

# SESGO INDUCIDO POR EL METODO DE MUESTREO EN EL ANALISIS DE DATOS DE DESOCUPACION

Elsa C. Servy, Daniel Wojdyla

*Instituto de Investigaciones Teóricas y Aplicadas en Estadística,  
Escuela de Estadística, Facultad de Ciencias Económicas y Estadística,  
Universidad Nacional de Rosario*

## 1 Introducción

La duración de la desocupación juega un papel crítico en la teoría económica, y una cuidadosa evaluación de la teoría requiere información fidedigna sobre el fenómeno. En Argentina, los datos sobre duración del desempleo se recogen periódicamente (antes, semestralmente, y a partir del año 2000 tres veces por año) mediante la Muestra Permanente de Hogares (EPH) del INDEC. La información sobre longitud del tiempo de desocupación se registra sólo en aquéllos individuos de la Población Económicamente Activa (PEA) que están desocupados en el momento de la encuesta. Esto significa que no se computa la duración de la desocupación de aquellos que están ocupados en el momento de la encuesta, pero que tuvieron un cierto tiempo de desocupación entre dos ondas sucesivas de la EPH. Supóngase que los puntos  $O_1$  y  $O_2$  de la Figura 1 representan los tiempos de dos ondas sucesivas de la EPH y que los vectores  $t_1$ ,  $t_2$ ,  $t_3$  y  $t_4$  representan las trayectorias de la desocupación de cuatro individuos de la PEA. Ni  $t_1$ , ni  $t_2$  serán captados por la EPH. Es decir, las desocupaciones cortas estarán subrepresentadas en la muestra.

Existen problemas adicionales en la captación de la información sobre desocupación. El carácter rotativo de la EPH determina que cada individuo permanezca en la muestra un número limitado de periodos.

El periodo de observación tiene una duración de dos años, que se denominará "ventana de observación" del individuo. Si la persona deja la encuesta cuando aún no ha conseguido trabajo, no se puede conocer la duración completa de su desempleo. Esa observación

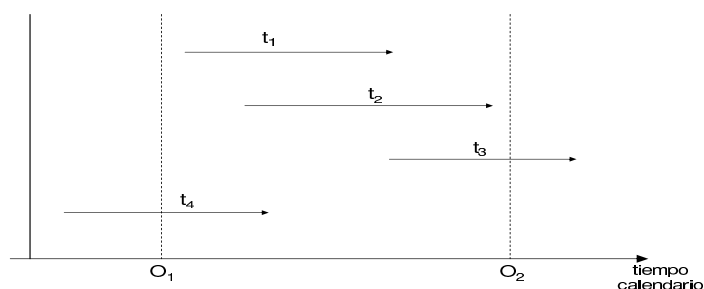


Figura 1: Ubicación de las trayectorias de desocupación en relación con las fechas de las encuestas

se denomina "censurada a la derecha". Hay otro problema, que no está vinculado al muestreo sino a la forma de registrar los datos. En la EPH, la manera de registrar el comienzo de la desocupación es a través de la pregunta: "Desde cuándo busca trabajo?". Esta información, como toda basada en la memoria del encuestado, debe ser tomada con cautela.

Se podría ignorar la respuesta, y considerar que la fecha de inicio es desconocida, y en consecuencia, considerar a la observación como censurada a la izquierda, pero esto sería desperdiciar información. Por otra parte, puede interesar estudiar la distribución de la longitud de la desocupación, para analizar en qué medida influyen sobre ella diversas variables socio-económicas. En este artículo, no se consideran modelos que incluyen variables explicativas dejando su estudio para más adelante. En esta primera etapa se consideran modelos paramétricos para la duración de la desocupación cuando i) la distribución está truncada por causa del método de muestreo utilizado y ii) además, está censurada a la derecha por causa de la rotación de la muestra. Antes de pasar a su discusión, son pertinentes algunas consideraciones. Debido a que las desocupaciones de los individuos de la muestra no transcurren simultáneamente, es necesario definir con cierta arbitrariedad el periodo al cual referir los indicadores del estudio. Por ejemplo, si los datos a analizar son las duraciones captadas por primera vez en mayo de 1996, el fenómeno no puede asociarse exclusivamente al periodo que va desde esa fecha en adelante, ya las duraciones captadas en ese momento se refieren a periodos anteriores a mayo de 1996. Además cada una tiene una fecha de iniciación diferente. Por ahora no se discutirá cuál es el punto de partida del estudio. Se supondrá que de alguna manera satisfactoria se logró convenir un tiempo inicial, llámese  $O$  (origen), de modo que será posible afirmar que el estudio se refiere a los individuos que estaban desocupados en mayo de 1996 y tuvieron como fecha

de iniciación de la desocupación una fecha entre ese origen y mayo de 1996.

## 2 Modelo para datos incompletos por truncamiento

Se debe considerar que existe un modelo estructural para las variables de interés, y otro modelo observacional para las variables observables, que no siempre coinciden con las primeras. Así, por ejemplo, si la duración de los episodios de desocupación siguen una distribución exponencial, el modelo estructural de la duración ( $l$ ) es,

$$f(l) = \theta e^{-\theta l} \quad l > 0 \quad (1)$$

Debido a que algunos valores pequeños de  $l$  no son observables, el modelo observacional es  $f_t(l)$ , una distribución truncada construida a partir de  $f(l)$ . La muestra de la EPH es un muestra de  $f_t(l)$ , pero con ella se desean estimar los parámetros de  $f(l)$ . Para poder definir  $f_t(l)$ , se considera el par de variables potencialmente observables  $(w, l)$ , donde  $w$  es el momento (medido a partir del origen  $O$ ) en que comienza la desocupación del individuo y  $l$  es la duración total de su desocupación. Las variables  $l$  y  $w$  se observan sólo si  $(w + l) > D$ , donde  $D$  es el momento de la encuesta. El modelo supone que  $w$  y  $l$  son variables independientes, es decir, el punto de iniciación de la desocupación es independiente de la longitud de la misma. Se supone que  $w$  tiene distribución uniforme entre  $O$  y  $D$ , y que la distribución de  $l$  es absolutamente continua en el eje real positivo, y depende de un parámetro  $\theta$ , que es un vector de una o más componentes reales. Luego,

$$f_\theta(l, w) = \frac{1}{D} f_\theta(l) \quad (2)$$

Si  $\Omega$  es el espacio muestral del par  $(w, l)$  de variables, el espacio muestral de las observables es el conjunto sombreado en la Figura 2.

## 3 Distribución de $f_t(t)$

Según el modelo de la sección anterior,

$$P(\{w + l > D\} \cap \{D < l\}) = \int_0^D \int_{D-l}^D \frac{1}{D} f(l) dw dl = \int_0^D f(l) l dl \quad (3)$$

$$P(\{w + l > D\} \cap \{D > l\}) = \int_D^\infty \int_0^D \frac{1}{D} f(l) dw dl = \int_D^\infty f(l) dl \quad (4)$$

$$P(\{w + l > D\}) = \frac{1}{D} \int_0^D f(l) l dl + \int_D^\infty f(l) dl \quad (5)$$

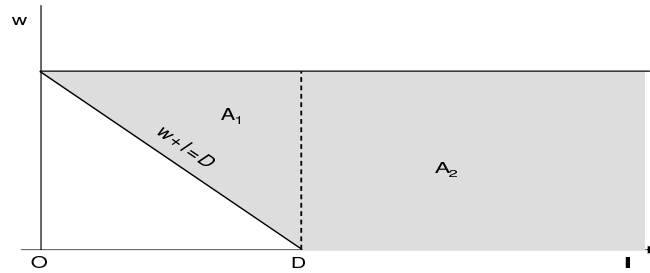


Figura 2: Espacio muestral observable

Sea  $0 < l_1 < l < l_2 < D$ ,

$$P(\{l_1 < l < l_2 \mid w + l > D\}) = \frac{\frac{1}{D} \int_{l_1}^{l_2} f(l) l dl}{\frac{1}{D} \int_0^D f(l) l dl} \quad (6)$$

De donde,

$$f_t(l) = f(l \mid w + l > D) = \frac{f(l) l}{\int_0^D f(l) l dl} \quad 0 < l < D \quad (7)$$

Sea  $l > D$ ,  $w + l > D$ . Luego,

$$f_t(l) = f(l \mid w + l > D) = \frac{f(l)}{(1 - F_l(D))} \quad (8)$$

donde  $F_l(D)$  es la función de distribución de  $l$ , valorizada en  $D$ . Luego, de (6) y (7),

$$f_t(l) = f(l \mid w + l > D) = \delta_{l,D} \frac{f(l) l}{\int_0^D f(l) l dl} + \delta_{D,l} \frac{f(l)}{1 - F_l(t)} \quad (9)$$

donde  $\delta_{xy} = 1$  si  $x < y$ , y  $\delta_{xy} = 0$  si  $x \geq y$ .

En cuanto al valor esperado de  $l$ , dado  $(w + l) > D$ ,

$$E(l \mid w + l > D) = \frac{\int_0^D f(l) l^2 dl}{\int_0^D f(l) l dl} \cdot \frac{P(\{w + l > D\} \cap \{D < l\})}{P(w + l > D)} + \quad (10)$$

$$+ \int_D^\infty \frac{f(l) l dl}{1 - F_l(D)} \cdot \frac{P(\{w + l > D\} \cap \{D < l\})}{P(w + l > D)}$$

Otra forma de expresar  $E(l|w+l > D)$  es,

$$E(l|w+l > D) = E(l|w+l, l \leq D) \cdot \frac{P(\{w+l > D\} \cap \{l \leq D\})}{P(w+l > D)} + \quad (11)$$

$$+ E(l|l > D) \frac{P(l > D)}{P(w+l > D)}$$

#### 4 Comparación de $E(l)$ y $E(l|w+l > D)$

Se puede escribir  $E(l)$  de la siguiente forma,

$$E(l) = E(l|l \leq D) P(l \leq D) + E(l|l > D) P(l > D) \quad (12)$$

$$= \frac{\int_0^D f(l) l dl}{\int_0^D f(l) dl} P(l \leq D) + \frac{\int_D^\infty f(l) l dl}{\int_D^\infty f(l) dl} P(l > D)$$

Tanto  $E(l)$  como  $E(l|w+l > D)$  son promedios ponderados de esperanzas condicionales. Se puede observar que,

$$E(l|w+l > D, l \leq D) > E(l|l \leq D) \quad (13)$$

O sea,

$$\frac{\int_0^D f(l) l^2 dl}{\int_0^D f(l) l dl} \geq \frac{\int_0^D f(l) l dl}{\int_0^D f(l) dl} \quad (14)$$

pues,

$$\frac{\int_0^D f(l) l^2 dl}{\int_0^D f(l) l dl} \geq \left( \frac{\int_0^D f(l) l dl}{\int_0^D f(l) dl} \right)^2 \quad (15)$$

porque la variancia de  $f(l|l \leq D)$  es positiva. Entonces, la diferencia

$$E(l|w+l > D) - E(l) = [E(l|w+l > D, l \leq D) - E(l|l \leq D)] \cdot P(l \leq D) + \quad (16)$$

$$+ [E(l|l > D) - E(l|w+l > D, l \leq D)] \cdot \left[ \frac{P(l > D)}{P(w+l > D)} - P(l > D) \right]$$

es positiva, porque el primer sumando es positivo según (15), y el segundo sumando también lo es, porque el campo de variación de la variable a la cual corresponde la

esperanza del minuendo es la semirecta  $l > D$ , y el de la variable de la esperanza del sustraendo está sujeto a  $l \leq D$ .

## 5 Estimación de los parámetros del modelo por el método de la máxima verosimilitud

La única información que proporciona la muestra se refiere a la distribución truncada por el evento  $w + l > D$ . Si  $f(l)$  depende del parámetro  $\theta$ ,  $f(l | w + l > D)$ , también depende de él, y es a partir de ella que  $\theta$  va a ser estimado. Se supone que la muestra consta de  $n_1$  valores de duración menores que  $D$  y  $n_2$  valores mayores que  $D$ . Entonces,

$$L = \prod_{i=1}^{n_1} \frac{f_{\theta}(l_i) l_i}{\left[ \int_0^D f_{\theta}(l) l dl \right]} \cdot \frac{\int_0^D f_{\theta}(l) l dl}{\left[ \int_0^D f_{\theta}(l) l dl + D \int_D^{\infty} f_{\theta}(l) dl \right]} \cdot \prod_{i=1}^{n_2} \frac{f_{\theta}(l_i)}{\left[ \int_D^{\infty} f_{\theta} l dl \right]} \cdot \frac{D \int_D^{\infty} f_{\theta}(l) dl}{\left[ \int_0^D f_{\theta}(l) l dl + D \int_D^{\infty} f_{\theta}(l) dl \right]} \quad (17)$$

Esta función se maximiza para  $\theta$ . El valor que la maximiza es el estimador máximo-verosímil de  $\theta$ .

Como ilustración del principio de la máxima verosimilitud en un problema concreto, se presenta el caso de una distribución exponencial con parámetro  $\theta$ . La distancia entre el origen y el momento de la encuesta se indica con el valor  $D$ .

La función de verosimilitud para este caso es,

$$L = \frac{\theta^{n_1} e^{-\theta \sum_{i=1}^{n_1} l_i} \prod_{i=1}^{n_1} l_i \theta^{n_2} e^{-\theta \sum_{i=1}^{n_2} l_i} D^{n_2}}{\left( \frac{1}{\theta D} \right)^n (1 - e^{-\theta D})^n} \quad (18)$$

Y su logaritmo, excluyendo los términos que no dependen de  $\theta$ ,

$$\log L = 2n \log \theta - \theta \sum_{i=1}^n l_i - n \log(1 - e^{-\theta D}) \quad (19)$$

Cuya derivada es,

$$\frac{\partial \log L}{\partial \theta} = \frac{2n}{\theta} - \sum_{i=1}^n l_i - n \frac{(De^{-\theta D})}{(1 - e^{-\theta D})} \quad (20)$$

La expresión (20) igualada a 0 debe resolverse iterativamente para hallar el estimador máximo-verosímil,  $\hat{\theta}$ , de  $\theta$ .

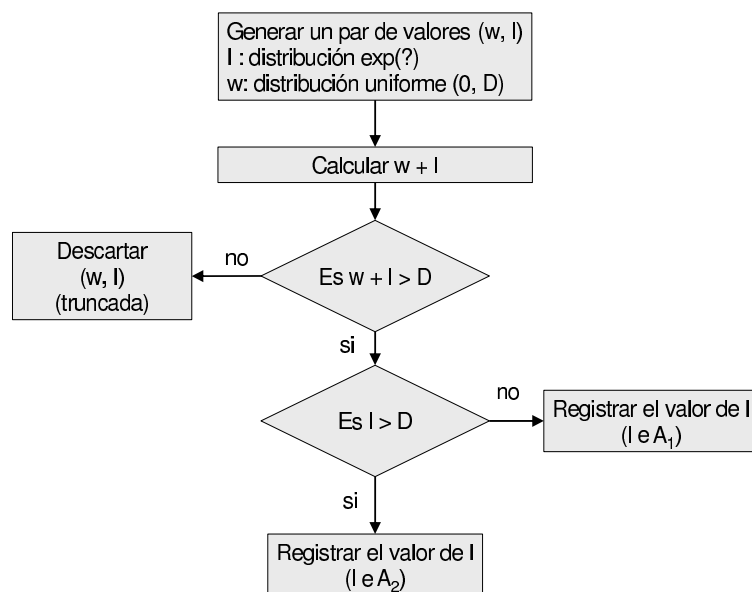


Figura 3: Generación de datos

Para evaluar los resultados anteriores se simularon 5000 valores pertenecientes a una distribución exponencial con  $\theta = \frac{1}{25} = 0.04$  y estableciendo el valor de  $D = 7$ . Los datos fueron generados de acuerdo con el diagrama presentado en la Figura 3.

Los valores obtenidos fueron,  $n_1 = 645$ ,  $n_2 = 4355$ , y  $\sum l_i = 123870.16$  que reemplazados en (20) producen el estimador  $\hat{\theta} = 0.0399$ .

Si las duraciones simuladas hubieran sido analizadas ignorando el truncamiento de  $f(l)$ , el estimador obtenido hubiese sido 0.0352. Se realizaron simulaciones similares a la descrita a partir de otros valores de  $\theta, D$  y  $n$ , hallándose siempre resultados coherentes con los esperados.

## 6 El caso de observaciones censuradas

En esta sección se introduce en el modelo la censura que padecen algunas observaciones suministradas por la EPH. La ventana de observación de cada individuo tiene una longitud de dos años, a partir del momento en que el mismo entra en la muestra. Si  $\Delta$  es el tamaño de la ventana de observación y  $D$  el momento en que el individuo fue captado por la encuesta, el momento en que el individuo sale de la misma es  $D_1 = D + \Delta$ , haya o no completado su periodo de desocupación. Si el individuo continúa desocupado más allá de  $D_1$ , la longitud de su desocupación se torna desconocida a los ojos del entrevistador. Sólo sabe que su desocupación es, al menos  $(D - w)$ . El entrevistador imputa a ese individuo

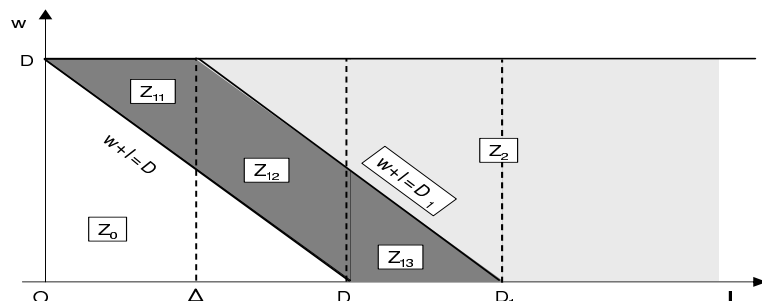


Figura 4: Zonas de truncamiento y censura

el valor  $(D - w)$ . El dato relativo a su tiempo de desocupación se denomina "censurado". La figura 4 describe los subconjuntos del espacio muestral afectados por el truncamiento y la censura. Se supondrá que  $\Delta < D$ .

Se observa que,

- i. No hay observaciones visibles en el triángulo  $DOD$ .
- ii. Las observaciones visibles y no censuradas están en la franja más oscura.
- iii. Las observaciones visibles, pero censuradas están en la zona gris claro.

Los conjuntos de observaciones invisibles, visibles y no censuradas, y censuradas se denominan  $Z_0$ ,  $Z_1$  y  $Z_2$  respectivamente.

### 6.1 Distribución de $l$ condicionada a $Z_1$

La distribución de  $l$  condicionada a  $Z_1$ , depende del subespacio de  $Z$  en que se mueve el punto muestral. Esos subespacios son:

- i.  $Z_{11} = \{D < w + l \leq D_1, 0 < l \leq \Delta\}$ ,
- ii.  $Z_{12} = \{D < w + l \leq D_1, \Delta < l \leq D\}$ ,
- iii.  $Z_{13} = \{D < w + l \leq D_1, D < l \leq D_1\}$ ,

Si  $(w, l) \in Z_{11}$ ,  $0 < l \leq \Delta$ ,

$$P(Z_{11}) = \int_0^\Delta \int_{D-l}^D \frac{1}{D} f(l) dw dl \quad (21)$$



y

$$P(l_1 < l \leq l_2 | Z_{11}) = \frac{\int_{l_1}^{l_2} \int_{D-l}^D \frac{1}{D} f(l) dw dl}{P(Z_{11})} \quad (22)$$

Entonces,

$$f(l | Z_{11}) = f_{11}(l) = \frac{\frac{1}{D} f(l) l}{P(Z_{11})}, \quad 0 < l < \Delta \quad (23)$$

Igualmente, si  $(w, l) \in Z_{12}$ ,

$$P(Z_{12}) = \int_{\Delta}^D \int_{D-l}^{D_1-l} \frac{1}{D} f(l) dw dl \quad (24)$$

y,

$$P(l_1 < l \leq l_2 | Z_{12}) = \frac{\int_{l_1}^{l_2} \int_{D-l}^{D_1-l} \frac{1}{D} f(l) dw dl}{P(Z_{12})} \quad (25)$$

Por cuanto,

$$f(l | Z_{12}) = f_{12}(l) = \frac{\frac{1}{D} f(l)}{P(Z_{12})}, \quad \Delta < l \leq D \quad (26)$$

Si  $(w, l) \in Z_{13}$ ,  $D < l \leq D_1$ ,

$$P(Z_{13}) = \int_D^{D_1} \int_0^{D_1-l} \frac{1}{D} f(l) dw dl \quad (27)$$

y,

$$P(l_1 < l \leq l_2 | Z_{13}) = \frac{\int_{l_1}^{l_2} \int_0^{D_1-l} \frac{1}{D} f(l) dw dl}{P(Z_{13})} \quad (28)$$

De donde,

$$f(l | Z_{13}) = f_{13}(l) = \frac{\frac{1}{D} (D-l) \cdot f(l)}{P(Z_{13})}, \quad D < l < D_1 \quad (29)$$

## 6.2 Distribución de $l$ condicionada a $Z_2$

$Z_2$  es el conjunto de puntos muestrales donde la observación está censurada. Allí, la variable aleatoria es  $(D_1 - w)$ , donde  $w$  es la variable que describe el momento inicial de la desocupación, pero condicionada a  $w + l > D$ .

$$P(Z_2) = \int_0^D \int_{D_1-w}^{\infty} \frac{1}{D} f(l) dw dl \quad (30)$$

$$P(w_1 < w \leq w_2 | (w + l) > D_1) = \frac{\int_{w_1}^{w_2} \int_{D_1-w}^{\infty} \frac{1}{D} f(l) dl dw}{P(Z_2)} \quad (31)$$

$$f(w | (w + l) > D) = \frac{\int_{D_1-w}^{\infty} \frac{1}{D} f(l) dl}{P(Z_2)} \quad (32)$$

Luego, la distribución de  $l^* = D_1 - w$ , condicionada a  $Z_2$  es,

$$f_2(l^*) = \frac{\int_{l^*}^{\infty} \frac{1}{D} f(l) dl}{P(Z_2)} \quad D_1 < l^* \leq D + D_1 \quad (33)$$

## 7 La función de verosimilitud

Sólo intervienen en la función de verosimilitud las probabilidades valorizadas en las observaciones visibles. Según se vio en el párrafo anterior, estas probabilidades difieren según el subconjunto del espacio muestral donde se encuentran los puntos muestrales. Cada conjunto contiene un número de observaciones  $n_{11}$ ,  $n_{12}$ ,  $n_{13}$  y  $n_2$  y sea  $n_1 = n_{11} + n_{12} + n_{13}$ ;  $n = n_1 + n_2$ . Sea  $\delta_j$  ( $j \in \{11, 12, 13, 2\}$ ) la función indicadora de los conjuntos  $Z_{11}$ ,  $Z_{12}$ ,  $Z_{13}$  y  $Z_2$ . Entonces,

$$f(l) = \delta_{11} f_{11}(l) + \delta_{12} f_{12}(l) + \delta_{13} f_{13}(l) + \delta_2 f_2(l) \quad (34)$$

Sujeto a que la observación sea visible,

$$P(\delta_j = 1) = \frac{P(Z_j)}{[P(Z_1) + P(Z_2)]} \quad j \in \{11, 12, 13, 2\} \quad (35)$$

donde  $Z_1 = Z_{11} \cup Z_{12} \cup Z_{13}$ .

Considerando los diferentes aportes de los puntos muestrales,

$$L = \prod_{i=1}^{n_{11}} \frac{f_{11}(l_i) P(Z_{11})}{[P(Z_1) + P(Z_2)]} \cdot \prod_{i=n_{11}+1}^{n_{11}+n_{12}} \frac{f(l_i) P(Z_{12})}{[P(Z_1) + P(Z_2)]} \cdot \prod_{i=n_{11}+n_{12}+1}^{n_1} \frac{f_{13}(l_i) P(Z_{13})}{[P(Z_1) + P(Z_2)]} \cdot \prod_{i=n_1+1}^n \frac{f(l_i) P(Z_2)}{[P(Z_1) + P(Z_2)]} \quad (36)$$

Según la fórmula (5),

$$P(Z_1) + P(Z_2) = P(w + l > D) = \frac{1}{D} \int_0^D f(l) l dl + \int_D^{\infty} f(l) dl \quad (37)$$

Realizando las sustituciones y simplificaciones correspondientes,

$$L = \frac{\prod_{i=1}^{n_1} f(l_i) l_i \prod_{i=n_{11}+1}^{n_{11}+n_{12}} \Delta f(l_i) \prod_{i=n_{11}+n_{12}+1}^{n_1} (D - l_i) f(l_i) \prod_{i=n_1+1}^n \int_{l_i}^{\infty} f(l) dl}{\left[ \int_0^D f(l) l dl + D \int_D^{\infty} f(l) dl \right]^n} \quad (38)$$

### 7.1 Una aplicación a la distribución exponencial

Si se supone que la variable latente es exponencial, de parámetro  $\theta$ ,

$$L = \frac{\theta^{n_1} \exp \left[ -\theta \sum_{i=1}^{n_1} l_i \right] \prod_{i=n_1+1}^n \exp [-\theta l_i^*] \prod_{i=1}^{n_{11}} l_i \Delta^{n_{12}} \prod_{i=n_{11}+n_{12}+1}^{n_{13}} (D - l_i)}{\theta^{-n} (1 - \exp(-\theta D))^n} \quad (39)$$

Su logaritmo luego de descartar los términos que no dependen de  $\theta$  es

$$\log L = (n_1 + n) \log \theta - \theta \left( \sum_{i=1}^{n_1} l_i + \sum_{i=n_1+1}^n l_i^* \right) - n \log(1 - \exp(-\theta D)) \quad (40)$$

Derivando es igualando a 0 se obtiene el estimador máximo-verosímil.

### 7.2 Estudio de simulación

Se generaron datos a partir de una distribución con parámetro  $\theta = \frac{1}{25}$ . Estos datos fueron truncados y censurados de acuerdo al modelo descrito anteriormente con valores  $D = 7$  y ventana de observación  $\Delta = 4$ . La Figura 5 presenta el algoritmo de generación.

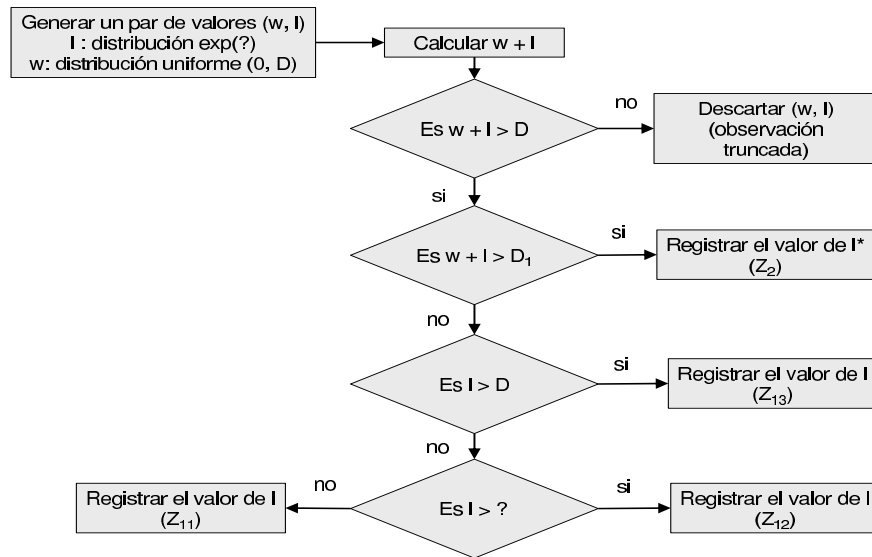


Figura 5: Algoritmo de generación de datos

El proceso de simulación se realizó para cuatro tamaños de muestra distintos considerando el modelo sin truncamiento: 100, 500, 1000 y 5000. En cada una de las 1000 repeticiones que se generaron para cada tamaño de muestra se obtuvo el estimador máximo-verosímil de acuerdo con lo expresado en la sección 7.1. Los tamaños de muestra

Tabla 1: Estadísticas Descriptivas

Tamaño de Muestra	$\hat{\lambda}$	Desvío Estándar
100	27.05	8.80
500	25.38	3.32
1000	25.25	2.18
5000	25.08	0.94

anteriores se redujeron debido a la exclusión de observaciones por truncamiento. A continuación se describe, a través estadísticas resumen y gráficos, la distribución de los estimadores obtenidos.

La Tabla 1 presenta el valor medio y la desviación estándar de los estimadores obtenidos en las 1000 repeticiones para cada tamaño de muestra ( $\lambda = \frac{1}{\theta}$ ).

La tabla muestra como los valores medios de los valores estimados se acercan al valor del parámetro de la exponencial, 25, y como la precisión de los estimadores aumentan con el tamaño de la muestra. Los estudios de simulación serán continuados en un futuro tomando como puntos de referencia otros valores de  $\lambda$ , de  $D$  y  $D_1$ , así como distribuciones diferentes (Weillbul, Gamma, etc.).

## Bibliography

- Klein, J., Moeschberger, M., (1997), *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.

**Tamaño Muestral – n = 100**

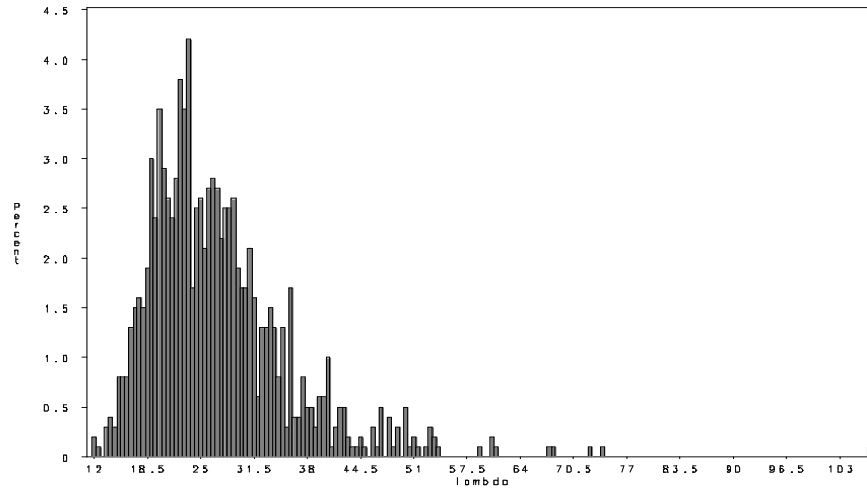


Figura 6: Tamaño Muestral n = 100

**Tamaño Muestral – n = 500**

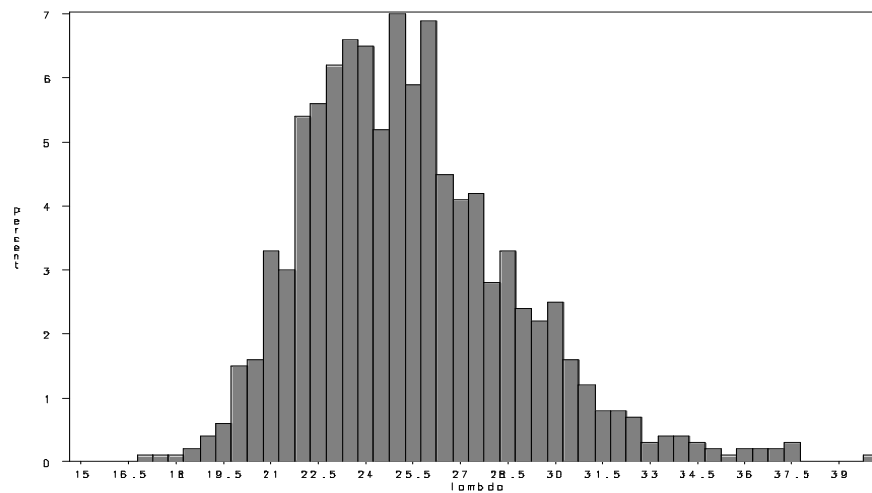


Figura 7: Tamaño Muestral n = 500

**Tamaño Muestral – n = 1000**

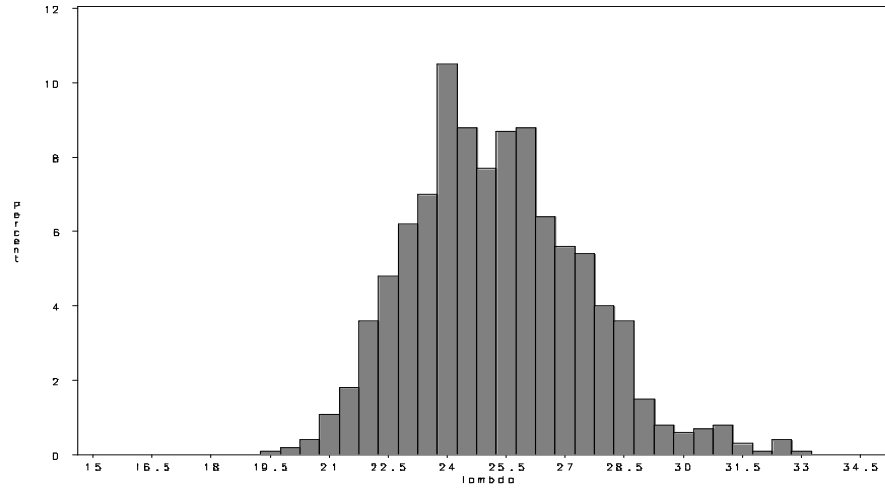


Figura 8: Tamaño Muestral n = 1000

**Tamaño Muestral – n = 5000**

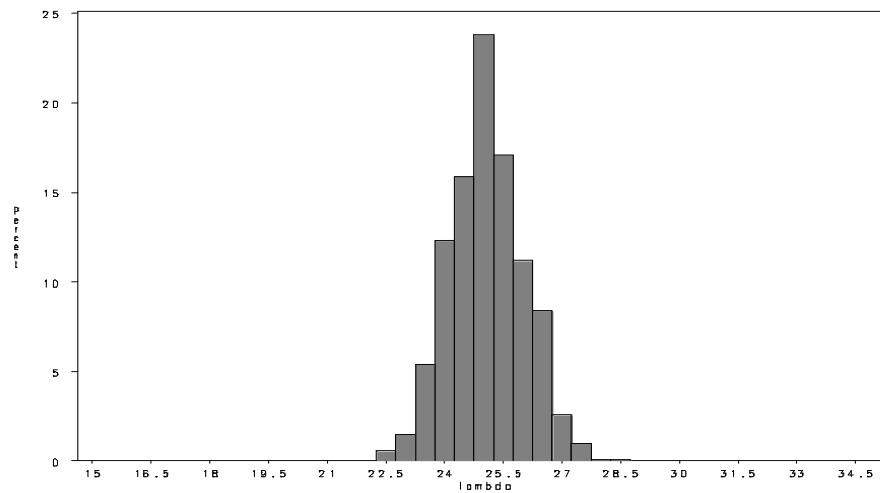


Figura 9: Tamaño Muestral n = 5000