Revista de Matemática: Teoría y Aplicaciones 2010 **17**(1) : 41–52
cimpa – ucr issn: 1409-2433

SELECCIÓN AUTOMÁTICA DEL p-VALOR EN LA COMPARACIÓN DE CURVAS DE SUPERVIVENCIA

AUTOMATIC SELECTION OF THE p-VALUE IN SURVIVAL CURVES COMPARISON

Pablo Martínez–Camblor*

Received: 13 Nov 2008; Revised: 27 Oct 2009; Accepted: 30 Oct 2008

Keywords: Curvas de supervivencia, Tests para k-muestras, Algoritmo doble mínimo, familia de Fleming y Harrington.

Palabras clave: Survival curves, k-sample tests, Double minimum algorithm, Fleming and Harrington family.

Mathematics Subject Classification: 62N03, 62G09.

---

*CAIBER, Oficina de Investigación Biosanitaria del Principado de Asturias. Rosal 7 Bis. 33009 Oviedo. Asturias, España. E-mail: pmcamblor@hotmail.com

Resumen

En este trabajo se desarrolla un algoritmo para la selección automática de tests para la comparación de curvas de supervivencia. El procedimiento introducido es una adaptación del algoritmo *doble mínimo* para la selección del parámetro de suavizado en *tests suaves* para la comparación de k-muestras independientes. El estudio de simulación realizado, sugiere que el método propuesto, sin llegar a igualar los resultados de una elección óptima en ninguna de las situaciones consideradas, es el más regular entre todos los tests estudiados.

Abstract

In this paper, an algorithm for the automatic selection of an adequate test for the survival curves comparison is developed. The introduced procedure is an adaptation of the *double minimum* algorithm for the bandwidth selection in the smoothed nonparametric k-sample tests. The simulation study which was carried out, suggests us that the proposed method, although never is as good as the best one of the considered tests, is the most regular of them.

1 Introducción

En numerosos estudios, en especial en los de carácter biomédico, la variable objetivo es un tiempo de fallo que, de forma usual, suele no ser conocida en su totalidad para una porción de los individuos de la muestra. En este tipo de estudios, el estimador más popular para la función de distribución, F, es el bien conocido estimador de Kaplan-Meier, $\hat{F}_{K_M}$. La comparación de dos o mas curvas de supervivencia $(1 - F)$ ha sido ampliamente estudiada. Los tests no paramétricos más populares, también conocidos como *tests ponderados* o *linear rank tests* (Harrington; 2005), están basados en una suma ponderada de las diferencias entre las supervivencias observadas y esperadas a lo largo del tiempo de estudio. Para el tiempo t_i la expresión de estos estadísticos viene dada por la fórmula

$$L_R(w) = \sum_{j=1}^{k} \sum_{i=1}^{D} w_i \left(d_{ij} - Y_{ij} \left(\frac{d_j}{Y_j} \right) \right)^2 \tag{1}$$

donde d_{ij} y d_i denotan, respectivamente, el número de eventos en el grupo j, $(1 \leq j \leq D)$ y en la muestra conjunta inmediatamente antes de t_i e Y_{ij} e Y_i son el número de sujetos a riesgo en la muestra j y en la muestra conjunta inmediatamente antes de t_i, respectivamente y donde w_i con

$1 \leq i \leq D$ son los pesos. Probablemente, la familia de tests más popular sea la conocida como G^ρ, introducida por Fleming y Harrington (2002) y, en la cual, los pesos vienen dados por $w_i(\rho) = \hat{S}(t_{i-1})^\rho$, donde $\hat{S}(t_{i-1})$ es el estimador de Kaplan-Meier para la curva de supervivencia conjunta en el punto t_{i-1}, resultando tests de la forma,

$$G^\rho = \sum_{j=1}^{k} \sum_{i=1}^{D} \hat{S}(t_{i-1})^\rho \left(d_{ij} - Y_{ij} \left(\frac{d_j}{Y_j} \right) \right)^2 \quad \rho \geq 0 \tag{2}$$

Sin duda, en la práctica, el test más usado para comparar curvas de supervivencia es el *log-rank* test, también conocido como test de Mantel-Haenszel ($\rho = 0$). Sin embargo, dependiendo del lugar en el que se localicen las diferencias entre las curvas existen tests más potentes. Así las cosas, es sabido (Letón y Zuluaga; 2002) que existe gran variabilidad en la potencia obtenida por cada test en función del escenario que se presente. A grandes rasgos, se puede decir que; en escenarios de diferencias proporcionales, el test más potente es el log-rank, si las diferencias son tempranas, el test de Peto-Peto ($\rho = 1$) es el más potente, si estamos en el escenario de diferencias tardías, el test log-rank vuelve a ser el ganador y, si las diferencias se presentan en la mitad de la curva, el test de Tarone-Ware es el más potente. El test de Tarone-Ware ($\rho = 1/2$) parece ser un test con un comportamiento intermedio en todos los escenarios.

A pesar de que existen distintas recomendaciones de cuando usar cada test (ver, por ejemplo, Léton y Zuluaga; 2006) esta elección no puede hacerse *a priori*, lo que hace que no se pueda garantizar que el tamaño inicial del test está siendo respetado.

En este trabajo, se propone un procedimiento que, para una familia determinada de tests, devuelve una significación final basada en corregir el menor de los P-valores obtenidos por los tests considerados. Su potencia se estudia mediante simulaciones de Monte Carlo (Sección 4) observándose que, este algoritmo, adaptación del método *doble mínimo* (DM) introducido por Martínez-Camblor y de Uña-Álvarez (2009) en el contexto de la selección del parámetro ventana en tests suaves para k-muestras, obtiene buenos resultados en todas las situaciones anteriormente descritas. Finalmente, en la Sección 5, se aplica el procedimiento a un caso real en el que se observa lo útil que puede resultar el procedimiento descrito en algunas situaciones.

2 Adaptación del algoritmo *doble mínimo*

Un problema usual en las técnicas suaves, es la selección del parámetro de suavizado o parámetro ventana. Con este propósito, y en el contexto de la comparación de k densidades desde poblaciones independientes, han sido propuestos diversos métodos, si bien, todos ellos (Martínez-Camblor; 2008), tratan de elegir el valor más adecuado entre una malla de posibles valores. Este tipo de algoritmos, se pueden aplicar directamente a la selección de otro tipo de parámetros, en concreto, a la selección del parámetro ρ en la familia de estadísticos dada en (2). Si se considera el método DM anteriormente citado, la adaptación de este algoritmo al caso considerado quedaría de la forma:

A. Se selecciona una malla de valores del parámetro, $\mathcal{P} = \{P_1, \ldots, P_t\}$.

B. Se calculan los P-valores del test para todos los elementos de $\mathcal{P}$, p_i $(1 \leq i \leq t)$.

C. Se elige el p_M de modo que $p_M = min\{p_1, \ldots, p_t\}$.

D. Se generan B muestras bootstrap y, para cada una de ellas, se repiten los pasos B y C obteniéndose B P-valores mínimos: $p_M^1, \ldots, p_M^B$.

E. Se computa el P-valor *doble mínimo* final de la forma,

$$p_F = \frac{1}{B} \sum_{i=1}^{B} I\{p_M > p_M^i\}$$

Con este método, se hace una corrección final del P-valor teniendo en cuenta no solamente el número de tests (t) considerados sino también las relaciones existentes entre estos. Teóricamente, el P-valor devuelto por el algoritmo no será inferior al menor de los P-valores obtenidos por los tests considerados (en la práctica, por efecto del remuestreo, pueden aparecer P-valores sensiblemente menores), no siendo necesario aplicarlo en aquellos casos en los que el no rechazo de la hipótesis nula sea claro.

3 Estudio de simulación

En esta sección se estudia la potencia obtenida por el método propuesto a partir de un estudio de simulación. Se consideran las siguientes funciones,

(i) $f_0(x) = Exp(1/4)$.

(ii) $f_1(x) = Exp(1/5)$.

(iii) $f_2(x) = (2/5)Exp(1/2) + (3/5)Exp(1/20)$.

(iv) $f_3(x) = \chi^2_5$.

(v) $f_4(x) = (7/10)Exp(1/3) + (3/10)N(8, 3/2)$.

donde $Exp(\lambda)$ denota la función de densidad de una distribución exponencial de parámetro λ, χ^2_f es la densidad de una distribución χ-cuadrado con f grados de libertad y $N(\mu, \sigma)$ es la densidad de una distribución normal con media μ y varianza σ^2. Tomando f_0 como función de referencia, las funciones consideradas representan cuatro situaciones diferentes: f_1; diferencias proporcionales (PH), f_2; diferencias tardías (LDH), f_3; diferencias tempranas (EDH) y f_4; diferencias en la parte media de las curvas (MDH). En la Figura 1 puede verse una representación gráfica de las distintas formas de las curvas de supervivencia obtenidas desde las funciones de densidad consideradas.

Una vez generados los tiempos de fallo, se simularon los tiempos de censura desde una distribución uniforme en el intervalo $[0, u]$, considerándose los casos $u = 10$, y $u = 15$. El nivel de confianza considerado fue $\alpha = 0.05$. El porcentaje de rechazos fue estimado desde 5,000 simulaciones de Monte Carlo y los P–valores obtenidos por el método DM fueron aproximados desde 501 iteraciones ($B = 501$). Se consideró la malla $\mathcal{P} = \{0, 1/5, 2/5, 3/5, 4/5, 1\}$. Los P-valores para cada test fueron calculados desde su distribución asintótica (χ-cuadrado con $k - 1$ grados de libertad). Se ha incluido una condición de parada de modo que, en el caso en que todos los P-valores sean superiores a 0.1, el algoritmo no se ejecuta, la hipótesis nula no se rechaza. El algoritmo empleado, codificado en R (R Development Core Team; 2006) se adjunta como apéndice.

Ejemplo 1. En primer lugar, consideraremos problemas con dos muestras independientes. En este caso, las muestras aleatorias, (X_1, X_2), de tamaño $n = (n_1, n_2) = (50, 50)$, son generadas desde las funciones siguientes:

MD 0-I. $X_1 \sim f_0, X_2 \sim f_0$.

MD 1-I. $X_1 \sim f_0, X_2 \sim f_1$.

MD 2-I. $X_1 \sim f_0, X_2 \sim f_2$.

MD 3-I. $X_1 \sim f_0, X_2 \sim f_3$.

MD 4-I. $X_1 \sim f_0, X_2 \sim f_4$.

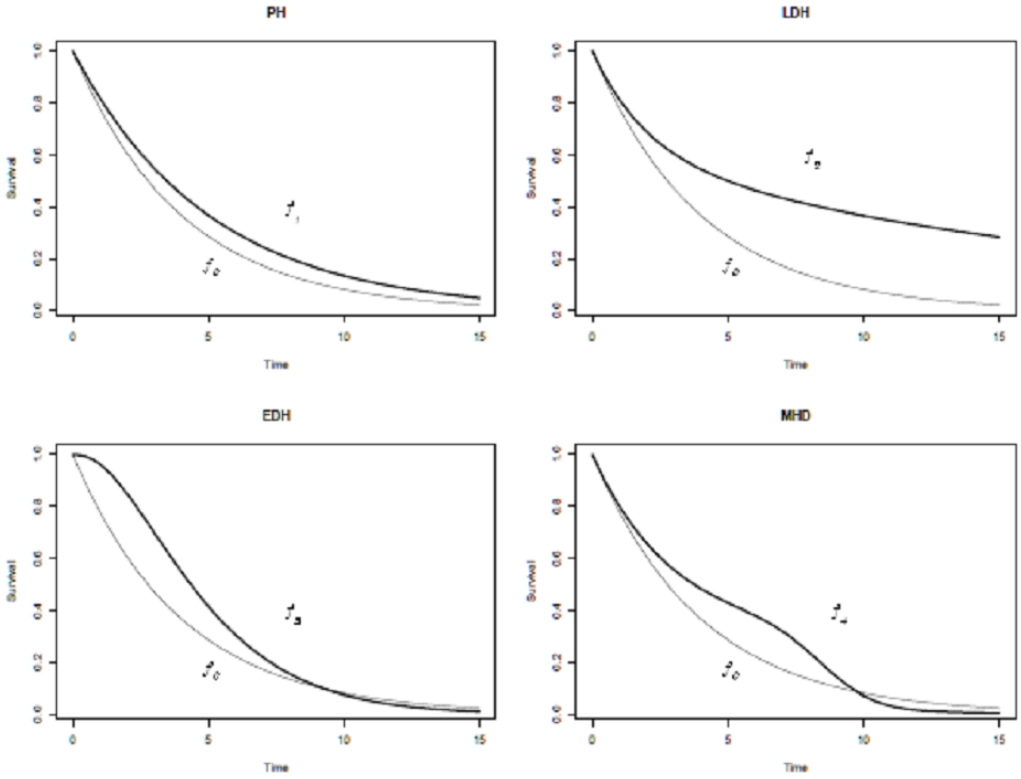

Figura 1: Representación gráfica de las funciones consideradas.

		$\boldsymbol{p}$							$\boldsymbol{P}_M$	
MD	u	0.0	0.2	0.4	0.6	0.8	1.0	DM	Media	SD
0-I	10	0.055	0.053	0.054	0.055	0.053	0.054	0.054	0.486	0.475
0-I	15	0.049	0.048	0.048	0.047	0.048	0.048	0.049	0.479	0.473
1-I	10	0.132	0.133	0.130	0.127	0.126	0.123	0.131	0.455	0.472
1-I	15	0.155	0.153	0.149	0.145	0.140	0.133	0.148	0.439	0.465
2-I	10	0.477	0.440	0.409	0.371	0.345	0.314	0.440	0.099	0.286
2-I	15	0.680	0.629	0.576	0.525	0.472	0.422	0.640	0.034	0.169
3-I	10	0.481	0.545	0.599	0.649	0.691	0.724	0.683	0.955	0.196
3-I	15	0.384	0.467	0.545	0.609	0.655	0.700	0.651	0.958	0.189
4-I	10	0.186	0.176	0.170	0.160	0.151	0.142	0.167	0.383	0.452
4-I	15	0.142	0.151	0.155	0.154	0.154	0.148	0.152	0.496	0.436

Tabla 1: Proporción de rechazos observada en 5,000 simulaciones de Monte Carlo. Se considera la malla $\mathcal{P} = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, α =0.05 y $n = (50, 50)$. Las dos últimas columnas indican la media y la desviación standard (SD) para P_M.

En el Cuadro 1 pueden verse las potencias obtenidas tanto por los seis tests considerados como por el algoritmo DM propuesto. En primer lugar destacar que el nivel de significación es respetado por el procedimiento en estudio. El porcentaje de rechazos cuando la hipótesis nula es cierta (MD 0-I), 5.5%, no es el mayor de los observados. Por otro lado, reseñar que los resultados obtenidos por el método DM son siempre intermedios, su porcentaje de rechazos esta siempre más próximo al mejor que al peor de los tests considerados. Este hecho se ve acentuado cuando las diferencias entre estos es mayor (modelos 2-I y 3-I).

Ejemplo 2. A continuación, se aborda un problema con tres muestras independientes. Las muestras aleatorias, (X_1, X_2, X_3), de tamaño $n = (n_1, n_2, n_3) = (50, 50, 50)$, son generadas desde las funciones siguientes:

MD 0-II. $X_1 \sim f_0, X_2 \sim f_0, X_3 \sim f_0$.

MD 1-II. $X_1 \sim f_0, X_2 \sim f_0, X_3 \sim f_1$.

MD 2-II. $X_1 \sim f_0, X_2 \sim f_0, X_3 \sim f_2$.

MD 3-II. $X_1 \sim f_0, X_2 \sim f_0, X_3 \sim f_3$.

MD 4-II. $X_1 \sim f_0, X_2 \sim f_0, X_3 \sim f_4$.

MD 5-II. $X_1 \sim f_0, X_2 \sim f_1, X_3 \sim f_2$.

		$\boldsymbol{p}$							$\boldsymbol{P}_M$	
MD	u	0.0	0.2	0.4	0.6	0.8	1.0	DM	Media	SD
0-II	10	0.054	0.054	0.052	0.051	0.051	0.050	0.053	0.486	0.474
0-II	15	0.057	0.058	0.058	0.056	0.056	0.056	0.055	0.496	0.474
1-II	10	0.127	0.127	0.125	0.121	0.117	0.116	0.126	0.455	0.424
1-II	15	0.152	0.146	0.142	0.139	0.134	0.130	0.137	0.420	0.463
2-II	10	0.500	0.463	0.426	0.385	0.344	0.312	0.458	0.089	0.273
2-II	15	0.706	0.649	0.592	0.526	0.469	0.410	0.661	0.028	0.157
3-II	10	0.481	0.552	0.615	0.669	0.716	0.751	0.701	0.957	0.187
3-II	15	0.387	0.483	0.570	0.631	0.684	0.728	0.667	0.963	0.178
4-II	10	0.175	0.171	0.164	0.154	0.146	0.135	0.161	0.363	0.448
4-II	15	0.145	0.155	0.160	0.157	0.151	0.144	0.153	0.488	0.437
5-II	10	0.397	0.365	0.333	0.306	0.280	0.260	0.360	0.139	0.333
5-II	15	0.592	0.536	0.481	0.431	0.378	0.344	0.540	0.053	0.211
6-II	10	0.377	0.434	0.484	0.531	0.575	0.616	0.563	0.928	0.248
6-II	15	0.282	0.342	0.404	0.471	0.525	0.576	0.513	0.917	0.269
7-II	10	0.159	0.156	0.151	0.143	0.138	0.134	0.147	0.407	0.461
7-II	15	0.158	0.155	0.153	0.151	0.145	0.139	0.153	0.453	0.455

Tabla 2: Proporción de rechazos observada en 5,000 simulaciones de Monte Carlo. Se considera la malla $\mathcal{P} = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, α =0.05 y $n = (50, 50, 50)$. Las dos últimas columnas indican la media y la desviación standard (SD) para P_M.

MD 6-II. $X_1 \sim f_0, X_2 \sim f_1, X_3 \sim f_3$.

MD 7-II. $X_1 \sim f_0, X_2 \sim f_1, X_3 \sim f_4$.

El Cuadro 2 muestra la proporción de rechazos observada tanto en los seis tests considerados como en el algoritmo propuesto. En primer lugar, destacar que, nuevamente, el nivel de significación es respetado. Por otro lado, los resultados obtenidos por el DB son notables. Al igual que en el ejemplo anterior, a pesar de que en algunos modelos ocupa posiciones bajas (hay cuatro tests que consiguen mejores potencias a la obtenida por el método DB en el modelo 1-II para $u = 15$), las pequeñas diferencias que se observan en las potencias cuando esto sucede, hace que el método propuesto sea, globalmente, el más regular entre todos los considerados.

4 Análisis de datos reales

Como ejemplo, y con el fín de ilustrar algunas ventajas en el análisis de datos biomédicos del método propuesto, se considerará un conjunto de

datos previamente analizados por Martínez-Camblor et al. (2009). Los datos proceden de un estudio sobre cáncer de mama. Como suele ser usual en este tipo de estudios, se está interesado en conocer la posible influencia que una serie de variables tiene sobre la supervivencia de mujeres diagnosticadas con esta patología. Se consideran un total de 418 mujeres diagnosticadas con cáncer de mama en el periodo de 1990 a 1995 en Gipuzkoa (norte de España). El seguimiento realizado fue de 10 años censurándose un total de 243 casos (60.5%).

Entre los factores considerados, en este ejemplo nos centraremos en los efectos de la Edad (agrupada en: 15-39 años, 40-54 años, 55-69 años y más de 70 años) y una variable que recoge la respuesta a dos Receptores Hormonales; Progesterona y Estrógeno, (agrupada en: Ambos positivos, Ambos negativos, Al menos uno positivo, Otros). En la Figura 2, pueden verse las curvas de supervivencia con respecto a la Edad (arriba) y a los Receptores Hormonales (abajo).

Aunque, con independencia del peso utilizado, es clara la influencia de estas variables sobre la supervivencia (todos los P-valores están próximos a cero) surgen problemas cuando se consideran pares de curvas. En concreto, si se estudian las diferencias entre el grupo de edad de 15-39 años frente al grupo de 55-69 años (líneas negras en la gráfica superior de la Figura 2) se tiene que el test log-rank rechaza la hipótesis de igualdad, el mismo resultado se obtiene para ρ =0.2, para ρ =0.4 el P-valor es no significativo al 5% aunque *borderline* y, la hipótesis nula no se rechaza para valores de ρ mayores (ver Cuadro 3). Se encuentra una situación similar si se pretende comparar el grupo de *Ambos Receptores Hormonales Negativos* frente a *Otros* (este grupo recoge, principalmente, situaciones en las que se desconoce la clasificación correcta). En este caso, se obtienen P-valores inferiores a 0.05 para valores de ρ superiores a 0.2 y diferencias no significativas al 5% para ρ =0.0 y ρ =0.2 (ver Cuadro 3). Así las cosas, resultaría difícil justificar, *a priori*, porque se utilizan tests diferentes en cada caso (notar que ningún test entre los seis considerados rechaza ambas hipótesis nulas simultáneamente) y, *a posteriori* (una vez vistas las curvas), no se podría garantizar que el nivel de significación es el fijado de antemano.

Aplicando el algoritmo descrito en la Sección 2 de este trabajo (se ha considerado $B = 5,001$), se obtienen P-valores de 0.048 y de 0.045, respectivamente. Esto permite afirmar que, en ambos casos, las diferencias son significativas al 5%.

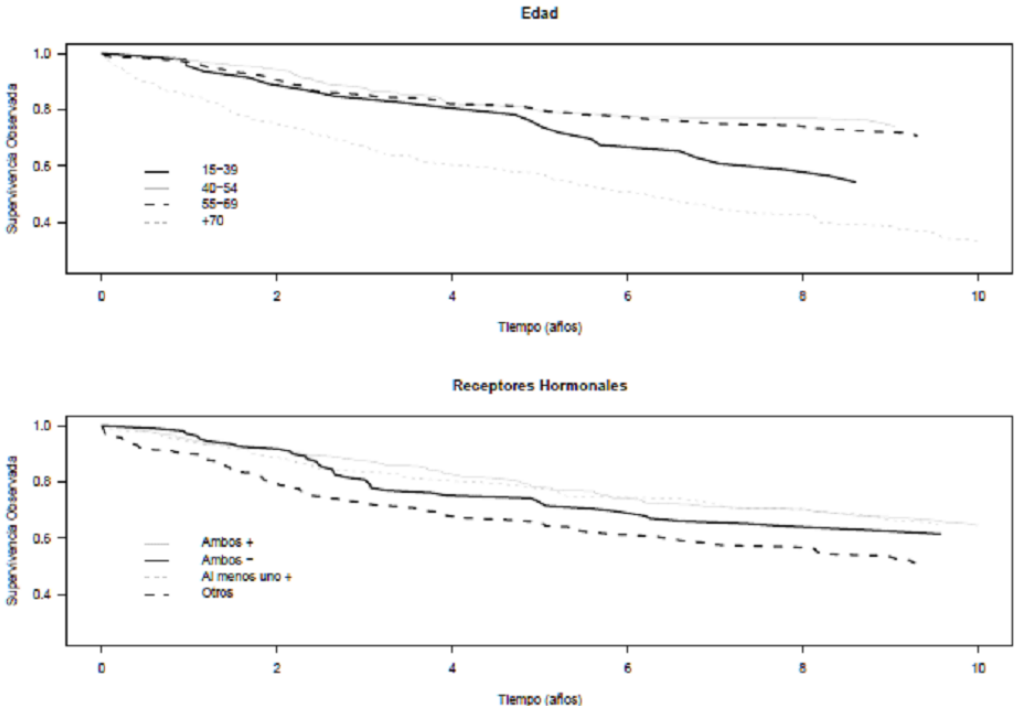

Figura 2: Curvas de Supervivencia por grupos de Edad (superior) y por Receptores Hormonales (inferior).

Variables	p						
	0.0	0.2	0.4	0.6	0.8	1.0	*DM*
Edad:							
14-39 vs. 55-69.	0.044	0.048	0.052	0.057	0.062	0.068	0.048
RH[1]: Al menos 1+							
vs. Otros	0.057	0.053	0.049	0.045	0.041	0.038	0.045

[1] Receptores Hormonales

Tabla 3: P-valores obtenidos por los distintos tests considerados y por el método DM. se considera la malla $\mathcal{P} = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

5 Principales conclusiones

En este trabajo se ha presentado un algoritmo para la toma de decisiones en contrastes de k-curvas de supervivencia en presencia de información incompleta. El método desarrollado, basado en $\mathcal{M} = min_p\{\mathcal{P}\{G^p > g^p\}\}$ no es solamente aplicable a una familia de tests (dependientes de un parámetro) sino que también se puede aplicar a grupos de tests (con diferentes naturalezas) para contrastar la misma hipótesis nula. Mediante remuestreo, se aproxima la distribución de $\mathcal{M}$. Un estudio de simulación muestra que, en los casos considerados, los resultados obtenidos por este método son muy regulares y altamente competitivos en una amplia gama de situaciones. Si bien, en alguno de los escenarios considerados la potencia estadística conseguida por el DM no está entre las mejores, la diferencia nunca es excesiva, y el método propuesto se muestra como el más regular entre los estudiados. Por otro lado, el método DB tiene la indiscutible ventaja de no tener que tomar una decisión *a priori* sobre el test a utilizar, lo que garantiza que se trabaja al tamaño inicialmente fijado.

Agradecimientos

El autor quiere expresar su agradecimiento al Registro de Cáncer de Gipuzkoa y, en especial, a Nerea Larrañaga por permitir el uso de los datos de cáncer de mama empleados en la Sección 5 del presente trabajo.

Referencias

[1] Harrington D.P.; Fleming, T.R. (1982) "A class of rank test procedures for censored survival data", *Biometrika* **69**: 553–566.

[2] Letón, E.; Zuluaga, P. (2006) "Cómo elegir el test adecuado para comparar curvas de supervivencia", *Medicina Clínica* **127**(3): 96–99.

[3] Letón, E.; Zuluaga, P. (2002) "Survival tests for r groups", *Biometrical Journal* **44**: 15–27.

[4] Martínez–Camblor, P.; Larrañaga, N.; Sarasqueta, C.; Basterretxea, M. (2009) "Esa corporeidad mortal y rosa. Análisis del cáncer de mama en Gipúzkoa en presencia de riesgos competitivos", por aparecer en *Gaceta Sanitaria*.

[5] Martínez-Camblor, P. (2008) "Estudio sobre los efectos del parámetro de suavizado en contrastes no-paramétricos para k– Muestras", *Revista Colombiana de Estadística* **31**(2): 157–168.

[6] Martínez-Camblor, P.; de Uña-Álvarez, J. (2009) "Nonparametric k-sample tests: density functions vs. distribution functions", *Computational Statistics & Data Analysis* **53**(9): 3344–3357.

[7] R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.r-project.org`

Apéndice: Una rutina en R

Para ejecutar las simulaciones presentadas en este trabajo se ha desarrollado una rutina en el paquete estadístico de libre distribución R 2.6.1 (R Development Core Team; 2006). Se ha usado el paquete *survival* (disponible en el CRAN de `www.r-project.org`), en concreto, la función *survdiff* que implementa la familia de test propuesta por Harrington y Fleming (1982). En la siguiente rutina, la matriz *(s,c)* contine el tiempo hasta el fallo ($c = 1$) o el tiempo hasta la censura ($c = 0$), el vector n, contiene los diferentes tamaños muestrales, x es el vector que indica a que grupo pertenece cada individuo, p contine los valores de la malla considera y, finalmente, B_0 es el número de iteraciones a realizar. La rutina devuelve un vector con los P-valores obtenidos por cada uno de los tests considerados, el P-valor obtenido por el método DM y el valor que ha obtenido el P-valor más bajo.

```
BM<-function(s,c,n,x,p,B0)
{l<-length(p);sig<-rep(0,l);f<-rep(-1,(B0+1));
sf<-rep(-1,l);k<-x[length(x)];pf<-␣-1;
nn<-c(0,cumsum(n))
for(j␣in␣1:l)
{sf[j]<-1-pchisq(survdiff(Surv(s,c)~x,rho=p[j])[[5]],k-1)}
f[1]<-min(sf);for(j␣in␣1:l){pf<-ifelse(f==sf[j],p[j],pf)};
if(f[1]>0.1){return(c(sf,f[1],pf))}
else{
for(b␣in␣1:B0)
{pos<-sample(1:sum(n),replace=T)
for(i␣in␣1:l)
{sig[i]<-
1-pchisq(survdiff(Surv(s[pos],c[pos])~x,rho=p[i])[[5]],k-1)}
f[b+1]<-min(sig)}
return(c(sf,(rank(f)[1]-1)/B0,pf))}};
```