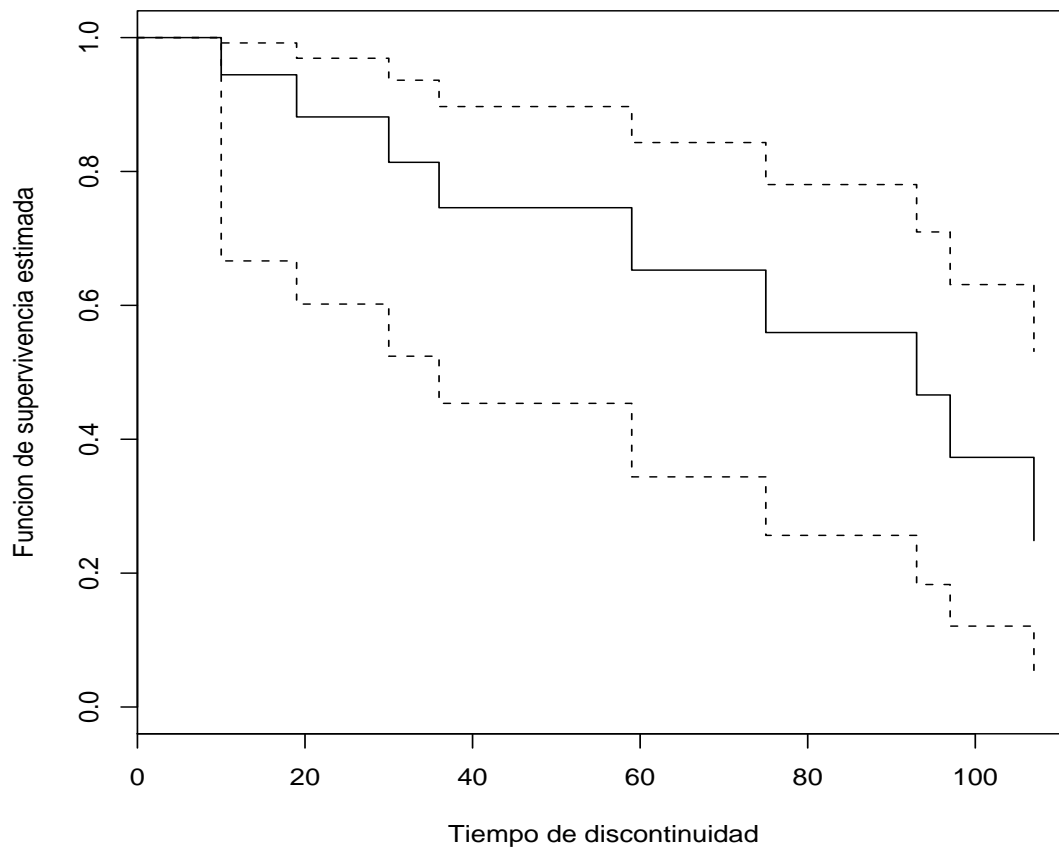


Figura 2.2: Estimador Kaplan-Meier y bandas de confianza al 95 % de la función de supervivencia para los datos del DIU.

```
# usando la formula de Greenwood para el error estandar de la curva
# y bandas de confianza del tipo "log-log".
diu.surv <- survfit(Surv(tiempo, estatus), data=diu,
error="greenwood", conf.type="log-log")

# Muestra los resultados
summary(diu.surv)

#Grafica el estimador Kaplan-Meier con sus bandas de confianza:
plot(diu.surv, conf.int=TRUE, mark.time=FALSE)
title(xlab="Tiempo de discontinuidad",
      ylab="Funcion de supervivencia estimada")
```