UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2017, Vol. 10, No. 1, pp. 145–167

Distribuciones en la clase *(a, b)*: estimación y generación de números aleatorios

(a, b) class of distributions: estimation and generation of random numbers

César Escalante Coterio[a]
cesar.escalante@epm.com.co

Resumen

Se presenta la estimación *detallada* de los parámetros de las distribuciones discretas de probabilidad clase (a, b) (Klugman et al. 2004, Escalante 2006) por los métodos de momentos y máxima verosimilitud. Se propone un algoritmo general para generar números aleatorios de las distribuciones clase (a, b). Los resultados se presentan de forma tal que puedan ser implementados en cualquier lenguaje de programación adecuado. Los ejemplos se realizaron en `R` con datos reales tomados de diferentes disciplinas.

***Palabras clave*:** Modelos aleatorios de frecuencia, estimación de distribuciones clase (a, b), teoría de riesgos, modelo de riesgo colectivo, algoritmo de Panjer.

Abstract

The estimate of the parameters of discrete probability distributions class (a, b) (Klugman et al. 2004, Escalante 2006) is presented in *detail* by methods of moments and maximum likelihood studied. A general algorithm is proposed to generate random numbers of distributions class (a, b). The results are presented so it can be implemented in any suitable programming language. Examples were made in `R` with real data taken from different disciplines.

***Keywords*:** Random models of frequency, estimation of distribution class (a, b), risk theory, the collective risk model, Panjer's algorithm.

1. Introducción

En el tratado de distribuciones discretas de probabilidad univariadas de Johnson et al. (2005, pg.82) se presenta la familia de distribuciones de Katz estudiada en su

[a]MSc. Matemáticas Aplicadas. Gerencia Integral de Riesgos. Empresas Públicas de Medellín, EPM.

tesis doctoral en 1945 y desarrollada en trabajos ulteriores que van hasta 1 965. En 1 981 Sundt & Jewell (1981), y luego Willmot (1988) en 1 988, estudiaron la familia de distribuciones de Sundt y Jewell, relacionada con la de Katz, con intereses actuariales. En Panjer & Willmot (1992) y en Klugman et al. (2004) se denomina a esta última, familia distribuciones clase (a, b). Estas distribuciones son usadas para modelar la frecuencia de fenómenos aleatorios y son básicas para modelos más complejos de frecuencia a través de las variables aleatorias mixtas y variables aleatorias compuestas. Junto con el algoritmo de Panjer, son de importancia capital en actuaría y modelos para riesgo operativo.

El primer objetivo del artículo es presentar la estimación detallada de los parámetros de las distribuciones clase (a, b) por los métodos de momentos y máxima verosimilitud que se estudian de manera general en Klugman et al. (2004). Los cálculos de los ejemplos se obtuvieron en `R` (R Development Core Team 2008) y algunas de sus librerías.

El segundo objetivo del artículo es proponer un algoritmo de tipo genético para generar números aleatorios de las distribuciones clase (a, b). Luego de su presentación, se indica una idea para que el lector mejore el algoritmo y disminuya el tiempo de ejecución. El autor no encontró referencias bibliográficas con algoritmos específicos para generar este tipo de números aleatorios.

La motivación principal del trabajo es promover la aplicación de estas importantes distribuciones en la empresa e industria, y en postgrado y pregrado de las asignaturas de probabilidad, estadística y métodos cuantitativos en análisis de riesgos. Los resultados presentados aún no están incorporados en forma automática en los programas de cálculo estadístico más populares.

Para que el artículo sea autocontenido se presentan primero las definiciones y proposiciones principales necesarias. El detalle de éstos, junto con las demostraciones, se encuentran en Klugman et al. (2004) y Escalante (2006).

Las distribuciones clase (a, b) son un conjunto de distribuciones discretas usadas en modelos de probabilidad aplicada en general y de modelos de frecuencia de pérdidas en matemática de los riesgos en particular. Véanse Escalante (2006), Klugman et al. (2004) y Johnson et al. (2005). Sea N un látice con soporte $\mathbb{N}_0 = \{0,\ 1,\ 2, \ldots\}$ y función de probabilidad (fp) $p : \mathbb{N}_0 \to [0,\ 1]$ tal que $p_k = \Pr(N = k)$. Una distribución $\{p_k\}$ pertenece a la clase $(a, b; 0)$ si existen las constantes reales a y b tales que

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k = 1,\ 2,\ 3,\ \ldots \tag{1}$$

En la figura 1 de Escalante (2006) y en la tabla 3 se caracterizan los valores de las constantes a y b de acuerdo con las distribuciones específicas. En Panjer & Willmot (1992), y con mayor detalle en Escalante (2006), se demuestra que las únicas distribuciones no degeneradas cuyas fp verifican la fórmula recursiva (1) son la Poisson $Poi(\lambda)$, $\lambda > 0$, la binomial $Bin(q, m)$, $0 < q < 1$, $m \in \mathbb{N}$, y la binomial negativa $BN(r, \beta)$, $r > 0$, $\beta > 0$, que incluye su caso particular a la distribución geométrica $Geo(\beta)$, cuando $r = 1$. En la tabla 1 se indica la fp, valor

Tabla 1: *Distribuciones clase* $(a, b; 0)$*: Poisson, binomial negativa y binomial. Fuente: elaboración propia.*

Distribución	fp p_k	$E(N)$	$Var(N)$
$Poi(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ
$BN(r, \beta)$	$\binom{k+r-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k$	βr	$\beta r(1+\beta)$
$Geo(\beta)^i$	$\frac{\beta^k}{(1+\beta)^{k+1}}$	β	$\beta(1+\beta)$
$Bin(q, m)^{ii}$	$\binom{m}{k}q^k(1-q)^{m-k}$	mq	$mq(1-q)$

i. Caso particular de la binomial negativa con $r = 1$.
ii. Soportes: $\{0, 1, \ldots, m\}$ para la $Bin(q, m)$ y $\mathbb{N}_0$ para la $Poi(\lambda)$ y $BN(r, \beta)$.

esperado y varianza para cada una de estas distribuciones. La varianza es mayor que la media para la distribución $BN(r, \beta)$, igual para la $Poi(\lambda)$ y menor para la $Bin(q, m)$.

La distribución $BN(r, \beta)$ con parámetro r entero positivo se denomina distribución de Pascal. En la actualidad algunos programas, *risk* de *palisade*, por ejemplo, denominan distribución binomial negativa a la distribución de Pascal y no brindan la posibilidad de trabajar la distribución *binomial negativa* con valores no enteros de r, que es más versátil.

Una distribución $\{p_k\}$ pertenece a la clase $(a, b; 1)$ si existen las constantes a y b tales que

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k = 2,\ 3,\ 4,\ \ldots \tag{2}$$

La única diferencia con las distribuciones clase $(a, b; 0)$ es que en éstas la recursión inicia en $k = 1$, mientras que en la clase $(a, b; 1)$ comienza en $k = 2$, y es claro que toda distribución clase $(a, b; 0)$ es también clase $(a, b; 1)$. Cuando la forma (no las probabilidades) de una distribución coincide, salvo en cero, con la de un látice clase $(a, b; 0)$ con fp $\{p_k\}$, se construye una distribución cero-modificada (ZM) $\{p_k^M\}$ con $p_0^M \in [0, 1)$ —valor arbitrario— y

$$p_k^M = \frac{1 - p_0^M}{1 - p_0} p_k,\ k = 1,\ 2, \ldots. \tag{3}$$

Una distribución así construida es clase $(a, b; 1)$, con lo cual obtenemos las distribuciones $ZM - Poi(\lambda)$, ZM-$BN(r, \beta)$ (que incluye a la ZM-$Geo(\beta)$) y ZM-$Bin(q, m)$. En el caso extremo, cuando $p_0^M = 0$, se obtiene la subclase especial de miembros de $(a, b; 1)$ denominada distribuciones Cero-Truncadas (ZT), con fp $\{p_k^T\}$: ZT-$Poi(\lambda)$, ZT-$BN(r, \beta)$ (que incluye a la ZT-$Geo(\beta)$) y ZT-$Bin(q, m)$. De (3),

$$p_k^T = \frac{p_k}{1 - p_0},\ k = 1,\ 2,\ 3 \ldots \tag{4}$$

Y de (3) y (4)

$$p_k^M = (1 - p_0^M)\, p_k^T;\ k = 1,\ 2,\ 3 \ldots \tag{5}$$

Tabla 2: *Distribuciones ETNB y logarítmica. Fuente: elaboración propia.*

Distribución con soporte $\mathbb{N}$	Función de probabilidad	Media	Varianza
ETNB, $\beta > 0$, $r > -1$, $r \neq 0$	$\frac{\binom{k+r-1}{k}\left(\frac{\beta}{1+\beta}\right)^k}{(1+\beta)^r-1}$	$\frac{\beta r}{1-(1+\beta)^{-r}}$	$\frac{\beta r\left[(1+\beta)-(1+\beta+\beta r)(1+\beta)^{-r}\right]}{\left[1-(1+\beta)^{-r}\right]^2}$
Logarítmica, $\beta > 0$	$\left(\frac{\beta}{1+\beta}\right)^k \cdot \frac{1}{k\ln(1+\beta)}$	$\frac{\beta}{\ln(1+\beta)}$	$\frac{\beta\left[1+\beta-\frac{\beta}{\ln(1+\beta)}\right]}{\ln(1+\beta)}$

Tabla 3: *Distribuciones clase* $(a, b; 1)$. *Fuente: elaboración propia.*

Distribución$^{(i)}$	Parámetros	a	b	p_0
$Poi(\lambda)$	$\lambda > 0$	0	λ	$e^{-\lambda}$
ZT-$Poi(\lambda)$	$\lambda > 0$	0	λ	0
ZM-$Poi(\lambda)$	$\lambda > 0$	0	λ	(ii)
$BN(r, \beta)$	$\beta > 0$, $r > 0$	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$(1+\beta)^{-r}$
ETNB	$\beta > 0$, $r > -1$, $r \neq 0$	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	0
ZM-ETNB	$\beta > 0$, $r > -1$, $r \neq 0$	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	(ii)
$Geo(\beta)$	$\beta > 0$	$\frac{\beta}{1+\beta}$	0	$(1+\beta)^{-1}$
ZT-$Geo(\beta)$	$\beta > 0$	$\frac{\beta}{1+\beta}$	0	0
ZM-$Geo(\beta)$	$\beta > 0$	$\frac{\beta}{1+\beta}$	0	(ii)
$Bin(q, m)$	$0 < q < 1$, $m \in \mathbb{N}$	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$(1-q)^m$
ZT-$Bin(q, m)$	$0 < q < 1$, $m \in \mathbb{N}$	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	0
ZM-$Bin(q, m)$	$0 < q < 1$, $m \in \mathbb{N}$	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	(ii)
$Log(\beta)$	$\beta > 0$	$\frac{\beta}{1+\beta}$	$-\frac{\beta}{1+\beta}$	0
ZM-$Log(\beta)$	$\beta > 0$	$\frac{\beta}{1+\beta}$	$-\frac{\beta}{1+\beta}$	(ii)

i. ZT: Cero truncada. ZM: Cero modificada.
ii. Arbitrario en $[0, 1)$. Con $p_0 = 0$, se obtiene la versión ZT.

A partir de la distribución ZT-$BN(r, \beta)$ se obtiene la distribución de la clase $(a, b; 1)$ llamada *Binomial Negativa Truncada Extendida* o ETNB (*Extended Truncated Negative Binomial*). En esta distribución el espacio de los parámetros a y b se amplía para admitir casos en los que $-1 < r < 0$. Un caso límite de la distribución ETNB, cuando $r \to 0$, es la distribución logarítmica. En la tabla 2 se indican los aspectos básicos de estas distribuciones.

En la tabla 3 se resumen los valores de las constantes a y b y la probabilidad en cero de las distribuciones clase $(a, b; 1)$, necesaria para iniciar la definición recursiva de éstas. En el resto del artículo se denominan modelos o distribuciones regulares a las distribuciones más conocidas $(a, b; 0)$.

La media y el segundo momento al origen de una v.a. cero modificada N^M son, respectivamente,

$$E\left(N^M\right) = \frac{1-p_0^M}{1-p_0} E\left(N\right) \tag{6}$$

y

$$E\left(N^{M2}\right) = E\left[\left(N^M\right)^2\right] = \frac{1-p_0^M}{1-p_0} E\left(N^2\right). \tag{7}$$

Respecto de la v.a. N^T, su media y segundo momento al origen se obtienen de (6) y (7) con $p_0^T = p_0^M = 0$, por definición.

Agradecimientos. El autor agradece a los árbitros del artículo, sus importantes observaciones y recomendaciones. Dr. Gerardo Iván Arango Ospina (In memoriam. Carolina del Príncipe, 11 de febrero de 1944 - Medellín, 19 de mayo de 2012), persona excelsa, invaluable amigo, matemático y ejecutivo de gran estatura intelectual.

2. Estimación de parámetros

2.1. Introducción

Para estimar los parámetros de las distribuciones $(a, b; 1)$ se tabulan los datos en la forma indicada en la tabla 4, esto es, en la muestra de tamaño n el número k se repite n_k veces, para $k = 0,\ 1,\ 2, \ldots,\ j$, y $n = \sum_{k=0}^{\infty} n_k = \sum_{k=0}^{j} n_k$, pues siempre es posible tabular los datos de forma tal que $n_k = 0$ para $k = j+1,\ j+2, \ldots$. Si en lugar de j (la frecuencia más alta) se escribe $j+$, significa que la frecuencia mayor o igual a j es de n_j.

La media y segundo momento al origen muestrales para los datos agrupados, son

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{\infty} k\ n_k \tag{8}$$

y

$$m_2' = \frac{1}{n} \sum_{k=1}^{\infty} k^2\ n_k. \tag{9}$$

Si los datos muestrales están organizados en la forma de la tabla 4 y en la última fila se indica que hay una frecuencia de n_j para valores de la v.a. iguales o superiores a j, se puede asumir que n_j es la frecuencia del valor exacto j para los cálculos de los estadísticos muestrales anteriores.

La función de verosimilitud L^M para las observaciones de la tabla 4 de N^M está dada por

$$L^M = \prod_{k=0}^{\infty} \left(p_k^M\right)^{n_k} = \left(p_0^M\right)^{n_0} \prod_{k=1}^{\infty} \left(p_k^M\right)^{n_k} = \left(p_0^M\right)^{n_0} \prod_{k=1}^{\infty} \left[\left(1-p_0^M\right) p_k^T\right]^{n_k}, \text{ por (5)},$$

Tabla 4: *Notación para los datos muestrales. Fuente: elaboración propia.*

k	n_k
0	n_0
1	n_1
2	n_2
$\vdots$	$\vdots$
j ($j+$)	n_j
Total de datos	$n = \sum_{k=0}^{j} n_k$

y su logaritmo es, para $0 < p_0^M < 1$,

$$\begin{aligned} l^M &= n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k \left[\ln\left(1-p_0^M\right) + \ln p_k^T\right] \\ &= n_0 \ln p_0^M + \ln\left(1-p_0^M\right) \sum_{k=1}^{\infty} n_k + \sum_{k=1}^{\infty} n_k \left[\ln p_k - \ln\left(1-p_0\right)\right], \text{ por (4)} \\ &= \underbrace{n_0 \ln p_0^M + (n-n_0) \ln\left(1-p_0^M\right)}_{l_0} + \underbrace{\sum_{k=1}^{\infty} n_k \ln p_k - (n-n_0)\ln\left(1-p_0\right)}_{l_1}. \end{aligned} \quad (10)$$

Las funciones L^M y l^M se expresan en términos de los parámetros desconocidos p_0^M, a y b. Para los modelos regulares, $p_0^M = p_0$, y se obtiene el logaritmo de la función de verosimilitud de las distribuciones regulares. Se observa que l_0 depende solo de p_0^M, mientras que l_1 es independiente de p_0^M, dependiendo de a y b. Esta separación simplifica la maximización de l^M, dado que

$$\frac{\partial l^M}{\partial p_0^M} = \frac{\partial l_0}{\partial p_0^M} = \frac{n_0}{p_0^M} - \frac{n-n_0}{1-p_0^M} = 0,$$

y por tanto,

$$\hat{p}_0^M = \frac{n_0}{n}, \quad (11)$$

o sea la proporción de observaciones del cero en la muestra. Por ser tan natural, en el resto del artículo se estima p_0^M con (11) también para la estimación por momentos; no vale la pena en este caso abandonar este resultado simple y natural por la ortodoxia. Dado que por definición $p_0^T = 0$, por (4), la función de verosimilitud L^T está dada por

$$L^T = \prod_{k=1}^{\infty} \left(p_k^T\right)^{n_k} = \prod_{k=1}^{\infty} \left(\frac{p_k}{1-p_0}\right)^{n_k} = \frac{1}{\left(1-p_0\right)^{n-n_0}} \prod_{k=1}^{\infty} p_k^{n_k},$$

y su logaritmo es

$$l^T = \sum_{k=1}^{\infty} n_k \ln p_k - (n-n_0)\ln\left(1-p_0\right) = l_1. \quad (12)$$

Por tanto, si a y b optimiza l^M, entonces también optimiza l^T para la misma fp $\{p_k\}$ perteneciente a la clase (a, b). Para evitar la posibilidad de obtener valores de los parámetros a y b por fuera de sus espacios (véase la tabla 3), estudiaremos las familias de distribuciones por separado.

En el resto de la sección, cuando se indica que una ecuación de una variable (como (13) o (14) para $\hat{\lambda}$) se resuelve por métodos numéricos, significa que la solución se obtiene aplicando métodos como iteración de punto fijo o Newton-Raphson, entre otros; véase el capítulo 2 de Burden & Faires (1998), por ejemplo.

2.2. Parámetro λ de la familia Poisson

De la tabla 3, para la familia de distribuciones Poisson, $a = 0$ y $b = \lambda$; por lo tanto, el objetivo se centra en la estimación de λ.

- **Estimación por momentos**
  - **ZM-**$Poi(\lambda)$. Se resuelve para λ, por métodos numéricos, la ecuación obtenida al igualar (6) con (8), esto es, $\hat{\lambda}$ tal que
$$\left(1 - \hat{p}_0^M\right) \hat{\lambda} - \bar{x} \left(1 - e^{-\hat{\lambda}}\right) = 0. \tag{13}$$
  - **ZT-**$Poi(\lambda)$. De (13) con $p_0^T = p_0^M = 0$:
$$\hat{\lambda} - \bar{x} \left(1 - e^{-\hat{\lambda}}\right) = 0. \tag{14}$$
  - $Poi(\lambda)$. Con $p_0^M = p_0 = e^{-\lambda}$ en (13), $\hat{\lambda} = \bar{x}$. Resultado bastante conocido.
- **Estimación por máxima verosimilitud**
  - **ZM-**$Poi(\lambda)$ **y ZT-**$Poi(\lambda)$. De (12) y (8) se obtiene $\hat{\lambda}$ de
$$\begin{aligned} l_1 &= \sum_{k=1}^{\infty} n_k \ln \frac{e^{-\lambda}\lambda^k}{k!} - (n - n_0) \ln\left(1 - e^{-\lambda}\right) \\ &= n\bar{x} \ln \lambda - (n - n_0) \left[\lambda + \ln\left(1 - e^{-\lambda}\right)\right] + c, \end{aligned}$$
    donde c es independiente de λ. Por tanto,
$$\frac{\partial l_1}{\partial \lambda} = \frac{n\bar{x}}{\lambda} - \frac{n - n_0}{1 - e^{-\lambda}} = \frac{\bar{x}}{\lambda} - \frac{1 - \frac{n_0}{n}}{1 - e^{-\lambda}} = \frac{\bar{x}}{\lambda} - \frac{1 - \hat{p}_0^M}{1 - e^{-\lambda}} = 0,$$
    que coincide con la estimación del parámetro por momentos, ecuación (13).
  - $Poi(\lambda)$. De manera similar sucede con la distribución de Poisson regular con
$$l = \sum_{k=0}^{\infty} n_k \ln p_k = -\lambda n + \bar{x} n \ln \lambda - \sum_{k=0}^{\infty} n_k \ln k!.$$

Resumen. Para la familia de distribuciones de Poisson la estimación por momentos coincide con la estimación por máxima verosimilitud.

- $Poi(\lambda)$. $\hat{\lambda} = \bar{x}$.
- **ZM-**$Poi(\lambda)$. Se calcula $\hat{p}_0^M$ de (11) y $\hat{\lambda}$ de (13).
- **ZT-**$Poi(\lambda)$. Se calcula $\hat{\lambda}$ de (14).

2.3. Parámetros r y β de la familia binomial negativa

De la tabla 3, para la familia $BN(\beta, r)$, $a = \beta/(1+\beta)$ y $b = a(r-1)$; por tanto, se estiman los parámetros r y β que determinan las distribuciones.

- **Estimación por momentos**
  - **ZM-**$ETNB(r, \beta)$**.** Se resuelve el sistema de ecuaciones obtenido al igualar las ecuaciones (6) con (8) y (7) con (9).

$$\bar{x} = \frac{(1-\hat{p}_0^M)\beta r}{1-(1+\beta)^{-r}},\ m_2' = \frac{(1-\hat{p}_0^M)\beta r}{1-(1+\beta)^{-r}}\left[\beta(r+1)+1\right].$$

    Al dividir lado a lado ambas ecuaciones y operar, se expresa β en función de r:

$$\hat{\beta} = \frac{m_2' - \bar{x}}{\bar{x}(\hat{r}+1)}. \tag{15}$$

    que al reemplazarse en la primera de las ecuaciones de los momentos, se obtiene por métodos numéricos $\hat{r}$ tal que

$$\frac{\bar{x}^2(\hat{r}+1)}{\hat{r}} = \frac{(1-\hat{p}_0^M)(m_2'-\bar{x})}{1-\left[1+\frac{m_2'-\bar{x}}{\bar{x}(\hat{r}+1)}\right]^{-\hat{r}}}, \tag{16}$$

    y con este valor $\hat{r}$, se calcula $\hat{\beta}$ de (15).
  - **ETBN(**r, β**).** De (16) con $\hat{p}_0^M = p_0^T = 0$, se obtiene $\hat{r}$ de

$$\frac{\bar{x}^2(\hat{r}+1)}{\hat{r}} = \frac{m_2' - \bar{x}}{1-\left[1+\frac{m_2'-\bar{x}}{\bar{x}(\hat{r}+1)}\right]^{-\hat{r}}}. \tag{17}$$

    Con $\hat{r}$ de (17) y (15), se obtiene $\hat{\beta}$.
  - **BN(**r, β**).** Con un poco de álgebra se verifica que (16) también es válida para la distribución $BN(r, \beta)$. De la ecuación (16) con

$$\hat{p}_0^M = \hat{p}_0 = \left(1+\hat{\beta}\right)^{-r} = \left[1+\frac{m_2'-\bar{x}}{\bar{x}(r+1)}\right]^{-r},$$

se calcula $\hat{r}$ en forma directa con

$$\hat{r} = \frac{\bar{x}^2}{m_2' - \bar{x} - \bar{x}^2}. \tag{18}$$

Nota. $\hat{r}$ es positivo, pues en la distribución binomial negativa se verifica que la varianza muestral $m_2 = m_2' - \bar{x}^2$ es mayor que la media $\bar{x}$: $m_2' - \bar{x} - \bar{x}^2 = \left(m_2' - \bar{x}^2\right) - \bar{x} > 0$.
Con $\hat{r}$ de (18) y (15), se obtiene

$$\hat{\beta} = \frac{m_2' - \bar{x}^2}{\bar{x}} - 1. \tag{19}$$

- **Estimación por máxima verosimilitud**

  - **ETNB**(r, β). Con los datos en la forma de la tabla 4, el logaritmo de la función de verosimilitud de la distribución ETNB, es

$$l = l_1 = \sum_{k=1}^{\infty} n_k \ln p_k = \sum_{k=1}^{j} n_k \ln \left\{ \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} \left(\frac{\beta}{1+\beta}\right)^r \left[(1+\beta)^r - 1\right]^{-1} \right\}. \tag{20}$$

  Los parámetros $\hat{\beta}$ y $\hat{r}$ se obtienen maximizando l_1 de la ecuación (20) numéricamente, lo que puede hacerse con el método simplex en la presentación de Klugman et al. (2004). Este método no está relacionado con el método simplex de la investigación de operaciones.

  - $BN(r, \beta)$. Con los datos en la forma de la tabla 4, se sabe que el logaritmo de su función de verosimilitud, es

$$l = \sum_{k=0}^{\infty} n_k \ln p_k = \sum_{k=0}^{j} n_k \ln \left[\binom{r+k-1}{k} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^k \right]. \tag{21}$$

  Los parámetros que optimizan (21) se obtienen al resolver numéricamente para $\hat{r}$ (por el método de Newton-Raphson, por ejemplo) la ecuación

$$\sum_{k=1}^{j} n_k \left(\sum_{i=0}^{k-1} \frac{1}{\hat{r}+i} \right) = n \ln \left(1 + \frac{\bar{x}}{\hat{r}}\right), \tag{22}$$

  y calculando $\hat{\beta}$ de

$$\hat{\beta} = \frac{\bar{x}}{\hat{r}}. \tag{23}$$

  En efecto, de (21)

$$\frac{\partial l}{\partial \beta} = \sum_{k=0}^{j} n_k \left(\frac{k}{\beta} - \frac{r+k}{1+\beta} \right).$$

Al igualar la anterior ecuación a cero, se obtiene (23). También de (21),

$$\begin{aligned}\frac{\partial l}{\partial r} &= \sum_{k=0}^{j} n_k \frac{\partial}{\partial r} \ln\left[\frac{(r+k-1)!}{(r-1)!k!}\right] - \sum_{k=0}^{j} n_k \ln(1+\beta) \\ &= \sum_{k=0}^{j} n_k \frac{\partial}{\partial r} \ln \prod_{s=0}^{k-1} (r+s) - n\ln(1+\beta) \\ &= \sum_{k=0}^{j} n_k \left(\sum_{s=0}^{k-1} \frac{1}{r+s}\right) - n\ln(1+\beta)\,.\end{aligned}$$

Al igualar a cero la anterior ecuación y reemplazar a β por $\bar{x}/r$, se obtiene (22).

Resumen. Familia de distribuciones binomial negativa. Para la optimización de la función de verosimilitud por el método simplex, se pueden usar como valores iniciales los estimados de los parámetros obtenidos por momentos.

- $BN(r,\beta)$
  - **Momentos**. Se calcula $\hat{r}$ de (18) y $\hat{\beta}$ de (19).
  - **Máxima verosimilitud**. Se calcula por métodos numéricos $\hat{r}$ de (22) y $\hat{\beta}$ de (23).
- **ZM-ETNB**(r,β)
  - **Momentos**. Se calcula $\hat{p}_0^M$ con (11). $\hat{r}$ de obtiene de (16) y con éste, $\hat{\beta}$ de (15).
  - **Máxima verosimilitud**. Se calcula $\hat{p}_0^M$ con (11). Se obtienen por métodos numéricos $\hat{r}$ y $\hat{\beta}$ que maximizan (20).
- **ETNB**(r,β)
  - **Momentos**. $\hat{r}$ de obtiene de (17) y con éste, $\hat{\beta}$ de (15).
  - **Máxima verosimilitud**. Se obtienen por métodos numéricos $\hat{r}$ y $\hat{\beta}$ que maximizan (20).

2.4. Parámetro β de la familia geométrica

De la tabla 3, para $Geo(\beta)$, $a = \beta/(1+\beta)$ y $b = 0$; por tanto, se estima el parámetro β que determina la distribución.

- **Estimación por momentos**

  - **ZM-***Geo*(β). Al igualar las ecuaciones (6) con (8), se obtiene

$$\hat{\beta} = \frac{\bar{x}}{1 - \hat{p}_0^M} - 1. \tag{24}$$

  - **ZT-***Geo*(β). De (24) con $\hat{p}_0^M = p_0^T = 0$, $\hat{\beta} = \bar{x} - 1$.
  - *Geo*(β). $\hat{\beta} = \bar{x}$.

- **Estimación por máxima verosimilitud**

  - **ZM-***Geo*(β) **y ZT-***Geo*(β). De (12), con los datos en la forma de la tabla 4,

$$l_1 = \sum_{k=1}^{j} n_k \ln \left[\frac{\beta^k}{(1+\beta)^{k+1}} \right] - (n - n_0) \ln \left(\frac{\beta}{1+\beta} \right).$$

    Al resolver $dl_1/d\beta = 0$ para β, se obtiene

$$\hat{\beta} = \frac{n\bar{x} - n + n_0}{n - n_0} = \frac{\bar{x}}{1 - \hat{p}_0^M} - 1,$$

    que coincide con la estimación por momentos (24).

  - *Geo*(β). $\hat{\beta} = \bar{x}$, igual que por momentos.

Resumen. Para la familia de distribuciones geométrica la estimación por momentos coincide con la estimación por máxima verosimilitud.

- *Geo*(β). $\hat{\beta} = \bar{x}$.
- **ZM-***Geo*(β). $\hat{p}_0^M$ de (11) y $\hat{\beta}$ de (24).
- **ZT-***Geo*(β). $\hat{\beta} = \bar{x} - 1$.

2.5. Parámetros m y q de la familia binomial

De la tabla 3, para $Bin(m, q)$, $a = -q/(1-q)$ y $b = -a(m+1)$; por lo tanto, se estiman los parámetros q y m que determinan las distribuciones.

En la práctica se aborda esta estimación con dos enfoques: considerando que m es conocido, $m = j$, el máximo valor observado en la muestra según la notación de la tabla 4, o bien, estimando ambos parámetros, m desconocido.

2.5.1. Parámetro m conocido

- **Estimación por momentos**
  - **ZM-**$Bin(m,q)$. Se resuelve numéricamente para q la ecuación obtenida al igualar la ecuación (6) con (8).

$$\bar{x} = \frac{1-\hat{p}_0^M}{1-(1-q)^m} mq. \tag{25}$$

  - **ZT-**$Bin(m,q)$. Con $\hat{p}_0^M = p_0^T = 0$ en (25), se obtiene $\hat{q}$ al resolver numéricamente

$$\bar{x} = \frac{mq}{1-(1-q)^m}. \tag{26}$$

  - $Bin(m,q)$. Con $\hat{p}_0^M = \hat{p}_0 = (1-\hat{q})^m$ en (25), $\hat{q} = \bar{x}/m$.

- **Estimación por máxima verosimilitud**
  - **ZM-**$Bin(m,q)$ **y ZT-**$Bin(m,q)$. De (12), con los datos en la forma de la tabla 4,

$$l_1 = \sum_{k=1}^{m} n_k \ln\left[\binom{m}{k} q^k (1-q)^{m-k}\right] - (n-n_0)\ln\left[1-(1-q)^m\right]. \tag{27}$$

    Como acá m es conocido, al derivar (27) respecto de q e igualar a cero, se obtiene

$$\frac{dl_1}{dq} = \frac{n\bar{x}}{q} + \frac{n\bar{x} - m(n-n_0)}{1-q} - \frac{m(n-n_0)(1-q)^{m-1}}{1-(1-q)^m} = 0. \tag{28}$$

    El parámetro $\hat{q}$ se obtiene resolviendo (28) numéricamente.
  - $Bin(m,q)$. Con los datos en la forma de la tabla 4, se sabe que el logaritmo de la función de verosimilitud es

$$l = \sum_{k=0}^{m} n_k \ln\left[\binom{m}{k} q^k (1-q)^{m-k}\right].$$

    Con m conocido, al derivar l respecto de q e igualar a cero, se obtiene la misma estimación para q que por el método de los momentos.

Resumen 1. Para la familia de distribuciones binomial con m conocido, máximo valor de la variable en la muestra.

- $Bin(m,q)$. La estimación por momentos coincide con la estimación por máxima verosimilitud: $\hat{q} = \bar{x}/m$.
- **ZM-**$Bin(m,q)$
  - **Momentos**. Se calcula $\hat{p}_0^M$ con (11), $\hat{q}$ se obtiene de (25).

- **Máxima verosimilitud**. Se calcula $\hat{p}_0^M$ con (11), $\hat{q}$ se obtiene de (28).

▪ **ZT-**$Bin(m, q)$

- **Momentos**. $\hat{q}$ se obtiene de (26).
- **Máxima verosimilitud**. $\hat{q}$ se obtiene de (28).

2.5.2. Parámetro m desconocido

▪ **Estimación por momentos**

- **ZM-**$Bin(m, q)$. Se resuelve el sistema de ecuaciones obtenido al igualar las ecuaciones (6) con (8) y (7) con (9).

$$\bar{x} = \frac{(1-\hat{p}_0^M)mq}{1-(1-q)^m},\ m_2' = \frac{(1-\hat{p}_0^M)mq}{1-(1-q)^m}\left(1-q+mq\right).$$

Al dividir lado a lado ambas ecuaciones y operar, se expresa q en función de m:

$$\hat{q} = \frac{m_2' - \bar{x}}{\bar{x}(\hat{m}-1)}, \tag{29}$$

que al reemplazarse en la primera de las ecuaciones de los momentos, se obtiene por métodos numéricos $\hat{m}$ tal que

$$\frac{\bar{x}^2(\hat{m}-1)}{m_2'-\bar{x}} = \frac{(1-\hat{p}_0^M)\hat{m}}{1-\left[1-\frac{m_2'-\bar{x}}{\bar{x}(\hat{m}-1)}\right]^{\hat{m}}}, \tag{30}$$

y con este valor se calcula $\hat{q}$ de (29).

- **ZT-**$Bin(m, q)$. De la ecuación (30) con $\hat{p}_0^M = p_0^T = 0$, se obtiene $\hat{m}$ de

$$\frac{\bar{x}^2(\hat{m}-1)}{m_2'-\bar{x}} = \frac{\hat{m}}{1-\left[1-\frac{m_2'-\bar{x}}{\bar{x}(\hat{m}-1)}\right]^{\hat{m}}}. \tag{31}$$

Con $\hat{m}$ de (31) y (29), se obtiene $\hat{q}$.

- $Bin(m, q)$. De la ecuación (30) con

$$\hat{p}_0^M = \hat{p}_0 = (1-\hat{q})^m = \left[1-\frac{m_2'-\bar{x}}{\bar{x}(\hat{m}-1)}\right]^m,$$

se calcula $\hat{m}$ en forma directa con

$$\hat{m} = \frac{\bar{x}^2}{\bar{x}^2+\bar{x}-m_2'}. \tag{32}$$

Nota. $\hat{m}$ es positivo, pues en la distribución binomial se verifica que la varianza muestral $m_2 = m_2' - \bar{x}^2$ es menor que la media $\bar{x}$: $\bar{x}^2 + \bar{x} - m_2' = \bar{x} - \left(m_2' - \bar{x}^2\right) > 0$.
Con $\hat{m}$ de (32) y (29), se obtiene

$$\hat{q} = 1 - \frac{m_2' - \bar{x}^2}{\bar{x}}. \tag{33}$$

Es claro que $0 < \hat{q} < 1$ es positivo por la misma razón citada en la nota anterior.

- **Estimación por máxima verosimilitud ZM-Bin y ZT-Bin**. La función de verosimilitud a maximizar se indicó en (27). Se presentan a continuación dos formas para hacerlo, siendo la primera más intuitiva. La notación $x \leftarrow 3$ significa que a la variable x se le asigna el valor 3.

  Forma 1. Los parámetros $\hat{q}$ y $\hat{m}$ se obtienen con el siguiente algoritmo:

  Paso 1 $\hat{m} \leftarrow j$, donde j es el máximo valor de la muestra.

  Paso 2 Se obtiene $\hat{q}$ de (29)

  Paso 3 Asignaciones:
  $i \leftarrow 2$
  TOL $\leftarrow 10^{-5}$, o la tolerancia elegida
  $y_1 \leftarrow -\infty$

  Paso 3 Se calcula l_1 de (27)

  Paso 4 $y_2 \leftarrow l_1$

  Paso 5 Mientras $y_i - y_{i-1} >$ TOL
  $i \leftarrow i + 1$
  $\hat{m} \leftarrow \hat{m} + 1$
  Se obtiene $\hat{q}$ de (29)
  Se calcula l_1 de (27)
  $y_i \leftarrow l_1$

  Los valores $\hat{m}$ y $\hat{q}$ en la salida del **Paso 5** maximizan (27).

  Forma 2. Los parámetros $\hat{q}$ y $\hat{m}$ se obtienen maximizando l_1 de la ecuación (27) numéricamente, por ejemplo usando el método simplex.

  Para la distribución Bin(m, q), con los datos en la forma de la tabla 4, se sabe que el logaritmo de la función de verosimilitud es

$$l = \sum_{k=0}^{j} n_k \ln \left[\binom{m}{k} + k \ln q + (m-k) \ln (1-q) \right]. \tag{34}$$

Los parámetros $\hat{q}$ y $\hat{m}$ de la distribución binomial se obtienen por cualquiera de las dos formas indicadas para l_1 en (27) realizando los cambios evidentes.

Resumen 2. Para la familia de distribuciones binomial con parámetro m desconocido.

- $Bin(m, q)$**.**

  Momentos. Se calcula $\hat{m}$ de (32) y $\hat{q}$ de (33).

  Máxima verosimilitud. Se estiman $\hat{m}$ y $\hat{q}$ usando cualquiera de los algoritmos denominados Forma 1 y Forma 2, cuidando de usar la función de verosimilitud l de (34) en lugar de la l_1 de (27).

- **ZM-**$Bin(m, q)$**.**

  Momentos. Se calcula $\hat{p}_0^M$ con (11). $\hat{m}$ de obtiene de (30) y con éste, $\hat{q}$ de (29).

  Máxima verosimilitud. Se estiman $\hat{m}$ y $\hat{q}$ usando cualquiera de los algoritmos denominados Forma 1 y Forma 2 para l_1 de (27).

- **ZT-**$Bin(m, q)$**.**

  Momentos. $\hat{m}$ se obtiene de (31) y con éste, $\hat{q}$ de (29).

  Máxima verosimilitud. Se estiman $\hat{m}$ y $\hat{q}$ usando cualquiera de los algoritmos denominados Forma 1 y Forma 2 para l_1 de (27).

2.6. Parámetro β de la familia logarítmica

Para la distribución $Log(\beta)$, $p_0 = 0$. De la tabla 3, $a = \beta/(1+\beta)$ y $b = -a$; por tanto, se estima el parámetros β.

- **Estimación por momentos ZM-**$Log(\beta)$. Al igualar las ecuaciones (6) con (8), se obtiene $\hat{\beta}$ numéricamente de

$$\bar{x} \ln\left(1+\hat{\beta}\right) = \left(1-\hat{p}_0^M\right)\hat{\beta}. \tag{35}$$

  $Log(\beta)$. Con $\hat{p}_0^M = p_0 = 0$ en (35), se obtiene $\hat{\beta}$ numéricamente de

$$\bar{x} \ln\left(1+\hat{\beta}\right) = \hat{\beta}. \tag{36}$$

- **Estimación por máxima verosimilitud**. De (12), con los datos en la forma de la tabla 4,

$$\begin{aligned} l = l_1 &= \sum_{k=1}^{j} n_k \ln p_k = \sum_{k=1}^{j} n_k \left\{ k \ln\left(\frac{\beta}{1+\beta}\right) - \ln\left[k \ln\left(1+\beta\right)\right] \right\} \\ &= n\bar{x}\left[\ln\beta - \ln\left(1+\beta\right)\right] - \sum_{k=1}^{j} n_k \ln\left[k \ln\left(1+\beta\right)\right]. \end{aligned}$$

Al resolver $dl_1/d\beta = 0$ para β, se obtiene $\hat{\beta}$ tal que verifica (36), igual estimación que por momentos.

Resumen. Para la familia de distribuciones logarítmicas la estimación por máxima verosimilitud coincide con la estimación por momentos.

- $Log(\beta)$**.** $\hat{\beta}$ se obtiene de (36).
- **ZM-**$Log(\beta)$**.** $\hat{p}_0^M$ se obtiene de (11) y $\hat{\beta}$ numéricamente de (35).

2.7. Ejemplos numéricos

A continuación se analizan cuatro conjuntos de datos reales de origen diverso:

1. De riesgo operativo: número de torres de energía dañadas en cada día por atentados terroristas en parte de la infraestructura del sistema de transmisión eléctrica de un país suramericano entre los años 2 000 y 2 002. Fuente: parte de los datos de un trabajo de consultoría del autor.

2. De seguros/actuaría: frecuencia de accidentes de autos en una flota. Fuente: Beard et al. (1984).

3. De salud ocupacional/actuaría de vida: número de muertes causadas por patadas de caballo en 14 cuerpos de caballería prusiana en el siglo XIX. Fuente: Härdle & Vogt (2014).

4. De riesgos de la naturaleza/riesgo operativo: número anual de huracanes intensos en la cuenca del océano Atlántico. Fuente: *NOAA* (`http://www.aoml.noaa.gov/hrd/`).

Todos los valores p de los ejemplos se refieren a la prueba estadística Ji-Cuadrada.

2.7.1. Frecuencia de daños por terrorismo en un sistema de transmisión eléctrica

En 244 días, entre enero de 2 000 y noviembre de 2 002, hubo atentados terroristas en una importante parte del sistema de transmisión eléctrica de un país suramericano. La tabla 5 muestra el número de torres de energía dañadas en cada uno de tales días: en 182 días dañaron una sola torre, 41 días dañaron 2 torres, y así sucesivamente.

Como el valor mínimo de la frecuencia es 1, se estudian las versiones cero truncadas de las distribuciones clase $(a, b; 0)$ y en la distribución logarítmica, como posibles modelos.

Tabla 5: *Daños en infraestructura eléctrica por terrorismo. Fuente: elaboración propia.*

N° Torres, k	N° de días, n_k	ZT-$Geo(\beta)$	$Log(\beta)$
1	182	177.72	183.18
2	41	48.28	41.51
3	16	13.11	12.54
4	3	3.56	4.26
5	1	0.97	1.55
6	1	0.36	0.96
Total	$n = 244$	244.00	244.00
Parámetros		$\hat{\beta} = 0.3730$	$\hat{\beta} = 0.8288$
$l = \ln(L)$		-195.6195	-195.0311
Valor p		0.5478	0.8195

Las dos columnas de la derecha de la tabla 5 contienen los dos mejores ajustes. Basados en el valor p y el logaritmo de la función de verosimilitud, el fenómeno se puede modelar con la distribución logarítmica $Log(0.8288)$.

Sean las variables aleatorias discretas X con distribución $Log(0.8288)$ ($X \sim Log(0.8288)$) y $N \sim Poi(82.2)$. Si N es el número de días en un año con atentados terroristas (un supuesto *a priori*) y X el número de torres dañadas en uno de tales días, entonces $Y = X_1 + X_2 + \cdots + X_N$ (y $Y = 0$ para $N = 0$), representa el número anual de torres dañadas por terrorismo, donde las variables aleatorias X_j, para $j = 1, 2, \ldots, N$ son independientes con distribución común $Log(0.8288)$.

La distribución de probabilidad de la *variable compuesta* Y se puede calcular con el conocido *algoritmo de Panjer*. Además, si se tiene un modelo para la severidad de la pérdida monetaria del daño de cada torre afectada, entonces, con el *modelo de riesgo colectivo*, se obtiene la distribución de las pérdidas agregadas anuales por atentados terroristas en las torres de energía de la empresa de tal país. Véanse Klugman et al. (2004) y Escalante (2006).

2.7.2. Frecuencia de reclamos en una póliza de autos

Los datos de la tabla 6 son tomados de Beard et al. (1984). La primera columna de la izquierda indica el número de accidentes y la segunda el número de pólizas de una cartera con $n = 421\,240$ vehículos asegurados. Así, $k = 1$ y $n_1 = 46\,545$ significa que en el período en estudio 46 545 vehículos/pólizas presentaron un único reclamo.

Para estos datos, $\bar{x} = 0.13174$, $m_2' = 0.15588$, por tanto, la varianza muestral $s^2 = 0.13852$.

Las dos columnas de la derecha del mismo tabla 6 resumen los resultados de los mejores ajustes de los datos a los modelos acá estudiados, a saber: ZM-$ETNB(p_0^M =$

Tabla 6: *Frecuencia de accidentes en una póliza de autos. Fuente: elaboración propia.*

N° de accidentes, k.	N° de observaciones, n_k	ZM-$ETNB$	ZM-Geo
0	370 412	370 412.00	370 412.00
1	46 545	46 547.79	46 555.16
2	3 935	3 926.84	3 913.64
3	317	324.49	329.00
4	28	26.53	27.66
5+	3	2.35	2.54
Total de datos	$n = 421\ 240$	421 240.00	421 240.00
Parámetros		$\hat{p}_0^M = 0.8793$	$\hat{p}_0^M = 0.8793$
		$\hat{r} = 1.1310$	$\hat{\beta} = 0.0918$
		$\hat{\beta} = 0.0860$	
$l = \ln(L)$		$-171\ 133.00$	$-171\ 133.10$
Valor p		0.7985	0.8872

$0.8793, r = 1.1310, \beta = 0.0860$) y ZM-$Geo$(0.0918). Este último modelo es más simple que el primero y tiene mejor ajuste, como lo muestra su valor p = 0.8872.

2.7.3. Frecuencia de muertes por patadas de caballo en Prusia

Las dos columnas de la izquierda de la tabla 7 son tomadas de Härdle & Vogt (2014) y corresponden a una muestra del registro llevado durante 20 años consistente en el número anual de muertes en 14 cuerpos de la caballería prusiana entre los años 1 875 y 1 894. El análisis de estos datos, con resultados similares a los que acá se presentan, fue publicado por el científico Ladislaus von Bortkiewicz en 1 898.

Las dos columnas de la derecha del mismo tabla 7 resumen los resultados de los ajustes de los datos a los modelos $Pois$(0.6181) y ZM-$Pois$(0.6100). Se observa que el modelo $Pois$(0.6181) es más simple y tiene mejor ajuste de acuerdo con el valor p. No es difícil pensar en análisis similares para accidentes industriales o laborales en las empresas, incluyendo los accidentes de tránsito que ocasionan tantas muertes en las actuales sociedades industrializadas.

2.7.4. Frecuencia anual de huracanes intensos en la cuenca del Atlántico

Desde 1 966, las imágenes de satélite del Centro Nacional de Huracanes (NHC por sus siglas en inglés) sirven para registrar los huracanes en la cuenca del Atlántico. Las dos columnas de la izquierda de la tabla 8 resumen la frecuencia anual de huracanes intensos (con intensidades 3, 4 o 5 en la escala Saffir-Simpson) entre los años $1\ 968 - 2\ 014$. Así, en $n_2 = 14$ de los $n = 47$ años se presentaron 2 huracanes intensos, 5 de los 47 años no presentaron huracanes intensos, y en 3 años se presentaron 6 o más de tales huracanes. Véase el registro anual en *NOAA* (`http://www.aoml.noaa.gov/hrd/`).

Tabla 7: *Muertes entre* $1\,875 - 1\,894$ *por patadas de caballo. Fuente: elaboración propia.*

N° de muertes, k.	N° de observaciones, n_k	ZM-$Pois(\lambda)$	$Pois(\lambda)$
0	109	109.00	108.67
1	65	65.76	66.29
2	22	20.32	20.22
3	3	4.19	4.11
4	1	0.74	0.71
Total	$n = 200$	200.00	200.00
Parámetros		$\hat{p}_0^M = 0.5450$ $\hat{\lambda} = 0.6181$	$\hat{\lambda} = 0.6100$
$l = \ln(L)$		-205.9738	-205.9796
Valor p		0.7483	0.8964

La media muestral anual de huracanes intensos es $\bar{x} = 2.3617$ y la varianza muestral es $s^2 = 2.8266$. La tabla 8 contiene los dos mejores ajustes entre las distribuciones estudiadas, a saber, $Poi(2.3617)$ con valor p $= 0.23911$ y $BN(r = 11.5516, \beta = 0.2044)$ (estimación por máxima verosimilitud) con valor p $= 0.28936$. Por tanto, un buen modelo del número anual de huracanes intensos en la cuenca del Atlántico es la distribución binomial negativa indicada.

3. Generación de números aleatorios

Se sabe que si una variable aleatoria X tiene función de distribución F, entonces $F(X)$ es una variable aleatoria con distribución uniforme en el intervalo $[0, 1]$, esto es, $F(X) \sim U[0, 1]$. Este hecho se usa para generar números aleatorios de X a partir de la inversión de su función de distribución F, técnica conocida como el método de la transformada inversa ((Ross 2013), sección 4.1 para variables aleatorias discretas y sección 5.1 para variables aleatorias continuas).

Así, para generar un número aleatorio x de la variable aleatoria X se genera un número aleatorio u de $U[0, 1]$ y se resuelve para x la ecuación $F(x) = u$.

Sea N una variable aleatoria discreta con media finita μ y función de probabilidad clase (a, b) $\pi_k = \Pr(N = k)$, $k = 0, 1, \ldots$. El n-vector r del siguiente algoritmo contiene los números aleatorios generados a partir de la función de probabilidad π_k.

En las líneas del algoritmo que tienen expresiones de la forma [*código*] se indica el código de la instrucción para el caso en el que $\pi_0 = p_0^M \in [0, 1)$, esto es, cuando π_k corresponde a la versión cero modificada o cero truncada de la distribución. En tal caso la instrucción se tiene en cuenta, y se ignora en caso contrario.

En el **Paso 2** del algoritmo, en las filas donde aparece una instrucción [*código*] al lado derecho de otra instrucción, ésta última se ignora cuando $\pi_0 = p_0^M \in [0, 1)$;

Tabla 8: *Frecuencia anual de huracanes intensos*$^{(i)}$ *en el Atlántico entre* 1 968 − 2 014. *Fuente: elaboración propia.*

N° anual, k.	N° de años, n_k	$Pois(\lambda)$	$BN(r,\beta)$
0	5	4.43	108.67
1	11	10.46	66.29
2	14	12.36	20.22
3	7	9.73	4.11
4	2	5.74	0.71
5	5	2.71	0.71
6+	3	1.57	0.71
Total	$n = 47$	47.00	47.00
Parámetros		$\hat{\lambda} = 2.3617$	$\hat{r} = 11.5516$
			$\hat{\beta} = 0.2044$
$l = \ln(L)$		-86.7285	-85.9068
Valor p		0.2391	0.2894

i. Huracanes con intensidades 3, 4 o 5 en la escala Saffir-Simpson.

por ejemplo, en

Paso 2 línea 4. $k \leftarrow 1$ $[k \leftarrow 2]$,

solo se tiene en cuenta la instrucción $k \leftarrow 2$.

Algoritmo para generar n números aleatorios de distribuciones clase (a, b).

Paso 1 Datos de entrada:

$[p_0^M]$, a y b, parámetros de la distribución.

n, cantidad de números aleatorios a generar.

Paso 2 Para i desde 1 hasta n:

1. Se genera un número aleatorio u de la distribución uniforme en $[0, 1]$.
2. $p \leftarrow \pi_0$ $[p \leftarrow \pi_1]$.
3. $F \leftarrow \pi_0$ $[F \leftarrow p_0^M + p]$.
4. $k \leftarrow 1$ $[k \leftarrow 2]$.
5. [Si $u \leq p_0^M$, entonces $r_i \leftarrow 0$].
6. [Si $u > p_0^M$, entonces]
   Mientras $u > F$
   $p \leftarrow p\left(a + \frac{b}{k}\right)$

$F \leftarrow F + p$
$k \leftarrow k + 1$

7. $r_i \leftarrow k - 1$.

Paso 3 Se imprime el n-vector r.

El algoritmo puede ser poco eficiente si μ tiene un valor alto, $\mu \geq 50$, por ejemplo, pues cada vez se inicia en 0 (o en 1, si $\pi_0 = 0$). El lector puede mejorar el algoritmo iniciando F en μ o en la moda o mediana de N. Tanto el algoritmo como la idea de mejoramiento iniciando a F en la moda de N fueron motivadas por Ross (2013, sec.4.2) al exponer el algoritmo para generar números aleatorios Poisson sacando ventaja de su presentación recursiva por ser clase (a, b).

Como en la generación de números aleatorios a partir de una distribución discreta empírica (Ross 2013, sec. 4.1), el algoritmo se construye acumulando probabilidades provistas por los números aleatorios u (tomados de la distribución continua uniforme en $[0, 1]$) y comparándolos con la función de distribución F. Como se observa, F se construye *a saltos* usando la recursividad $\pi_{k+1} = \pi_k\,(a + b/k)$ en el bucle Mientras del **paso 2**.

Ejemplo de generación de n números aleatorios ZM-$Poi(\lambda)$. A continuación, se define la función `rzmpoi` en R con el algoritmo de la sección 3 para generar n números aleatorios ZM-$Poi(\lambda)$.

```
rzmpoi <- function(n, param){
    p0M <- param[1]
    lambda <- param[2]
    a <- 0
    b <- lambda
    x <- Inf
    for(i in 1:n){
        u <- runif(1)
        if(u <= p0M){x[i] <- 0}
        if(u > p0M){
            p <- (1-p0M)*exp(-lambda)*lambda/(1-exp(-lambda))
            EFE <- p0M + p
            k <- 2
            while(u>EFE){
                p <- p*(a+b/k)
                EFE <- EFE + p;
                k <- k + 1
            }
            x[i] <- k - 1
        }
    }
    x}
```

Por ejemplo, para $n = 40$, $p_0^M = 0.3$ y $\lambda = 2.63$:

```
> set.seed(620)
> x <- rzmpoi(n = 40, param = c(0.3, 2.63))
> x
 [1] 0 1 0 0 1 4 1 3 0 3 0 4 3 4 1 2 3 3 0 2 2 0 0 0 0 0 3 1 4 1 0 1 4 5 2 0
[37] 5 3 5 4
> table(x)
x
 0  1  2  3  4  5
13  7  4  7  6  3
```

El código siguiente muestra la comparación entre la frecuencia relativa de los números simulados `prob.sim` y las probabilidades teóricas `prob.teo`:

```
> ca <- 0:max(x)
> data.frame("k"=ca, "prob.sim"=round(as.numeric(table(x)/40), 3),
+ "prob.teo"=c(0.3, round((1-0.3)*dpois(ca[-1],2.63)/(1-dpois(0,2.63)),3)))
  k prob.sim prob.teo
1 0    0.325    0.300
2 1    0.175    0.143
3 2    0.100    0.188
4 3    0.175    0.165
5 4    0.150    0.108
6 5    0.075    0.057
```

Es claro que la función `rzmpoi` del presente ejemplo puede usarse para generar números aleatorios $Poi(\lambda)$ y ZT-$Poi(\lambda)$ con $p_0^M = \exp(-\lambda)$ y $p_0^M = 0$, respectivamente.

4. Conclusiones

Es evidente la versatilidad de las distribuciones clase (a, b) para modelar la frecuencia de modelos aleatorios de orígenes disímiles. No hay razón para no usarlas por parte de los analistas de riesgos que aplican modelos actuariales, ingenieros de confiabilidad, estudios de bioestadística, salud pública, seguridad industrial, entre otras disciplinas.

El algoritmo propuesto para generar números aleatorios es de tipo genético en términos de la generación de números aleatorios de distribuciones discretas generales que sacan ventaja de recursividad común de las distribuciones (a, b). Al final del algoritmo se dan ideas para mejorar su eficiencia.

Como se mencionó en la introducción, las variables aleatorias compuestas y la mixtura de distribuciones discretas son extensiones naturales de las distribuciones acá estudiadas. El algoritmo de Panjer, que es un *buen amigo* del analista de riesgos, y en él las distribuciones clase (a, b) son fundamentales. Se invita al lector para que acometa su estudio y aplicación.

Recibido: 7 de abril de 2016
Aceptado: 9 de noviembre de 2016

Referencias

Beard, R., Pentikainen, T. & Pesonen, E. (1984), *Risk Theory*, Chapman & Hall.

Burden, R. & Faires, J. D. (1998), *Análisis Numérico*, 6 edn, Thomson.

Escalante, C. (2006), 'Distribuciones clase (a, b) y algoritmo de Panjer', *Matemáticas: Ensenanza Universitaria.* **16**(2), 3–17.

Härdle, W. & Vogt, A. (2014), Ladislaus von Bortkiewicz: statistician, economist, and a European intellectual.

Johnson, N. L., Kemp, A. W. & Kotz, S. (2005), *Univariate Discrete Distributions*, 3 edn, Wiley.

Klugman, S. A., Panjer, H. H. & Willmot, G. E. (2004), *Loss Models: From Data to Decisions*, 3 edn, Wiley.

Panjer, H. & Willmot, G. (1992), *Insurance Risk Models*, Society of Actuaries.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Ross, S. (2013), *Simulation*, 5 edn, Academic Press.

Sundt, B. & Jewell, W. (1981), 'Further results on recursive evaluation of compound distributions.', *ASTIN Bulletin* **18**, 27–39.

Willmot, G. E. (1988), 'Sundt and Jewell's of discrete distributions', *ASTIN Bulletin* **18**, 17–29.