Trabajo Fin de Máster

Estimación presuavizada para el problema de truncamiento

Ana Nicieza Iglesias

Máster en Técnicas Estadísticas

Curso 2014-2015

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación presuavizada para o problema de truncamento
Título en español: Estimación presuavizada para el problema de truncamiento
English title: Presmoothed estimation under random truncation
Modalidad: A
Autora: Ana Nicieza Iglesias, Universidade de Santiago de Compostela
Director: César A. Sánchez Sellero, Universidade de Santiago de Compostela
Breve resumen del trabajo: El propósito de este trabajo es proponer una corrección del estimador máximo verosímil de la función de distribución para datos truncados, que permita solventar el problema de los conjuntos a riesgo con pocos individuos, o en el caso más grave con un solo individuo. En primer lugar se estudiarán las propiedades del estimador límite-producto, después se introducirá el problema de los conjuntos a riesgo pequeños, se propondrá un mecanismo de corrección por umbrales fijos y por un umbral escogido mediante bootstrap, y por último se estudiarán las propuestas presentadas mediante simulación.
Recomendaciones:
Otras observaciones:

Don César A. Sánchez Sellero, profesor titular de universidad de la Universidade de Santiago de Compostela, , informa que el Trabajo Fin de Máster titulado

Estimación presuavizada para el problema de truncamiento

fue realizado bajo su dirección por doña Ana Nicieza Iglesias para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, da su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 7 de septiembre de 2015.

El director:

Don César A. Sánchez Sellero

La autora:

Doña Ana Nicieza Iglesias

Índice general

Resumen

Resumen en español

En este trabajo se estudia el problema que padece el estimador límite-producto bajo truncamiento aleatorio, por el riesgo de que los conjuntos a riesgo sean pequeños, o incluso que posean un único individuo. Un conjunto a riesgo unitario da lugar a resultados anómalos del estimador límite-producto, mientras que conjuntos a riesgo con pocos individuos elevan considerablemente la varianza del estimador.

Posteriormente se propone un método de corrección del estimador límite-producto para superar estos inconvenientes, que consiste en fijar un umbral mínimo para los conjuntos a riesgo. Para la selección del umbral se propone un método bootstrap. Se estudian mediante simulaciones las propiedades de estimadores corregidos mediante umbrales fijos, así como del selector bootstrap. Se obtiene que la corrección por un umbral mínimo proporciona estimaciones más eficientes, en términos del error cuadrático medio, que el estimador límite-producto ordinario.

English abstract

The weakness of the product-limit estimator, due to possibly small risk sets, or even risk sets with only one observation, is studied. A risk set with only one observation leads to anomalous outcomes from the product-limit estimator, while risk sets with few observations lead to an increased variance of the product-limit estimator.

Then, a correction to the product-limit estimator is proposed, which consists of setting a lower threshold to the risk sets. A bootstrap method is proposed to choose the appropriate threshold. The behavior of the corrected estimator both with fixed thresholds and with a bootstrap threshold, is studied by a simulation study. The correction by a lower threshold is shown to be more efficient, in terms of mean squared error, than the ordinary product-limit estimator.

Capítulo 1

Estimación bajo truncamiento aleatorio

1.1. Introducción al análisis de supervivencia.

El análisis de supervivencia consiste en un conjunto de técnicas estadísticas que se emplean con el objetivo de analizar datos en los que la variable de interés es el tiempo que transcurre desde un instante inicial, bien definido, hasta la ocurrencia de un evento de interés o instante final. Es decir, el análisis de supervivencia trata de obtener información sobre la variable tiempo de vida, por ejemplo, tiempo desde el nacimiento hasta la muerte o desde que se contrae una enfermedad hasta su curación. A la ocurrencia del evento de interés se le suele denominar fallo o muerte.

Vamos a introducir una serie de conceptos previos, necesarios para la comprensión del resto del trabajo:

- Tiempo de vida: es el tiempo que transcurre desde un suceso inicial hasta la ocurrencia de un suceso final. Lo representaremos mediante la variable aleatoria X y será una variable continua y no negativa, $X \geq 0$.

- Tiempo de observación o seguimiento: es el tiempo que transcurre desde la fecha de entrada en el estudio hasta la fecha registrada en la última observación del individuo.

- Función de supervivencia: se define como la probabilidad de que el tiempo de vida estudiado sea mayor que un tiempo dado t. Es decir, si X es la variable aleatoria que representa el tiempo de vida, con función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$. La función de supervivencia $S(t)$ se define como:

$$S(t) = P[X > t] = 1 - F(t)$$

  Sus propiedades son:

- Monótona, decreciente y continua.
- $S(0) = 1$, $\lim_{t \to +\infty} S(t) = 0$.
- $S(t) = \mathrm{P}[X > t] = \int_t^{\infty} f(x)dx$, $0 \leq t \leq \infty$ donde f es la función de densidad asociada a X.

- Función de riesgo: la función de razón de riesgo o tasa instantánea de fallos $\lambda(t)$ se define como el cociente entre la función de densidad y la función de supervivencia:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

Se interpreta como la probabilidad de que el evento de interés ocurra en la siguiente unidad de tiempo Δt dado que no ha ocurrido hasta el tiempo t.

La función de riesgo acumulada $\Lambda(t)$ se define como:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$$

Ahora bien, en la práctica es muy probable y lo más frecuente, encontrarse con situaciones en las que no es posible la observación completa del tiempo desde el evento inicial hasta el evento final. Las dos causas principales de que esto ocurra son la presencia de censura y truncamiento. El análisis de supervivencia en caso de censura y/o truncamiento cobra una importancia primordial y también se le conoce como análisis de datos censurados y/o truncados. En este trabajo nos centraremos en el problema de truncamiento.

1.2. Modelo de truncamiento

Cuando se estudian los tiempos de vida de la variable de interés pueden aparecer peculiaridades en torno a la recogida de la información muestral. Esto es lo que ocurre precisamente en un esquema muestral con truncamiento, donde no se observarán todos los individuos del estudio sino únicamente aquellos que verifiquen cierta condición que se impone de antemano. En concreto, el tiempo de fallo se observa solamente cuando se excede cierto umbral aleatorio, denominado tiempo de truncamiento. Existen dos tipos de truncamiento, el truncamiento por la izquierda y el truncamiento por la derecha.

El truncamiento por la izquierda es el tipo de truncamiento más habitual. Ocurre cuando los individuos entran en distintos momentos al estudio y son observados desde ese 'tiempo retrasado de entrada' hasta que ocurre el evento final, de manera que aquellos individuos que experimenten el suceso de interés antes de incorporarse al estudio, serán excluidos de éste. Veamos un ejemplo que ilustre este tipo de truncamiento.

Supongamos que se está estudiando el tiempo de duración de una enfermedad vírica en los habitantes de un pueblo, y lo representamos con la variable X. En el estudio, solo se dispondrá de los datos de aquellos pacientes que acuden al médico del pueblo padeciendo

dicha enfermedad. Si representamos con la variable T al tiempo que pasa desde que la enfermedad comienza en el individuo hasta que éste decide acudir al médico, como solo estamos observando a aquellos individuos que acuden a la consulta, solo observamos a aquellos individuos con X mayor o igual que T. Y si el individuo cura antes de acudir al médico ($X < T$) el valor de X para ese individuo queda fuera del estudio.
Cualquier estudio que no tenga en cuenta esta situación proporcionará resultados sesgados. Ilustremos el ejemplo mediante un gráfico:

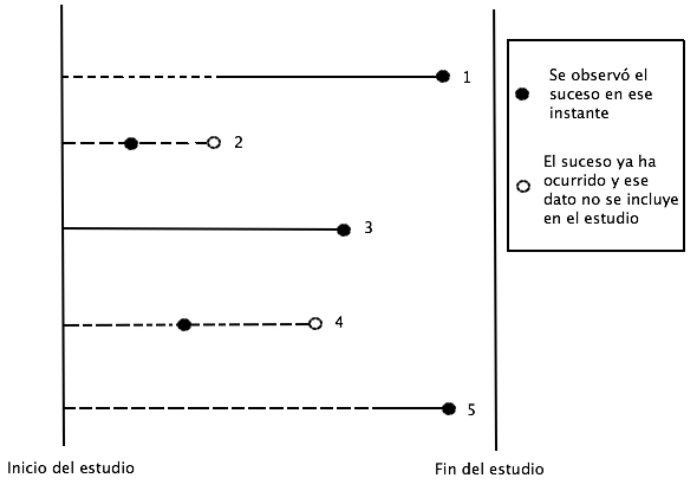

Figura 1.1: Truncamiento por la izquierda

Cada línea horizontal de la figura 1.1 corresponde a un individuo. El trazo continuo representa el tiempo de vida observado, es decir, la duración de la enfermedad observada, mientras que el discontinuo hace referencia a porciones de tiempo que no se observan, es decir, al tiempo de duración de la enfermedad que no se observa porque el individuo aún no ha acudido al médico. En el gráfico vemos que los individuos 2 y 4 han presentado el evento final antes de que comience su estudio, se curan antes de acudir al médico, es decir su tiempo de vida X es menor que el tiempo de truncamiento T, por lo que dichos individuos no serán incluidos en el estudio. Sin embargo, los individuos 1 y 5 comienzan a ser observados una vez empieza el estudio, pero el evento final aún no se produjo y tiene lugar tiempo después, llegando a ser observado. Es decir que acuden al médico antes de que se curen. El tiempo de vida X de estos individuos es mayor que el tiempo de truncamiento T y por lo tanto estos individuos sí que se incluyen en el estudio; son datos truncados por la izquierda. Por último el individuo 3 se observa desde el inicio del estudio hasta que se produce el evento final, por lo que este dato no posee truncamiento y es una observación exacta.

Con el ejemplo anterior se ilustra *el modelo de truncamiento aleatorio por la izquierda.* Planteado formalmente, se considera (X_i,T_i) con $i = 1, ..., N$ una muestra aleatoria simple de un cierto vector aleatorio formado por (X,T), dos variables positivas, aleatorias e independientes, donde X respresenta la variable de interés, tiempo de fallo, y T la variable de truncamiento. Entonces si suponemos que solo se observan aquellos pares que cumplen que $T_i \leq X_i$, la muestra observada es de la forma:

(X_i, T_i), tal que $T_i \leq X_i$ y con $i = 1, ...n$ donde n es aleatorio, N desconocido y $n \leq N$.

Podríamos decir que la variable T llamada de truncamiento impide la observación de la variable de interés X cuando esta última se sitúa a su izquierda $(X < T)$.

De forma análoga al truncamiento por la izquierda, se puede presentar el mecanismo de truncamiento por la derecha, que proporcionará observaciones solamente cuando la variable de interés X no supere la variable de truncamiento T. Entonces la muestra de datos observados será (X_i,T_i) con $X_i \leq T_i$.

Un ejemplo de este tipo de truncamiento por la derecha es el de los enfermos del SIDA. Supongamos que estudiamos el tiempo que transcurre entre la infección por el virus (VIH) hasta el desarrollo de la enfermedad, lo que se denomina periodo de latencia del virus del SIDA y lo definimos como X. Dado que solo se van a registrar individuos que han desarrollado el SIDA, solo se tendrán en cuenta aquellos cuyo tiempo de incubación sea menor o igual que la duración total del estudio (T), es decir solo entran en el estudio aquellos individuos que cumplan $X \leq T$, por tanto las observaciones son truncadas por la derecha. En estos estudios el tiempo de infección se averigua retrospectivamente para individuos que ya han desarrollado la enfermedad (era el caso de los primeros enfermos de SIDA infectados por transfusiones de sangre). Estos estudios retrospectivos constituyen una fuente de información importante sobre el tiempo de latencia del virus, pero presentan el problema de que al ser de individuos que han desarrollado la enfermedad dentro de un intervalo de tiempo, se desconoce el número total de personas que han sido infectadas por esa vía hasta el momento actual.

Si este problema se estudia considerando el tiempo cronológico de manera invertida, se transforma en un problema de análisis de truncamiento por la izquierda, lo que permite que nos centremos solo en este tipo de truncamiento.

Volviendo al tema que nos ocupa, que es el análisis del tiempo de vida, la distribución empírica no es un buen estimador de la función de distribución cuando los datos están truncados. Trabajar con la distribución empírica construída con las observaciones X_i proporcionaría valores sesgados.

El primero que aborda este problema es Lynden-Bell (1971), en un artículo del campo de la Astronomía, en el cual propone un estimador de la función de distribución para datos truncados, que resulta ser el estimador máximo verosímil de la función de distribución con truncamiento. Este estimador se conoce como *el estimador de Lynden-Bell o estimador límite producto* y en la siguiente sección expondremos con detalle en qué consiste. También analizaremos la primera investigación matemática que se realizó sobre este estimador y que se puede atribuir a Woodroofe (1985), quien también examinó algunos ejemplos de datos truncados en el campo de la Astronomía y la Economía. Véase también Wang (1989) para aplicaciones en el análisis de los datos del SIDA.

1.3. Estimador Límite Producto

El estimador límite producto, o estimador de Lynden-Bell, es el estimador máximo verosímil de la función de distribución en base a datos truncados y surge como alternativa a la distribución empírica, que no es un estimador adecuado para este tipo de datos. Sus propiedades son muy similares a las del estimador de Kaplan-Meier para datos censurados, y entre ellas cabe destacar la consistencia, debida a Woodroofe (1985), y distintas propiedades sobre la convergencia asintótica que se estudian en Woodroofe (1985), Wang et al. (1986) y Keiding y Gill (1990).

Recordamos que el modelo de truncamiento por la izquierda surge en el campo de la Astronomía, donde tiene interés estudiar la luminosidad de los objetos astronómicos. Supongamos que X mide el brillo de un objeto observado desde la Tierra. Nosotros solo podremos observar aquellos objetos que son lo suficientemente brillantes, es decir observaremos objetos con $T_i \leq X_i$, donde T_i representa el brillo mínimo detectable desde la Tierra.

La razón de que el estimador de Lynden-Bell sea un estimador límite producto se debe a que estima la probabilidad de supervivencia en un intervalo $(0, x]$ mediante el producto de supervivencias estimadas en cada uno de los subintervalos de una partición de $(0, x]$, construidas de forma que en cada uno de dichos intervalos solo haya una observación (intervalos límite).

En las siguientes líneas vamos a definir el estimador de Lynden-Bell siguiendo a Woodroofe (1985).

Para el modelo de truncamiento por la izquierda, al igual que ocurre para el modelo de censura, hay varias funciones importantes que están relacionadas con las variables de tiempo de vida y de truncamiento, y que es necesario definir antes de poder llegar a la expresión del estimador límite producto. Veámos cuales son:

Sean F y G las funciones de distribución de X y T, respectivamente, con X y T independientes y positivas, y sean N el tamaño muestral total y $n \leq N$ el número de individuos que se observan porque cumplen que $T_i \leq X_i$ $i = 1, ..., n$. Entonces por la Ley Fuerte de los Grandes Números,

$$\frac{n}{N} \to \alpha \equiv P(T \leq X), \text{ cuando } N \to \infty \text{ , con probabilidad 1.}$$

Asumimos sin pérdida de generalidad que $\alpha > 0$, pues en caso contrario no habría ninguna observación.

Denotamos por H_* a la distribución conjunta de (X, T) y por F_* y G_* a las distribuciones marginales de X y T, en ambos casos condicionadas a que $X \leq T$. Entonces

$$H_*(x, t) = \alpha^{-1} \int_0^x G(t \wedge z) dF(z), \tag{1.1}$$

y además $F_*(x) = H_*(x, \infty)$ y $G_*(t) = H_*(\infty, t)$ para $0 \leq x, t < \infty$, donde $\alpha \equiv P(T \leq X) = \int_0^\infty G(z)dF(z) = \int_0^\infty [1 - F(z^-)]dG(z)$, $t \wedge z$ denota el mínimo entre t y z; y $F(z^-) = lim_{x \uparrow z} F(x)$.

Una vez se definen estas funciones estamos más cerca de poder encontrar los estimadores consistentes de F y G, pero antes hay que demostrar que efectivamente F_* y G_* determinan F y G.

Supongamos que K es una función de distribución en $[0, \infty)$ y sea

$$a_K = \text{ínf}\{z > 0 : K(z) > 0\} \geq 0$$

y

$$b_K = \sup\{z > 0 : K(z) < 1\} \leq \infty.$$

Entonces (a_K, b_K) es el interior del soporte convexo de K.

A partir de estas definiciones, si $a_G < b_F$ se verificará que $\alpha > 0$ en la ecuación (1.1) y por otro lado mientras no se cumpla que $a_G \leq b_F$ $\alpha = 0$. De modo que si $\alpha > 0$ y si F_* y G_* están relacionadas con F y G por las ecuaciones (1.1), entonces $a_{F_*} = max\{a_F, a_G\}$, $b_{F_*} = b_F$, $a_{G_*} = a_G$ y $b_{G_*} = min\{b_F, b_G\}$. Es adecuado tomar la siguiente notación y definir los siguientes conjuntos $\mathfrak{K}$ y $\mathfrak{K}_o$ como:

$$\mathfrak{K} = \{(F, G) : F(0) = 0 = G(0), \alpha(F, G) > 0\}$$

$$\mathfrak{K}_o = \{(F, G) \in \mathfrak{K} : a_G \leq a_F, b_G \leq b_F\},$$

$$R(F, G) = H_*, \quad (F, G) \in \mathfrak{K}.$$

Esto nos permite definir dos lemas que dicen que si $(F, G) \in \mathfrak{K}$ y F_0 y G_0 son las funciones de distribución condicionales de X y T cumpliendo que $X \geq a_G$ y $T \leq b_F$. Entonces $(F_0, G_0) \in \mathfrak{K}_o$ y $R(F_0, G_0) = R(F, G)$. Y como por definición $\mathfrak{K}_o \subset \mathfrak{K}$ se verifica que $R(\mathfrak{K}) = R(\mathfrak{K}_o)$.

Con estos conjuntos definidos, podemos volver al análisis de supervivencia y definir la función de razón de fallo acumulada de la función de distribución F (con $F(0) = 0$) como:

$$\Lambda(x) = \int_0^x \frac{dF(z)}{[1 - F(z^-)]}, \qquad 0 \leq x < \infty,$$

a partir de ella podemos determinar de forma unívoca la función de distribución F, por la relación definida en la siguiente ecuación:

$$S_F(t) = 1 - F(t) = exp[-\Lambda_F(t)]$$

.

Y en el que caso que Λ_F presente discontinuidades, la relación anterior se expresa del siguiente modo:

$$1 - F(t) = exp[-\Lambda_F^c(t)] \prod_{a_i \in A/a_i \leq t} (1 - \Lambda_F\{a_i\}) \tag{1.2}$$

Si D denota el conjunto de valores x para los cuales $0 \leq x < b_F$ y $\lambda(x) = \Lambda(x) - \Lambda(x^-) > 0$; entonces

$$1 - F(x) = \left\{ \prod_{x \in D, z \leq x} [1 - \lambda(z)] \right\} exp[-\lambda_c(x)], \qquad 0 \leq x < b_F. \tag{1.3}$$

Y en base a los conjuntos que definimos anteriormente, si suponemos que $H_* \in R(\mathfrak{K})$. Entonces existe un único par $(F, G) \in \mathfrak{K}$ para el cual $R(F, G) = H_*$. El par (F, G) se determina por las siguientes condiciones:

$$\Lambda(x) = \int_0^x \frac{dF_*(z)}{C(z)}, \qquad 0 \leq x < \infty,$$

y

$$\int_t^\infty \frac{dG(z)}{G(z)} = \int_t^\infty \frac{dG_*(z)}{C(z)}, \qquad 0 \leq t < \infty,$$

donde $C(\cdot)$ es una función crucial en este contexto que se define como:

$$C(z) = P(T \leq z \leq X | T \leq X),$$

y denota el conjunto de individuos a riesgo en el modelo de truncamiento aleatorio por la izquierda. Siguiendo la notación, $C(\cdot)$ también es igual a:

$$C(z) = G_*(z) - F_*(z), \qquad 0 \leq z < \infty. \tag{1.4}$$

Todas estas condiciones dependen de una simple identidad, y es que $C(z) = \alpha^{-1} G(z)[1 - F(z^-)]$ para $z \geq 0$, que se deduce de lo siguiente:

Ya que α tal y como está definido en (1.1) representa la $P(T \leq X)$, condición que cumplen las observaciones del modelo, y X y T son independientes, entonces:

$$\begin{gathered} \alpha C(z) = P(T \leq X) P(T \leq z \leq X | T \leq X) = \\ = P(T \leq z \leq X, T \leq X) = P(T \leq X, T \leq z) - P(T \leq X, X < z) = \\ = P(T \leq X, T \leq z \leq X) = P(T \leq z) - P(X < z, T \leq z) = \\ = G(z) - G(z) F(z^-) = G(z)[1 - F(z^-)]. \end{gathered}$$

Por último, la función C puede estimarse consistentemente por:

$$C_n(z) = 1/n \sum_{i=1}^{n} 1_{T_i \leq z \leq X_i}$$

Llegados a este punto y con todas las condiciones planteadas, ya podemos estimar la función de riesgo acumulada Λ y definir el estimador de Lynden-Bell de la función de distribución para datos truncados.

Sean F y G las funciones de distribución de X y T con $(F, G) \in \Re$; X y T variables aleatorias independientes y $(X_1, T_1), ..., (X_N, T_N)$ realizaciones independientes e idénticamente distribuidas como (X, T). Suponemos que solo observamos los pares (X_i, T_i) tales que $i \leq N$ y $T_i \leq X_i$, garantizando que existe al menos uno. Y sean $(x_1, t_1), ..., (x_n, t_n)$ las observaciones de esos pares, de modo que $(x_1, t_1), ..., (x_n, t_n)$ son condicionalmente independientes e idénticamente distribuidas dado n.

Tomamos las funciones de distribución empírica de $x_1, ..., x_n$ y $t_1, ..., t_n$, F_n^* y G_n^*:

$$F_n^*(z) = \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \leq z},$$

$$G_n^*(z) = \frac{1}{n} \sum_{j=1}^{n} 1_{t_j \leq z}, \qquad 0 \leq z < \infty,$$

que estiman las funciones de distribución condicionales F_* y G_*. Si sustituimos F_* y G_* por sus estimadores en la ecuación (1.4), obtenemos el conjunto de individuos a riesgo denotado por C_n:

$$C_n(z) = G_n^*(z) - F_n^*(z^-), \qquad 0 \leq z < \infty,$$

que por como está definido y la condición de que $T_i \leq X_i$, cumple que $C_n(x_i) \geq \frac{1}{n}$ para todo $i \leq n$.

Entonces, el estimador de la función de razón de fallo se define como:

$$\hat{\Lambda}_n(z) = \int_0^z \frac{dF_n^*(x)}{C_n(x)} = \sum_{i:x_i \leq z} \frac{1}{nC_n(x_i)}.$$

Observamos que $\hat{\Lambda}_n$ es una función escalonada con discontinuidades (solamente) en los puntos $x_1, ..., x_n$.

Finalmente, la relación entre Λ y F dada por la ecuación (1.3) sugiere definir *el estimador límite producto de la función de distribución F propuesto por Lynden-Bell (1971)* como:

$$\hat{F}_n(z) = 1 - \prod_{i:x_i \leq z} \left[1 - \frac{r(x_i)}{nC_n(x_i)}\right], \qquad 0 \leq z < \infty,$$

donde $r(x_i) = \sum_{k \leq n} 1_{x_k = x_i}$ para $1 \leq i \leq n$, el producto se extiende sobre distintos valores de $x_1, ..., x_n$ y un producto nulo se interpreta como uno. La cantidad $r(x_i)$ es el número de individuos cuyo tiempo de vida coincide con x_i, por lo que en distribuciones continuas $r(x_i) = 1$. El factor $\frac{1}{nC_n(x_i)}$, se corresponde entonces con el riesgo en sentido ordinario al que está sometido cada individuo cuando alcanza el tiempo x_i.

De forma similar se puede definir el estimador de G:

$$\hat{G}_n(z) = \prod_{j:t_j>z} \left[1 - \frac{s(t_j)}{nC_n(t_j)}\right], \qquad 0 \leq z < \infty,$$

donde $s(t_j) = \sum_{k\leq n} 1_{t_k=t_j}$ para $1 \leq j \leq n$.

Una vez conseguidas las expresiones explícitas de ambos estimadores, podemos utilizar las siguientes representaciones gráficas para explicar cómo se estiman las funciones de distribución F y G mediante estos estimadores:

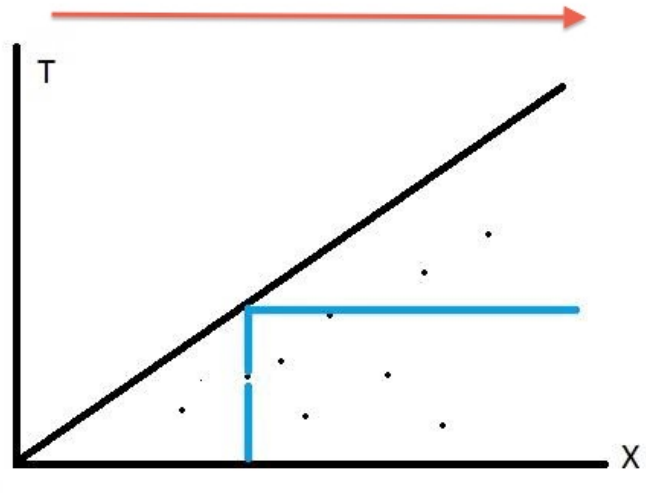

Figura 1.2: Estimación de F

Esta primera gráfica representa la estimación de F. El eje de abscisas corresponde a la variable tiempo de vida X y el eje de ordenadas a la variable de truncamiento T, los puntos representan las observaciones (x_i, t_i) que están por debajo de la diagonal $X = T$ porque en caso de truncamiento por la izquierda solo se observan los individuos que cumplen que $t_i \leq x_i$. El procedimiento que se lleva a cabo consiste en partir del individuo que tiene menor tiempo de vida X y desde él trazar una vertical hasta la recta $T = X$ y después una horizontal perpendicular a la primera, de modo que se van dibujando rectángulos tal y como se ve en el gráfico. El número de individuos cuyo tiempo de vida queda enmarcado en el interior del rectángulo, corresponde con el valor $nC_n(x_i)$, que es el número de individuos a riesgo en el instante x_i.

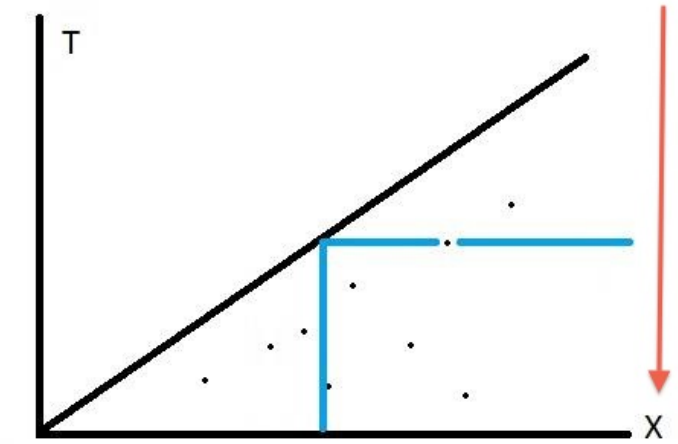

Figura 1.3: Estimación de G

Para el caso de la estimación de la función de distribución G, el procedimiento sería análogo al de F, pero comenzando por los individuos que tienen mayor tiempo de vida. Se trata de una estimación en la que el riesgo presente es un riesgo en tiempo invertido.

A partir del estimador de F que acabamos de obtener para el modelo de truncamiento también podemos obtener la estimación de la función de supervivencia S, que recordamos se define como la probabilidad de que el tiempo de vida estudiado sea mayor a un tiempo dado t, entonces:

$$\hat{S}_n(t) = 1 - \hat{F}_n(t) = \prod_{i:x_i \leq t} \left[1 - \frac{r(x_i)}{nC_n(x_i)}\right], \qquad 0 \leq t < \infty, \tag{1.5}$$

A continuación y una vez definido el estimador de Lynden-Bell, vamos a exponer algunas de las propiedades que cumple y en la siguiente sección las ilustraremos mediante un estudio de simulación sencillo.

Propiedades:

- El estimador de Lynden-Bell coincide con la función de distribución empírica en ausencia de truncamiento.

- El estimador de Lynden-Bell admite una representación gráfica mediante una función escalonada con saltos en las observaciones X_i.
- El estimador de Lynden-Bell es el estimador no paramétrico de máxima verosimilitud de F.
- El estimador de Lynden-Bell es un estimador consistente de la función de distribución; Si las funciones de distribución de X y T, F y G son continuas y los extremos superior e inferior del soporte convexo de G son menores o iguales a los de F, entonces los estimadores convergen a las distribuciones verdaderas F y G cuando el número de datos de la muestra tiende a infinito $N \to \infty$.
- El estimador de Lynden-Bell converge débilmente sobre intervalos compactos, propiedad que fue probada por Woodroofe (1985).
- La estimación del error para F converge en distribución a un proceso Gaussiano, siempre y cuando se verifique que $\int_0^\infty (1/G)dF < \infty$.

1.4. Ilustración mediante datos simulados

Vamos a realizar una simulación para evaluar, en una muestra de tamaño $n = 10$ observaciones, las propiedades del estimador propuesto.

En la simulación, el tiempo de vida X y el tiempo de truncamiento T se generan como mixturas de dos uniformes. En concreto,

$$\begin{aligned} X &\sim U[0,1] \text{ con probabilidad } 0,75 \text{ y } U[2,3] \text{ con probabilidad } 0,25 \\ T &\sim U[0,1] \text{ con probabilidad } 0,25 \text{ y } U[2,3] \text{ con probabilidad } 0,75 \end{aligned}$$

De este modo, la probabilidad de (no)-truncamiento, $P(T \leq X)$, vale tan solo $0,25$, por lo que hay un considerable fenómeno de truncamiento, que posteriormente producirá conjuntos a riesgo pequeños.

Una vez esté generada la muestra truncada, cumpliendo que X y T son independientes y que $T_i \leq X_i$ para todo $i = 1, ..., n$, calculamos los conjuntos a riesgo muestrales C_n en los x_i, $i = 1, ..., n$.

Recordamos que estos conjuntos se definen como:

$$C_n(z) = (1/n) \sum_i^n 1_{T_i \leq z \leq X_i}$$

A continuación se calcularía el estimador límite-producto $\hat{F}_n$:

$$\hat{F}_n(z) = 1 - \prod_{i:x_i \leq z} \left[1 - \frac{r(x_i)}{nC_n(x_i)} \right]$$

La Figura 1.4 contiene una representación de los conjuntos a riesgo, $nC_n(x_i)$, para los $n = 10$ datos generados. Para el valor de x_i más pequeño, que denotaremos como $x_{(1)}$, el conjunto a riesgo contiene 4 observaciones, lo cual es un valor pequeño, pues da lugar a una probabilidad de fallo estimada de $1/4 = 0,25$, que en este caso se convierte en la probabilidad asociada a ese individuo por el estimador límite-producto. De modo que el primer punto ya se lleva la cuarta parte de toda la probabilidad, cuando en la muestra observada hay $n = 10$ individuos. Nótese que si la muestra no estuviera truncada, la distribución empírica otorgaría la misma probabilidad $1/10$ a todos los datos, lo cual se correspondería con una probabilidad de fallo de $1/10$ en el dato menor y un conjunto a riesgo de 10 individuos, pues todos los individuos pertenecerían al conjunto a riesgo, en ausencia de truncamiento.

El estimador límite-producto produce entonces un efecto de desplazamiento de la probabilidad hacia los valores observados de X más pequeños, lo cual es razonable para compensar el efecto del truncamiento, que impide observar a muchos de estos individuos. Para ello, disminuye el valor de los conjuntos a riesgo, que son $(n - i + 1)$ para el dato i de X en la muestra ordenada sin tener en cuenta el truncamiento (los valores serían $\{n, n-1, n-2, \ldots, 3, 2, 1\}$), y se reducen a los valores $nC_n(x_{(i)})$ en la construcción del estimador límite-producto.

En el Cuadro 1.1 se presentan los valores numéricos de los conjuntos a riesgo, el estimador límite-producto y las probabilidades asociadas. En las dos primeras columnas se encuentran los pares (x, t) ordenados de forma ascendente según los valores x_i. En la columna tercera se muestran los valores asociados de C_n, en la cuarta los valores de $\hat{F}_n$ y en la quinta y última columna figuran las probabilidades que otorga el estimador límite-producto a cada observación $x_{(k)}$.

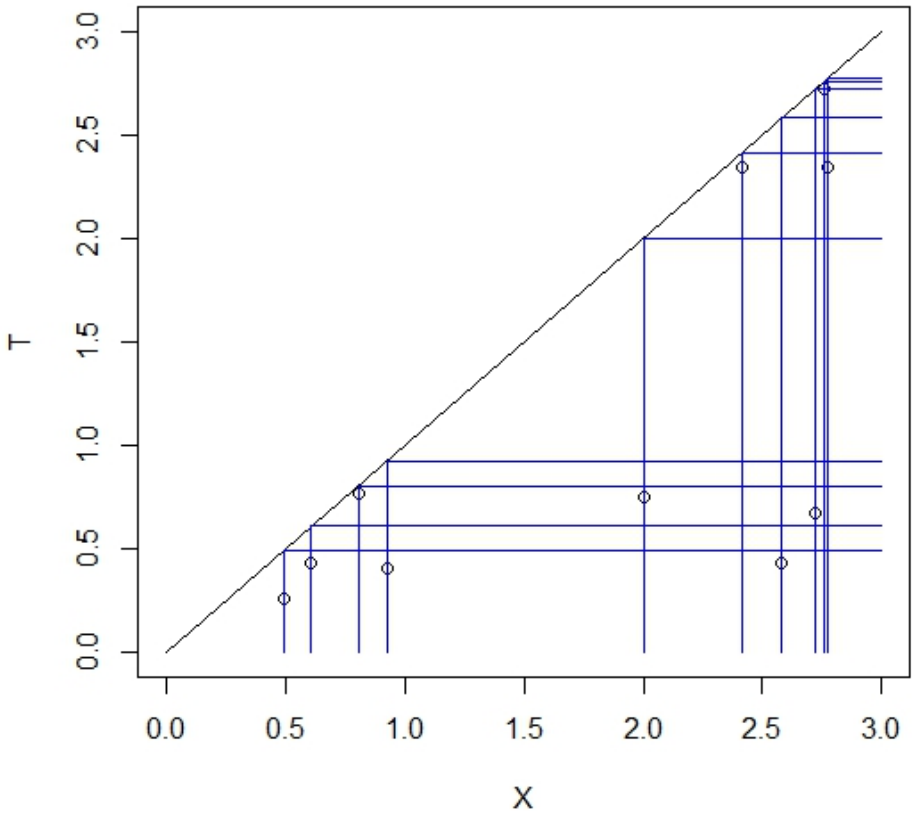

Figura 1.4: Conjuntos a riesgo, $C_n(x_i)$

Para completar la ilustración, vamos a comparar la media teórica de X, que sería $\mu = 0{,}75 * 0{,}5 + 0{,}25 * 2{,}5 = 1$, con sus estimaciones a través de la media muestral $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, y a través de la media ponderada $\hat{\mu}$, que calculamos a partir del estimador límite-producto de la siguiente manera:

$$\hat{\mu} = x_{(1)}\hat{F}_n[x_{(1)}] + \sum_{k=2}^{n} x_{(k)}(\hat{F}_n[x_{(k)}] - \hat{F}_n[x_{(k-1)}]) \tag{1.6}$$

$x_{(k)}$	$t_{[k]}$	$C_n(x_{(k)})$	$\hat{F}_n(x_{(k)})$	p_k
0.4887383	0.2547663	4	0.250	0.250
0.6070602	0.4275162	3	0.500	0.250
0.8039725	0.7650698	5	0.600	0.100
0.9238390	0.4041118	4	0.700	0.100
2.0028204	0.7462951	3	0.800	0.100
2.4168627	2.3467749	4	0.850	0.050
2.5834644	0.4322259	3	0.900	0.050
2.7224856	0.6667858	2	0.950	0.050
2.7595620	2.7264683	2	0.975	0.025
2.7743821	2.3464818	1	1.000	0.025

Cuadro 1.1: Valores obtenidos de C_n, $\hat{F}_n$ y de las probabilidades asociadas.

Según los resultados obtenidos, la media muestral de los tiempos de vida es $\bar{x} = 1{,}809$, mientras que la media ponderada calculada según (1.6) es $\hat{\mu} = 1{,}171$. Si comparamos el valor de la media ponderada 1,17 y el de la media que obtenemos de la muestra 1,809 vemos que el primer valor es notablemente menor que el segundo, esto es debido a que el estimador $\hat{F}_n$ tiende a corregir el sesgo de selección que se produce al haber por ejemplo un único punto en el intervalo $(0, 1/2]$ y cuatro en el intervalo $(5/2, 3/2]$. Para corregirlo el estimador asigna pesos mayores a los valores pequeños de $x_1, ..., x_n$, entonces el valor de la media se compensa. Ahora bien, la asignación de los pesos grandes a los valores más pequeños provoca que incremente la variabilidad, dando lugar a un comportamiento erróneo del estimador para tiempos de vida menores o iguales que $1/2$.

Podemos observar en la última columa del Cuadro 1.1, los pesos que adjudica el estimador límite-producto a cada una de las observaciones. En general en las muestras con gran presencia de truncamiento y con conjuntos a riesgos demasiado pequeños, como en nuestro caso, estos pesos serán muy desequilibrados.

Aunque en este estudio podamos controlar ambas adversidades, ante una muestra de datos reales no ocurrirá lo mismo. No podremos manejar la presencia de truncamiento de los datos, pues no depende de nosotros, pero en cambio sí podremos vigilar el tamaño de los conjuntos a riesgo, para evitar que sean muy pequeños. En los siguientes capítulos del trabajo propondremos procedimientos para garantizar un número mínimo de individuos

en estos conjuntos.

En vista a estos resultados, el estimador de Lynden-Bell no siempre proporciona los resultados correctos. Los conjuntos a riesgo, que son una sucesión determinista, $\{n, n-1, n-2, \ldots, 3, 2, 1\}$, con datos completos, e incluso con datos censurados, se convierten en cantidades aleatorias en presencia de truncamiento. Además, si un conjunto a riesgo vale uno, entonces la probabilidad estimada de fallo en ese punto será uno, lo cual impide otorgar probabilidad a datos posteriores. Esto constituye una anomalía grave del estimador, que veremos con más detalle en el próximo capítulo de este trabajo.

Además de esta anomalía, los conjuntos a riesgo reducidos como consecuencia del truncamiento, incrementan la varianza del estimador, pues la estimación de la probabilidad de fallo se realiza con menos observaciones. Este hecho también se estudiará con más detalle en el próximo capítulo. Por último, debemos observar que los conjuntos a riesgo se hacen pequeños en el extremo inferior de la distribución (esquina inferior izquerda de los diagramas de dispersión, como el que muestra la Figura 1.4), especialmente si las distribuciones de X y T tienen el mismo extremo inicial, $a_G = a_F$. Esto genera un problema particular de estimación de la distribución de X en su extremo inicial (cola de la izquierda).

En el siguiente capítulo explicaremos con más detalle todas estas dificultades del estimador de Lynden-Bell y presentaremos varias propuestas para solventarlas.

Capítulo 2

Corrección del estimador con truncamiento

2.1. Introducción

Vamos a mostrar que los estimadores $\hat{F}_n$ y $\hat{G}_n$ podrían llegar a admitir como soporte subconjuntos propios de $\{x_1, ..., x_n\}$ y $\{t_1, ..., t_n\}$. Sean $x_{(1)} < ... < x_{(n)}$ los valores ordenados de $\{x_1, ..., x_n\}$. Si

$$nC_n[x_{(k)}] = 1, \qquad \text{para algún } k,\ 1 \leq k < n,$$

entonces

$$\hat{F}_n[x_{(k)}] = 1.$$

Esta estimación no es correcta, pues significaría que toda la probabilidad residual recae sobre un individuo, mientras que la probabilidad del resto de individuos en estudio sería cero. Esto ocurre cuando el número de indiviuos a riesgo en un determinado instante es solo uno.

El hecho de que el conjunto de individuos a riesgo pueda contener pocos elementos es el principal inconveniente con el que nos podemos encontrar a la hora de construir el estimador de Lynden-Bell, y se agrava en el caso de que este conjunto contenga un único individuo, ya que se obtiene una estimación incorrecta.

Utilizando lo que sería un análogo de la fórmula de Greenwood con truncamiento, dada por Wang, Jewell & Tsai (1987), obtenemos una estimación práctica de la varianza asintótica de la función de supervivencia $\hat{S}_n = 1 - \hat{F}_n$, en la cual podemos ver el efecto que producen los conjuntos a riesgo con un sólo individuo:

$$\widehat{Var}(\hat{S}_n(x)) = (\hat{S}_n(x))^2 \prod_{x_j < x} \frac{d_j}{n_j(n_j - d_j)}, \tag{2.1}$$

donde d_j es el número de fallos en el instante x_j y n_j es el número de individuos a riesgo en el instante x_j, es decir, el valor de $nC_n(x_j)$.

En esa fórmula se ve que si los n_j son pequeños, la varianza es muy grande. De hecho, si $n_j = d_j$ entonces la varianza se hace infinita, porque hay un agujero en los datos; luego efectivamente la existencia de conjuntos a riesgo pequeños provoca un aumento en la varianza del estimador. En el caso de censura este problema no es tan grave (solo surge en el extremo derecho de los datos), porque los n_j siguen una secuencia determinista.

Gráficamente, en la Figura 1.2, se puede observar este suceso. Si el número de individuos cuyo tiempo de vida queda enmarcado en el interior del rectángulo que se define representa el valor $nC_n(x_i)$; que $nC_n(x_i) = 1$ equivale a que en el interior del rectángulo solamente hay un sujeto.

Los agujeros en los datos son otro grave problema de los conjuntos a riesgo y se deben a que estos conjuntos C_n pueden llegar a ser cero entre los puntos de los datos que se sitúan en la parte central. Keiding y Gill (1990) fueron los que reconocieron el peligro de estos conjuntos y llamaron a estos agujeros "empty inner risk sets", que se traduce como "conjuntos de riesgo internos vacíos".

Para intentar paliar el problema de los conjuntos a riesgo con un sólo individuo Stute y Wang (2008) propusieron una corrección sencilla del estimador; que consiste en lo siguiente:

Sea k_n una función decreciente para la cual $k_n(x) > k_n[x_{(n)}] = 1/n$ para todo $x < x_{(n)}$. Si reemplazamos C_n por

$$C_n^*(z) = max\{C_n(z), k_n(z)\}, \qquad 0 \leq z \leq x_{(n)}$$

en la definición del estimador $\hat{F}_n$, resultará otro estimador $\hat{F}_n^*$ que no podrá tener como soporte ningún subconjunto propio de $\{x_1, ..., x_n\}$.

De hecho, $1/nk_n\{[x_i]\}$ es la proporción máxima de la probabilidad estimada $1 - F_n^*(x_{(i)}^-)$, que el experimentador está dispuesto a asignar a $x_{(i)}$ para $i = 1, ..., n$. En grandes rasgos, esta propuesta se basa en robustificar con un recorte de conjuntos a riesgo pequeños, garantizando un umbral mínimo de individuos.

En base a esta modificación, obtendríamos un estimador asintóticamente equivalente a $\hat{F}_n$:

$$\hat{F}_n^*(z) = 1 - \prod_{i:x_i \leq z} \left[1 - \frac{r(x_i)}{nC_n(x_i) + 1}\right], \qquad 0 \leq z < \infty.$$

Con el sumando 1 en el denominador se garantiza que en los conjuntos a riesgo haya al menos algún individuo y permite contribuciones a la masa de $\hat{F}_n$ con los datos que tienen agujeros en su izquierda. Stute y Wang (2008) estudiaron el error cuadrático medio que provocaba esta modificación y comprobaron mediante simulaciones que en caso de muestras pequeñas el error resultante era menor, pero en muestras grandes apenas se notaba diferencia entre el estimador de Woodroofe (1985) y el propuesto por Stute y Wang (2008).

Basándonos en estos resultados, nuestro objetivo va a ser encontrar un umbral mínimo de individuos en los conjuntos a riesgo, que nos permita obtener un estimador robustificado más eficiente en comparación con el estimador límite-producto y que no proporcione estimaciones anómalas.

Mediante simulaciones para varios *umbrales mínimos fijos* que elegiremos previamente y para un *umbral escogido por bootstrap* aproximaremos los sesgos, varianzas y errores cuadráticos medios del estimador límite-producto para cada umbral de corte. El umbral mínimo con el que se obtenga un menor error cuadrático medio será el que garantice un estimador límite-producto más eficiente.

El error cuadrático medio es uno de los criterios de error más empleados y se define como:

$$ECM(\hat{\theta}) = Var(\hat{\theta}) + sesgo(\hat{\theta})^2$$

siendo $\hat{\theta}$ un estimador de θ. El ECM mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. Además, es igual a la suma de la varianza y el cuadrado del sesgo del estimador.

En las dos secciones siguientes exponemos dos propuestas novedosas: la corrección mediante umbrales mínimos fijos y con un umbral escogido por bootstrap.

2.2. Umbral mínimo de los conjuntos a riesgo

En esta sección proponemos corregir el estimador límite-producto mediante un umbral fijo que garantice un mínimo de individuos en los conjuntos a riesgo.

Recordamos que el estimador límite-producto de la función de distribución F para una muestra de datos truncados se define como:

$$\hat{F}_n(z) = 1 - \prod_{i:x_i \leq z} \left[1 - \frac{r(x_i)}{nC_n(x_i)}\right], \qquad 0 \leq z < \infty.$$

donde el factor $nC_n(x_i)$ hace referencia al número de individuos que se consideran en riesgo en un cierto instante y es el que nos interesa corregir.

La propuesta de un umbral mínimo fijo consiste simplemente en escoger un número, que denotaremos por u_f, que será la cantidad mínima que deben tomar todos los conjuntos a riesgo. De este modo, los conjuntos a riesgo corregidos quedarían definidos como el máximo de u_f y el propio valor de cada conjunto a riesgo.

Así, si $u_f = 2$, esta corrección supone incrementar de 1 a 2 los conjuntos a riesgo que tengan un solo individuo. Con esta sencilla corrección se suprime el problema de los conjuntos a riesgo unitarios y se suaviza levemente la varianza del estimador. Por supuesto, nos podemos plantear umbrales mayores que 2.

Ahora vamos a añadir un detalle adicional: no sería razonable incrementar un conjunto a riesgo más allá del valor que tendría en una muestra no truncada, que es $(n - i + 1)$ para el dato i-ésimo de la muestra ordenada de los x_i, que venimos denotando por $x_{(i)}$. Nótese que $nC_n(x_{(i)}) \leq n - i + 1$ para todo $i \in \{1, \ldots, n\}$. Finalmente, la propuesta de

conjuntos a riesgo corregidos por umbral fijo quedaría formulada de la siguiente manera:

$$nC_n^{r,u_f}(x_{(i)}) = \text{mín}\{n - i + 1, \text{máx}\{u_f, nC_n(x_{(i)})\}\}$$

Estos conjuntos a riesgo corregidos, únicamente podrán producir incrementos respecto del valor de los conjuntos a riesgo originales. La consecuencia será una disminución de la varianza del estimador de la distribución, como parece deducirse de la fórmula de Greenwood dada en (2.1). Sin embargo, cabe esperar un sesgo del estimador ocasionado por esta corrección, pues los incrementos de los conjuntos a riesgo suponen una disminución de las probabilidades estimadas de fallo. Al disminuir las probabilidades de fallo, se desplazarán las probabilidades otorgadas por el estimador de la distribución hacia valores mayores de la variable. Esto producirá un sesgo negativo en el estimador límite-producto de la distribución con conjuntos a riesgo corregidos, que definimos así

$$\hat{F}_n^{r,u_f}(z) = 1 - \prod_{i:x_i \leq z} \left[1 - \frac{r(x_i)}{nC_n^{r,u_f}(x_i)}\right]$$

Nótese que el estimador límite-producto ordinario puede presentar sesgo en sí mismo, pero es asintóticamente insesgado como cabe esperar de un estimador de máxima verosimilitud.

En este trabajo vamos a evaluar las propuestas de corrección mediante el estimador de la media de X, que ya hemos considerado en la Sección 1.4. En concreto, el estimador de la media mediante los pesos de un estimador límite-producto con conjuntos a riesgo corregidos se definiría así:

$$\hat{\mu} = x_{(1)}\hat{F}_n^{r,u_f}[x_{(1)}] + \sum_{k=2}^{n} x_{(k)}(\hat{F}_n^{r,u_f}[x_{(k)}] - \hat{F}_n^{r,u_f}[x_{(k-1)}])$$

El desplazamiento de las probabilidades hacia valores superiores de la variable X, que hemos mencionado anteriormente, producirá un sesgo positivo en el estimador de la media, $\hat{\mu}$.

La consecuencia de todo esto es que umbrales u_f pequeños presentan poco sesgo pero conservan mucha varianza, mientras que umbrales grandes reducen mucho la varianza a costa de un gran sesgo. La selección del umbral adecuado se parece de esta manera a un problema de encontrar el equilibrio entre sesgo y varianza, semejante al de selección de un parámetro de suavizado en Estadística no paramétrica.

A pesar de esta similitud con los problemas que surgen en los métodos de suavización, debemos indicar que la propuesta finalmente presentada en este trabajo se adecúa más a un propósito de robustización del estimador límite-producto. A su vez, el método de umbrales también se asemeja más a un método robusto que a un método de suavización. Aunque se han estudiado procedimientos de suavización entre los conjuntos a riesgo, en coherencia con el título de la propuesta de trabajo fin de máster, el mejor funcionamiento del método de umbrales fijos ha hecho que nos decantáramos por un método de este tipo.

2.3. Selección del umbral mínimo mediante bootstrap

A continuación se plantea un método bootstrap para la selección del umbral de corrección de los conjuntos a riesgo. En este caso en vez de con un umbral mínimo fijo, los conjuntos se corregirán mediante un umbral escogido por bootstrap.

El método bootstrap, que explicaremos a continuación, seleccionará, para cada muestra simulada, el mejor umbral mínimo de entre varios, escogidos a tanteo, que corregirán los conjuntos a riesgo de remuestras bootstrap de la muestra original.

El método bootstrap fue aplicado por Efron (1981) a datos censurados, mientras que Wang (1991) extendió el algoritmo bootstrap obvio para datos truncados. Aquí aplicaremos el método bootstrap obvio para estimar el error cuadrático medio del estimador de la media $\hat{\mu}^{ru_f}$ para distintos valores de u_f. Después se escogerán como umbral bootstrap aquél que haga mínimo el error cuadrático medio estimado. Los pasos del procedimiento bootstrap serían los siguientes:

1. Se calculan los estimadores límite-producto de las distribuciones de la variable de interés, $\hat{F}_n^{r,2}(z)$, y de la variable de truncamiento, $\hat{G}_n^{r,2}(z)$ con los conjuntos a riesgo corregidos según un pequeño valor mínimo (en las simulaciones hemos empleado el valor 2).

2. Se arrojan observaciones bootstrap independientes, x_i^* con distribución $\hat{F}_n^{r,2}$ y t_i^* con distribución $\hat{G}_n^{r,2}$. Si $t_i^* \leq x_i^*$ se toma el par (t_i^*, x_i^*), en caso contrario se sigue extrayendo hasta conseguir n pares que cumplan la desigualdad.

3. Se considera la remuestra bootstrap $(t_1^*, x_1^*), ..., (t_n^*, x_n^*)$.

4. Se calculan los conjuntos a riesgo de las observaciones bootstrap $\{x_1^*, ..., x_n^*\}$.

5. Se corrigen los conjuntos a riesgo obtenidos según varios recortes tanteados y se estima la función de distribución y la media de X^* con los conjuntos corregidos. Resultarán los valores bootstrap del estimador:

$$\hat{\mu}^{r,u_f,*} \qquad \text{para distintos valores de } u_f$$

Se repiten los pasos $2-5$ B veces y se calcula la estimación bootstrap del error cuadrático medio para cada umbral u_f de la siguiente manera:

$$E\hat{C}M(\hat{\mu}^{r,u_f}) = \frac{1}{b}\sum_{b=1}^{B}\left(\hat{\mu}^{r,u_f,*b} - \hat{\mu}^{r,2}\right)^2$$

siendo $\hat{\mu}^{r,2}$ el estimador de la media asociado a la distribución $\hat{F}_n^{r,2}$ de la cual se generan las muestras bootstrap.

Finalmente el selector bootstrap del umbral u_f será el valor que haga mínima la expresión anterior, esto es:

$$u_b = \arg\min_{u_f} E\hat{C}M(\hat{\mu}^{r,u_f})$$

2.4. Estudio de simulación

En esta sección presentaremos un estudio de simulación para analizar las propiedades de sesgo, varianza y error cuadrático medio del estimador de la media con umbrales fijos para los conjuntos a riesgo y con un umbral escogido por bootstrap. Además, se van a comparar con el estimador original, basado en el estimador límite-producto ordinario.

El tiempo de vida X y el tiempo de truncamiento T se van a generar a partir de una mixtura de dos uniformes de manera similar a los datos presentados en la Sección 1.4, esto es:

$$\begin{aligned} X &\sim U[0,1] \text{ con probabilidad } (1-p) \text{ y } U[2,3] \text{ con probabilidad } p \\ T &\sim U[0,1] \text{ con probabilidad } p \text{ y } U[2,3] \text{ con probabilidad } (1-p) \end{aligned}$$

De este modo, la probabilidad de (no)-truncamiento, $P(T \leq X)$ coincide con p, lo cual nos permite considerar valores diferentes de esta probabilidad. Tomaremos los valores $p = 0{,}25$ y $p = 0{,}6$.

Consideraremos también modelos con distribuciones exponenciales para X y para T, donde el parámetro de las dos distribuciones exponenciales permite controlar la probabilidad de (no)-truncamiento, que también tomarán los valores 0,25 y 0,6.

Finalmente, los cuatro modelos de truncamiento que vamos a considerar son:

* **Modelo 1**: Mixtura de dos uniformes con probabilidad de (no)-truncamiento 0,25.
* **Modelo 2**: Mixtura de dos uniformes con probabilidad de (no)-truncamiento 0,6.
* **Modelo 3**: Exponenciales $X \sim Exp(1)$ y $T \sim Exp(0{,}33)$, y probabilidad de (no)-truncamiento 0,25.
* **Modelo 4**: Exponenciales $X \sim Exp(1)$ y $T \sim Exp(1{,}5)$ y probabilidad de (no)-truncamiento 0,6.

Para llevar a cabo la simulación se generan 1000 muestras aleatorias de tamaño $n = 20, 50, 100$ bajo cada modelo y con los siguientes parámetros fijos:

- El número de réplicas bootstrap es $B = 100$, por lo general es un número de réplicas suficiente para obtener una buena estimación del error cuadrático medio.
- Los umbrales fijos mínimos que se utilizan son 0, 2, 3, 4, 6, 8, 10, 15 y 20. Cuando el umbral mínimo es 0, los conjuntos a riesgo no se corrigen. Por lo tanto, el estimador obtenido en ese caso es el estimador límite-producto ordinario.
- El umbral mínimo para los conjuntos a riesgo en las distribuciones bootstrap es 2.
- Los recortes tanteados por bootstrap son 0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20.

En los Cuadros 2.1 y 2.2 de esta sección se muestran el sesgo, la varianza y el error cuadrático medio para cada uno de los estimadores-límite producto corregidos según los umbrales mínimos fijos y según el umbral mínimo escogido por bootstrap en los cuatro escenarios de truncamiento. El valor 0 del umbral mínimo da lugar al estimador límite-producto ordinario, por lo que estos resultados permiten comparar las correcciones propuestas con el estimador sin corrección.

		Modelo 1			*Modelo 2*		
n	Umbral	*Sesgo*	*Varianza*	*ECM*	*Sesgo*	*Varianza*	*ECM*
20	0	0.0485	0.1307	0.1330	0.0157	0.2032	0.2034
	2	0.0643	0.1093	0.1134	0.0359	0.1559	0.1572
	3	0.0879	0.0947	0.1024	0.0540	0.1335	0.1364
	4	0.1189	0.0824	0.0965	0.0742	0.1164	0.1219
	6	0.2217	0.0610	0.1102	0.1194	0.0912	0.1054
	8	0.3730	0.0492	0.1884	0.1697	0.0728	0.1016
	10	0.5324	0.0443	0.3278	0.2274	0.0566	0.1084
	15	0.7984	0.0421	0.6796	0.3884	0.0326	0.1834
	20	0.8748	0.0447	0.8100	0.4633	0.0299	0.2445
	bootstrap	0.0829	0.1186	0.1254	0.1007	0.1613	0.1714
50	0	0.0131	0.0718	0.0720	0.0112	0.0953	0.0954
	2	0.0218	0.0597	0.0602	0.0180	0.0784	0.0787
	3	0.0321	0.0520	0.0530	0.0247	0.0700	0.0707
	4	0.0418	0.0471	0.0489	0.0322	0.0640	0.0651
	6	0.0612	0.0405	0.0443	0.0470	0.0559	0.0581
	8	0.0811	0.0365	0.0430	0.0630	0.0500	0.0540
	10	0.1044	0.0329	0.0438	0.0789	0.0457	0.0519
	15	0.1986	0.0230	0.0625	0.1215	0.0378	0.0526
	20	0.3661	0.0179	0.1519	0.1707	0.0306	0.0597
	bootstrap	0.0512	0.0609	0.0636	0.0697	0.0700	0.0748

100	0	0.0171	0.0352	0.0355	0.0022	0.0515	0.0515
	2	0.0201	0.0309	0.0313	0.0050	0.0446	0.0447
	3	0.0225	0.0288	0.0293	0.0086	0.0401	0.0402
	4	0.0256	0.0272	0.0279	0.0113	0.0377	0.0378
	6	0.0328	0.0249	0.0260	0.0178	0.0339	0.0342
	8	0.0410	0.0229	0.0246	0.0247	0.0309	0.0315
	10	0.0490	0.0216	0.0240	0.0318	0.0287	0.0297
	15	0.0710	0.0190	0.0240	0.0505	0.0250	0.0276
	20	0.0963	0.0170	0.0262	0.0708	0.0222	0.0272
	bootstrap	0.0449	0.0291	0.0312	0.0361	0.0364	0.0377

Cuadro 2.1: Cálculo de sesgos, varianzas y ECMs en los modelos 1 y 2.

		Modelo 3			*Modelo 4*		
n	Umbral	*Sesgo*	*Varianza*	*ECM*	*Sesgo*	*Varianza*	*ECM*
20	0	0.0282	0.1441	0.1449	0.0239	0.1044	0.1050
	2	0.0546	0.1120	0.1150	0.0350	0.0887	0.0899
	3	0.0784	0.0978	0.1040	0.0478	0.0801	0.0824
	4	0.1065	0.0875	0.0989	0.0604	0.0744	0.0781
	6	0.1826	0.0709	0.1042	0.0904	0.0658	0.0740
	8	0.2895	0.0588	0.1426	0.1321	0.0584	0.0759
	10	0.4178	0.0566	0.2312	0.1890	0.0524	0.0882
	15	0.6628	0.0715	0.5108	0.3488	0.0538	0.1754
	20	0.7340	0.0778	0.6166	0.4071	0.0583	0.2241
	bootstrap	0.0788	0.1196	0.1258	0.0624	0.0929	0.0968
50	0	-0.0024	0.0760	0.0760	0.0036	0.0483	0.0483
	2	0.0086	0.0610	0.0611	0.0082	0.0416	0.0417
	3	0.0190	0.0536	0.0540	0.0133	0.0380	0.0382
	4	0.0294	0.0488	0.0497	0.0185	0.0355	0.0359
	6	0.0489	0.0429	0.0453	0.0283	0.0326	0.0334
	8	0.0708	0.0390	0.0440	0.0381	0.0307	0.0321
	10	0.0950	0.0360	0.0450	0.0483	0.0291	0.0314
	15	0.1691	0.0296	0.0582	0.0781	0.0262	0.0323
	20	0.2796	0.0234	0.1016	0.1169	0.0236	0.0372
	bootstrap	0.0382	0.0626	0.0641	0.0350	0.0406	0.0418

100	0	0.0026	0.0399	0.0399	0.0039	0.0274	0.0274
	2	0.0073	0.0333	0.0333	0.0068	0.0230	0.0231
	3	0.0106	0.0306	0.0307	0.0091	0.0211	0.0212
	4	0.0142	0.0286	0.0288	0.0111	0.0201	0.0202
	6	0.0223	0.0256	0.0261	0.0154	0.0187	0.0189
	8	0.0309	0.0236	0.0245	0.0198	0.0176	0.0180
	10	0.0402	0.0221	0.0237	0.0242	0.0167	0.0173
	15	0.0654	0.0194	0.0237	0.0357	0.0152	0.0165
	20	0.0938	0.0177	0.0265	0.0485	0.0141	0.0164
	bootstrap	0.0326	0.0313	0.0324	0.0247	0.0203	0.0209

Cuadro 2.2: Cálculo de sesgos, varianzas y ECMs en los modelos 3 y 4.

En todos los modelos, se observa que *la varianza* del estimador límite-producto sin corregir (valor del umbral mínimo 0) es mayor que cuando se corrige con cualquier umbral. Esto se debe a que al poner los umbrales mínimos se produce una disminución de la varianza. Sin embargo, se crea un sesgo adicional, porque se está disminuyendo artificialmente la razón de fallo. En los datos de las tablas también se puede ver cómo *el sesgo* es mayor para los estimadores corregidos que para el estimador sin corregir.

En referencia a los tamaños muestrales $n = 20, 50$ y 100 el comportamiento es similar para los cuatro modelos. A medida que el tamaño muestral aumenta, los umbrales mínimos que garantizan el menor error cuadrático medio también aumentan. Y si para tamaño $n = 20$ los umbrales más grandes son los que provocan un error cuadrático mayor, si el tamaño muestral es 100, para los cuatro modelos el peor error cuadrático medio se obtiene con el estimador sin corregir (umbral 0).

Además en cada modelo por separado, también observamos que al aumentar el tamaño muestral, las varianzas disminuyen en general y los sesgos disminuyen con umbrales fijos. Ambos fenómenos son conformes a lo esperado.

Indicamos a continuación los umbrales fijos que dan lugar al menor error cuadrático medio:

- Modelo 1 y tamaño $n = 20 \longrightarrow$ Umbral mínimo fijo 4.
- Modelo 1 y tamaño $n = 50 \longrightarrow$ Umbral mínimo fijo 8.
- Modelo 1 y tamaño $n = 100 \longrightarrow$ Umbral mínimo fijo 10.

- Modelo 2 y tamaño $n = 20 \longrightarrow$ Umbral mínimo fijo 8.
- Modelo 2 y tamaño $n = 50 \longrightarrow$ Umbral mínimo fijo 10.
- Modelo 2 y tamaño $n = 100 \longrightarrow$ Umbral mínimo fijo 20.

- Modelo 3 y tamaño $n = 20 \longrightarrow$ Umbral mínimo fijo 4.
- Modelo 3 y tamaño $n = 50 \longrightarrow$ Umbral mínimo fijo 8.
- Modelo 3 y tamaño $n = 100 \longrightarrow$ Umbral mínimo fijo 15.

- Modelo 4 y tamaño $n = 20 \longrightarrow$ Umbral mínimo fijo 6.
- Modelo 4 y tamaño $n = 50 \longrightarrow$ Umbral mínimo fijo 10.
- Modelo 4 y tamaño $n = 100 \longrightarrow$ Umbral mínimo fijo 20.

Si recordamos que los modelos 1 y 3 tienen la misma probabilidad de (no)-truncamiento 0,25 y los modelos 2 y 4 probabilidad 0,6, puede entenderse por qué los resultados son tan similares para estas dos parejas de modelos.

Al observar los umbrales mínimos que se obtienen para cada tamaño muestral nos damos cuenta de que la mayoría de estos valores son aproxidamente el 20 % de dicho tamaño. Esto supondría afirmar que el umbral mínimo de individuos con el que se deben corregir los conjuntos a riesgo para obtener el estimador límite-producto de la función de distribución más eficiente para X debe ser aquel valor en torno al 20 % del tamaño muestral.

Por último es importante mencionar que con los umbrales mínimos escogidos por bootstrap, en ningún modelo se obtiene el menor error cuadrático medio, aunque tampoco el peor. Corregir los conjuntos a riesgo con este umbral no nos proporcionaría el estimador límite-producto más eficiente, objetivo por otra parte inalcanzable, pues el umbral óptimo es en la práctica desconocido y cualquier selector únicamente puede intentar aproximarse a él. En todo caso, los resultados del selector bootstrap son más que aceptables.

Destacamos finalmente que el estimador límite-producto sin corrección presenta peor error cuadrático medio que la mayor parte de los umbrales considerados, y es claramente peor que el estimador corregido mediante un umbral bootstrap.

Capítulo 3

Conclusiones y problemas abiertos.

El estimador de Lynden-Bell (1971) surge como alternativa a la función de distribución empírica en situaciones de muestras truncadas. En este trabajo hemos propuesto métodos para corregir este estimador, con objeto de obtener mejores resultados en caso de que los conjuntos a riesgo se queden con pocos datos o incluso con un único dato.

Para ello, en primer lugar hemos dado la definición de los principales conceptos del análisis de supervivencia y hemos explicado en qué consiste el modelo de truncamiento aleatorio por la izquierda. Hemos visto ejemplos de este tipo de truncamiento y hemos definido el estimador límite-producto, mencionando algunas de sus principales propiedades, que posteriormente hemos ilustrado mediante datos simulados.

Una vez claro el funcionamiento del estimador de Lynden-Bell, hemos detectado el problema del estimador cuando los conjuntos a riesgo son demasiado pequeños. Para intentar solventar este inconveniente hemos propuesto dos correcciones de los conjuntos a riesgo, mediante umbrales mínimos fijos y mediante un umbral corregido por bootstrap. Hemos hecho un estudio de simulación considerando varios modelos de truncamiento y hemos visto para distintos tamaños muestrales cuáles son los umbrales mínimos que proporcionan el menor error cuadrático medio del estimador. Quedándonos con umbrales mínimos que garanticen en torno a 4 y 8 individuos como mínimo en los conjuntos a riesgo para muestras de tamaño $n = 20$, entre 8 y 10 individuos para muestras de tamaño $n = 50$ y en torno a 10 y 20 individuos para muestras de tamaño 100.

Es decir, hemos comprobado que garantizando un mínimo de individuos en los conjuntos a riesgo de en torno al 20 % del tamaño de la muestra se consigue el estimador límite-producto de la función de distribución de la variable tiempo de vida X más eficiente; con el menor error cuadrático medio. En cualquier caso, esta recomendación de un umbral en torno al 20 % del tamaño de la muestra está basada en una experiencia limitada a unos pocos modelos y carece de un aval teórico sólido.

Por lo demás, el método bootstrap proporciona una solución al dilema de selección

del umbral para los conjuntos a riesgo, y presenta un comportamiento razonable en comparación con los umbrales de riesgo fijos próximos al óptimo.

Finalmente, destacamos que la corrección de los conjuntos a riesgo mediante umbrales mínimos fijos o escogidos por bootstrap mejora las propiedades del estimador límite-producto, solventando el problema de un posible conjunto a riesgo con un único elemento y disminuyendo el error cuadrático medio del estimador.

Como problemas abiertos de este trabajo desearíamos señalar los siguientes:

- Estudiar las propiedades teóricas del método de corrección con umbral fijo y con umbral escogido por bootstrap.
- Diseñar nuevas propuestas para corregir el estimador límite-producto, que mejoren las propiedades del estimador corregido. Por ejemplo, introducir elementos de suavización entre conjuntos a riesgo, como manera de disminuir el sesgo que acarrea un incremento neto de los conjuntos a riesgo. Hemos estudiado un procedimiento basado en la suavización general de los conjuntos a riesgo, que no produjo buenos resultados, en comparación con los umbrales. Sin embargo, un procedimiento más sutil de suavización, combinado con los umbrales, podría arrojar mejores resultados.
- Considerar un procedimiento jackknife para la selección del umbral mínimo de los conjuntos a riesgo.
- Establecer alguna relación entre el problema de los conjuntos a riesgo pequeños y el truncamiento dependiente de la variable de interés. Nótese que el truncamiento dependiente puede dar lugar a conjuntos a riesgo claramente más pequeños. Por tanto, la presencia de conjuntos a riesgo pequeños, especialmente en zonas centrales de la distribución, puede ser síntoma de un fenómeno de truncamiento dependiente. A su vez, las técnicas de Inferencia propias del truncamiento dependiente podrían verse reforzadas empleando los métodos de corrección que hemos presentado aquí.

Bibliografía

[1] Efron B (1981). Censored data and the bootstrap. Journal of the American Statistical Association 76: 312- 319.

[2] Cao R, López de Ullibarri I, Janssen P, Veraverbeke N. Presmoothed Kaplan-Meier and Nelson-Aalen estimators. Journal of Nonparametric Statistics 17: 31-56.

[3] Keiding N, Gill RD (1990). Random truncation models and Markov processes. The Annals of Statistics 18: 582-602.

[4] Lynden-Bell D (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. Monthly Notices Roy. Astronom. Soc. 155: 95-118.

[5] Stute W, Wang J-L (2008). The central limit theorem under random truncation. Bernoulli 14: 604-622.

[6] Tsai W-Y, Jewell NP, Wang M-C (1987). A note on the product-limit estimator under right censoring and left truncation. Biometrika 74: 883-886.

[7] Wang, M.C. (1989). A semiparametric model for randomly truncated data. J. Amer. Statist. Assoc. 84: 742? 748.

[8] Wang M-C (1991). Estimation from cross-sectional survival data. Journal of the American Statistical Association 86: 130- 143.

[9] Wang M-C, Jewell NP, Tsai W-Y (1986). Asymptotic properties of the product limit estimate under random truncation. The Annals of Statistics 14: 1597-1605.

[10] Woodroofe M (1985). Estimating a distribution function with truncated data. The Annals of Statistics 13: 163-177.