

2.2. Comparación de tiempos de supervivencia para dos grupos

La forma mas simple de comparar los tiempos de supervivencia de dos grupos de individuos es graficar los estimadores de las dos funciones de supervivencia correspondientes. Es posible que exista una diferencia real entre las dos curvas, lo que indica que un grupo tiene una supervivencia diferente a la del otro. Por otra parte, es posible ver pocas diferencias reales entre los grupos y que estas pequeñas diferencias sean resultado de variaciones sin mayor explicación.

Para ayudar a distinguir si existen en verdad diferencias significativas en los dos grupos, es posible llevar a cabo una prueba de hipótesis. En esta sección es pertinente concentrarse en la *prueba log-rank*. Para construir este procedimiento, se debe de comenzar por considerar los tiempos de supervivencia de los dos grupos por separado. Los grupos se etiquetarán como Grupo I y Grupo II.

Supóngase que hay r tiempos de muertes observadas en los dos grupos, $t_{(1)} < t_{(2)} < \dots < t_{(r)}$, y que en el tiempo $t_{(j)}$ hay d_{1j} muertes en el grupo I y d_{2j} muertes en el grupo II, $j = 1, 2, \dots, r$. Supóngase además que hay n_{1j} individuos en el primer grupo en riesgo de morir poco antes del tiempo $t_{(j)}$, y que hay n_{2j} en el segundo grupo. De esta forma, hay $d_j = d_{1j} + d_{2j}$ muertes en $t_{(j)}$ de un total de $n_j = n_{1j} + n_{2j}$ individuos en riesgo. Este escenario está resumido en la Tabla 2.6.

Tabla 2.6: Número de muertes en el j -ésimo tiempo de muerte en cada uno de los grupos de individuos.

Grupo	Número de muertes en $t_{(j)}$	Número de sobrevivientes después de $t_{(j)}$	Número en riesgo poco antes de $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Considérese la hipótesis nula de que no hay diferencia entre los tiempos de supervivencia de los dos grupos. Una forma de evaluar la validez de esta hipótesis es considerar las desviaciones entre el número observado de individuos en los dos grupos que mueren en cada tiempo de muerte, y el número esperado bajo la hipótesis nula. La información de estas desviaciones pueden entonces combinarse con todas los tiempos de muerte.

Si los totales marginales en la Tabla 2.6 se toman como fijos, y si la hipótesis nula es verdadera, entonces las cuatro entradas de la tabla son determinadas solamente por el número de muertes en el grupo I en $t_{(j)}$, d_{1j} . De esdta froma es posible tomar a d_{1j} como una variable aleatoria, la cual puede tomar valores entre cero y el mínimo de d_j y n_{1j} . Bajo estas circunstancias, d_{1j} se distribuye *hipergeométrica*, $d_{1j} \sim \text{HIP}(n_{1j}, d_j, n_j)$, y entonces la función de probabilidad correspondiente es la probabilidad de que el número de muertes en el primer grupo tome el valor d_{1j} de un total de n_j individuos en riesgo cuando hay $n_j - d_j$ individuos que sobreviven y n_{1j} individuos en el primer grupo que estan en riesgo de morir, que es

$$\frac{\binom{d_j}{d_{1j}} \binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}.$$

La media de la variable aleatoria hipergeométrica d_{1j} está dada por

$$e_{1j} = n_{1j}d_j/n_j,$$

que es el número esperado de individuos que mueren en el grupo I en $t_{(j)}$. Ahora, bajo la hipótesis nula de que la probabilidad de muerte en $t_{(j)}$ no depende del grupo al que un individuo pertenece, la probabilidad de muerte en $t_{(j)}$ es d_j/n_j , y multiplicando esta cantidad por n_{1j} se obtiene e_{1j} .

El siguiente paso es obtener un estadístico que resuma las desviaciones entre lo observado y lo esperado. La forma mas simple de hacerlo es calcular la suma de las

diferencias $d_{1j} - e_{1j}$ para $r = 1, 2, \dots, r$. El estadístico resultante es

$$U_L = \sum_{j=1}^r (d_{1j} - e_{1j}).$$

Este estadístico tiene esperanza cero ya que $E[d_{1j}] = e_{1j}$. La varianza del estadístico U_L es

$$V_L = \text{Var}[U_L] = \sum_{j=1}^r v_{1j},$$

donde v_{1j} es la varianza de d_{1j} dada por

$$v_{1j} = \text{Var}[d_{1j}] = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Cuando el número de muertes es relativamente grande U_L se distribuye aproximadamente normal y entonces $U_L/\sqrt{V_L}$ se distribuye aproximadamente normal estándar; esto es

$$\frac{U_L}{\sqrt{V_L}} \sim N(0, 1) \quad \text{cuando } n \rightarrow \infty.$$

El cuadrado de esta variable aleatoria se distribuye asintóticamente ji cuadrada con un grado de libertad:

$$W_L = \frac{U_L^2}{V_L} \sim \chi^2(1).$$

Por lo tanto, W_L resume las desviaciones que hay entre los tiempos de supervivencia observados en los dos grupos y los esperados bajo la hipótesis nula de no diferencias entre los dos grupos. Entre mas grandes sean los valores del estadístico W_L mayor será la evidencia en contra de la hipótesis nula. La hipótesis nula se rechaza si la W_L observada con los datos excede el valor crítico $\chi_{1-\alpha}$, donde α es el nivel de significancia y $\chi_{1-\alpha}$ es el percentil $1 - \alpha$ de una ji cuadrada con un grado de libertad. El *valor p* correspondiente se calcula como

$$\text{valor } p = \Pr\{\chi^2 > W_L\}, \quad \text{donde } \chi^2 \sim \chi^2(1).$$

Ejemplo 2.4. La Tabla 2.7 muestra tiempos de supervivencia (en meses) de mujeres que recibieron una mastectomía para tratar un tumor de grado II, III, IV, entre enero de 1969 y diciembre de 1971. Los tiempos de supervivencia se encuentran

Tabla 2.7: Tiempos de supervivencia de mujeres conlog.rank tumores que registraron marcas positivas y negativas de HPA.

Marca negativa	Marca positiva	
23	5	68
47	8	71
69	10	76*
70*	13	105*
71*	18	107*
100*	24	109*
101*	26	113
148	26	116*
181	31	118
198*	35	143
208*	40	154*
212*	41	162*
224*	48	188*
	50	212*
	59	217*
	61	225*

clasificados de acuerdo a si el cáncer fué o no marcado con *Helix pomatia agglutinin* (HPA), una marca que indica si el cáncer mamario primario tiene metastásis o no.

El interés principal en este estudio es el de determinar si hay una diferencia significativa entre los tiempos de supervivencia de los dos grupos. La Figura 2.3 muestra los estimadores Kaplan-Meier de las funciones de supervivencia para cada grupo. Es posible notar que existen ciertas diferencias, las mujeres con una marca negativa de HPA tienden a sobrevivir mas que las que tienen una marca positiva.

Las cantidades relevantes para construir el estadístico de prueba se muestran en la Tabla 2.8. En este caso se obtiene $\sum d_{1j} = 5$, $\sum e_{1j} = 9,565$, $V_L = 5,929$. Por lo tanto, el valor observado del estadístico W_L es de 3.515. El valor p correspondiente es de 0.0608, lo que inspira duda sobre la hipótesis nula de que no hay diferencia entre las funciones de supervivencia de los dos grupos de mujeres. Se puede, entonces, concluir que hay cierta evidencia de que el diagnóstico de una paciente de cáncer mamario depende del resultado del estatus de la marca de HPA.

En el lenguaje R de programación es posible llevar a cabo este análisis usando la función `survdif`. A continuación se ilustra el ejemplo de los datos en la Tabla 2.7.

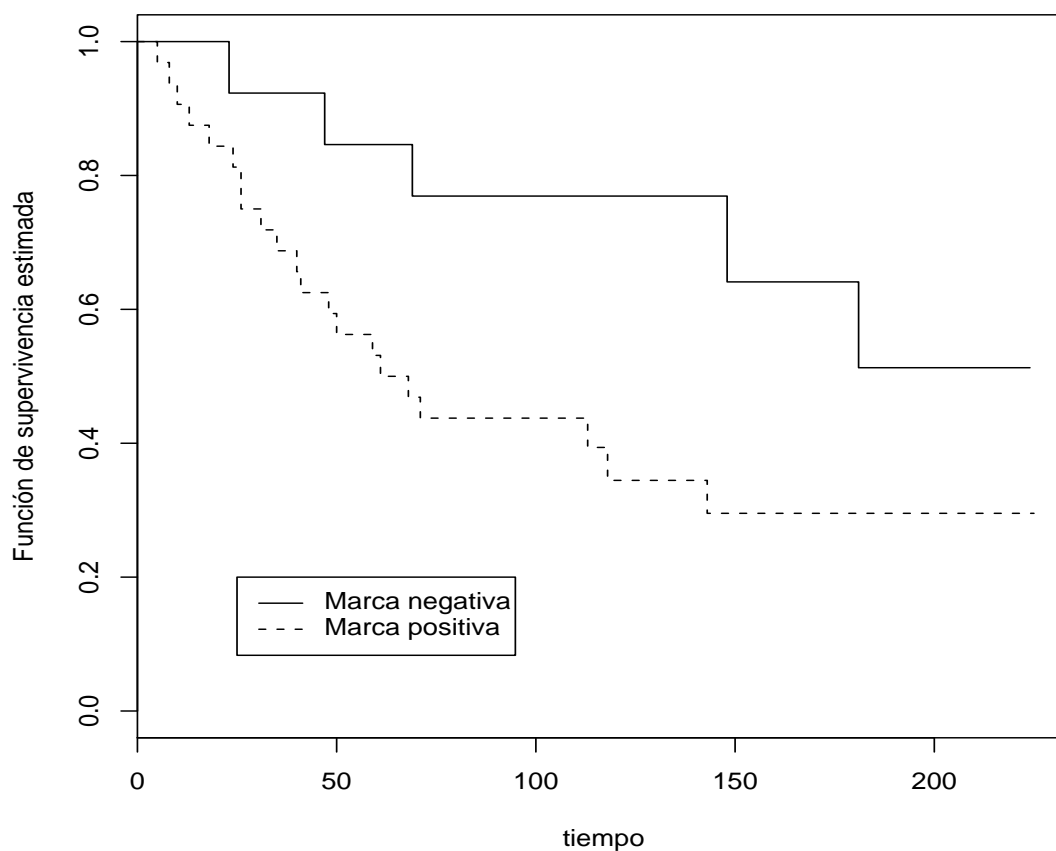


Figura 2.3: Estimador Kaplan-Meier de la función de supervivencia para los datos de cáncer de mujeres clasificado por marca de HPA.

```
#Escribe los datos en un objeto data.frame:
hpa <- data.frame(tiempo=c(23, 47, 69, 70, 71, 100, 101, 148, 181,
                          198, 208, 212, 224,
                          5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50,
                          59, 61, 68, 71, 76, 105, 107, 109, 113, 116, 118,
                          143, 154, 162, 188, 212, 217, 225),
                  censura = c(1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
                              rep(1, 16), 1, 1, 0, 0, 0, 0, 1, 0, 1, 1,
                              rep(0, 6)),
                  marca=c(rep(-1,13), rep(1, (16*2))))

library(survival)
```

2.2. COMPARACIÓN DE TIEMPOS DE SUPERVIVENCIA PARA DOS GRUPOS 33

Tabla 2.8: El cálculo del estadístico log-rank para los datos de cancer con marcas positivas o negativas de HPA.

Tiempo de muerte	d_{1j}	n_{1j}	d_{2j}	n_{2j}	d_j	n_j	e_{1j}	v_{1j}
5	0	13	1	32	1	45	0.2889	0.2054
8	0	13	1	31	1	44	0.2955	0.2082
10	0	13	1	30	1	43	0.3023	0.2109
13	0	13	1	29	1	42	0.3095	0.2137
18	0	13	1	28	1	41	0.3171	0.2165
23	1	13	0	27	1	40	0.3250	0.2194
24	0	12	1	27	1	39	0.3077	0.2130
26	0	12	2	26	2	38	0.6316	0.4205
31	0	12	1	24	1	36	0.3333	0.2222
35	0	12	1	23	1	35	0.3429	0.2253
40	0	12	1	22	1	34	0.3529	0.2284
41	0	12	1	21	1	33	0.3636	0.2314
47	1	12	0	20	1	32	0.3750	0.2314
48	0	11	1	20	1	31	0.3548	0.2289
50	0	11	1	19	1	30	0.3667	0.2322
59	0	11	1	18	1	29	0.3793	0.2354
61	0	11	1	17	1	28	0.3929	0.2385
68	0	11	1	16	1	27	0.4074	0.2414
69	1	11	0	15	1	26	0.4231	0.2441
71	0	9	1	15	1	24	0.3750	0.2344
113	0	6	1	10	1	16	0.3750	0.2344
118	0	6	1	8	1	14	0.4286	0.2449
143	0	6	1	7	1	13	0.4615	0.2485
148	1	6	0	6	1	12	0.5000	0.2500
181	1	5	0	4	1	9	0.5556	0.2469
Total	5						9.5652	5.9289

```
#Obten los estimadores Kaplan-Meier para los dos grupos
hpa.survfit <- survfit(Surv(tiempo, censura) ~ marca, data= hpa)

#Dibuja la grafica de las dos curvas
plot(hpa.survfit, mark.time=FALSE, lty=1:2)
title(xlab="tiempo", ylab="Funcin de supervivencia estimada")
legend(25, 0.2, legend=c("Marca negativa", "Marca positiva"), lty=1:2)

#Realiza la prueba log-rank
hpa.survdif <- survdiff(Surv(tiempo, censura) ~ marca, data= hpa, rho=0)
```

