Ciencia en Desarrollo, Vol. 9 No. 1
ISSN 0121 - 7488 - Enero a Junio de 2018

Una comparación de pruebas de igualdad de dos riesgos competitivos

A comparison test of equality of two competing risks

Liliana C. Molina-Blanco [a*]
Carlos M. Lopera-Gómez [b*]

Fecha de Recepción: 29 - nov. - 2016
Fecha de Aceptación: 12 - sep. - 2017.

Resumen

En este artículo se abordó la problemática de dos riesgos que están compitiendo para causar la falla de un sujeto; en particular determinar si los riesgos o probabilidad de falla asociada a cada tipo de falla son igualmente importantes o si un riesgo es más serio que el otro. Para este fin se hizo un estudio de la prueba de hipótesis para la igualdad de las dos funciones de incidencia acumulada asociadas a los riesgos. Se realizó un estudio de simulación donde se comparan algunos de los procedimientos de prueba que han sido propuestos para este fin; y así, poder determinar el comportamiento de estos procedimientos de prueba bajo varios escenarios que permitan evaluar el desempeño de los mismos. Se incluye una aplicación de los procedimientos de prueba usando datos reales de pacientes con linfoma.

***Palabras clave*:** Función de incidencia acumulada, tasa de riesgo de causa-específica, bootstrap, aproximación de simetrización aleatoria, remuestreo y supremo generalizado.

Abstract

In this paper, it is tackled the problematic of the risks that are competing to cause the failure from the subject; in particular whether the risks or likelihood of failure associated with each type of failure are equally important or whether a risk is more serious than the other. For this purpose will be made a study of hypothesis tests for equality of cumulative incidence functions of associated with risks. A comparative study of some of the test procedures that have been proposed for this purpose, and thus able to determine the behavior of the different tests in various scenarios to evaluate the performance of the same will be made. Test procedures are included using real data of patients with lymphoma.

***Key words*:** Cumulative incidence function, cause-specific hazard rates, quantile, bootstrap, random symmetrization approximation, resampling and generalized supremum.

a Magíster en Ciencias-Estadística, Universidad Nacional de Colombia, Medellín.
*Correo electrónico: licmolinabl@unal.edu.co
b Profesor asociado, Universidad Nacional de Colombia, Medellín.
*Correo electrónico: cmlopera@unal.edu.co

1. INTRODUCCIÓN

Los estudios de confiabilidad y supervivencia buscan analizar por medio de un conjunto de técnicas la variable "tiempo hasta que ocurre un evento", tales como el tiempo hasta la muerte o curación, la probabilidad de falla en cada instante de tiempo, el riesgo de falla, etc. El análisis de los modelos de riesgos competitivos es apropiado para estudiar el comportamiento de una unidad o sujeto que puede fallar por diferentes causas, donde se observa tanto el tiempo hasta la falla, como el tipo de falla.

En muchas situaciones prácticas es común analizar únicamente el evento de interés sin tener en cuenta los riesgos que están compitiendo; los procedimientos de análisis en este contexto solo responden preguntas en las que el objeto de estudio es estimar el efecto "puro" debido a una sola causa de falla [1]. En otros campos de investigación se desean estudiar los riegos que están actuando simultáneamente en una misma población, para determinar si los diversos riesgos que se consideran son igual de graves o si uno de los riesgos es más serio que el otro, para tal fin es necesario estudiar las diferentes pruebas que se han diseñado para análizar este tipo de situaciones, con fin de determinar el comportamiento de estas.

Recientemente en la literatura se han estudiado ampliamente las pruebas para la comparación de riesgos independientes. Sin embargo, en situaciones reales es frecuente que los riesgos sean dependientes y los tiempos de falla estén sujetos a censuras a derecha. En este contexto general, la referencia [2] propone pruebas de distribución asintóticamente libre para la comparación de las funciones de incidencia acumulada o equivalentemente de las tasas de riesgo de causa específica. La anterior propuesta fue ampliada en la investigación de la prueba del supremo generalizado para la igualdad de tasas de riesgo de causa-específica [3], donde se proponen tres clases de pruebas de hipótesis y por medio de un estudio de simulación se establece que una de las pruebas es menos sensible a la elección de la función de peso y es adecuada cuando no se tiene información suficiente en cuanto a la naturaleza de los datos, por tanto, esta sería una buena prueba para comparar dos riesgos competitivos; en general, cuando no se tiene información acerca de las características de los datos. Por otro lado, se han desarrollado varios tipos de pruebas para la comparación de funciones de incidencia acumulada a través de métodos de remuestreo propuesta [4]. En la referencia [5] se propone una prueba para la proporcionalidad de dos funciones de incidencia acumulada en una configuración de riesgos competitivos.

En este artículo se aborda la problemática de determinar si los riesgos o probabilidad de falla asociada a cada tipo de falla son igualmente importantes o si un riesgo es más serio que el otro. Para este fin se realizó un estudio comparativo de la prueba del supremo generalizado [3] y las pruebas basadas en métodos de remuestreo para la igualdad de las funciones de incidencia acumulada: el método bootstrap y la aproximación por simetrización aleatoría propuesta en [4]. En este último caso se propuso una modificación de estas pruebas, especifícamente en el método bootstrap; luego, a través de un estudio de simulación se evaluó el comportamiento de las diferentes pruebas en varios escenarios con el fin de estudiar el desempeño de las mismas.

La estructura del artículo es la siguiente: en la Sección 2 se dan algunos conceptos básicos de riesgos competitivos. Las pruebas para comparar la igualdad de las funciones de incidencia acumulada asociadas a dos riesgos competitivos son presentadas en la Sección 3. En la Sección 4 se presenta un estudio de simulación para evaluar el desempeño de las pruebas presentadas. Un análisis de los resultados obtenidos se muestra en la Sección 5. Posteriormente, en la Sección 6 se presenta un ejemplo con datos reales. Finalmente, en la Sección 7 se presentan las conclusiones más relevantes del trabajo y se describen algunos temas que pueden ser de interés para trabajo futuro.

2. RIESGOS COMPETITIVOS

En la referencia [6] se describen los riesgos competitivos como la situación en la que un individuo puede experimentar más de un tipo de evento. En la referencia [1] se presenta un enfoque matemático para tratar los riesgos competitivos como una variable bivariada, este enfoque muestra que en ausencia de riesgos competitivos, los datos de supervivencia se presentan usualmente

como variables aleatorias biva- riadas (T,δ), donde δ es una indicadora del estatus del individuo que toma el valor de 1 si se observa el evento de interés, o 0 si la observación es censurada. Cuando $\delta = 1$, el primer miembro del par, T, corresponde al tiempo en que se produce el evento y cuando $\delta = 0$, T es el tiempo en el que se censuró la observación.

Esta definición se puede extender a la situación de riesgos competitivos donde son posibles $j \geq 2$ tipos de fallas o eventos. Los datos son nuevamente representados por un par (T,δ), donde la variable de censura δ esta vez es discreta tomando el valor de 0 si la observación es censurada. En el caso de que la observación no este censurada, δ tomará el valor j, donde j es el tipo de falla o evento observado $(\delta = j)$, en este caso T es el tiempo en el que el evento o falla de tipo j se produjo; de lo contrario, es el tiempo de censura [6].

2.1 Subdistribución o Función de incidencia acumulada (CIF).

Sea T el tiempo de vida del sujeto, asumiendo que el tiempo es continuo, con función de distribución F y función de supervivencia S, y sea δ la causa de falla, es decir, $\{\delta = j\}$ es el evento cuya falla se debe al riesgo $j = 1,2$, la función de subdistribución está definida como [1]:

$$\begin{aligned} F_j(t) &= P(T \leq t, \delta = j) \\ &= \int_0^t S(u)\lambda_j(u)\,du \end{aligned}$$

La función de distribución acumulada de falla $F(t)$ estará dada por $F(t) = F_1(t) + F_2(t)$. Sea C un tiempo de censura independiente de T con función de supervivencia $S_C(t)$. Asuma que $S_C(t) > 0$ para todo t, entonces δ se convierte en una variable aleatoria discreta que representa el estatus del individuo que toma valores de $\delta = 0,\ 1,\ 2$, donde $\{\delta = 0\}$ es el evento en el que el sujeto fue censurado (censura a derecha) y $\{\delta = j\}$ cuando el sujeto falla debido a la causa j.

En muchas situaciones prácticas, es importante saber si en una población donde actúan simultáneamente dos riesgos, estos se pueden considerar igual de graves o si uno de los riesgos es más serio que el otro. Para dar respuesta a este problema se han diseñado pruebas estadísticas que usan los modelos de riesgos competitivos; cabe recordar que en presencia de los riesgos competitivos, la comparación de los riesgos de causa específica entre dos grupos no es equivalente a la comparación de las funciones de subdistribución (CIF), sin embargo, comparar los riesgos de causa específica de dos riesgos en una misma población es equivalente a comparar las subdistribuciones (CIF).

3. PRUEBAS PARA COMPARAR DOS RIESGOS COMPETITIVOS

Se tienen n observaciones independientes e identicamente distribuidas (i.i.d) de la forma (X_i, δ_i), $i = 1,2...,n$, donde $X_i = \min\{T_i, C_i\}$. De acuerdo a los datos mencionados, se formula el problema de poner a prueba las siguientes hipótesis

$$\mathbf{H_0} : F_1(t) = F_2(t), \text{ para } t \geq 0 \text{ vs } \mathbf{H_1} : F_1(t)\ F_2(t)$$

o equivalentemente

$$H_0 : \lambda_1(t) = \lambda_2(t) \quad \text{vs.} \quad H_1 : \lambda_1(t) \neq \lambda_2(t)$$

Para probar estas hipótesis, se utilizarán las siguientes pruebas:

3.1 Pruebas del supremo generalizado

En la referencia [3] se consideró un modelo de riesgos competitivos con dos causas de falla, proponiendo dos clases de pruebas basadas en la distribución asintóticamente libre y la tipo Renyi, para probar la igualdad de dos riesgos con posible censura. Para tal fin, plantean el estadístico del supremo que es una generalización de las pruebas propuestas en [2] mediante la adopción de diferentes funciones de peso w. El estadístico de prueba es

$$C_n^*(w) = \sup_{0 \leq s < t < \infty} \frac{|L_n(t) - L_n(s)|}{S_n(\infty)} \qquad (1)$$

este estadístico puede expresarse como funciones ponderadas de los estadísticos de tipo log-rank de la forma

$$L_n(t) = \int_0^t w(u)\, d(\widehat{\Lambda}_2 - \widehat{\Lambda}_1)(u) \qquad (2)$$

siendo $\Lambda_j(t) = \int_0^t \lambda_j(u)\,du$ la función de riesgo de causa-específica acumulada para el riesgo j, cuando $j = 1, 2$ y el estimador de Nelson-Aalen [7] de Λ_j es

$$\widehat{\Lambda}_j(t) = \sum_{i:X_i \le t} \frac{I(\delta_i = j)}{R_i} \tag{3}$$

donde, Ri = #{$k : Xk \geq Xi$} es el tamaño del riesgo fijado antes del tiempo Xi denotado como el tiempo X_i−. La función de peso w(u) refleja la importancia que se concede a las funciones de incidencia acumulada (CIF) en el tiempo *u.*

Bajo H_0, $n^{1/2}L_n(t)$ es un proceso martingala con varianza predecible $\sigma^2(t)$, que se puede estimar por

$$S_n^2(t) = \int_0^t \frac{w^2(u)}{\frac{\overline{Y}^2(u)}{n}}\, d\overline{N}(u) \tag{4}$$

donde $\overline{Y}(u) = \sum_{i=1}^n I(X_i \geq u)$ es el número total de elementos en riesgo hasta u−, y $\overline{N}(u)$ es el número total de fallas hasta el tiempo *u.* Para este trabajo se tomó de referencia la función de peso $w(u) = \overline{Y}(u)$; es importante mencionar que según [3] la elección de la función peso debe basarse de acuerdo al criterio que desee enfatizar el investigador; pero en sus estudios de simulación observaron que las pruebas basadas en la función de peso $w(u) = \overline{Y}(u)$ presentaban potencias más razonables en la mayoría de situaciones prácticas. Cabe resaltar que de acuerdo a sus investigaciones se encontró que la prueba C_n^* es menos sensible a la elección de la función de peso, por ende es considerada como una prueba adecuada cuando no se tiene información acerca de las características de los datos.

En la referencia [3] se demuestra que bajo la hipótesis nula (H_0) $n^{1/2}C_n^*(w)$ converge aproximadamente a

$$8\sum_{k=1}^{\infty} (-1)^{k-1}\, k^2 \phi(kx) \tag{5}$$

donde φ es la función de densidad de una normal estándar. Los detalles de esta distribución se encuentra en [3, 8]. Los cuantiles del 95% y 99%se encuentran en 2.497 y 3.023, respectivamente. Para tomar la decisión bajo H_0 con base a la prueba C_n^* se puede obtener un cuantil $1 - \alpha$ denotelo $q_{1-\alpha}$, se rechaza H_0 si $C_n^*(w)$ supera a $q_{1-\alpha}$, esto es, $C_n^*(w) > q_{1-\alpha}$.

3.2 Pruebas para determinar la igualdad de dos CIF's a través de métodos de remuestreo

En la referencia [4] se diseñaron estadísticos de prueba basados en vector de procesos relacionados con las funciones de incidencia acumulada para determinar la igualdad de las funciones de incidencia acumulada. Como las distribuciones asintóticas parecen muy complicadas y dependen de la distribución subyacente de los datos, se utilizan dos técnicas de remuestreo (método bootstrap y método de simetrización al azar), para aproximar los valores críticos de las pruebas. Sin hacer ninguna hipótesis sobre la naturaleza de dependencia entre los riesgos, las pruebas permiten comparar dos o más riesgos *(j ≥ 2)* simultáneamente bajo el modelo de censura aleatoria. Para probar H_0 vs H_1 consideran el estadístico de Cramér-Von Mises dado por:

$$D_n = n\int_0^{\infty} (\widehat{F}_{n,2}(t) - \widehat{F}_{n,1}(t))^2 \widehat{F}_n(dt) \tag{6}$$

donde $\widehat{F}_n(t)$ es el análogo muestral de $\widehat{F}(t) = P(T \le t)$ y $\widehat{F}_{n,j}$ s la contraparte empírica de la función de sub-distribución $(\widehat{F}_j)$ definida por

$$\widehat{F}_{n,j}(t) = n^{-1}\sum_{i=1}^{n} I(X_i \le t, \delta_i = j),\ j = 1, 2. \tag{7}$$

Se rechaza la hipótesis nula (H_0) cuando el valor D_n es demasiado grande. El estadístico (7) se puede reescribir como [4]

$$D_n = \int_0^{\infty} (W_n(t))^2 \widehat{F}_n(dt) \tag{8}$$

donde $W_n(t) = W_{n,2}(t) - W_{n,1}(t)$ siendo $W_{n\,j}(t) = \sqrt{n}\,(F_{n,\,j} - F_j)$; $j = 1, 2$. Los autores plantean que W_n puede ser usada como una herramienta para obtener la distribución asintótica del estadístico bajo H_0 y además demuestran que W_n converge en distribución a un vector de procesos gaussianos [4], los límites de la distribución y sus valores críticos no son analíticamente tratables, por tal razón se optaron por utilizar los siguientes métodos de remuestreo.

3.2.1 Aproximación bootstrap

Sea $\{(X_i^*, \delta_i^*); i = 1, 2, \ldots, n\}$ una muestra bootstrap que se extrae con reemplazo de la muestra original. Se denota las versiones bootstrap de $\widehat{F}_{n,j}$ y $\widehat{F}_n$ por $\widehat{F}_{n,j}^*$ y $\widehat{F}_n^*$ respectivamente. Entonces la versión bootstrap de W_n^* está dada por

$$W_n^*(t) = W_{n,2}^*(t) - W_{n,1}^*(t) \tag{9}$$

donde $W_{n,j}^*(t) = \sqrt{n}(\widehat{F}_{n,j}^* - \widehat{F}_{n,j})$. Por otra parte, el estadístico bootstrap está dado por

$$D_n^* = \int_0^\infty (W_n^*(t))^2 \widehat{F}_n^*(dt) \tag{10}$$

En la literatura se han encontrado investigaciones en las que se utilizan técnicas basadas en la aproximación bootstrap para construir pruebas relacionadas a los riesgos, entre estos, en la referencia [9] se construye una prueba de bondad de ajuste para el modelo de riesgos proporcionales. En [10] se utilizó un enfoque similar para las pruebas de un modelo de riesgo aditivo semiparamétrico.

El procedimiento de la aproximación bootstrap para D_n es presentado a continuación:

- Se extraen M muestras bootstrap de la forma $\{(X_i^*, \delta_i^*); i = 1, 2, \ldots, n\}$ con reemplazo de la muestra original.
- A cada conjunto de datos remuestrado se le calcula el estadístico de prueba, obteniendo $D_{n_1}^*, D_{n_2}^*, \ldots, D_{n_M}^*$
- Se obtienen la distribución de los D_n^*'s y se calcula el cuantil $1 - \alpha$ denotado por $q_{1-\alpha}$ de los valores D_n^*'s H_0 si $D_n > q_{1-\alpha}$.

3.2.2 Aproximación por simetrización aleatoria (RAS)

Sea $Z = \{Z_i; i = 1, 2, \ldots, n\}$ una muestra aleatoria independiente e idénticamente distribuida que toman signos ± 1 con igual probabilidad. Asuma que estas variables de permutaciones son independientes de la muestra. La versión **RAS** de $\widehat{F}_{n,j}$ está dada por

$$F_{n,j}^Z(t) = n^{-1} \sum_{i=1}^{n} Z_i I(T \leq t, \delta_i = j) \tag{11}$$

esto lleva al proceso $W_n^Z(t) = n^{\frac{1}{2}}(F_{n,2}^Z(t) - F_{n,1}^Z(t))$. Entonces la versión RAS del estadístico de prueba es

$$D_n^Z = \int_0^\infty (W_{n,1,2}^Z(t))^2 \widehat{F}_n(dt) \tag{12}$$

El procedimiento de la prueba RAS para D_n es presentado a continuación:

- Denote al valor observado de D_n dado en (6) como $D_n^{Z^0}$
- Genere M conjuntos de Z's, denotados {Z ^{1}, Z^2,...,Z^M , }
- Se obtiene $D_n^{Z^1}, D_n^{Z^2}, \ldots, D_n^{Z^M}$
- Denote como D_{M+1} al conjunto formado por $(D_n^{Z^0}, D_n^{Z^1}, \ldots, D_n^{Z^M})$
- Se rechaza la hipótesis nula H_0 cada vez que $\hat{p} \leq \alpha$, donde el valor p estimado es

$$\hat{p} = \frac{a}{M+1}$$

siendo a el número de valores del conjunto D_{M+1} que son mayores o iguales a $D_n^{Z^0}$.

En la referencia [4] se demuestra que en casi todas las secuencias de las muestras, el proceso bootstrap W_n^* y el proceso RAS de W_n^Z, convergen condicionalmente en distribución a un proceso gaussiano. El enfoque **RAS** está motivado por métodos de ponderación al azar que han sido técnicas básicas para abordar la convergencia de los procesos empíricos [11].

3.3 Método bootstrap alternativo usando los estimadores de las CIF's y Kaplan-Meier

Dado que en la referencia [4] se sugiere para la construcción del estadístico de prueba en los métodos de remuestreo el uso de las funciones empíricas para estimar, tanto las distribuciones como las subdistribuciones asociadas a las probabilidades de falla, una propuesta alternativa es cambiar los estadísticos D_n por una versión modificada donde se usen estimaciones de las CIF's y los estimadores de Kaplan-Meier en lugar de estas distribuciones empíricas. Luego, para probar H_0

vs. H_a , se considera el estadístico de Cramér-Von Mises alternativo

$$D'_n = n\int_0^{\infty} (\widehat{F}_2(t) - \widehat{F}_1(t))^2 (1 - \widehat{S}(dt)) \qquad (13)$$

Donde $\widehat{S}(t) = P(T > t)$ es la función de supervivencia estimada por medio del método de Kaplan-Meier y $\widehat{F}_j(t),\ j = 1,2$ son las CIF's estimadas para los dos modos de falla.

Sobre la modificación hecha al estadístico de Cramér-Von Mises, se aplica el mismo procedimiento de prueba que se usó en el método bootstrap, que consiste en encontrar la distribución bajo la hipótesis nula (H_0) del estadístico alternativo D'_n para finalmente tomar una decisión acerca de la hipótesis de interés, es decir, se decide rechazar H_0 si el valor D'_n resulta mayor al cuantil $(1 - \alpha)$ de la distribución empírica del estadístico bajo la hipótesis nula.

4. ESTUDIO DE SIMULACIÓN

Para evaluar el desempeño de los estadísticos de prueba se consideraron dos causas de falla. Sea (T_1, T_2) los tiempos potenciales de falla de dos componentes de un sistema en serie. El tiempo de falla X = mín{T_1, T_2} fue generado usando la distribución exponencial absolutamente continua de Block & Basu denotada por (ACBVE) usando las indicaciones presentadas en [12, 13, 14] donde la pareja (T_1, T_2) bajo esta distribución tiene una densidad dada por

$$f(t_1,t_2) = \begin{cases} \frac{\lambda\lambda_1\lambda_{02}}{\lambda_{12}} e^{-\lambda_1 t_1 - \lambda_{02} t_2} & \text{si} \quad t_1 < t_2 \\ \\ \frac{\lambda\lambda_2\lambda_{01}}{\lambda_{12}} e^{-\lambda_{01} t_1 - \lambda_2 t_2} & \text{si} \quad t_1 \geq t_2 \end{cases} \qquad (14)$$

donde λ12 = λ1 +λ2, λ01 = λ0 +λ1, λ02 = λ0 +λ2 y λ = λ0 + λ1 + λ2.

Siendo λ0 el parámetro que controla el grado de dependencia entre T1 y T2, donde λ0 = 0.0 corresponde a la independencia de los dos riesgos y λ0 = 1.0 la dependencia de los dos riesgos. En este caso, la tasa de riesgo de causa específica son proporcionales, y están dadas por

$$\lambda_j(t) = \frac{\lambda_j(\lambda_0 + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2};\ j = 1,2 \qquad (15)$$

Al usar la distribución ACBVE se replanteó las hipótesis H_0 versus H_1 como

$$H_0 : \lambda_1(t) = \lambda_2(t) \text{ vs. } H_1 : \lambda_1(t)\ \lambda_2(t)$$

Se fijó la tasa de falla del primer evento como λ_1 =1.0 y se varió la tasa de falla del segundo evento λ_2 tomando diferentes valores, tales como, λ_2 =0.5, 1.5, 2.0, 2.5 y en el caso de ser iguales las tasas λ_2 =1.0, para determinar el comportamiento de las pruebas a medida que este cambia. Los valores asignados al parámetro λ_0 y λ_1 se tomaron teniendo como referencia los estudios de simulación realizados en [2, 4, 3].

Para mirar el efecto de las censuras en cada una de las pruebas se consideraron tres niveles de censura, sin censura, censura moderada (18 % − 35 %) y censura alta (45 % − 60 %) denotadas por SC, CM y CA respectivamente. La variable censura C de acuerdo a lo planteado por [15], fue generada por una distribución exponencial con parámetro γ; para lo cual se usó la probabilidad de una censura P ($C < X$) tomando valores de 0 (sin censura), 0.2 (censura moderada) y

0.5 (censura alta), que bajo el modelo ACBVE tiene la forma

$$P(C < X) = \frac{\gamma}{\gamma + \lambda} \qquad (16)$$

de donde se despeja γ en cada uno de los escenarios. El nivel de significancia usado en las pruebas es de α =0.05. Para cada conjunto de parámetros asumi dos, se generaron muestras de n = 50, 100 y 250. Adicionalmente se generaron 5000 muestras boots- trap, tanto para el estadístico bootstrap (Boot) y el estadístico bootstrap alternativo (Alt), también se extrajeron 5000 conjuntos de datos **RAS** {Z_1, Z_2, . . . , Z_n}.

Se programaron en el software libre **R** las pruebas basadas en los métodos bootstrap, **RAS** y el método bootstrap alternativo usando los estimadores de las CIF's y Kaplan-Meier; para la prueba del supremo (Sup) se usó la programación realizada por [1].

En total se consideraron 90 escenarios de simulación, donde se evaluó el error tipo **I** (bajo $H_0 : \lambda_1(t) = \lambda_2(t)$) y la potencia de las diferentes pruebas presentadas (bajo $H_1 : \lambda_1(t) /= \lambda_2(t)$). La tasa el error tipo **I** está dado por:

$$\hat{\alpha} = \frac{1}{N}\sum I(\text{Rechazar } H_0 | H_0 \text{ es Cierta}),$$

donde I(Rechazar $H_0|H_0$ es Cierta) es una indicadora de rechazo de una hipótesis cierta para una prueba particular.

La potencia alcanzada está definida como la probabilidad de rechazar una hipótesis nula falsa, esto es

$$1 - \hat{\beta} = \frac{1}{N}\sum I(\text{Rechazar } H_0 | H_0 \text{ es Falsa}) \quad (17)$$

donde I (Rechazar $H_0|H_0$ es Falsa) es una indicadora de rechazo de una hipótesis falsa para una prueba particular y N es el número de simulaciones realizadas en cada una de las pruebas estudiadas, en cada uno de los escenarios de simulación considerados. El proceso de simulación se describe a continuación:

- **Paso 1.** Se genera una muestra de tamaño n de la forma (X_i, δ_i) con un porcentaje de censura especificado.
- **Paso 2.** A este conjunto de datos se le aplican las pruebas basadas en el supremo, los métodos de remuestreo: bootstrap, **RAS** y la alternativa, con un nivel de significancia fijo.
- **Paso 3.** Los pasos 1 y 2 son simulados N veces obteniendo la decisión para cada simulación.
- **Paso 4.** Se suman las indicadoras de las N simulaciones y se promedian para obtener el error tipo **I** o la potencia alcanzada.

5. ANÁLISIS DE RESULTADOS

La tabla 1 presenta la proporción de veces que se rechaza la hipótesis nula de manera que cuando λ_1 es igual a λ_2 (bajo H_0) estas proporciones corresponden a la tasa de error tipo **I**, mientras que en valores de λ_2 distintos a λ_1 (bajo H_1) tales porcentajes corresponden a las potencias empíricas; para los escenarios de independencia (λ_1 = 0.0) y dependencia (λ_1 = 1.0). Los valores de la tasa de error tipo **I** se presentan en negrilla para facilitar la lectura de la tabla.

Tabla 1. Tasas de error tipo **I** y potencias empíricas (en %) de la prueba $H_0 : \lambda_1(t) = \lambda_2(t)$ *vs* $H_1 : \lambda_1(t) =/ \lambda_2(t)$ con un nivel de significancia del 5 % y n = 50.

		$\lambda_0 = 0.0$ y $\lambda_1 = 1.0$				$\lambda_0 = 1.0$ y $\lambda_1 = 1.0$			
	λ_2	Sup	Boot	RAS	Alt	Sup	Boot	RAS	Alt
	0.5	48.3	50.7	47.3	49.6	44.8	50.7	47.5	51.0
	1.0	**5.0**	**7.1**	**7.1**	**7.6**	3.2	**4.3**	**3.9**	**5.0**
SC	1.5	17.5	22.1	19.0	21.3	21.6	23.8	22.8	24.5
	2.0	43.4	49.7	50.0	49.4	42.4	55.3	49.1	52.1
	2.5	68.6	73.2	73.0	72.1	68.8	72.3	72.3	70.6
	0.5	33.6	39.1	43.1	44.7	34.7	38.3	40.1	43.8
	1.0	**4.7**	**5.0**	**6.3**	**5.7**	**3.2**	**4.5**	**4.2**	**4.6**
CM	1.5	12.6	16.6	17.5	16.9	14.6	21.5	11.7	20.4
	2.0	37.2	40.4	42.4	42.6	31.1	40.5	40.9	41.7
	2.5	59.6	57.2	64.2	66.1	55.3	62.2	64.7	64.5
	0.5	20.8	25.1	26.8	26.8	20.2	22.9	23.7	25.4
	1.0	**2.9**	**5.0**	**5.6**	**3.9**	**1.7**	**3.9**	**3.5**	**2.7**
CA	1.5	7.5	10.5	10.5	10.6	8.9	12.8	14.0	13.7
	2.0	19.6	23.5	27.3	24.8	20.4	21.5	25.0	27.1
	2.5	41.0	38.3	42.7	42.8	33.5	38.5	42.5	41.3

Tabla 2. Tasas de error tipo **I** y potencias empíricas (en %) de la prueba $H_0 : \lambda_1(t) = \lambda_2(t)$ *vs* $H_1 : \lambda_1(t) =/ \lambda_2(t)$ con un nivel de significancia del 5 % y n = 100.

		$\lambda_0 = 0.0$ y $\lambda_1 = 1.0$				$\lambda_0 = 1.0$ y $\lambda_1 = 1.0$			
	λ_2	Sup	Boot	RAS	Alt	Sup	Boot	RAS	Alt
	0.5	80.2	77.8	77.5	79.2	79.2	79.6	78.4	79.0
	1.0	**5.0**	**7.5**	**8.4**	**6.2**	**3.6**	**4.3**	**4.2**	**4.5**
SC	1.5	35.0	34.3	37.0	35.7	39.7	38.5	38.6	38.4
	2.0	79.5	77.6	79.6	79.2	82.2	79.2	78.9	77.0
	2.5	95.4	95.4	94.3	95.4	95.1	94.8	89.0	93.3
	0.5	68.8	69.7	69.4	71.7	69.2	70.2	66.2	75.8
	1.0	**4.6**	**6.4**	**7.0**	**5.9**	**3.3**	**3.9**	**4.3**	**3.5**
CM	1.5	30.7	30.4	27.7	33.1	31.8	32.2	29.2	34.7
	2.0	68.2	71.4	67.9	72.9	68.0	68.3	68.3	73.0
	2.5	90.7	89.2	90.6	92.8	90.4	89.5	88.9	92.6
	0.5	45.5	44.7	48.5	45.9	43.6	45.1	48.9	46.9
	1.0	**3.1**	**5.6**	**4.9**	**4.0**	**3.2**	**4.2**	**4.2**	**2.9**
CA	1.5	15.1	20.8	20.5	19.0	20.7	19.1	23.8	25.3
	2.0	44.8	44.9	49.5	48.5	45.7	44.8	44.0	48.4
	2.5	68.3	70.4	68.0	70.7	70.0	67.6	70.4	70.1

Las tablas 2 y 3 presentan los resultados del estudio de simulación cuando se tienen tamaños de muestras de n = 100 y n = 250 con sus respectivos resultados de las potencias empíricas y las tasa de error tipo **I**, tomando como referencia en los estudios de simulación un nivel de significancia de α =0.05.

Tabla 3. Tasas de error tipo **I** y potencias empíricas (en %) de la prueba $H_0 : \lambda_1(t) = \lambda_2(t)$ *vs* $H_1 : \lambda_1(t) =/ \lambda_2(t)$ con un nivel de significancia del 5 % y $n = 250$.

		$\lambda_0 = 0.0$ y $\lambda_1 = 1.0$				$\lambda_0 = 1.0$ y $\lambda_1 = 1.0$			
	λ_2	Sup	Boot	RAS	Alt	Sup	Boot	RAS	Alt
	0.5	99.1	98.8	99.3	99.0	99.7	99.1	99.2	99.0
	1.0	**7.5**	**6.6**	**6.5**	**7.0**	**3.3**	**4.1**	**4.6**	**3.7**
SC	1.5	76.6	74.5	72.2	70.7	75.1	66.9	69.8	69.1
	2.0	99.3	99.0	99.5	99.2	99.7	99.4	98.6	99.5
	2.5	100	100	100	100	100	100	100	100
	0.5	98.1	96.7	96.5	98.7	98.3	97.4	97.9	99.2
	1.0	**6.0**	**6.1**	**7.4**	**7.3**	**3.8**	**4.1**	**5.3**	**4.3**
CM	1.5	67.3	63.5	61.1	67.7	64.6	61.3	63.7	65.0
	2.0	98.6	97.7	97.2	98.6	98.9	97.7	97.3	98.7
	2.5	100	99.9	100	100	100	99.9	99.9	100
	0.5	90.1	84.0	83.8	89.2	89.4	87.0	87.5	89.8
	1.0	**4.9**	**6.1**	**6.6**	**4.1**	**2.9**	**3.8**	**4.9**	**3.8**
CA	1.5	44.7	44.7	43.4	42.1	45.5	42.0	47.1	46.8
	2.0	88.1	85.1	86.7	89.3	89.6	87.2	85.5	89.5
	2.5	99.1	98.4	97.9	98.6	98.6	98.5	97.7	97.9

Para analizar los resultados obtenidos en las tablas 1, 2 y 3 se contruyeron gráficos de líneas simultáneos del nivel de significancia alcanzado con relación a los diferentes tamaños de muestras. Las potencias empíricas se relacionan con los distintos valores asignados a λ_2; con el fin de evaluar la capacidad de las pruebas de detectar diferencias entre los dos riesgos a medida que se le asignan valores de λ_2 cercanos y alejados de $\lambda_2 = 1.0$, tanto en presencia y ausencia de censura para los escenarios de independencia y dependencia en cada tamaño de muestra.

5.1 Gráficos de resultados de la simulación de diferentes tamaños de muestras

A continuación se presentan el análisis de las potencias empíricas para los diferentes tamaño de muestra tomados en el estudio de simulación a saber, $n = 50$, $n = 100$ y $n = 250$, a través de gráficos de líneas simultáneas diferenciadas con un número y color específicado, que muestra el comportamiento de las potencias alcanzadas bajo dependencia e independencia y diferentes valores de λ_2 por las pruebas del supremo generalizado (1), método bootstrap (2) y **RAS** (3) , y el bootstrap alternativo (4).

En la figura 1 se puede observar que en el caso de muestras pequeñas cuando no hay presencia de censura la prueba del supremo generalizado presenta errores Tipo **I** más cercanos al nivel de significancia fijado, tanto para el escenario de independencia como el de dependencia. En el caso de las pruebas **RAS**, bootstrap y en la prueba alternativa, se obser- va que aunque el error tipo **I** es pequeño en los tres casos, en el escenario de independencia los errores tipo **I** se alejan un poco del nivel de significancia establecido.

En las figuras 2, 3 y 4, se observa que en tamaños de muestras grandes hay mayor estabilidad de resultados en las potencias de las pruebas para los dos escenarios, ya sea en presencia de censura o en ausencia de esta. Cuando $\lambda_2 = 1.5$ en el caso de muestras grandes y en presencia de censura hay una disminución en las potencias empíricas pero no tan marcados como en el caso de muestras pequeñas, en el caso de la muestra $n = 250$ se puede ver que ya las potencias empíricas tienen mayor estabilidad tanto en presencia de censura como en ausencia de esta, en las cuatro pruebas. En todos los casos, las potencias empíricas convergen de manera constante al número uno cuando aumenta el tamaño de la muestra.

También se observan errores tipo I más cercanos al nivel de significancia en las pruebas del supremo generalizado y alternativa cuando la censura es alta para ambos escenarios, mientras que los métodos bootstrap y **RAS** siguen teniendo mejor comportamiento en en los escenarios de dependencia.

En general, las pruebas basadas en métodos de remuestreo bootstrap y **RAS** tienen un buen funcionamiento en ambos escenarios, pero especialmente en los casos de dependencia entre los tiempos de fallas, sus errores tipos I son más cercanos al nivel de significancia.

En los diferentes tamaños de muestra se observa que a medida que el valor de la tasa del segundo evento aumenta las potencias empíricas de las pruebas son muy altas, mientras que, cuando los valores de dicha tasa son cercanos a uno, las pruebas son un poco inestables alcanzando potencias muy bajas que se van estabilizando a medida que el tamaño de muestra va creciendo.

Por último en los diferentes gráficos, los resultados de la simulación para la propuesta bootstrap usando los estimadores de las CIF's y Kaplan-Meier, presentan comportamientos similares a las pruebas basadas en los métodos de remuestreo: bootstrap y **RAS**, donde se ve evidenciado nuevamente que en presencia de censura los potencias empíricas disminuyen pero a medida que el tamaño de la muestra crece sus potencias empíricas también; en cuanto a los errores tipo I, la prueba alternativa muestra un mejor comportamiento que los propuestos por [4].

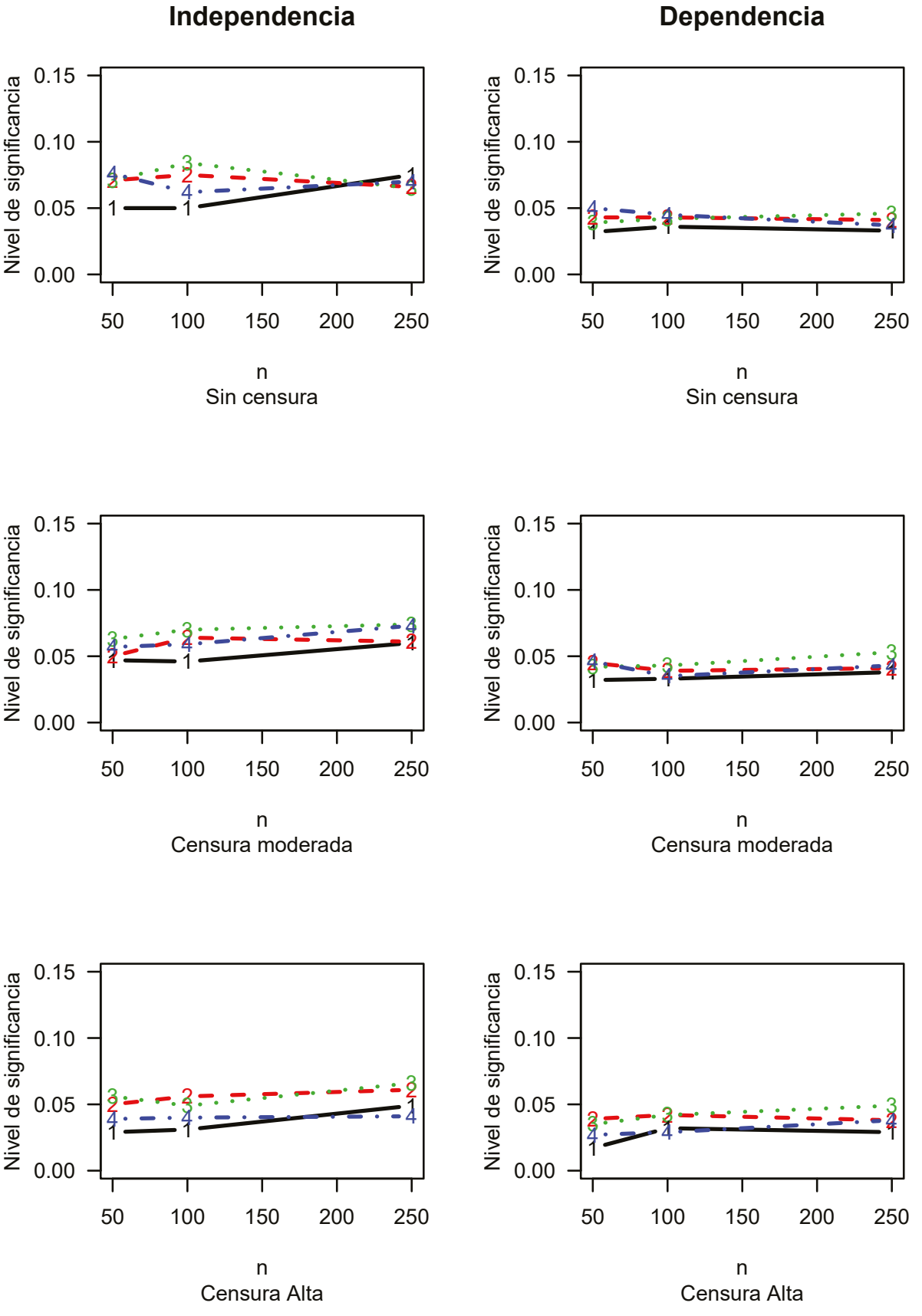

Figura 1. Significancia estadística alcanzada.

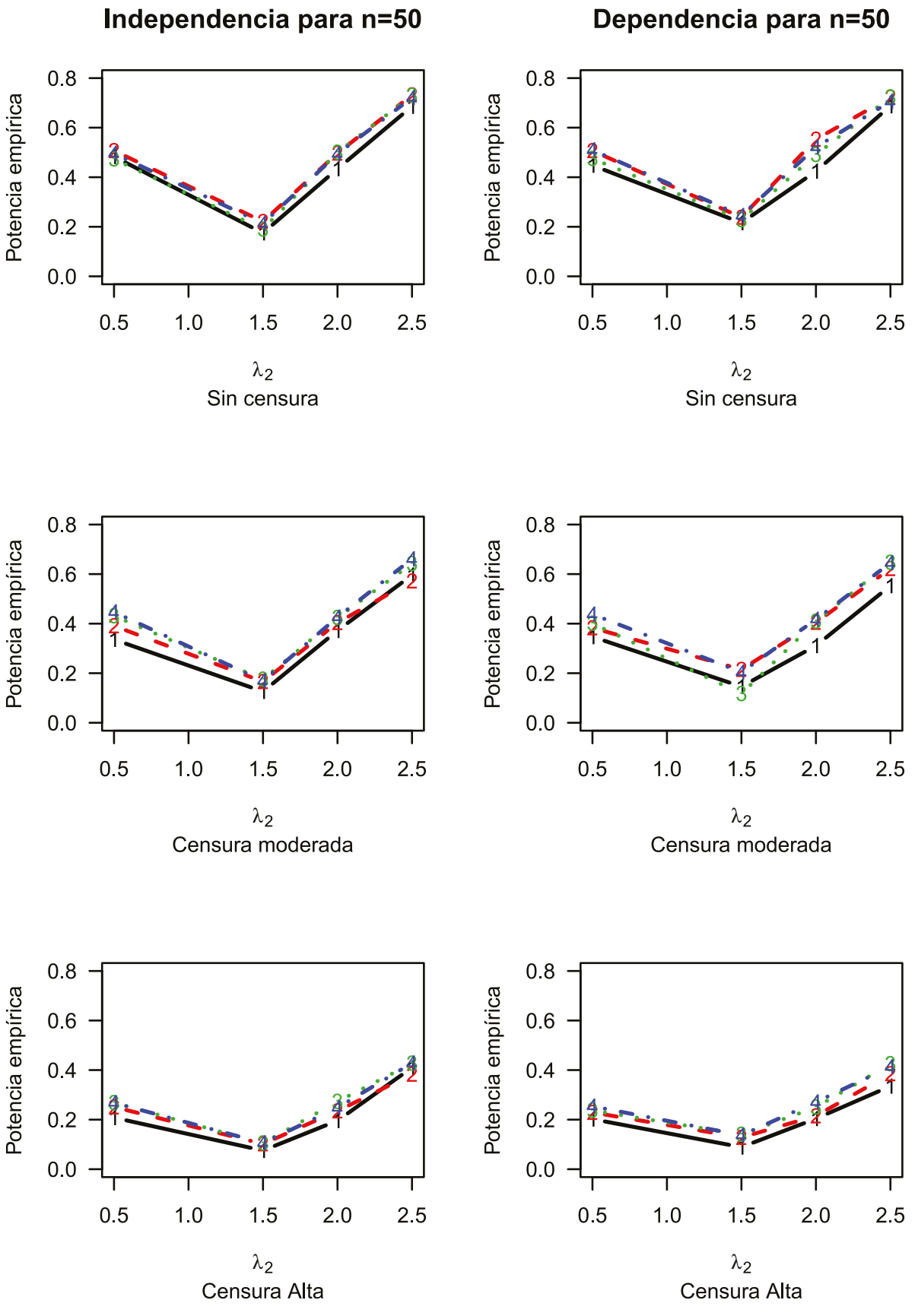

Figura 2. Potencia empírica alcanzada para tamaño de muestra de n=50.

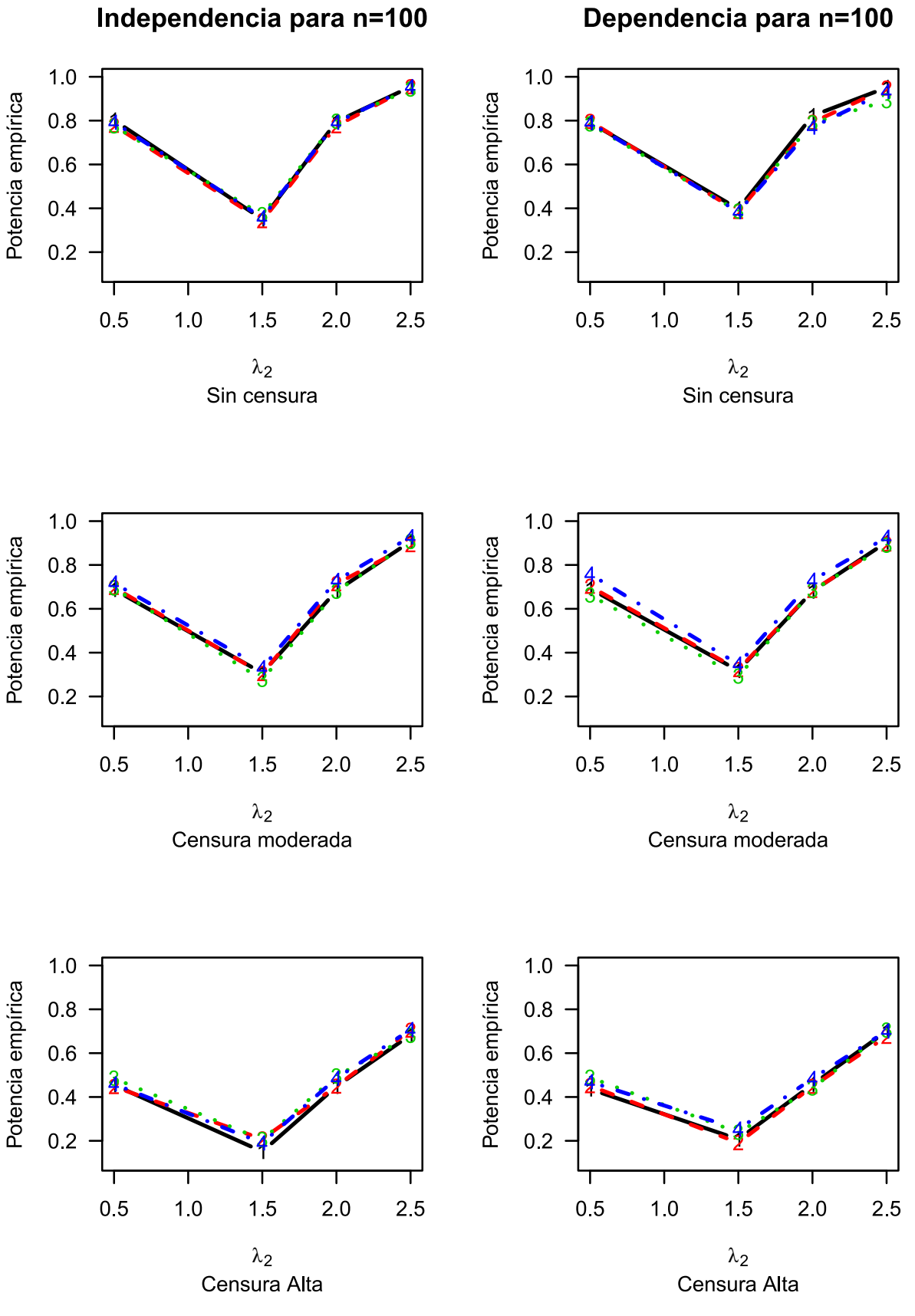

Figura 3. Potencia empírica alcanzada para tamaño de muestra de n=100

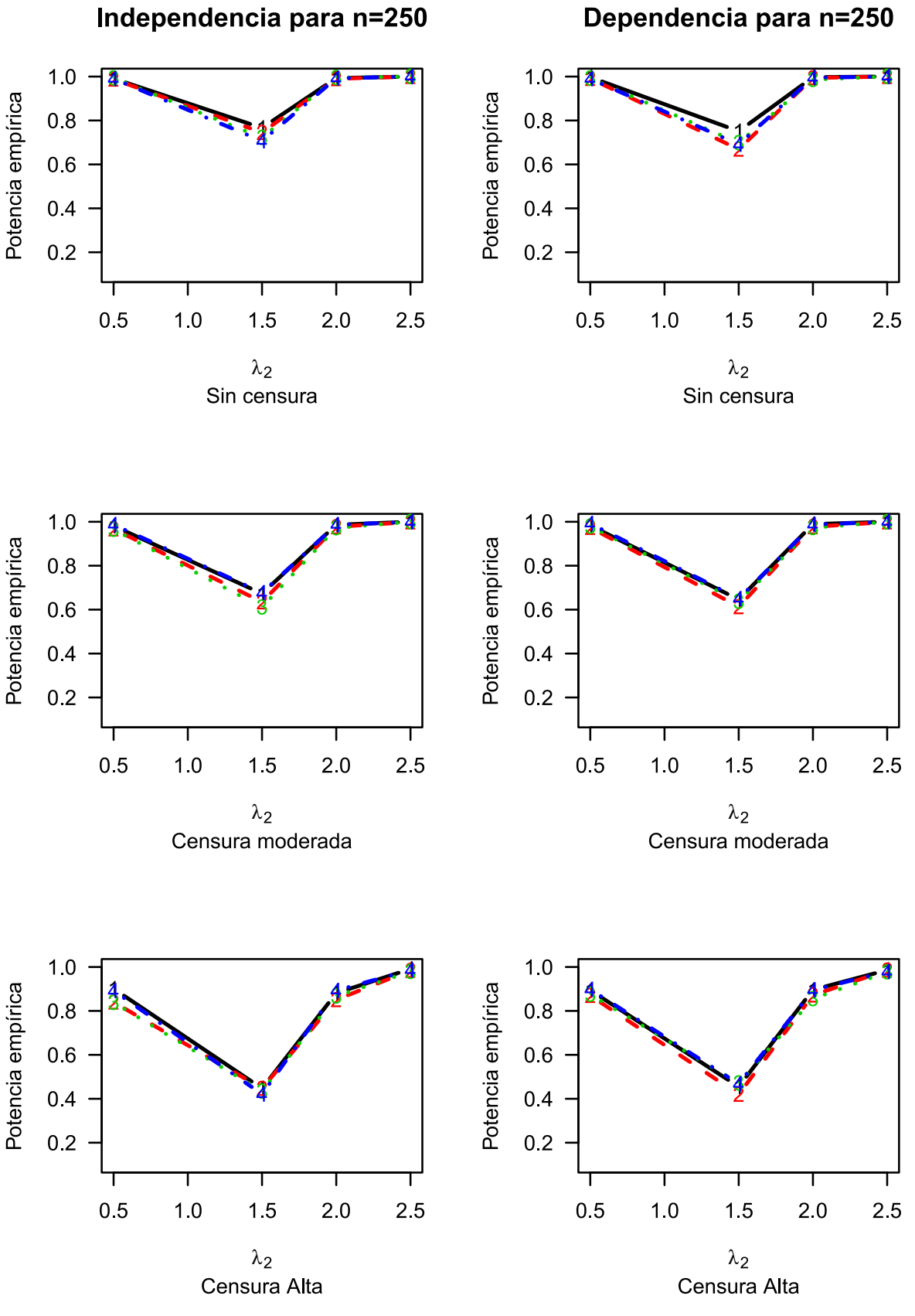

Figura 4. Potencia empírica alcanzada para tamaño de muestra de n=250.

6. EJEMPLO: PACIENTES CON LINFOMA

En la referencia [1] se presenta una base de datos de los pacientes con linfoma creada en el hospital Princess Margaret, Toronto, con registros que datan desde 1967. En la actualidad, la base de datos es de tipo prospectivo, pacientes que están siendo introducidos y registrados para recibir tratamiento en el hospital. Esta base de datos contiene un subgrupo de 541 pacientes de todos los pacientes identificados con linfoma tipo folicular, registrados para tratamiento en el hospital entre 1967 y 1996, con enfermedad en la etapa temprana y tratados con solo radioterapia o con radiación y quimioterapia. El objetivo de este estudio era mostrar los resultados a largo plazo en este grupo de pacientes.

El resultado registrado incluye la respuesta al tratamiento, la primera recaída y la muerte. El tiempo hasta la primera falla se calcula en años a partir de la fecha de diagnóstico. Un informe sobre una parte de este conjunto de datos se puede encontrar en [16]. De la anterior base de datos se tomó una muestra aleatoria de 74 pacientes, para determinar si hay diferencias entre los riesgo de recaída y muerte sin recaída, después del tratamiento. En la figura 5 se muestra las CIF´s de los riesgo de recaída y muerte sin recaída.

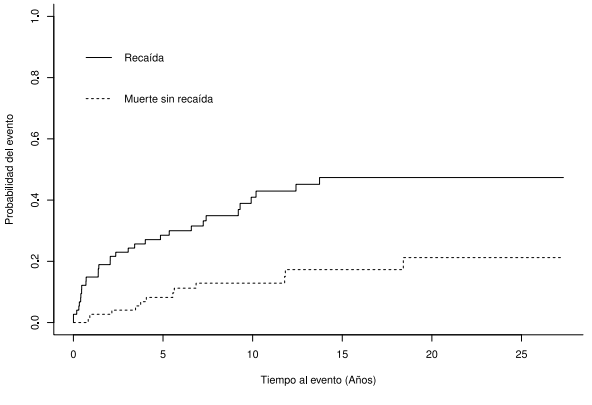

Figura 5. Comparación del riesgo por recaída con el riesgo por muerte sin recaída de los datos de pacientes con linfoma

A continuación se muestran en la tabla 4 los Valores p correspondientes a los estadísticos del supremo generalizado, los métodos bootstrap, **RAS** y la propuesta para comparar la incidencia de recaída con la incidencia de morir sin recaída después del tratamiento.

Tabla 4. Estadísticos de prueba (Valor-p) para los datos de pacientes con linfoma

Estadístico de prueba	Valor del estadístico	Valor - p
Supremo	2.7449	0.0242
RAS	1.4779	0.0010
Bootstrap	1.4779	0.0004
Propuesta	4.5625	0.0030

Tomando un nivel de significancia de $\alpha = 0.05$ los valores p de las pruebas son significativos, estos valores sugieren que hay una diferencia entre la incidencia de recaídas y la incidencia de morir sin recaída. En la tabla 5 la CIF de recaída está muy por encima de la CIF para la muerte sin recaída, apoyando la conclusión encontrada usando los estadísticos de prueba estudiados.

Aunque las tres pruebas arrojan la misma decisión, cabe mencionar que de acuerdo a la muestra que se tomó de 74 pacientes, esta posee un 41.89% de datos censurados, por tanto de acuerdo al estudio de simulación se está en un caso de censura moderada con una muestra relativamente pequeña; si existe alguna sospecha de que hay diferencias entre los dos riesgos de acuerdo al estudio de simulación en el caso de muestra pequeña, es recomendable usar la pruebas basadas en los métodos de remuestreo, dado que durante el estudio presentaron mejores comportamientos en comparación a la prueba del supremo generalizado en el momento en que hay presencia de censura.

7. CONCLUSIONES

En el estudio de simulación, se pudo evidenciar bajo los escenarios estudiados que las pruebas tienen un buen funcionamiento, cuando los tamaños de muestras son grandes. La prueba del supremo presenta errores tipo I más cercanos al nivel nominal fijado en comparación a las otras pruebas en todos los escenarios. En relación a la potencia cuando $n = 50$ esta prueba tiene un desempeño inferior a las pruebas basadas en remuestreo y para muestras grandes su desempeño depende de si se tiene o no censura.

Las pruebas basadas en métodos de remuestreo alcanzaron errores tipo I pequeños pero mostraron un mejor comportamiento cuando se encuen-

tra en un escenario de dependencia; computacionalmente el método **RAS** es más eficiente que el bootstrap. En relación a la potencia cuando n=50 los métodos de remuestreo funcionan mejor en comparación a la prueba del supremo, destacándose el método alternativo que está a la cabeza de sus pares y para muestras grandes cuando no hay censura el método del supremo alcanzó mayores potencias pero en escenarios con censura se observó que el método alternativo es muy buen competidor.

Una buena alternativa cuando hay presencia de censura es usar la propuesta del método de bootstrap alternativo usando las estimaciones de las CIF's y Kaplan-Meier, dado que estas metodologías están diseñadas especialmente para tratar datos en presencia de riesgos competitivos y eventos censurados, respectivamente. En los estudios de simulación realizados este método presentó un mejor comportamiento que los obtenidos por el método bootstrap y la aproximación **RAS**, e incluso supera en algunos escenarios al método del supremo.

Para trabajos futuros sería importante verificar que la prueba de proporcionalidad de funciones de incdencia acumulada propuesta en [5] es comparable con las pruebas estudiadas en este artículo con el fin de extender la comparación realizada. También es de interés evaluar el desempeño de las pruebas cuando hay más de dos riesgos competitivos presentes, usando como alternativa la distribución geométrica bivariada discreta para modelar los datos de tiempo de vida, lo cual podría traer algunas ventajas computacionales en comparación a la distribución de Basu [17]. Además, queda como trabajo futuro demostrar las propiedades de convergencia del método de boots- trap alternativo propuesto basado en las estimaciones de las CIF's y Kaplan-Meier.

REFERENCIAS

[1] M. Pintilie, *Competing risks: A practical perspective*. Canadá, John Wiley & Sons Ltd, 2006.

[2] A. Aly, S. C. Kochar, y I. W. McKeague, "Some tests for comparing cumulative incidence functions and cause-specific hazard rates", *Journal of the American Statistical Association*, Vol. 89, pp. 994-999, 1994.

[3] S. C. Kochar, K. F. Law y P. Yip, "Generalized Supremum Tests for the Equality of Cause Specific Hazard Rates", *Lifetime Data Analysis*, Vol.8, pp. 277-288, 2002.

[4] C. Y. Kam, Z. Lixing y Z. Dixin, "Comparing k Cumulative Incidence Functions Through Resampling Methods", *Lifetime Data Analysis*, Vol. 8, pp. 401-412, 2002.

[5] J.Y. Dauxois, S. N. U. A. Kirmani. "On testing the proportionality of two cumulative incidence functions in a competing risks setup". *Journal of Nonparametric Statistics*, Vol. 16(3-4), pa- ges 479-491, 2004.

[6] 1980 J. D Kalbfleisch, R. L. Prentice, *The Statistical Analysis of Failure Time Data*. New York, John Wiley & Sons Ltd, 1980.

[7] T. R. Fleming y D. P. Harrington, *Counting Processes and Survival Analysis*. New York, John Wiley & Sons Ltd, 1991.

[8] W. Feller, "The asymptotic distribution of the range of sums of independent random variables", *Annals of Mathematical Statistics*, Vol. 22, pp. 427-432, 1951.

[9] M. D. Burke y K. C. Yuen, "Goodness-of-fit tests for the Cox model via bootstrap method", *Journal of Statistical Planning Inference*, Vol. 47, pp. 237-256, 1995.

[10] K. C. Yuen, M. D. Burke, "A test of fit for a semiparametric additive risk model", *Biometrika*, Vol. 84, pp. 631-639, 1997.

[11] D. Pollard, *Convergence of Stochastic Processes*. New York, Springer-Verlag, 1984.

[12] H. Block y A. Basu, "A Continuous bivariate exponential extension", *Journal of the American Statistical Association*, Vol. 69, pp.1031- 1037, 1974.

[13] R. Leandro, y J. Achcar, "Generation of bivariate lifetime data assuming the Block & Basu exponential distribution", *Revista de matemática e estatística, Sao Paulo*, Vol. 14, pp. 43-52, 1996.

[14] D, S. Friday y G. P. Patil, "A Bivariate Exponential Model With Applications to Reliability and Computer Generation of Random Variables", *The Theory and Applications of Reliabi- lity With Emphasis on Bayesian and Nonpara- metric Methods* Vol.1, pp. 527-549, 1977. eds. C. P. Tsokos and I. N. Shimi, New York: Aca- demic Pres.

[15] Y. Sun y R. C. Tiwari,"Comparing Cause-Specific Hazard Rates of a Competing Risks Model with Censored Data", *Institute of Mathematical Statistics*, Vol. 27, pp. 225-270, 1995.

[16] P. M. Petersen, M. Gospodarowicz, R. Tsang, M. Pintilie,W. Wells, D. Hodgson, A. Sun,M. Crump, B. Patterson, y D. Bailey, "Long-term outcome in stage I and II follicular lymphoma following treatment with involved field radiation therapy alone", *Journal of Clinical Oncology*, vol. 22, pp. 563S, 2004.

[17] N. Davarzani, J. A. Achcar, y R. Peeters,"Bivariate lifetime geometric distribution in presence of cure fractions", *Journal of Data Science*, Vol. 13, pp. 755-770, 2015.