Universidad de Jaén

Escuela de Doctorado

TESIS DOCTORAL

•

NUEVOS MODELOS DE DISTRIBUCIONES PARA DATOS DE CONTEO

PRESENTADA POR:
VALENTINA CUEVA LÓPEZ

DIRIGIDA POR:
JOSÉ RODRÍGUEZ AVI
MARÍA JOSÉ OLMO JIMÉNEZ

JAÉN, 9 de marzo de 2023

ISBN

Índice general

Agradecimientos

Este trabajo no lo habría podido realizar sin el apoyo de mis directores de tesis, Pepe y María José. Muchas gracias por vuestra ayuda y consejos durante todos estos años. A todos los compañeros del Departamento de Estadística e Investigación Operativa de la Universidad de Jaén, los que están o han pasado por aquí en este periodo, por vuestro ánimo en momentos complicados. Gracias a mi familia y amigos, por no dejarme tirar la toalla en momentos difíciles.

Capítulo 1

Distribuciones discretas

1.1. Introducción

Un aspecto esencial del Cálculo de Probabilidades consiste en proponer modelos matemáticos teóricos que sirvan para describir los fenómenos aleatorios de modo que, si los datos se ajustan a los modelos, puedan ser interpretados a través de las propiedades inherentes a ellos. En este sentido, múltiples modelos de distribuciones han sido descritos. El caso probablemente más estudiado corresponde a las variables continuas, en donde la distribución normal es la principal referencia. Sin embargo, hay otro tipo de datos de especial interés que corresponde a variables discretas, referidas a aquellas variables procedentes de datos de conteo. Estas variables se caracterizan porque la probabilidad se acumula en puntos aislados, concretamente números naturales que comienzan en 0. Dado que en este trabajo se desarrollan modelos para este tipo de datos, se va a proceder a describir brevemente algunas de las distribuciones más utilizadas en el caso de variables de conteo, así como a mencionar algunas de las desarrolladas en los últimos años.

1.1.1. Breve recorrido histórico

El análisis de datos de conteo se inicia tempranamente en la historia del Cálculo de Probabilidades, debido en parte al interés por los juegos de azar, como era el caso del problema del lanzamiento de 3 dados y el por qué la probabilidad de que sumaran 9 era distinta de la que sumaran 10. En este sentido las primeras menciones se refieren a las distribuciones de Bernouilli y la distribución binomial. Su autoría se atribuye al matemático suizo Jakob (o James, o Jacques, todos ellos versiones del nombre de Santiago) Bernoulli ($1654-1705$) y aparece en 1713 en su obra póstuma *Ars Conjectandi* (libro publicado por su sobrino Nicholas), siendo una de las distribuciones más antiguas estudiadas en la literatura. En su desarrollo se utiliza el coeficiente binomial, que fue trabajado con anterioridad por Pascal. En el artículo de Boyer (1950) se pueden encontrar las primeras referencias a esta distribución, y aspectos históricos y filosóficos pueden verse en García-García et al. (2022) y Fernández Coronado et al. (2022).

Al mismo tiempo que se estudiaba la distribución binomial, se estudiaron formas específicas de la distribución de Pascal y la distribución binomial negativa (Fermat et al., 1679; Montmort, 1714). En 1837 Poisson publicó la obtención de la distribución que lleva su nombre (Poisson, 1837) a través de la aproximación de la distribución binomial. Para obtener dicha aproximación Poisson dedujo que las condiciones necesarias eran que el número de sucesos

debía ser muy elevado y la probabilidad de ocurrencia del suceso muy pequeña, añadiendo lo ya conocido (la independencia de sucesos, así como la igualdad de probabilidad de que ocurra el suceso en dos intervalos de tiempo o espacio cualesquiera). En 1898, Bortkiewicz llamó a esta situación "La ley de los números pequeños" (von Bortkiewicz, 1898). Para su aplicación, Bortkiewicz utilizó el número de muertes por patadas de mulas por año en el ejército prusiano, teniendo que la probabilidad de que ocurriese esta situación era muy pequeño y la cantidad de soldados muy grande, aunque las condiciones de independencia y de probabilidad constante son dudosamente satisfechas. En 1907, W. S. Gosset (*Student*) utilizó la distribución de Poisson como una primera aproximación del número de partículas caídas en un área muy pequeña cuando el número de estas se distribuye aleatoriamente dentro de un área muy grande (Student, 1907). Otro ejemplo es la representación de las variaciones del número de partículas emitidas por una fuente radioactiva en periodos de tiempo marcados (Rutherford et al., 1910, 1930).

Fue también *Student* quien a principios de siglo propuso la distribución binomial negativa para modelizar el número de glóbulos rojos en una gota de sangre, como alternativa a la distribución de Poisson para modelizar el número de ocurrencias de un suceso cuando los datos presentan sobredispersión (la varianza es mayor que la media, por lo que se incumple la propiedad que caracteriza a una distribución de Poisson, según la cual media y varianza coinciden). En 1920 la distribución binomial negativa se obtuvo también como consecuencia del uso de supuestos simples en la propensión de modelos de accidentes (Greenwood and Yule, 1920) y, por otro lado, como la distribución límite de un "modelo de urnas" (Eggenberger and Pólya, 1923).

1.2. Distribuciones de puro azar

Si bien todas las distribuciones de probabilidad se usan para modelizar fenómenos aleatorios, un caso especial es cuando esos fenómenos se deben fundamentalmente al azar (*pure chance*). En el caso continuo esa situación se modeliza mediante la distribución normal, mientras que en el caso discreto ese lugar le corresponde a la distribución binomial (caso finito) o Poisson (caso infinito numerable).

1.2.1. Distribución binomial

La distribución binomial aparece de forma natural cuando se realizan repeticiones independientes de un experimento cuyos resultados son "éxito", con probabilidad p o "fracaso", con probabilidad $q = 1-p$. Este experimento es conocido como experimento de Bernouilli. Así, sea X el número de éxitos obtenidos en los n ensayos independientes ($n > 1$) de Bernouilli. Se dice entonces que X sigue una distribución binomial con parámetros n y p, y se denota $B(n, p)$, y su función de masa de probabilidad (en adelante, fmp) está dada por:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

donde $\binom{n}{x} = n!/x!(n-x)!$ es el número combinatorio.

Las hipótesis de independencia y de probabilidad de éxito constante pueden no ser correctas, pero a menudo tienen una representación suficientemente adecuada.

Algunas características de la distribución binomial son:

- Esperanza: $E(X) = \mu = np$

- Varianza: $Var(X) = \sigma^2 = npq$
- Coeficiente de asimetría:
$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}.$$
Por esta razón, la distribución tiene una asimetría positva si $p > 1/2$ y negativa si $p < 1/2$.
- Coeficiente de curtosis:
$$\gamma_2 = 3 + \frac{1-6pq}{npq}.$$
- Función generatriz de probabilidad: $G(t) = (q+pt)^n, \quad t \in \mathbb{R}$.

Una propiedad importante de la distribución binomial es que es infradispersa, ya que su varianza es siempre menor que su media ($npq < np$).

La distribución binomial es de rango finito y ha sido ampliamente aplicada a lo largo de la historia, sobre todo, en los siglos XX y XXI. Por citar algunas publicaciones recientes, en el campo de la astronomía, Lu et al. (2022) utilizan la distribución binomial para el desarrollo de su modelo *Chocolate Chip Cookie* y el ajuste de la dependencia de la inclinación tanto del enrojecimiento efectivo del polvo de los componentes estelares como de las lineas de emisión caracterizadas por el decremento de Balmer para una gran muestra de galaxias; en Medicina, Singh et al. (2022) la utilizan para la remota monitorización de pacientes. Otros artículos recientes en los que se utiliza la distribución binomial son, por ejemplo, Irshad et al. (2022) y Shah et al. (2022), entre otros.

Si nos centramos en la implementación que tiene la distribución en diversos programas informáticos, se puede observar que está disponible en todos ellos. Concretamente, en R este modelo se encuentra en la librería básica `stats` (Team, 2023).

1.2.2. Distribución de Poisson

La distribución de Poisson es una distribución discreta para datos de conteo de rango infinito que expresa el número de ocurrencias en un intervalo de tiempo o espacio (volumen, distancias o áreas) siempre que estas ocurran con una razón media constante e independientemente del tiempo transcurrido desde el último evento. Se puede aplicar a situaciones con gran número de sucesos, así como a sucesos raros, es decir, con probabilidades pequeñas.

Se dice que X sigue una distribución de Poisson de parámetro λ, siendo λ el número medio de ocurrencias en un intervalo de tiempo o espacio, si su fmp es

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots \tag{1.1}$$

Una forma de obtener esta distribución de Poisson es a través de una aproximación de la distribución binomial (ver 1.2.1) para un gran tamaño de muestra, n, y una probabilidad, p, de ocurrencia muy pequeña. Matemáticamente hablando, cuando $n \to \infty$ y $p \to 0$, la media se puede aproximar por $\lambda = np$, entonces la distribución binomial converge a una distribución de Poisson, esto es,

$$P\left(X = x\right) = \binom{n}{x} p^x \left(1-p\right)^{n-x} \underset{n\uparrow\infty}{\longrightarrow} P(X = x) = \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda}$$

Algunas de las propiedades más importantes de dicha distribución son:

- $E(X) = \mu = \lambda$
- $Var(X) = \sigma^2 = \lambda$
- Coeficiente de asimetría: $\gamma_1 = \frac{1}{\sqrt{\lambda}} > 0$. Por esta razón, la distribución tiene asimetría a la derecha.
- Coeficiente de curtosis: $\gamma_2 = \frac{1}{\lambda} + 3$. Se observa entonces que la distribución es leptocúrtica.
- Función generatriz de probabilidad es $G(t) = e^{\lambda(t-1)}, \quad t \in \mathbb{R}$

Esta distribución es, en la actualidad, una de las más utilizadas en las investigaciones de diversas áreas. Así, entre los recientes estudios publicados podemos citar, en Ingeniería, el trabajo de Holmukhe et al. (2022) que la emplean para evaluar la energía potencial eólica en la India. Link et al. (2022) utilizan la distribución de Poisson en el campo de la Química, para la determinación de la encapsulación activa. Otras publicaciones recientes en las que se utiliza esta distribución son, por ejemplo, Lizama and Ponce (2023) y Santhiya et al. (2023).

En cuanto a la programación de esta distribución, está implementada en casi todos los programas estadísticos disponibles, tanto de libre acceso como comerciales. En R, está implementada en muchas librerías, la básica es `stats`.

La distribución de Poisson es utilizada como base para obtener otras distribuciones discretas, a través de mixturas, en donde se supone que λ es una variable aleatoria con alguna distribución de probbailidad. Pero además, esta distribución puede obtenerse también comomixtura. Así, si X sigue una distribución Binomial con parámetros (n, p) y n es una variable aleatoria que se fisitribuye según una Poisson de parámetro λ, entonces X sigue una distribución de Poisson de parámetro $n\lambda$.

La distribución de Poisson verifica que la media siempre coincide con la varianza. Esta relación, denominada *equidispersión*, se refleja en el *índice de agregación* (IA), que se define como:

$$IA = \frac{Var(X)}{E(X)} \tag{1.2}$$

Este coeficiente, que es constante e igual a 1 para cualquier distribución de Poisson sea cual sea λ, desempeña un papel, en cuanto a clasificación de distribuciones discretas, similar al que desempeña el coeficiente de curtosis de Fisher en el caso continuo, el cual es constante para la distribución normal, también de manera independiente de los parámetros. Así si ese coeficiente sirve para clasificar las distribuciones como mesocúrticas, platicúrticas o leptocúrticas, también el IA se utiliza para clasificar las distribuciones de conteo en sobredispersas, ($IA > 1$), equidispersas ($IA = 1$) o infradispersas ($IA < 1$), lo que tiene gran interés a la hora de proponer modelos para un conjunto concreto de datos. De hecho, cualquier mixtura de una Poisson es siempre sobredispersa (ver, por ejemplo, Willmot (1986).

1.3. Distribuciones sobredispersas

La equidispersión de la distribución de Poisson es una propiedad muy importante relacionada con el hecho de que esta distribución es la que modeliza el puro azar. Sin embargo,

en múltiples circunstancias, los datos de conteo presentan una varianza mayor que la media, debido a la presencia de otras causas que afectan a los resultados. En muchos casos, esa sobredispersión está relacionada con la presencia de una mayor probabilidad del 0 que lo que ocurre en la distribución de Poisson, que es, recordemos, $e^{-\lambda}$. En ocasiones, esas distribuciones pueden obtenerse como mixturas de la distribución de Poisson, lo que además proporciona la propiedad de que la varianza resultante pueda expresarse en términos debidos al puro azar (*randomness*) y la parte que responde a otras causas. A continuación describimos brevemente las distribuciones sobredispersas más utilizadas.

1.3.1. Distribución geométrica

En las mismas condiciones en las que se definió la distribución binomial, supongamos que el interés no reside en el número de éxitos obtenidos en las n repeticiones del experimento, sino en el número de repeticiones necesarias hasta encontrar el primer éxito. Así, si se define X como el número de fracasos antes de encontrar el primer éxito, se dice que X sigue una distribución geométrica de parámetro p, $Geo(p)$ (p la probabilidad de éxito) y su fmp está dada por

$$P(X = x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots \tag{1.3}$$

Algunas características destacadas de la distribución geométrica son:

- Función de distribución: $F(x) = 1 - (1-p)^{x+1}$
- Esperanza:
$$\mu = \frac{1-p}{p}$$
- Varianza:
$$\sigma^2 = \frac{1-p}{p^2} = \frac{\mu}{p}$$
- Índice de agregación:
$$IA = \frac{1}{p} > 1,$$
luego la distribución geométrica es sobredispersa.
- Coeficiente de asimetría:
$$\gamma_1 = \frac{2-p}{\sqrt{1-p}} > 0,$$
por tanto, la distribución presenta asimetría a la derecha.
- Coeficiente de curtosis:
$$\gamma_2 = \frac{p^2 - 6p + 6}{1-p} + 3,$$
de modo que la distribución es leptocúrtica.

La distribución geométrica se utiliza en la distribución de tiempos de espera, de manera que si las repeticiones se realizan a intervalos regulares de tiempo, esta variable aleatoria proporciona el tiempo que transcurre hasta el primer éxito. Esta distribución es la única discreta que presenta la propiedad denominada "falta de memoria", que implica que la probabilidad de tener que esperar un tiempo no depende del tiempo que ya haya transcurrido. Esta propiedad también la satisface la distribución exponencial (que es continua), de modo que la distribución geométrica se considera la análoga discreta de la exponencial.

1.3.2. Distribución binomial negativa

Una generalización directa de la distribución geométrica surge cuando se considera que el número de éxitos deseados es $k \geq 1$. La variable aleatoria X que mide el número de fracasos hasta obtener k éxitos sigue una distribución binomial negativa de parámetros k y p, $BN(k,p)$ y tiene fmp dada por:

$$P(X=x) = \binom{x+k-1}{x} p^k (1-p)^x, \quad x = 0,1,2,\ldots \tag{1.4}$$

Si en lugar de contar el número de fracasos se cuenta el número de repeticiones del experimento necesarias para conseguir k éxitos, aparece la distribución de Pascal. Así, si Y : Número de repeticiones necesarias para conseguir k éxitos, $Y \sim Pascal(k,p)$ su fmp tiene la expresión

$$P(Y=y) = \binom{y-1}{k-1} p^k (1-p)^{y-k}, \quad y = k, k+1, k+2,\ldots \tag{1.5}$$

En consecuencia, puede establecerse la siguiente relación entre las distribuciones binomial negativa y Pascal:

$$Y \sim Pascal(k,p) \Leftrightarrow Y - k \sim BN(k,p)$$

Las principales propiedades y características de la distribución binomial negativa son:

- Esperanza:
$$\mu = \frac{k(1-p)}{p}$$
- Varianza:
$$\sigma^2 = \frac{k(1-p)}{p^2} = \frac{\mu}{p}$$
- Índice de agregación:
$$IA = \frac{1}{p} > 1,$$
de modo que la distribución binomial negativa también es sobredispersa.
- Coeficiente de asimetría:
$$\gamma_1 = \frac{2-p}{\sqrt{k(1-p)}} > 0,$$
luego la distribución es asimétrica a la derecha.
- Coeficiente de curtosis:
$$\gamma_2 = \frac{p^2 - 6p + 6}{k(1-p)} + 3 > 3,$$
por tanto, la distribución es leptocúrtica.
- Si $k \to \infty, q = (1-p) \to 0$ y $kq \to \lambda$, entonces la distribucíon binomial negativa tiende a una distribución de Poisson.

La distribución binomial negativa puede obtenerse como una mixtura de Poisson con una distribución gamma. Concretamente, si $X|\lambda \sim \mathcal{P}(\lambda)$, $\lambda > 0$ y $\lambda \sim Gamma(\theta, \nu)$, en dónde $\theta, \nu > 0$, entonces $X \sim BN(\theta, p)$ con $p = 1/(1+\nu)$.

Esta distribución es un modelo de distribución contagiosa, en el sentido de que si unos datos de conteo por área siguen una binomial negativa, eso implica que la probabilidad de "éxito" no es la misma en todas las zonas, sino que cuantos más "éxitos" haya, más probable es que se incremente su número (efecto de contagio). En la actualidad, existen muchas aplicaciones de dicha distribución en diferentes áreas de conocimiento. Así, por ejemplo, Steutel et al. (1979) realizan un resumen de los modelos probabilísticos utilizados en Ecología, siendo la binomial negativa el eje de todas ellas; en el ámbito de la medicina, específicamente en el estudio de los datos generados por el SARS-CoV-2, Du et al. (2022) realizan un estudio de meta-análisis centrado en la modelización de estos datos a través de la distribución binomial negativa, así como de todas sus versiones truncadas o infladas, y en Ingeniería agrícola Vagelas and Leontopoulos (2022) modelizan la adhesión de esporas de *Pasteuria penetrans* in vitro. Otras publicaciones en dónde se utiliza este modelo son, por ejemplo, Liu et al. (2023), Li et al. (2023), entre otras.

Para su cálculo, en R está la funciones asociadas a nbinom (pnbinom, qnbinom, dnbinom, rnbinom).

1.3.3. Distribución de Waring generalizada univariante

Se dice que una variable aleatoria X sigue una distribución de Waring generalizada univariante (UGW) de parámetros $a > 0$, $k > 0$ y $\rho > 0$ si su fmp viene dada por:

$$P(X = x) = {}_2F_1(a, k; a + k + \rho; 1)^{-1} \frac{(a)_x (k)_x}{(a + k + \rho)_x} \frac{1}{x!}, \qquad x = 0, 1, 2, \ldots \tag{1.6}$$

donde

$$(\alpha)_r = \frac{\Gamma(\alpha + r)}{\Gamma(\alpha)}, \tag{1.7}$$

con $\alpha > 0$ y $r \in \mathbb{N}$, es el símbolo de Pochhammer (Abramowitz and Stegun, 1972).

El origen de esta distribución es la distribución de Waring clásica que surge del desarrollo de la serie de Waring:

$$\frac{1}{x - a} = \sum_{r=0}^{\infty} \frac{(a)_r}{(x)_{r+1}} \tag{1.8}$$

Realizando el cambio de variable $\rho = x - a$, Irwin (1963) obtuvo la fmp de la distribución relaciona con (1.8), cuya expresión es:

$$P(X = r) = \rho \frac{(a)_r}{(a + \rho)_{r+1}}, \quad r = 0, 1, \ldots,$$

siendo $a, \rho > 0$. La serie (1.8) puede generalizarse de la forma:

$$\frac{1}{(x - a)_k} = \sum_{r=0}^{\infty} \frac{(a)_r (k)_r}{(x)_{r+k}} \frac{1}{r!}, \tag{1.9}$$

con $k > 0$, de modo que considerando de nuevo $\rho = x - a$ se obtiene la correspondiente fmp

$$P(X = r) = \frac{(\rho)_k}{(a + \rho)_k} \frac{(a)_r (k)_r}{(a + k + \rho)_r} \frac{1}{r!}, \quad r = 0, 1, \ldots. \tag{1.10}$$

Esta expresión coincide con la de la fmp de la distribución $UGW(a,k,\rho)$ en (1.6) con $x=r$.

La función generatriz de probabilidad de la distribución UGW está dada por

$$G(t)=\frac{{}_2F_1(a,k;a+k+\rho;t)}{{}_2F_1(a,k;a+k+\rho;1)},\quad t\in\mathbb{R} \tag{1.11}$$

siendo ${}_2F_1(\alpha,\beta;\gamma;\lambda)$ la función hipergeométrica de Gauss. Esta serie de potencias converge cuando $|\lambda|<1$ y no está definida cuando γ es un entero negativo (Abramowitz and Stegun, 1972). Además, al poder definirse de esta forma, se dice que la distribución UGW pertenece a la familia de *distribuciones hipergeométricas gaussianas*, GHD (Johnson et al., 2005).

$${}_2F_1(\alpha,\beta;\gamma;\lambda)=\sum_{r=1}^{\infty}\frac{(\alpha)_r(\beta)_r}{(\gamma)_r}\frac{\lambda^r}{r!}. \tag{1.12}$$

Cuando $\lambda=1$, $\alpha,\beta,\gamma\notin\mathbb{Z}^-$ y $Re(\gamma-\alpha-\beta)>0$, el Teorema de Sumación de Gauss (Slater, 1966) permite calcular explícitamente el valor de esta suma en términos de la función gamma como

$${}_2F_1(\alpha,\beta;\gamma;1)=\frac{\Gamma(\gamma-\alpha-\beta)\Gamma(\gamma)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}, \tag{1.13}$$

de modo que otra expresión alternativa de (1.6) es:

$$P(X=x)=\frac{\Gamma(a+\rho)\Gamma(k+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)}\frac{(a)_x(k)_x}{(a+k+\rho)_x}\frac{1}{x!},\qquad x=0,1,2,\ldots \tag{1.14}$$

O equivalentemente, en términos de la función $\Gamma()$ y utilizando (1.7) y (1.13):

$$P(X=x)=\frac{\Gamma(\rho+a)\Gamma(\rho+k)\Gamma(a+x)\Gamma(k+x)}{\Gamma(a)\Gamma(k)\Gamma(\rho)\Gamma(a+k+\rho+x)\Gamma(x+1)},\qquad x=0,1,2,\ldots \tag{1.15}$$

Otra forma de obtener la distribución UGW es como mixtura en dos pasos de una distribución de Poisson. Específicamente, si

- $X|\Lambda=\lambda\sim\mathcal{P}(\lambda)$,
- $\Lambda|P=p\sim Gamma(a,v)$ con $v=(1-p)/p$ cuya función de densidad es

$$\frac{l^{a-1}e^{-l/v}}{\Gamma(a)v^a},\quad l>0$$

- $P\sim BetaI(\rho,k)$ con función de densidad

$$\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}p^{\rho-1}(1-p)^{k-1},\quad 0<p<1$$

entonces, $X\sim UGW(a,k,\rho)$.

Las propiedades de la distribución UGW han sido estudiadas en profundidad en numerosos artículos (Irwin, 1968a,b,c,d; Xelakaki, 1983b,a; Rodríguez-Avi et al., 2007). Destacamos aquí las principales:

- Esperanza:

$$\mu=\frac{ak}{\rho-1}, \tag{1.16}$$

que existe si $\rho>1$. En general, $E\left(X^k\right)<\infty$ si y solo si $\rho>k$

- Varianza:

$$\sigma^2 = \frac{ak(\rho+a-1)(\rho+k-1)}{(\rho-1)^2(\rho-2)} = \mu\frac{(\rho+a-1)(\rho+k-1)}{(\rho-1)(\rho-2)} \tag{1.17}$$

que existe si $\rho > 2$.

- Al tratarse de una mixtura de Poisson, es una distribución sobredispersa. Como consecuencia, la cola de esta distribución puede ser larga, fenómeno conocido como *efecto de cola pesada*. Además

$$IA = \frac{(\rho+a-1)(\rho+k-1)}{(\rho-1)(\rho-2)}.$$

- Como consecuencia de la obtención de la UGW como una mixtura en dos pasos de una Poisson, la varianza puede expresarse como suma de tres componentes (Irwin, 1968a):

$$\sigma^2 = \frac{ak}{\rho-1} + \frac{ak(k+1)}{(\rho-1)(\rho-2)} + \frac{a^2k(\rho+k-1)}{(\rho-1)^2(\rho-2)} \tag{1.18}$$

El primer término está relacionado con la *aleatoriedad*, el segundo con la variabilidad externa que afecta a la población (*riesgo*) y el último se debe a las diferencias en las condiciones internas de los individuos de la población (*predisposición*). Hay que resaltar que para la existencia de esta partición $\rho > 2$, puesto que la varianza debe ser finita.

Uno de los principales inconvenientes que presenta la distribución UGW es la intercambiabilidad de los parámetros a y k. Esto supone que las componentes de la varianza en (1.18) no quedan claramente identificadas, esto es, no se sabe qué parte corresponde al riesgo y cuál a la predisposición. Irwin (1968a), Xekalaki (1984) o Rodríguez-Avi et al. (2009) propusieron algunas soluciones parciales a dicho problema. En el Capítulo **??** de esta memoria se propone una solución alternativa al mismo.

Existen aplicaciones recientes de este modelo de probabilidad. Así por ejemplo, Huete-Morales and Marmolejo-Martín (2020) lo utilizan para la modelización del número de granjas ecológicas en Andalucía, mientras que Ariza-López and Rodríguez-Avi (2015) lo emplean para determinar el número de errores de completitud en conjuntos de datos geográficos. Otras aplicaciones recientes aparecen en Mitov and Nadarajah (2023), Gning et al. (2022) o Rivas and Galea (2020), entre otros.

Desde el punto de vista computacional, la librería `GWRM` de R permite ajustar esta distribución (Vílchez-López et al., 2016; Sáez-Castillo et al., 2021).

1.4. Distribuciones infra y sobredispersas

Existen distribuciones cuyas características hacen posible la modelización tanto de datos sobredispersos ($\mu > \sigma^2$) como infradispersos ($\mu < \sigma^2$).

1.4.1. Distribución de Poisson generalizada

Se dice que X sigue una distribución de Poisson generalizada con parámetros θ y λ, $X \sim GP(\theta, \lambda)$, si

$$P(X = x) = \frac{\theta(\theta+\lambda x)^{x-1}}{x!}, \quad x = 0, 1, \ldots$$

cuando $\lambda > 0$. Si $\lambda < 0$, $P(X = x) = 0$ para valores de $x > m$, siendo $m \geq 4$ el mayor entero positivo tal que $\theta + m\lambda > 0 \Leftrightarrow \lambda > -\theta/m$. Esta cota inferior para el parámetro λ se impone para que haya al menos 5 puntos en el espacio muestral con probabilidades positivas. Como consecuencia de esta definición, cuando $\lambda < 0$ la distribución GP tiene rango finito (desde 0 hasta m) y sus probabilidades no suman 1, con lo que deben ser normalizadas, si bien este error es inferior al 5 %, lo que no supone grandes diferencias en las aplicaciones prácticas (Consul and Shoukri, 1985; Consul and Famoye, 1989).

Cuando $\lambda = 0$, la distribución GP coincide con la distribución de Poisson de parámetro θ.

Consul and Jain (973a,b) definieron, estudiaron y analizaron algunas aplicaciones de la distribución GP, si bien las propiedades y aplicaciones de este modelo se analizan en profundidad en el libro de Consul (1989).

A continuación enumeramos las principales características de esta distribución:

- Esperanza:
$$\mu = \frac{\theta}{1-\lambda}$$
- Varianza:
$$\sigma^2 = \frac{\theta}{(1-\lambda)^3} = \frac{\mu}{(1-\lambda)^2}$$
- Índice de agregación:
$$IA = \frac{1}{(1-\lambda)^2}.$$
Por tanto, si $\lambda < 0$ la distribución es infradispersa, si $\lambda = 0$ es equidispersa (coincide con la Poisson) y si $\lambda < 0$ la distribución es infradispersa.
- Coeficiente de asimetría:
$$\gamma_1 = \frac{1+2\lambda}{\sqrt{\theta(1-\lambda)}}$$
- Coeficiente de curtosis:
$$\gamma_2 = 3 + \frac{1+8\lambda+6\lambda^2}{\theta(1-\lambda)}$$
- Función generatriz de probabilidad:
$$G(t) = e^{-\frac{\lambda}{\theta}[W(-\theta t e^{-\theta})+\theta]}$$
donde W es la *función de Lambert* definida como
$$z = W(z)e^{W(z)}, \quad z \in \mathbb{C}.$$

Este modelo ha sido aplicado en casi todos las áreas de estudio, sobre todo el modelo de regresión. Recientemente, Ser (2022) emplean esta distribución para modelizar la variablidad extrema en los experimentos de germinación sobredispersos.

Computacionalmente, las funciones asociadas a esta distribución (fmp, función de distribución y generación de números aleatorios) están implementadas en la librería `HMMpa` de R (Witowski and Foraita, 2018). La librería `VGAM` (Yee, 2015) también permite la estimación de los parámetros del modelo de regresión basado en la distribución generalizada de Poisson utilizando las funciones `genpoisson0()` para una $GP(\theta, \lambda)$ o `genpoisson1()` para una $GP(\mu, \phi)$, siendo ϕ el parámetro de dispersión (Yang et al., 2009).

1.4.2. Distribución de Poisson ponderada

A partir del concepto de distribución ponderada introducido por Rao (1965), se han obtenido distintas familias de distribuciones ponderadas. La más estudiada ha sido aquella en la que la variable aleatoria de interés, X, tiene como distribución subyacente la distribución de Poisson.

En general, una variable aleatoria X se distribuye según una distribución de Poisson ponderada si su fmp se escribe de la forma

$$P(X = x) = \frac{e^{-\lambda}\lambda^x w_x}{Wx!}, \quad x = 0, 1, 2, \ldots; \quad \lambda > 0 \tag{1.19}$$

donde $W = \sum_{s=0}^{\infty} e^{-\lambda}\lambda^s w_s/s!$ es una constante de normalización.

En general, los pesos son proporcionales a x o a x elevado a un cierto valor. Castillo and Pérez-Casany (1998) utilizan los pesos

$$w_x = (x + a)^r, \quad a > 0$$

y denotan la distribución $WP(\lambda, r, a)$. Además, estudian sus principales propiedades utilizando la constante de normalización, $C(\lambda, r, a)$, definida como[1]

$$C(\lambda, r, a) = e^{\lambda} E_{\lambda}\left[(X + a)^r\right] = \sum_{k=0}^{\infty} \frac{\lambda^k (k + a)^r}{k!}$$

y que resumimos en:

- Esperanza:
$$\mu = \lambda \frac{C(\lambda, r, a + 1)}{C(\lambda, r, a)}$$
- Varianza:
$$\sigma^2 = \frac{C(\lambda, r + 2, a)C(\lambda, r, a) - C^2(\lambda, r + 1, a)}{C^2(\lambda, r, a)}$$
- Índice de agregación:
$$IA = \frac{C(\lambda, r + 2, a)C(\lambda, r, a) - C^2(\lambda, r + 1, a)}{C(\lambda, r, a)C(\lambda, r + 1, a) - aC^2(\lambda, r, a)}$$

  de modo que la distribución WP es infradispersa sii $r > 0$, equidispersa sii $r = 0$ y sobredispersa sii $r < 0$. Así, el parámetro r del modelo se interpreta como un *parámetro de repulsión*. Por su parte, el parámetro a es considerado una *medida de aproximación* a la distribución de Poisson.

Posteriormente, los mismos autores extienden esta distribución considerando pesos exponenciales (Castillo and Pérez-Casany, 2005)

$$w_x = e^{rt(x)}$$

obteniendo una familia exponencial biparamétrica con parámetros $\theta, r \in \mathbb{R}$ y donde $(x, t(x))$ es un estadístico suficiente. La distribución WP aparece tomando $t(x) = \ln(x + a)$. Estos

[1] Esta serie converge si $\lambda > 0, a > 0$ y $r \in \mathbb{R}$ o si $\lambda > 0, a = 0$ y $r \in \mathbb{R}^+$. Para valores enteros positivos de r esta constante tiene una expresión explícita; en caso contrario, debe calcularse por métodos numéricos.

autores demuestran que si t es una función convexa y $r \neq 0$, la distribución es infradispersa si $r > 0$, equidispersa si $r = 0$ (se obtiene la distribución de Poisson de parámetro e^{θ}) y sobredispersa si $r < 0$.

Cameron and Johansson (1997) utilizan pesos polinomiales de la forma

$$w_x = \left(1 + \sum_{j=1}^{p} \alpha_j x^j\right), \quad \alpha_j \in \mathbb{R}$$

(los cuadrados se emplean para evitar valores negativos) obteniendo así la distribución polinomial de Poisson de orden p, $PPp(\lambda, \alpha_1, \ldots, \alpha_p)$.

Así, por ejemplo para una distribución $PP2(\lambda, \alpha_1, \alpha_2)$ se tienen las siguientes propiedades:

- Esperanza:

$$\mu = \frac{m_1 + 2\alpha_1 m_2 + (\alpha_1^2 + 2\alpha_2)m_3 + 2\alpha_1\alpha_2 m_4 + \alpha_2^2 m_5}{\eta_2(\lambda, \alpha_1, \alpha_2)}$$

- Varianza:

$$\sigma^2 = \frac{m_2 + 2\alpha_1 m_3 + (\alpha_1^2 + 2\alpha_2)m_4 + 2\alpha_1\alpha_2 m_5 + \alpha_2^2 m_6}{\eta_2(\lambda, \alpha_1, \alpha_2)}$$

  donde m_i representa el momento no centrado de orden i de una variable aleatoria de Poisson con parámetro λ y $\eta_2(\lambda, \alpha_1, \alpha_2) = 1 + 2\alpha_1 m_1 + (\alpha_1^2 + 2\alpha_2)m_2 + 2\alpha_1\alpha_2 m_3 + \alpha_2^2 m_4$.

- Índice de agregación:

$$IA = \frac{m_1 + 2\alpha_1 m_2 + (\alpha_1^2 + 2\alpha_2)m_3 + 2\alpha_1\alpha_2 m_4 + \alpha_2^2 m_5}{m_2 + 2\alpha_1 m_3 + (\alpha_1^2 + 2\alpha_2)m_4 + 2\alpha_1\alpha_2 m_5 + \alpha_2^2 m_6}$$

  que puede ser menor o mayor que 1, dependiendo de los valores de los parámetros, permitiendo infra y sobredispersión.

Por su parte, Ridout and Besbeas (2004) consideran los pesos

$$w_x = \begin{cases} e^{-\beta_1(\lambda - k)} & \text{si } k \leq \lambda \\ e^{-\beta_2(k - \lambda)} & \text{si } k > \lambda \end{cases}$$

que, para valores positivos de β_1 y β_2 "empujan" las probabilidades de Poisson hacia la media. La distribución obtenida se denomina Poisson ponderada exponencialmente triparamétrica, $EWP_3(\lambda, \beta_1, \beta_2)$. Si $\beta_1 = \beta_2 = \beta$ se obtiene la distribución de Poisson ponderada exponencialmente biparamétrica, $EWP_2(\lambda, \beta)$, con pesos

$$w_x = e^{-\beta|k - \lambda|}.$$

Estas distribuciones son infradispersas cuando $\beta_1, \beta_2 > 0$ ($\beta > 0$), equidispersas cuando $\beta_1 = \beta_2 = \beta = 0$ (ya que se reducen a la distribución de Poisson) y sobredispersas cuando $\beta_1, \beta_2 < 0$ ($\beta < 0$). Sin embargo, no existen expresiones explícitas para los momentos de las distribuciones EPW, sino que estos han de calcularse numéricamente truncando la suma que determina la constante de normalización W para un valor suficientemente elevado.

Recientemente Balakrishnan et al. (2018) han desarrollado otro modelo perteneciente a la familia de distribuciones de Poisson ponderadas utilizando la función peso $w_x(\phi, v) =$

$1-\phi p(x;\lambda)^v$, con $\phi \leq 1, v \geq 0$ y $p(x;\lambda)$ la fmp de la Poisson de parámetro λ, y la emplean en la modelización de la tasa de curación en escenarios con diferentes causas. Berger et al. (2022) aplican el modelo de regresión basado en una distribución Poisson ponderada a la inseguridad alimentaria asociada con interrupciones educativas durante la pandemia de COVID-19. Otras aplicaciones de estos modelos pueden verse, por ejemplo, en Tang et al. (2020), Pagliara and Mauriello (2020), Tyas et al. (2023) o Esmailian et al. (2023), entre otros.

1.4.3. Distribución Conway-Maxwell-Poisson

La distribución Conway-Maxwell-Poisson, también conocida como CMP o $COM-$Poisson, fue propuesta por Conway and Maxwell (1962) como la solución a sistemas de colas con razón de servicio dependiente del estado. Sin embargo, el desarrollo de las propiedades estadísticas de la distribución y su uso en aplicaciones más generales se debe a Shmueli et al. (2005) y Sellers et al. (2011).

Se dice que una variable X sigue una distribución CMP con parámetros $\lambda > 0$ y $v \geq 0$, $CMP(\lambda, v)$, si su fmp tiene la expresión:

$$P(X = x) = \frac{\lambda^x}{(x!)^v Z(\lambda, v)}, \quad x = 0, 1, 2, \ldots \tag{1.20}$$

siendo $Z(\lambda, v) = \sum_{s=0}^{\infty} \lambda^s/(s!)^v$ una constante de normalización.

Esta distribución puede considerarse como una generalización, no sólo de la distribución de Poisson, $(v = 1)$, sino también de la distribución Geométrica, $(v = 0, \lambda < 1)$, y de la distribución de Bernouilli $(v \to \infty$, con $p = \lambda/(1 + \lambda))$.

La distribución CMP no tiene forma explícita de los momentos, pero sí recursiva:

$$E\left(X^{r+1}\right) = \begin{cases} \lambda E(X+1)^{1-\nu} & r = 0 \\ \lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} E\left(X^r\right) + E(X)E\left(X^r\right) & r > 0 \end{cases}$$

Shmueli et al. (2005) establecieron una aproximación asintótica para $Z(\lambda, v)$ especialmente precisa cuando $\nu \leqslant 1$ o $\lambda > 10^{\nu}$, que permite calcular los momentos de forma aproximada. Así, las principales características de esta distribución son:

- Esperanza:
$$\mu = \lambda \frac{\mathrm{d}[\log\{Z(\lambda, \nu)\}]}{\mathrm{d}\lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \tag{1.21}$$

- Varianza:
$$\sigma^2 \approx \frac{1}{v} \lambda^{1/v} \tag{1.22}$$

- Índice de agregación:
$$IA \approx \frac{\frac{1}{v}\lambda^{1/v}}{\lambda^{1/\nu} - \frac{\nu-1}{2\nu}} \tag{1.23}$$

  La distribución es sobredispersa cuando $v < 1$, equidispersa cuando $v = 1$ (en cuyo caso coincide con la Poisson de parámetro θ) e infradispersa cuando $v > 1$.

- Función generatriz de probabilidad:
$$G(t) = \frac{Z(\lambda t, v)}{Z(\lambda, v)}, \quad t \in \mathbb{R}.$$

La distribución CMP también puede verse como una distribución de Poisson ponderada (véase Sección 1.4.2 con pesos $w_x = (x!)^{1-v}$ (Ridout and Besbeas, 2004; Rodrigues et al., 2009). En Sellers (2023) aparece más información sobre esta distribución. Aplicaciones recientes pueden verse en Bedbur and Kamps (2023), Adeoti et al. (2022) o Abdella et al. (2019), entre otros.

En cuanto a su implementación computacional, las principales funciones asociadas a la distribución CMP están incluidas en las librerías de R `COMPoissonReg` (Sellers et al., 2019) y `DGLMExtPois` (Sáez-Castillo et al., 2022; Sáez-Castillo et al., 2023).

1.4.4. Distribución hiper-Poisson

La distribución de Crow-Bardwell o hiper-Poisson de parámetros $\gamma > 0$ y $\lambda > 0$, $hP(\gamma, \lambda)$, fue propuesta por Bardwell and Crow (1964) y su fmp tiene la expresión:

$$P(X = x) = \frac{1}{{}_1F_1(1;\gamma;\lambda)} \frac{\lambda^x}{(\gamma)_x}, \quad x = 0, 1, 2, \ldots, \tag{1.24}$$

siendo $(a)_r$ el símbolo de Pochhammer y

$${}_1F_1(a;c;z) = \sum_{r=0}^{\infty} \frac{(a)_{\mathrm{r}}}{(c)_{\mathrm{r}}} \frac{z^{\mathrm{r}}}{r!}$$

la función hipergeométrica confluente (Johnson et al., 2005). Esta distribución puede considerarse como una distribución de Poisson ponderada con pesos $w_x = e^{\lambda} x!/(\gamma)_x$ (Ridout and Besbeas, 2004). No existen expresiones explícitas de sus momentos, ya que estos están dados en términos de la función ${}_1F_1$. No obstante, las principales características de la distribución HP son:

- Esperanza:

$$\mu = \lambda - (\gamma - 1)(1 - f_0) = \lambda - (\gamma - 1)\frac{{}_1F_1(1;\gamma;\lambda) - 1}{{}_1F_1(1;\gamma;\lambda)} = \frac{\lambda}{\gamma}\frac{{}_1F_1(2;\gamma+1;\lambda)}{{}_1F_1(1;\gamma;\lambda)}$$

- Varianza:

$$\sigma^2 = \lambda + [\lambda - (\gamma - 1)]\,\mu - \mu^2$$

- Índice de agregación:

$$IA = \frac{\lambda}{\mu} + \lambda - (\gamma - 1) - \mu$$

  La distribución HP es infradispersa si $\gamma < 1$, equidispersa si $\gamma = 1$ (ya que coincide con la distribución de Poisson de parámetro λ) y sobredispersa si $\gamma > 1$.

- Función generatriz de probabilidad:

$$G(t) = \frac{{}_1F_1(1;\gamma;\lambda t)}{{}_1F_1(1;\gamma;\lambda)}, \quad t \in \mathbb{R}$$

  En consecuencia, la HP también pertenece a la familia de distribuciones hipergeométricas confluentes (Johnson et al., 2005).

Algunas aplicaciones recientes pueden verse en Satheesh Kumar and Ramachandran (2020), Bogdanov et al. (2020) o Santos et al. (2019), entre otros. Desde un punto de vista computacional, las funciones asociadas a esta distribución están implementadas en la librería de R `DGLMExtPois` (Sáez-Castillo et al., 2023).

1.4.5. Generalización infra y sobredispersa de la distribución geométrica

Gómez-Déniz (2010) propone una generalización biparamétrica de la distribución geométrica, GG, cuya fmp está dada por:

$$P(X=x)=\frac{\alpha\theta^x(1-\theta)}{[1-(1-\alpha)\theta^{x+1}]\,[1-(1-\alpha)\theta^x]},\quad x=0,1,\ldots \tag{1.25}$$

donde $\alpha>0, 0<\theta<1$. Si $\alpha=1$ se obtiene la distribución geométrica de parámetro θ. Cuando $1-\alpha=\theta$ se obtiene la distribución de generaciones en procesos de ramificación subcrítica cuando se comienza con una distribución geométrica de descendientes (Johnson et al., 2005, p. 473).

Algunas propiedades destacables de esta distribución son:

- Para $0<\alpha<1$, la fmp de la distribución GG puede escribirse como una mixtura infinita de distribuciones geométricas.
- Momento no centrado de orden r:

  $$E\left(X^r\right)=\sum_{x=1}^{\infty}\left(x^r-(x-1)^r\right)S(x-1).$$

  con $S(x)=\dfrac{\alpha\theta^{x+1}}{1-(1-\alpha)\theta^{x+1}}$ la función de supervivencia.
- Esperanza:

  $$\mu=\sum_{x=1}^{\infty}\frac{\alpha\theta^x}{1-(1-\alpha)\theta^x}=\alpha\frac{d}{d\alpha}\log{}_0\Phi_0\left(;\frac{1}{\theta},1-\alpha\right),$$

  con ${}_A\Phi_B(;,;)$, la $q-$serie, definida como

  $${}_A\Phi_B\left(a_1,\ldots,a_A;b_1,\ldots,b_B;q,z\right)=\sum_{j=0}^{\infty}\frac{\left(a_1;q\right)_j\ldots\left(a_A;q\right)_j z^j}{\left(b_1;q\right)_j\ldots\left(b_B;q\right)_j}$$

  siendo $(a;q)_\infty=\prod_{k=0}^{\infty}(1-aq^k)$, $\quad 0<q<1$, el símbolo $q-$Pochhammer (Andrews et al., 2013; Johnson et al., 2005). Otra expresión alternativa para la media cuando $0<\alpha<1$ es

  $$\mu=\frac{\alpha}{(1-\alpha)}\sum_{j=0}^{\infty}\frac{[(1-\alpha)\theta]^{j+1}}{1-\theta^{j+1}}.$$

- Varianza:

  $$\sigma^2=\sum_{x=1}^{\infty}\frac{(2x-1)\alpha\theta^x}{1-\bar{\alpha}\theta^x}-\mu^2$$

- Los autores muestran empíricamente que para valores $0<\theta<0.5$ y $\alpha>2$ la distribución es infradispersa y la media crece más rápido que la varianza.
- La mediana de la distribución es:

  $$x_{\text{med}}=\left[-1-\frac{\ln(1+\alpha)}{\ln\theta}\right],$$

  con $[\cdot]$ es la parte entera.

- Si $\alpha < (1+\theta)/\theta^2$, la moda es 0; en caso contrario, la moda está en el punto $[-\log(-1+\alpha)/\log\theta - 1]$, salvo que $-\log(-1+\alpha)/\log\theta - 1$ sea un número entero, en cuyo caso presenta dos modas consecutivas en dicho valor y el siguiente.
- Función generatriz de probabilidad ($0 < \alpha < 1$):

$$G(t) = \alpha \sum_{j=0}^{\infty} (1-\theta)^j \frac{1-\theta^{j+1}}{1-t\theta^{j+1}}, \quad t \in \mathbb{R}$$

1.4.6. La distribución de Poisson infradispersa

Singh et al. (2021) proponen una versión biparamétrica infradispersa de la distribución de Poisson y la denominan distribución de Poisson infradispersa, $UDPD-I$. Así pues, se dice que la variable aleatoria X sigue una distribución $UDPD-I$ con parámetros $\lambda > 0$ y $\theta > 0$ si su fmp es:

$$P(X=x) = \frac{e^{-\lambda}\lambda^{x-1}(\lambda+\theta x)}{(1+\theta)x!} \quad x = 0,1,2,\ldots \tag{1.26}$$

Si $\theta = 0$ se obtiene la distribución de Poisson clásica de parámetro λ.

La distribución $UDPD-I$ puede verse como una distribución de Poisson ponderada con pesos $w_x = 1 + \theta x/\lambda$.

Esta distribución es una mixtura de una distribución de Poisson y una distribución de Poisson sesgada de la forma:

$$p(x) = \alpha p_1(x) + (1-\alpha)p_2(x), \quad x = 0,1,\ldots$$

donde $\alpha = \dfrac{1}{1+\theta}$, $p_1(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $\quad x = 0,1,\ldots$ y $p_2(x) = \dfrac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}$, $\quad x = 1,2,\ldots, \lambda > 0$.

Las principales características de esta distribución se resumen a continuación:

- Esperanza:

$$\mu = \lambda + \frac{\theta}{1+\theta}$$

- Varianza:

$$\sigma^2 = \lambda + \frac{\theta}{(1+\theta)^2}$$

- Coeficiente de asimetría:

$$\gamma_1 = \frac{\lambda(1+\theta)^3 + \theta(1-\theta)}{[\lambda(1+\theta)^2+\theta]^{3/2}}$$

- Coeficiente de curtosis:

$$\gamma_2 = \frac{(1+\theta)^2\left[\lambda(1+\theta)^2+\theta\right] - 6\theta^2}{[\lambda(1+\theta)^2+\theta]^2}$$

- Índice de agregación:

$$IA = \frac{\lambda(1+\theta)^2+\theta}{(1+\theta)[\lambda(1+\theta)+\theta]}$$

  Como $IA < 1$, la distribución $UPDP-I$ es infradispersa.

- Función generatriz de probabilidad:

$$G(t) = \frac{1+\theta t}{1+\theta} e^{\lambda(t-1)}, \quad t \in \mathbb{R}.$$

Los autores no especifican el programa utilizado en sus ejemplos, por lo que para usar dicha distribución se procederá a su implementación en R.

1.5. Otros modelos recientes

En los últimos años se han ido generando nuevos modelos de probabilidad para el tipo de dato en el que se está trabajando. Así, Ahsan-ul Haq and Zafar (2023) desarrollan una nueva distribución para datos aplicados a situaciones reales utilizando el modelo neutrosófico. Erbayram and Akdoğan (2023) proponen el modelo transmutado de Poisson exponencial, que es una mixtura entre la distribución de Poisson y la distribución exponencial transmutada. Acu and Rasa (2023) establecen un modelo de probabilidad basado en los números $a_{n,j}$. Otros modelos son la distribución bimodal desplazada de Poisson (*bimodal shifted Poisson model*) introducida por Gómez-Déniz et al. (2020), un modelo discreto análogo a la distribución de Lindley (Al-Babtain et al., 2020) o una generalización de la distribución hipergeométrica del tipo Conway-Maxwell-Poisson (Roy et al., 2023).

Todos estos ejemplos y otros más demuestran que el tema de generación de modelos discretos, su interpretación y su aplicación a datos de conteo es de total actualidad. Dentro de esta línea de investigación es donde se encuadra el trabajo que presentamos.

1.6. Objetivos y desarrollo de la memoria

Como hemos comentado, el principal objetivo de esta memoria es la descripción de un modelo de distribución discreta para datos de conteo, la interpretación y estimación de sus parámetros, la comparación con otros modelos similares y la explicación de las ventajas que proporciona su uso respecto a otros modelos, lo que permite justificar su estudio. Por último, se demuestra la versatilidad de esta distribución en la modelización de diferentes conjuntos de datos reales.

Para ello, esta memoria se estructura en 6 capítulos. En el primero, se describen los modelos de conteo más utilizados y se realiza un resumen del estado del arte. En el segundo capítulo se resumen resultados de ecuaciones en diferencias y funciones hipergeométricas, así como otros resultados matemáticos necesarios para el desarrollo del resto de la memoria. Entre ellos, se presenta la distribución CTP en cuyo desarrollo la doctoranda ha colaborado (Olmo-Jiménez et al., 2018). En el tercer capítulo se define el modelo de distribución extendida biparamétrica de Waring, objeto de esta memoria, y avalada por publicaciones de la doctoranda (Cueva-López et al., 2019, 2021). En el Capítulo 4 se realiza una comparación exhaustiva de esta nueva distribución con otros modelos discretos, incluyendo estudios de simulación. El Capítulo 5 está dedicado a la modelización de diversos conjuntos de datos reales mediante este modelo, donde también se incluyen otras investigaciones de la doctoranda (Cueva López et al., 2022). Por último, en el Capítulo 6 se reflejan algunas consideraciones y se desarrollan futuras vías de investigación.

Capítulo 2

Familia de distribuciones de Pearson

2.1. Introducción

Otra vía tradicional para proponer modelos de conteo se basa en el estudio de familias de distribuciones que verifican una ecuación funcional, diferencial o en diferencias. El primer paso lo dio Pearson, en 1895, para el caso de distribuciones continuas cuyas funciones de densidad $f(x)$ verifican la ecuación diferencial (Pearson, 1895):

$$f'(x) = \frac{x-a}{b_0 + b_1 x + b_2 x^2} f(x),$$

donde a, b_0, b_1 y b_2 son parámetros reales.

En el mismo trabajo, para el caso discreto, la condición se tradujo a verificar la ecuación en diferencias:

$$\Delta f_{x-1} = \frac{x-a}{b_0 + b_1 x + b_2 x^2} f_{x-1}, \quad x \in T \subseteq \mathbb{Z},$$

donde a y b_i son parámetros reales y f_x es la función masa de probabilidad.

A través de estos dos métodos de generación de distribuciones, se han propuesto múltiples modelos para datos de conteo (Johnson et al., 2005) y su aplicación en diversos campos se incrementa día a dia. Así, por ejemplo, en la base de datos *Scopus* se muestran 4772 entradas referentes a "count data variables" sólo entre 2020 y 2022, y más de 170 en enero de 2023. A continuación procedemos a describir brevemente algunos de los modelos más utilizados y relacionados con el contenido de esta memoria.

2.2. Familia de distribuciones discretas univariantes de Pearson

La familia de distribuciones discretas univariantes de Pearson puede obtenerse como solución de la ecuación en diferencias

$$G(x) f_{x+1} - L(x) f_x = 0, \quad x \in \mathbb{Z}^+ \tag{2.1}$$

donde $L : \mathbb{Z}^+ \longrightarrow \mathbb{R}$ y $G : \mathbb{Z}^+ \longrightarrow \mathbb{R} \setminus \{0\}$ son funciones, en principio, cualesquiera. Esta ecuación en diferencias propuesta por Fajardo Caldera (1985) generaliza la familia expuesta

por Ord (1972). La solución de dicha ecuación (Guelfond, 1963; Jordan, 1965) viene dada por:

$$f_x = \begin{cases} f_0 \prod_{t=0}^{x-1} \frac{L(t)}{G(t)} & x \geq 1 \\ f_0 & x = 0 \end{cases} \tag{2.2}$$

donde f_0 es una constante no nula.

El siguiente teorema nos indica qué condiciones debe verificar una solución de (2.2) para que sea, verdaderamente, una función masa de probabilidad.

Teorema 2.2.1. *Dado el conjunto $\mathcal{H} = \{x \in \mathbb{Z}^+; L(x) = 0\}$, una condición necesaria y suficiente para que la función $f : \mathbb{Z}^+ \longrightarrow \mathbb{R}$, solución de la ecuación en diferencias (2.1), sea una función masa de probabilidad, es que verifique las siguientes condiciones:*

(i) Condición de Positividad

$$\begin{array}{ll} L(x)\,G(x) > 0, & \forall x \in \mathbb{Z}^+ \qquad \text{si } \mathcal{H} = \emptyset \\ L(x)\,G(x) \geq 0, & x = 0, 1, \ldots, m = \min \mathcal{H} \quad \text{si } \mathcal{H} \neq \emptyset \end{array}$$

(ii) Condición de Convergencia

$$\sum_{x=1}^{\infty} \prod_{t=0}^{x-1} \frac{L(t)}{G(t)} < \infty$$

(iii) Condición de Normalización

$$f_0 = \frac{1}{1 + \sum_{x=1}^{\infty} \prod_{t=0}^{x-1} \frac{L(t)}{G(t)}}$$

Nota 2.2.1. *Estas condiciones imponen ciertas restricciones sobre las funciones L y G.*

A partir de (2.1) pueden obtenerse las probabilidades de la distribución de forma recursiva como se indica

$$P(X = x+1) = f_{x+1} = \frac{L(x)}{G(x)} f_x = \frac{L(x)}{G(x)} P[X = x]$$

de donde el cociente entre probabilidades es

$$\frac{f_{x+1}}{f_x} = \frac{L(x)}{G(x)} \tag{2.3}$$

En lo que sigue nos centramos en el caso particular en el que L y G sean polinomios, de forma que la función generatriz de probabilidad asociada se caracteriza en términos de una ecuación diferencial.

Teorema 2.2.2. *Sean las funciones L y G polinomios en x y $x+1$, respectivamente, de cualquier orden, y supongamos que se verifican las condiciones dadas en el Teorema 2.2.1. Entonces la función generatriz de probabilidad asociada a la solución (2.2), $g(t)$, verifica la ecuación diferencial*

$$G(\theta)\,g(t) - tL(\theta)\,g(t) = b_0 f_0 \tag{2.4}$$

para $|t| < 1$, siendo $\theta = t\dfrac{d}{dt}$ con $\theta^0 = 1$, el operador identidad, y b_0 el término independiente del polinomio G expresado en $x+1$.

Análogamente, la función generatriz de momentos, función característica y función generatriz de cumulantes satisfacen una ecuación diferencial similar a (2.4) sin más que realizar un cambio de variable.

Corolario 2.2.1. *Bajo las condiciones del Teorema 2.2.1,*

(*i*) *la función generatriz de momentos,* $M(t)$*, verifica la ecuación diferencial*

$$G(D)M(t) - e^t L(D)M(t) = b_0 f_0 \tag{2.5}$$

con $D = \dfrac{d}{dt}$,

(*ii*) *la función característica,* $\phi(t)$*, verifica la ecuación diferencial*

$$G(\theta_i)\phi(t) - e^{it} L(\theta_i)\phi(t) = b_0 f_0 \tag{2.6}$$

con $\theta_i = \dfrac{1}{i}D$,

(*iii*) *la función generatriz de cumulantes,* $k(t) = \ln \phi(t)$*, verifica la ecuación diferencial*

$$G(\theta_i)e^{k(t)} - e^{it} L(\theta_i)e^{k(t)} = b_0 f_0 \tag{2.7}$$

Cualquiera de estas funciones generatrices nos permite obtener los momentos no centrados de la distribución derivando sucesivamente y tomando $t = 1$ (en la función generatriz de probabilidad) o $t = 0$ (en las restantes), siempre que existan las correspondientes derivadas y sean finitas.

Teorema 2.2.3. *En las condiciones del Teorema 2.2.2 para* L *y* G*, y suponiendo que los momentos no centrados de orden* k *existen, con* $k \geq q$ *(*q *orden del polinomio* G*), se verifica la siguiente relación de recurrencia entre momentos,*

$$\sum_{j=0}^{q} b_j \mu'_{j+h} - \sum_{i=0}^{p} a_i \left(\sum_{m=0}^{h} \binom{h}{m} \mu'_{i+m} \right) = b_0 \theta^h f_0 \tag{2.8}$$

para $h = 0, 1, \ldots, k - q$ *y* a_i, b_j *coeficientes de los polinomios* L *y* G *respectivamente.*

Observemos que el segundo miembro de la ecuación (2.8) será 0 si y sólo si $b_0 = 0$, en cuyo caso la relación será únicamente de momentos, mientras que si $b_0 \neq 0$ se incluye también la probabilidad del valor 0, f_0.

La relevancia de este resultado radica en su utilización para la estimación de los parámetros de la distribución perteneciente a esta familia utilizando el método de los momentos. Este método no proporciona estimadores con alta eficiencia, si bien su uso está bastante extendido por su simplicidad frente a otros métodos tales como el de máxima verosimilitud que supone un gran número de cálculos y, además, de elevada complejidad. Así pues, para la aplicación del método de los momentos, se consideran tantas ecuaciones como parámetros, de modo que se obtenga un sistema compatible y determinado en el que se sustituyen los momentos poblacionales, siempre que éstos existan y sean finitos, por los muestrales.

Si en la ecuación en diferencias (2.1) se consideran como funciones L y G sendos polinomios de grados q y p, respectivamente, ambos con raíces reales y donde una de las raíces de G

Distribuciones	fgp
Binomial	$C \cdot {}_1F_0(-n;;\lambda t), \quad \lambda = -p/(1-p)$
Poisson	$C \cdot \lim_{\substack{n\to\infty \\ p\to 0}} {}_1F_0(-n;;\lambda t), \quad \lambda = -p/(1-p)$
Binomial negativa	$C \cdot {}_1F_0(k;;(1-p)t)$
Hiper-Poisson	$C \cdot {}_1F_1(1;\gamma;\lambda t)$
Hipergeométrica	$C \cdot {}_2F_1(-n,-Np;N-Np-n+1;t)$
Hipergeométrica negativa	$C \cdot {}_2F_1(-n,Np;N-Np-n+1;t)$
Beta-Pascal	$C \cdot {}_2F_1(-n,Np;Np-N-n+1;t)$
Pòlya	$C \cdot {}_2F_1(-n,p/c;-b/c-n+1;t)$
Beta-Binomial	$C \cdot {}_2F_1(-n,\alpha;\beta;t)$
Waring	$C \cdot {}_2F_1(1,k;k+\rho+1;t)$
Waring generalizada	$C \cdot {}_2F_1(a,k;k+\rho+a;t)$
CBP	$C \cdot {}_2F_1(bi,-bi;\gamma;t)$
CTP	$C \cdot {}_2F_1(a+bi,a-bi;\gamma;t)$

Tabla 2.1: Distribuciones discretas pertenecientes a la familia de distribuciones hipergeométricas generalizadas (C es la correspondiente constante de normalización)

es igual a -1, en virtud de (2.2) la solución de dicha ecuación viene dada en términos de la función hipergeométrica generalizada ${}_qF_p$

$$ {}_qF_p\left(\alpha_1,\ldots,\alpha_q;\gamma_1,\ldots,\gamma_p;\lambda\right) = \sum_{r=0}^{\infty} \frac{(\alpha_1)_r \cdots (\alpha_q)_r}{(\gamma_1)_r \cdots (\gamma_p)_r} \frac{\lambda^r}{r!}, \tag{2.9} $$

con α_i las raíces reales del polinomio L y γ_j las restantes raíces reales de G. Por su parte, $(a)_r = a(a+1)\cdots(a+r-1)$, $a \in \mathbb{C}$ y $r \in \mathbb{N}$, es el *factorial ascendente*, también conocido como *símbolo de Pochhammer* cuando $a \notin \mathbb{Z}^-$, en cuyo caso puede calcularse equivalentemente como $\Gamma(a+r)/\Gamma(a)$, con $\Gamma(\cdot)$ la función gamma (Abramowitz and Stegun, 1972). Si $a \in \mathbb{Z}^-$, $(a)_r = 0, r = n+1, n+2, \ldots$ Además, cuando $r = 0$ se toma $(a)_0 = 1$.

Notemos que esta serie es finita cuando alguna de las raices α_i es un entero negativo ya que, como hemos indicado, el factorial ascendente correspondiente se anula. Cuando dicha serie es infinita:

- converge para $|x| < \infty$ si $q \leq p$,
- converge para $|x| < 1$ si $q = p+1$, y
- diverge para todo $x \neq 0$ si $q > p+1$.

Además, si llamamos

$$ w = \sum_{i=1}^{p} b_i - \sum_{i=1}^{q} a_i $$

se tiene que para $|x| = 1$, la serie es

- absolutamente convergente si $w > 0$ y $q = p+1$,
- condicionalmente convergente si $-1 < w \leq 0$ y $x \neq 1$, y
- divergente si $w \leq -1$.

Las distribuciones así generadas se conocen como *distribuciones de probabilidad hipergeométricas generalizadas* (Johnson et al., 2005) y pueden verse como distribuciones en series de

potencias. A esta familia pertenecen algunas de las distribuciones discretas más conocidas tal y como se recoge en la Tabla 2.1.

Estudios de estas distribuciones con parámetros reales son los realizados por Bowman et al. (1991), Gutiérrez-Jaimez and Rodríguez-Avi (1997) cuando $p = 2$ y $q = 3$, Rodríguez Avi et al. (1999) cuando $p = 3$ y $q = 4$ y Rodríguez-Avi et al. (2001) cuando $q = p + 1$, en general, entre otros. En la siguiente sección de este capítulo nos centramos en el caso $p = 1$ y $q = 2$ por su especial relación con el núcleo de esta memoria.

2.3. Familia de distribuciones hipergeométricas gaussianas

El caso más estudiado es cuando los polinomios $L(x)$ y $G(x)$ son cuadráticos de la forma:

$$L(x) = \lambda(\alpha + x)(\beta + x) \tag{2.10}$$

$$G(x) = (\gamma + x)(x + 1) \tag{2.11}$$

donde α, β y γ son números reales cualesquiera tales que la función f_x sea una verdadera fmp. Así, la solución de (2.1) es

$$f_x = P(X = x) = f_0 \frac{(\alpha)_x (\beta)_x}{(\gamma)_x} \frac{\lambda^x}{x!}, \quad x = 0, 1, \ldots \tag{2.12}$$

con $f_0 = {}_2F_1(\alpha, \beta; \gamma; \lambda)^{-1}$ la constante de normalización.

Las distribuciones de esta familia están generadas por la función hipergeométrica de Gauss ${}_2F_1(\alpha, \beta; \gamma; \lambda)$, ya que su fgp está dada por

$$G(t) = \frac{{}_2F_1(\alpha, \beta; \gamma; \lambda t)}{{}_2F_1(\alpha, \beta; \gamma; \lambda)}, \quad t \in \mathbb{R}.$$

y se conocen como distribuciones hipergeométricas Gaussianas, GHD (Kemp and Kemp, 1975; Johnson et al., 2005).

Para obtener valores explícitos de la fmp es necesario disponer de un resultado de sumación de la función ${}_2F_1(\alpha, \beta; \gamma; \lambda)$. El único resultado general es el teorema de sumación de Gauss cuando $\lambda = 1$

$${}_2F_1(\alpha, \beta; \gamma; 1) = \frac{\Gamma(\gamma - \alpha - \beta)\Gamma(\gamma)}{\Gamma(\gamma - \alpha)\Gamma(\gamma - \beta)}. \tag{2.13}$$

En este caso, la fmp dada en (2.12) puede reescribirse como:

$$f_x = P(X = x) = \frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\beta){}_2F_1(\alpha, \beta; \gamma; \lambda)} \frac{\Gamma(\alpha + x)\Gamma(\beta + x)}{\Gamma(\gamma + x)} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots. \tag{2.14}$$

A partir de (2.8) se deduce que los momentos no centrados μ'_i de las distribuciones de la familia GHD satisfacen la siguiente relación de recurrencia:

$$\mu'_{h+2} + b_1 \mu'_{h+1} = \sum_{m=0}^{h} \binom{h}{m} \left[a_2 \mu'_{m+2} + a_1 \mu'_{m+1} + a_0 \mu'_m \right] \tag{2.15}$$

donde a_i y b_i son los coeficientes de los polinomios $L(x) = a_0 + a_1 x + a_2 x^2$ y $G(x + 1) = b_0 + b_1(x + 1) + b_2(x + 1)^2$, respectivamente, es decir,

$$a_0 = \lambda\alpha\beta$$
$$a_1 = \lambda(\alpha + \beta)$$
$$a_2 = \lambda$$

y

$$\begin{aligned} b_0 &= 0 \\ b_1 &= \gamma - 1 \\ b_2 &= 1 \end{aligned}$$

Así, por ejemplo, para $h = 0, 1, 2$ se obtienen las relaciones:

$$\begin{aligned} \mu_2' (1 - a_2) - \mu (b_1 - a_1) - a_0 &= 0 \\ \mu_3' (1 - a_2) - \mu_2' (b_1 - a_1 - a_2) - \mu (a_1 + a_0) - a_0 &= 0 \\ \mu_4' (1 - a_2) - \mu_3' (b_1 - a_1 - 2a_2) - \mu_2' (2a_1 + a_0 + a_2) - \mu (a_1 + 2a_0) - a_0 &= 0 \end{aligned}$$

Si particularizamos al caso $\lambda = 1$ el coeficiente a_2 es igual a 1, de forma que las relaciones anteriores se expresan en términos de los parámetros de la función ${}_2F_1$ y los tres primeros momentos no centrados:

$$\begin{aligned} \mu (b_1 - a_1) - a_0 &= 0 \\ \mu_2' (b_1 - a_1 - 1) - \mu (a_1 + a_0) - a_0 &= 0 \\ \mu_3' (b_1 - a_1 - 2) - \mu_2' (2a_1 + a_0 + 1) - \mu (a_1 + 2a_0) - a_0 &= 0 \end{aligned}$$

Resolviendo el este sistema se calculan los tres primeros momentos no centrados conocidos los parámetros de la distribución. Así,

$$\begin{aligned} \mu &= \frac{a_0}{b_1 - a_1} = \frac{\alpha\beta}{\gamma - \alpha - \beta - 1} \\ \mu_2' &= \frac{\alpha\beta(\alpha\beta + \gamma - 1)}{(\gamma - \alpha - \beta - 1)(\gamma - \alpha - \beta - 2)} \\ \mu_3' &= \frac{\alpha\beta \left\{ \alpha^2\beta^2 + (\gamma - 1)(\gamma + \beta - 1) + \alpha \left[\gamma - 1 + \beta(3\gamma - 4) \right] \right\}}{(\gamma - \alpha - \beta - 1)(\gamma - \alpha - \beta - 2)(\gamma - \alpha - \beta - 3)} \end{aligned} \tag{2.16}$$

de donde se deduce el valor de la varianza:

$$\sigma^2 = \frac{\alpha\beta(\gamma - \alpha - 1)(\gamma - \beta - 1)}{(\gamma - \alpha - \beta - 1)^2(\gamma - \alpha - \beta - 2)} = \frac{\mu(\mu + \gamma - 1)}{\gamma - \alpha - \beta - 2}. \tag{2.17}$$

Sibuya (1979) y Sibuya and Shimizu (1981) realizan un primer análisis de la familia GHD cuando $\lambda = 1$. Rodríguez-Avi et al. (2003b) abordan el problema de la estimación en la familia GHD con $\lambda = 1$ mediante métodos de frecuencias, método de los momentos, métodos mixtos (de frecuencias y momentos) y método de máxima verosimilitud y los comparan entre sí. Rodríguez-Avi et al. (2007a) estudian diversos aspectos probabilísticos relacionados con la familia GHD con $\lambda > 0$ y desarrollan métodos de estimación mixtos basados en momentos y frecuencias, mínima χ^2 y máxima verosimilitud y lo aplican a conjuntos de datos procedentes de diferentes áreas, mostrando así la versatilidad de estos modelos.

La Tabla 2.2, extraída de Rodríguez-Avi et al. (2007), contiene una clasificación exhaustiva de las distribuciones pertenecientes a la familia GHD en términos de sus parámetros, incluyendo las condiciones que estos deben satisfacer (véase Teorema 2.2.1).

Parámetros			Condiciones	Rango	Tipo
$\gamma > 0$	$0 < \lambda \leq 1$	$\alpha, \beta > 0$	$\gamma > \alpha + \beta$ if $\lambda = 1$	$[0, \infty)$	I
		$\alpha, \beta \in \mathbb{C}$, $\alpha = \bar{\beta}$	$\gamma > \alpha + \beta$ if $\lambda = 1$	$[0, \infty)$	II
		$\alpha, \beta < 0$, $\alpha, \beta \notin \mathbb{Z}^-$	$[\alpha] = [\beta]$	$[0, \infty)$	III
	$0 < \lambda$	$\alpha, \beta < 0$, $\alpha \in \mathbb{Z}^-$	$\lvert\beta\rvert > \lvert\alpha\rvert - 1$	$[0, \lvert\alpha\rvert)$	IV
	$\lambda < 0$	$\alpha \in \mathbb{Z}^-, \beta > 0$		$[0, \lvert\alpha\rvert)$	V
$\gamma < 0$	$0 < \lambda \leq 1$	$\alpha < 0, \alpha \notin \mathbb{Z}^-$, $\beta > 0$	$[\alpha] = [\gamma]$, $\gamma > \alpha + \beta$ if $\lambda = 1$	$[0, \infty)$	VI
	$0 < \lambda$	$\alpha \in \mathbb{Z}^-, \beta > 0$	$\lvert\gamma\rvert > \lvert\alpha\rvert - 1$	$[0, \lvert\alpha\rvert)$	VII
	$\lambda < 0$	$\alpha, \beta < 0$, $\alpha \in \mathbb{Z}^-$	$\lvert\gamma\rvert, \lvert\beta\rvert > \lvert\alpha\rvert - 1$	$[0, \lvert\alpha\rvert)$	VIII

Tabla 2.2: Clasificación de la familia GHD (α and β son intercambiables). $[\cdot]$ representa la parte entera

2.3.1. La distribución $GHDI$

En Rodríguez-Avi et al. (2007) se estudia con detalle el caso $\gamma > 0$ y $0 < \lambda \leq 1$, cuando las raices del polinomio $L(x)$ son reales positivas, correspondiente a la distribución GHD tipo I, y que se considera una nueva generalización de la distribución de Waring, ya que una $UGW(a, k, \rho)$ es una $GHDI(a, k, \rho, 1)$. De esta forma, la UGW se obtiene como un caso límite de una $GHDI(\alpha, \beta, \gamma, \lambda)$ cuando la cola decrece lo más lentamente posible, ya que - aplicando (2.3) - se tiene que

$$\lim_{x\to\infty} \frac{f_{x+1}}{f_x} = \lim_{x\to\infty} \lambda \frac{(\alpha + x)(\beta + x)}{(\gamma + x)(x + 1)} = \lambda.$$

Entre las propiedades más detacables de la distribución $GHDI(\alpha, \beta, \gamma, \lambda)$ se encuentra el hecho de puede expresarse como una doble mixtura de una distribución de Poisson, con una distribución gamma y una generalización de la distribución beta, cuando $\alpha > 0$ y $\gamma > \beta$.

Teorema 2.3.1. *Si $\alpha > 0$ y $\gamma > \beta$, la distribución $GHDI(\alpha, \beta, \gamma, \lambda)$ es la mixtura*

$$Poisson(\Lambda) \underset{\Lambda}{\wedge} Gamma\left(\alpha, \frac{\lambda(1-P)}{1-\lambda(1-P)}\right) \underset{P}{\wedge} Gbeta(\gamma - \alpha - \beta, \beta, \alpha, \lambda),$$

siendo $Gbeta(\gamma - \alpha - \beta, \beta, \alpha, \lambda)$ una generalización de la distribución beta con función de densidad

$$f(p) = {}_2F_1(\alpha, \beta; \gamma; \lambda) \frac{\Gamma(\gamma)}{\Gamma(\gamma - \beta)\Gamma(\beta)} \frac{p^{\gamma-\beta-1}(1-p)^{\beta-1}}{(1-\lambda(1-p))^{\alpha}}, \quad 0 \leq p \leq 1.$$

Este resultado permite descomponer la varianza en tres componentes al igual que en la UGW. La ventaja de este modelo es que si $\lambda < 1$ no hay restricciones paramétricas para esta descomposición (salvo las exigidas en el teorema), ya que existen todos sus momentos y, por tanto, no se produce el efecto de varianza infinita que presenta la UGW cuando $\rho < 2$.

Corolario 2.3.1. *Si* $X \sim GHDI(\alpha, \beta, \gamma, \lambda)$ *con* $\alpha > 0$ *y* $\gamma > \beta$*:*

$$Var(X) = \alpha E_P(V) + \alpha E_P(V^2) + \alpha^2 Var_P(V),$$

donde $V = \dfrac{\lambda(1-P)}{1-\lambda(1-P)}$ *y* $P \sim Gbeta(\gamma - \alpha - \beta, \beta, \alpha, \lambda)$.

El inconveniente de esta partición es que, al igual que para la UGW la identificación de las componentes riesgo y predisposición no está clara.

Un ejemplo de aplicación de esta distribución aparece en Rodríguez-Avi et al. (2007) en donde se utiliza para describir datos procedentes del número de entidades bancarias por municipio, de modo que se puede obtener la descomposición de la varianza. Además, es utilizada en otros trabajos, como por ejemplo Gning et al. (2022), Satheesh Kumar and Harisankar (2020) y Kuznetsov et al. (2022), Vasquez et al. (2022), por citar los más recientes.

2.4. Caso con raíces complejas

El polinomio cuadrático $L(x)$ con coeficientes reales también puede presentar dos raíces complejas conjugadas. Irwin (1968b) menciona la posibilidad teórica de obtener una versión con parámetros complejos de la distribución de Waring generalizada univariante, pero no llega a estudiarla. Como precedente del estudio de distribuciones generadas a partir de la ecuación (2.1) con parámetros complejos, tenemos un caso particular de la familia de Ord, denominado Tipo IV, que aparece cuando se contempla que el polinomio G tenga dos raíces complejas. Sin embargo, en este caso, el rango de la distribución es todo $\mathbb{Z}$. En este escenario se inscriben las distribuciones CTP y CBP que se describen a continuación.

2.4.1. Distribución de Pearson triparamétrica compleja

La distribución de Pearson triparamétrica compleja, abreviadamente CTP, es una distribución de rango infinito desarrollada por Rodríguez-Avi et al. (2004) y Olmo-Jiménez et al. (2018) como solución de la ecuación en diferencias 2.1 cuando L y G dos polinomios cuadráticos, el primero con dos raices complejas conjugadas y el segundo con dos raices reales, una de ellas -1; específicamente,

$$L(x) = (\alpha + x)(\overline{\alpha} + x) \quad y \quad G(x) = (\gamma + x)(x + 1)$$

$\alpha = a + ib$, $a, b \in \mathbb{R}$, i la unidad imaginaria y $\gamma > 0$.

Así pues, se define la distribución CTP de la siguiente forma:

Definición 2.4.1. *Se dice que una variable aleatoria* X *sigue una distribución* CTP *con parámetros* $a, b \in \mathbb{R}$ *y* $\gamma > 0$*,* $CTP(a, b, \gamma)$*, si su fmp se expresa como*

$$P(X = x) = f_0 \frac{(a+bi)_x (a-bi)_x}{(\gamma)_x} \frac{1}{x!}, \quad x = 0, 1, \ldots \tag{2.18}$$

donde i *es la unidad imaginaría,* $(\alpha)_r$ *el símbolo de Pochhammer (*$\alpha \in \mathbb{C} \setminus \mathbb{Z}^-$*,* $r \in \mathbb{N}$*) y* f_0 *la constante de normalización dada por*

$$f_0 = \frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(\gamma)\Gamma(\gamma - 2a)}. \tag{2.19}$$

Teniendo en cuenta (2.19) y que $(\alpha)_r = \Gamma(\alpha + r)/\Gamma(\alpha)$ cuando $\alpha \notin \mathbb{R}^-$, la fmp dada en (2.18) puede escribirse equivalentemente como:

$$f(x) = C \cdot \frac{\Gamma(a + x + ib)\Gamma(a + x - ib)}{\Gamma(\gamma + x)\Gamma(x + 1)}, \quad x = 0, 1, \ldots \tag{2.20}$$

donde la constante de normalización C es

$$C = \frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(a + ib)\Gamma(a - ib)\Gamma(\gamma - 2a)}.$$

La función generatriz asociada a la distribución CTP es proporcional a la función hipergeométrica de Gauss:

$$G(t) = f_0 \cdot {}_2F_1(a + bi, a - bi; \gamma; t), \quad t \in \mathbb{R}, \tag{2.21}$$

por ello, se dice que este modelo pertenece a la familia GHD (Johnson et al., 2005).

En el Capítulo 4 compararemos esta distribución con otras infra y sobredispersas existentes en la literatura. Por esta razón, creemos conveniente detallar de forma más exhaustiva las propiedades de este modelo.

Momentos

Existe una relación de recurrencia entre los momentos no centrados de la distribución:

$$\mu'_{h+2} + (\gamma - 1)\mu'_{h+1} = \sum_{m=0}^{h} \binom{h}{m} [\mu'_{m+2} + 2a\mu'_{m+1} + (a^2 + b^2)\mu'_m]$$

donde μ'_j representa el momento no centrado de orden j. En general, para garantizar la existencia del momento de orden k es necesario que $\gamma > 2a + k$. Claramente, los momentos pueden calcularse de forma explícita resolviendo un sistema de ecuaciones, ya que únicamente dependen de los parámetros a, b y γ de la distribución. Así, por ejemplo, para $h = 1$ se obtiene un sistema de ecuaciones de donde se obtienen los dos primeros momentos no centrados:

$$\mu'_1 = \mu = \frac{a^2 + b^2}{\gamma - 2a - 1} \quad y \quad \mu'_2 = \frac{(a^2 + b^2)(a^2 + b^2 + \gamma - 1)}{(\gamma - 2a - 1)(\gamma - 2a - 2)}.$$

En consecuencia, la varianza de dicha distribución es

$$\sigma^2 = \frac{(a^2 + b^2)[(a + \gamma - 1)^2 + b^2]}{(\gamma - 2a - 1)^2(\gamma - 2a - 2)}, \tag{2.22}$$

que puede expresarse en términos de la media como

$$\sigma^2 = \mu \frac{\mu + \gamma - 1}{\gamma - 2a - 2}. \tag{2.23}$$

Análogamente, se obtiene el momento centrado de orden 3 dado por:

$$\mu_3 = \frac{(a^2 + b^2)[(a + \gamma - 1)^2 + b^2][4b^2 + (\gamma - 1)^2]}{(\gamma - 2a - 1)^3(\gamma - 2a - 2)(\gamma - 2a - 3)}. \tag{2.24}$$

Como es positivo, el coeficiente de asimetría de Fisher también lo es y, en consecuencia, la distribución presenta asimetría a la derecha.

Dispersión

Proposición 2.4.1. *La distribución CTP es infradispersa si* $a < -(\mu+1)/2$*, equidispersa si* $a = -(\mu+1)/2$ *y sobredispersa si* $a > -(\mu+1)/2$*.*

Demostración. Teniendo en cuenta la expresión (2.23), el IA correspondiente a la distribución CTP es

$$\frac{\mu+\gamma-1}{\gamma-2a-2}.$$

Así, $IA < 1$ sii $a < -(\mu+1)/2$, $IA = 1$ sii $a = -(\mu+1)/2$ y $IA > 1$ sii $a > -(\mu+1)/2$. □

Corolario 2.4.1. *Dada* $X \sim CTP(a,b,\gamma)$ *infradispersa, entonces* $a < -0.5$*.*

Demostración. $X \sim CTP(a,b,\gamma)$ es infradispersa sii $a < -(\mu+1)/2 \Leftrightarrow a < -0.5$, ya que $\mu > 0$. □

Moda

La distribución es unimodal con moda en

$$\left[\frac{(a-1)^2+b^2}{\gamma-2a+1}\right]$$

si $\frac{(a-1)^2+b^2}{\gamma-2a+1} \notin \mathbb{Z}$, siendo $[\cdot]$ la parte entera; en caso contrario, la distribución tiene dos modas consecutivas en los valores

$$\frac{(a-1)^2+b^2}{\gamma-2a+1} - 1 = \frac{a^2+b^2-\gamma}{\gamma-2a+1} \quad y \quad \frac{(a-1)^2+b^2}{\gamma-2a+1}$$

En particular, si $a^2+b^2 < \gamma$ solo hay una moda en 0. Esto implica que la distribución CTP tiene un perfil en forma de $J-$traspuesta o acampanado.

Divisibilidad infinita

Una condición suficiente para que $X \sim CTP(a,b,\gamma)$ sea infinitamente divisible (i.d.) es que $a > -0.5$ y $\gamma > (a^2+b^2)/(1+2a)$. En virtud del Corolario 2.4.1, si la distribución CTP es infradispersa, no puede ser i.d.

La propiedad de i.d. implica que X se puede expresar como una suma arbitraria de variables aleatorias independientes e idénticamente distribuidas. Este tipo de distribuciones tienen un papel muy importante en el contexto de los teoremas límite.

Convergencia

Teorema 2.4.1. *La distribución* $CTP(a,b,\gamma)$ *converge a la distribución de Poisson con parámetro* $\lambda = (a^2+b^2)/(\gamma-2a-1)$ *cuando* γ *y* $a^2+b^2 \to$ *tienden a infinito con el mismo orden de convergencia.*

Teorema 2.4.2. *Dada* $X \sim CTP(a,b,\gamma)$ *con* $E(X) = \mu$ *y* $Var(X) = \sigma^2$*, entonces,* $(X-\mu)/\sigma$ *converge a la distribución normal estándar cuando* γ *y* $\sqrt{a^2+b^2}$ *tienen el mismo orden de convergencia.*

Desde el punto de vista computacional, la librería `cpd` de R permite el cálculo de las principales funciones, así como el ajuste del modelo a un conjunto de datos (Olmo-Jiménez et al., 2022). Algunas palicciones en economía pueden verse en Rodríguez-Avi and Olmo-Jiménez (2016).

2.4.2. Distribución de Pearson biparamétrica compleja

La distribución de Pearson biparamétrica compleja con parámetros $b \in \mathbb{R}$ y $\gamma > 0$, abreviadamente $CBP(b, \gamma)$, desarrollada por Rodríguez-Avi et al. (2003a) y Rodríguez-Avi and Olmo-Jiménez (2017), también pertenece a la familia de distribuciones generadas por la función hipegeométrica de Gauss o GHD (Johnson et al., 2005), en concreto, por la función ${}_2F_1(bi, -bi; \gamma; 1)$, siendo i la unidad imaginaria. Esta distribución tiene como fmp:

$$P(X = x) = {}_2F_1(bi, -bi; \gamma; 1)^{-1} \frac{(bi)_x(-bi)_x}{(\gamma)_x x!}, \quad x = 0, 1, 2, \ldots$$

que, teniendo en cuenta (1.13), equivale a

$$P(X = x) = \frac{\Gamma(\gamma - bi)\Gamma(\gamma + bi)}{\Gamma(\gamma)^2} \frac{(bi)_x(-bi)_x}{(\gamma)_x x!}, \quad x = 0, 1, 2, \ldots$$

donde $(\alpha)_r = \Gamma(\alpha + r)/\Gamma(\alpha)$, con $\alpha > 0$ y $r \in \mathbb{N}$, es el símbolo de Pochhammer.

Cabe resaltar que dicha distribución puede verse como un caso particular de la distribución CTP cuando $a = 0$ (véase Sección 2.4.1). A continuación, resumimos algunas de las características más importantes de esta distribución:

- Esperanza:
$$\mu = \frac{b^2}{\gamma - 1}$$
que existe[1] si $\gamma > 1$.

- Varianza:
$$\sigma^2 = \frac{b^2\left[(\gamma - 1)^2 + b^2\right]}{(\gamma - 1)^2(\gamma - 2)} = \frac{\mu(\mu + \gamma - 1)}{\gamma - 2}.$$

- Índice de agregación:
$$IA = \frac{\mu + \gamma - 1}{\gamma - 2} > 1.$$
En consecuencia, la distribución CBP es siempre sobredispersa.

- Coeficiente de asimetría:
$$\gamma_1 = \frac{(4\mu + 1)\sigma^2 - 3\mu^2}{\sqrt{\sigma^2(\mu^2 + 2\mu - \sigma^2)}}$$

- Coeficiente de curtosis:
$$\gamma_2 = \frac{\mu^4(13 + 3\sigma^2) - \mu^3(1 + 27\sigma^2) + 3\mu^2\sigma^2(-1 + 7\sigma^2) + \mu\sigma^2(-1 + 9\sigma^2)}{\sigma^2(\mu^2 + 3\mu - 2\sigma^2)(\mu^2 + 2\mu - \sigma^2)} + \frac{2\sigma^4}{\sigma^2(\mu^2 + 3\mu - 2\sigma^2)(\mu^2 + 2\mu - \sigma^2)} - 3.$$

[1] $E\left(X^k\right) < \infty$ si y solo si $\gamma > k$.

- Función generatriz de probabilidad:

$$G(t) = \frac{{}_2F_1(bi, -bi; \gamma; t)}{{}_2F_1(bi, -bi; \gamma; 1)}, \quad t \in \mathbb{R}$$

La distribución CBP resulta útil para modelizar datos que presentan sobredispersión, pero en los que la probabilidad del 0 no es demasiado elevada sino similar a la de la distribución de Poisson con la misma media.

Desde el punto de vista computacional, la librería `cpd` de R permite el cálculo de las principales funciones, así como el ajuste del modelo a un conjunto de datos (Vilchez-Lopez et al., 2022; Olmo-Jiménez et al., 2022).

2.5. Extensiones

Se han considerado también extensiones de la familia de distribuciones hipergeométricas generalizadas al caso de argumentos complejos con $q = p + 1 \geq 3$ (Olmo-Jiménez, 2002). Destaca el trabajo de Rodríguez-Avi et al. (2008) donde se analiza la familia de distribuciones generadas por la función hipergeométrica ${}_3F_2$ con argumentos complejos y $\lambda = 1$. Concretamente, si

$$\begin{aligned} L(x) &= (\alpha_1 + x)(\alpha_2 + x)(\alpha_3 + x) \\ G(x) &= (\gamma_1 + x)(\gamma_2 + x)(x + 1) \end{aligned}$$

con $\alpha_i, i = 1, 2, 3, \gamma_j, j = 1, 2$ números complejos y $\gamma_j \notin \mathbb{Z}^-$, la solución de la ecuación en diferencias (2.1) está dada por

$$f_x = f_0 \frac{(\alpha_1)_x (\alpha_2)_x (\alpha_3)_x}{(\gamma_1)_x (\gamma_2)_x} \frac{1}{x!}, \quad x \in \mathbb{Z}^+,$$

donde $f_0 = {}_3F_2(\alpha_1, \alpha_2, \alpha_3; \gamma_1, \gamma_2; 1)^{-1}$ es la constante de normalización.

En el caso de rango infinito (esto es, cuando $\alpha_i \notin \mathbb{Z}^-$, $i = 1, 2, 3$), la función f_x será una verdadera fmp si

$$\gamma_1 + \gamma_2 > \alpha_1 + \alpha_2 + \alpha_3.$$

Puesto que los polinomios cúbicos pueden tener tres raices reales o dos complejas y una real, los autores consideran los siguientes casos con raices puramente complejas:

- Caso 1: $\alpha_i \in \mathbb{R}, \alpha_j = \overline{\alpha}_k \in \mathbb{C}\backslash\mathbb{R}, i, j, k = 1, 2, 3, i \neq j \neq k, \gamma_l \in \mathbb{R}, l = 1, 2.$
- Caso 2: $\alpha_i \in \mathbb{R}, \gamma_j \in \mathbb{C}\backslash\mathbb{R}, i = 1, 2, 3, j = 1, 2, \gamma_1 = \overline{\gamma}_2.$
- Caso 3: $\alpha_i \in \mathbb{R}, \alpha_j = \overline{\alpha}_k \in \mathbb{C}\backslash\mathbb{R}, i, j, k = 1, 2, 3, i \neq j \neq k, \gamma_l \in \mathbb{C}\backslash\mathbb{R}, l = 1, 2, \gamma_1 = \overline{\gamma}_2.$

Asimismo, establecen 8 tipos distintos de distribuciones pertenecientes a esta familia, algunas de ellas con hasta 2 modas locales. Además, demuestran que esta distribución puede verse como una mixtura de una distribución GHD y una generalización de la distribución beta.

Por otra parte, la familia de distribuciones generadas por funciones hipergeométricas se extiende al caso multivariante considerando extensiones de la ecuación (2.1) a sistemas de

ecuaciones en diferencias parciales, de forma que la solución está generada por la función

$$F_D^{(p)}(\alpha, \beta_1, \ldots, \beta_p; \gamma; \lambda_1, \ldots, \lambda_p) = \sum_{m_1,\ldots,m_p=0}^{\infty} \frac{(\alpha)_{m_1+\ldots+m_p}(\beta_1)_{m_1} \cdot (\beta_p)_{m_p}}{(\gamma)_{m_1+\ldots+m_p}} \frac{(\lambda_1)^{m_1}}{m_1!} \cdots \frac{(\lambda_p)^{m_p}}{m_p!}$$

Las distribuciones discretas mutivariantes más conocidas, tales como la multinomial, Poisson $k-$variante y multivariante de Waring pertenecen a esta familia.

En el caso bivariante, el sistema de ecuaciones en diferencias parciales está dado por:

$$G(x,y) f_{x+1,y} - L(x,y) f_{x,y} = 0 \tag{2.25}$$
$$H(x,y) f_{x,y+1} - N(x,y) f_{x,y} = 0 \tag{2.26}$$

donde

$$L, N : \mathbb{Z}^+ \times \mathbb{Z}^+ \longrightarrow \mathbb{R}$$
$$G, H : \mathbb{Z}^+ \times \mathbb{Z}^+ \longrightarrow \mathbb{R}\backslash\{0\}$$

son funciones, en principio, cualesquiera, y $f : \mathbb{Z}^+ \times \mathbb{Z}^+ \longrightarrow \mathbb{R}$, la función desconocida.

La solución de este sistema no está garantizada. Una condición necesaria para que la función f sea solución de dicho sistema es que se verifique la igualdad:

$$\frac{L(x,y+1) N(x,y)}{G(x,y+1) H(x,y)} = \frac{N(x+1,y) L(x,y)}{H(x+1,y) G(x,y)} \tag{2.27}$$

Así, la solución viene dada por el siguiente resultado:

Teorema 2.5.1. *Si la función $f : \mathbb{Z}^+ \times \mathbb{Z}^+ \longrightarrow \mathbb{R}$ es solución del sistema (2.25), entonces puede escribirse de la forma*

$$f_{x,y} = \begin{cases} f_{0,0} \prod_{t=0}^{x-1} \prod_{t'=0}^{y-1} \frac{L(t,y) N(0,t')}{G(t,y) H(0,t')} & x \geq 1, y \geq 1 \\ f_{0,0} \prod_{t=0}^{x-1} \frac{L(t,0)}{G(t,0)} & x \geq 1, y = 0 \\ f_{0,0} \prod_{t'=0}^{y-1} \frac{N(0,t')}{H(0,t')} & x = 0, y \geq 1 \\ f_{0,0} & x = 0, y = 0 \end{cases} \tag{2.28}$$

fijada $f_{0,0}$ que supondremos distinta de cero.

Seguidamente se establecen las condiciones para que la función $f_{x,y}$ sea la fmp de un vector aleatorio discreto (X, Y).

Teorema 2.5.2. *Sea el conjunto $\mathcal{H} = \{(r,s) \in \mathbb{Z}^+ \times \mathbb{Z}^+; f_{r,s} \neq 0\}$. La condición necesaria y suficiente para que la función $f : \mathbb{Z}^+ \times \mathbb{Z}^+ \longrightarrow \mathbb{R}$, solución del sistema (2.25), dada por la expresión (2.28), sea una función masa de probabilidad, es que se verifiquen las siguientes condiciones:*

(*i*) *Condición de positividad*

$$\begin{matrix} L(x,y)G(x,y) \geq 0 \\ N(x,y)H(x,y) \geq 0 \end{matrix} \qquad \forall x, y \in \mathcal{H}$$

(*ii*) *Condición de convergencia*

$$\sum_{\substack{x,y=0\\x+y\neq 0\\(x,y)\in\mathcal{H}}}^{\infty}\prod_{t=0}^{x-1}\prod_{t'=0}^{y-1}\frac{L\left(t,y\right)}{G\left(t,y\right)}\frac{N\left(0,t'\right)}{H\left(0,t'\right)}<\infty$$

(*iii*) *Condición de normalización*

$$f_{0,0}=\frac{1}{1+\sum\limits_{\substack{x,y=0\\x+y\neq 0\\(x,y)\in\mathcal{H}}}^{\infty}\prod\limits_{t=0}^{x-1}\prod\limits_{t'=0}^{y-1}\frac{L(t,y)}{G(t,y)}\frac{N(0,t')}{H(0,t')}}$$

Si las funciones G, L, H y N son polinomios cualesquiera, las fgp y fgm se caracterizan vía los sistemas de ecuaciones diferenciales parciales que verifican.

Teorema 2.5.3. *Sean las funciones L y N polinomios en las variables (x,y), G un polinomio en $(x+1,y)$ y H un polinomio en $(x,y+1)$, de órdenes cualesquiera y tal que verifiquen las condiciones dadas en el Teorema 2.5.2. Entonces su función generatriz de probabilidad, $g(t_1,t_2)$, verifica el sistema de ecuaciones diferenciales en derivadas parciales*

$$\begin{aligned}G(\theta_1,\theta_2)g(t_1,t_2)-t_1L(\theta_1,\theta_2)g(t_1,t_2)&=G\left(\theta_1,\theta_2\right)\sum_{y=0}^{\infty}f_{0,y}t_2^y\\H(\theta_1,\theta_2)g(t_1,t_2)-t_2N(\theta_1,\theta_2)g(t_1,t_2)&=H\left(\theta_1,\theta_2\right)\sum_{x=0}^{\infty}f_{x,0}t_1^x\end{aligned}\tag{2.29}$$

para $|t_1|<1$, $|t_2|<1$ y siendo $\theta_i=t_i\dfrac{\partial}{\partial t_i}\quad i=1,2$.

Realizando un cambio de variable se obtienen los sistemas de ecuaciones diferenciales que satisfacen las funciones generatrices de momentos y función característica.

Al igual que en el caso univariante, los momentos no centrados de orden r y s, $\mu'_{r,s}$, pueden obtenerse a partir de cualesquiera de las funciones anteriores, siempre que existan sus derivadas y sean finitas para $t_i=1$, $i=1,2$, en el caso de la fgp y para $t_i=0$, $i=1,2$ en las restantes. Así, a través de los operadores D_i, θ_i y θ'_i se pueden definir los momentos de la forma

$$\mu'_{r,s}=\left[\theta_1\theta_2g(t_1,t_2)\right]_{t_1=t_2=1}=\left[D_1D_2M(t_1,t_2)\right]_{t_1=t_2=0}=\left[\theta'_1\theta'_2\phi(t_1,t_2)\right]_{t_1=t_2=0}$$

lo que permite obtener también relaciones de recurrencia para el cálculo de estos momentos.

Considerando distintas expresiones para los polinomios L,G,N y H surgen las familias de distribuciones generadas por las funciones hipergeométricas F_1,F_2,F_3 y F_4 de Appell (Appell and Kampe de Feriet, 1926). Un ejemplo de distribución generada de esta forma puede verse en Rodríguez-Avi et al. (2006).

Sólo es posible extender al caso de argumentos complejos las distribuciones generadas por las funciones F_3 y F_4, estudio realizado en Olmo-Jiménez (2002).

Otra vía para extender la familia GHD al caso multivariante es considerar la función hipergeométrica de argumento matricial ${}_2F_1(\alpha,\beta;\gamma;X)$ como generadora de distribuciones

multivariantes discretas. Esta función se define de la siguiente forma

$$ {}_2F_1(\alpha, \beta; \gamma; X) = \sum_{k=0}^{\infty} \sum_{\kappa} \frac{(\alpha)_k (\beta)_k}{(\gamma)_k} \frac{C_{\kappa(X)}}{k!}, $$

donde $\sum_\kappa$ denota la suma sobre todas las particiones $\kappa = (k_1, \ldots, k_p)$ de k, $C_\kappa(X)$ es el polinomio zonal de X in k, $(\alpha)_\kappa = \prod_{i=1}^{p} \left(\alpha - \frac{1}{2}(i-1)\right)_{k_i}$ es el coeficiente hipergeométrico generalizado, X es una matriz simétrica $m \times m$ y α, β y γ son números complejos arbitrarios. La serie converge si el mayor valor propio de X es, en valor absoluto, mayor que 1.

En este sentido, destacan los trabajos de Gutiérrez-Jáimez et al. (1999), Sáez-Castillo (1997) y Sáez-Castillo (2001) donde, además, se estudian los correspondientes métodos de estimación que permiten la modelización de datos reales.

Capítulo 3

La distribución de Waring biparamétrica extendida

3.1. Introducción

Como se ha comentado en el Capítulo 2, una vía de generación de distribuciones discretas viene dada por la verificación de la ecuación en diferencias (2.1) en donde los coeficientes son polinomiales. El caso más estudiado es el que corresponde a polinomios de grado 2. La clasificación de tales distribuciones viene dada por la naturaleza de las raíces del polinomio $L(r)$, en donde los modelos más estudiados corrrsponden principalmente al caso en que las dos raíces son reales (binomial, binomial negativa, hipergeométrica, UGW, $GHDI$, etc). Posteriormente se ha desarrollado el caso en que las dos raíces son complejas conjugadas (CBP y CTP). Si embargo, queda otro modelo en donde el polinomio $L(r)$ es un cuadrado perfecto del tipo $L(r) = (r - \alpha)^2$, por lo que presenta una raíz doble, y que puede verse como un caso límite entre los modelos con raices reales (por ejemplo, UGW) y complejas (CTP), que es el que nos proponemos estudiar en este capítulo.

Este modelo con dos raíces idénticas presenta varias ventajas. Así, como veremos, produce una distribución biparamétrica que, sin embargo, puede reproducir propiedades de distribuciones con tres parámetros. En este sentido, recordemos que la distribución UGW, que se obtiene como una mixtura en dos pasos de la distribución de Poisson, permite descomponer la varianza en tres partes: aleatoriedad (o variabilidad no explicable), riesgo (o variabilidad debida a factores de exposición externa) y predisposición (o variabilidad debida a factores internos). Un ejemplo claro de esto aparece en Xekalaki (1984) en donde se aplica la distribución UGW al estudio del número de accidentes por conductor. Así, la parte de variabilidad aleatoria es la que corresponde a la "mala suerte" o a factores que incidentalemente pueden afectar a la conducción; el factor riesgo corresponde a factores explicables por motivos externos o ajenos al conductor (por ejemplo, mal estado de la carretera, condiciones climatológicas adversas, etc) y la parte de la predisposición se correponde con la actitud interna de cada conductor (más o menos prudente, atento, responsable, etc). Esta información adicional, en principio muy interesante a efectos del análisis de la variabilidad de los datos, encuentra un importante inconveniente relacionado con la expresión de la fmp de la distribución UGW (véase la ecuación (1.10)), ya que, en ella, sus dos primeros parámetros son intercambiables, lo que implica que las componentes riesgo y predisposición son difíciles de identificar. En la literatura se han planteado distintas soluciones. Así, por ejemplo, Irwin (1968a) sugiere que sea el investigador, según su experiencia, quien

determine qué componente se refiere al riesgo y cuál a la predisposición; Xekalaki (1984), por su parte, propone una versión bivariante de la distribución de Waring; mientras que Rodríguez-Avi et al. (2009) resuelven el problema utilizando la información adicional que proporcionan las covariables a través de un modelo de regresión. En cualquier caso, es necesario disponer de información externa que no siempre está disponible. Otras extensiones de la distribución UGW que tampoco consiguen solucionar el problema de identificación mencionado son la distribución $GHDI$, considerada como una generalización de la distribución de Waring (Rodríguez-Avi et al., 2007) o la distribución de Waring generalizada de Sttutering (Panaretos and Xekalaki, 1986).

Por otra parte, en el caso de soluciones complejas de $L(r)$, la CTP tiene la propiedad de poder modelizar tanto datos sobre como infradispersos, siendo una de las más cómodas de utilizar para tal tipo de datos. Además, presenta un caso particular biparamétrico ya estudiado, que corresponde al caso $a = 0$ (es decir, $L(r) = (x + \alpha)^2$ con $\alpha > 0$), el cual no hereda esa propiedad de posible infradispersión. En este caso, el modelo que aquí se presenta también puede verse como un caso particular biparamétrico de la CTP pero que, como veremos, sí hereda la posibilidad de ser utilizada en caso de datos con infradispersión.

Por último, tanto la UGW como la CTP son distribuciones de conteo de rango infinito. En cambio, el nuevo modelo propuesto también permite distribuciones de rango finito uniparamétricas e infradispersas, que también se desarrollarán en este capítulo.

Gran parte de los resultados que se muestran han sido publicados en las revistas *Mathematics* Cueva-López et al. (2021) y *Computational and Mathematical Methods* Cueva-López et al. (2019).

3.2. Definición

En este capítulo nos centramos en el caso en el que el polinomio $L(x)$ tiene una raiz doble y $\lambda = 1$, es decir, $\alpha = \beta$. Este caso se enmarca dentro de las distribuciones GHD tipo I y II. Así pues, la solución de la ecuación en diferencias (2.1), viene dada en términos de una ${}_2F_1(\alpha, \alpha; \gamma; 1)$ que, utilizando el teorema de sumación de Gauss dado en 2.13, se expresa como

$$ {}_2F_1(\alpha, \alpha; \gamma; 1) = \frac{\Gamma(\gamma)\Gamma(\gamma - 2\alpha)}{\Gamma(\gamma - \alpha)^2}. \tag{3.1} $$

En consecuencia, podemos realizar la siguiente definición.

Definición 3.2.1. *Una variable aleatoria X sigue una distribución de Waring biparamétrica extendida, EBW, con parámetros α y γ, si su fmp es:*

$$ P(X = x) = \frac{\Gamma(\gamma - \alpha)^2}{\Gamma(\gamma - 2\alpha)} \frac{(\alpha)_x^2}{\Gamma(\gamma + x)} \frac{1}{x!}, \quad x = 0, 1, \ldots \tag{3.2} $$

donde $\alpha \in \mathbb{R}$ y $\gamma > \max(0, 2\alpha)$.

Proposición 3.2.1. *Para garantizar la existencia de la media, μ, y de la varianza, σ^2, de $X \sim EBW(\alpha, \gamma)$ es condición necesaria que $\gamma > 2\alpha + 1$ y $\gamma > 2\alpha + 1$, respectivamente, ya que*

$$ \mu = \frac{\alpha^2}{\gamma - 2\alpha - 1} \tag{3.3} $$

y

$$\sigma^2 = \frac{\alpha^2(\gamma-\alpha-1)^2}{(\gamma-2\alpha-1)^2(\gamma-2\alpha-2)} = \mu\frac{\mu+\gamma-1}{\gamma-2\alpha-2}. \tag{3.4}$$

Demostración. Basta tener en cuenta las expresiones de la media y la varianza dadas en (2.16) y (2.17) y considerar $\alpha = \beta$. □

Nota 3.2.1. *En general, para que exista el momento no centrado de orden m es necesario que $\gamma > 2\alpha + m$.*

Cuando $\alpha > 0$, el parámetro γ ha de ser mayor que 2α, con lo que puede considerarse otra parametrización alternativa para la distribución EBW en términos de $\alpha > 0$ y un nuevo parámetro ρ definido como $\gamma - 2\alpha > 0$. Así, utilizando que $(\alpha)_x = \Gamma(\alpha+x)/\Gamma(x)$, la fmp dada en (3.2) se reescribe como:

$$P(X=x) = \frac{\Gamma(\alpha+\rho)^2}{\Gamma(\alpha)^2\Gamma(\rho)}\frac{\Gamma(\alpha+x)^2}{\Gamma(2\alpha+\rho+x)}\frac{1}{x!}, \quad x = 0,1,2,\ldots \tag{3.5}$$

Del mismo modo, la media y la varianza dadas en (3.3) y (3.4) se reducen a

$$\mu = \frac{\alpha^2}{\rho-1} \tag{3.6}$$

y

$$\sigma^2 = \frac{\alpha^2(\alpha+\rho-1)^2}{(\rho-1)^2(\rho-2)} = \mu\frac{\mu+2\alpha+\rho-1}{\rho-2}. \tag{3.7}$$

Un caso de especial interés que conviene estudiar de forma específica es cuando $\alpha \in \mathbb{Z}^-$. En esta situación, la distribución EBW se convierte en una distribución de rango finito. Por este motivo, se define un nuevo parámetro $n = -\alpha$, y se parametriza la distribución EBW en términos de n y γ, quedando la fmp definida de la siguiente forma:

$$P(X=x) = \frac{\Gamma(\gamma+n)^2}{\Gamma(\gamma+2n)}\frac{(-n)_x^2}{\Gamma(\gamma+x)}\frac{1}{x!}, \quad x = 0,1,\ldots,n \tag{3.8}$$

Teniendo en cuenta que $(-n)_x = (-1)^x(n-x+1)_x$ y $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ (3.8) puede expresarse equivalentemente como:

$$\begin{aligned} P(X=x) =& \frac{\Gamma(\gamma+n)^2}{\Gamma(\gamma+2n)}\frac{(-n)_x^2}{\Gamma(\gamma+x)}\frac{1}{x!} = \frac{\Gamma(\gamma+n)^2}{\Gamma(\gamma+2n)}\frac{(-1)^{2x}(n-x+1)_x^2}{\Gamma(\gamma+x)}\frac{1}{x!} \\ =& \frac{\Gamma(\gamma+n)^2}{\Gamma(\gamma+2n)}\frac{\Gamma(n+1)^2}{\Gamma(n-x+1)^2}\frac{1}{\Gamma(\gamma+x)}\frac{1}{x!} = \frac{\Gamma(\gamma+n)^2}{\Gamma(\gamma+2n)}\frac{n!^2}{(n-x)!^2\Gamma(\gamma+x)}\frac{1}{x!} \\ =& \binom{n}{x}^2\frac{\Gamma(\gamma+n)^2}{\Gamma(\gamma+2n)}\frac{x!}{\Gamma(\gamma+x)}, \quad x = 0,1,\ldots,n \end{aligned} \tag{3.9}$$

Asimismo, las probabilidades de la EBW finita también pueden obtenerse de forma iterativa:

$$P(X=x) = P(X=x-1)\frac{(n-x+1)^2}{(\gamma-x+1)x}, \quad x = 1,2,\ldots,n, \tag{3.10}$$

donde

$$P(X=0)=\frac{\Gamma(\gamma+n)^2}{\Gamma(\gamma+2n)\Gamma(\gamma)}.$$

Por último, la media, μ, y la varianza, σ^2, de $X\sim EBW(n,\gamma)$ tienen las siguientes expresiones:

$$\mu=\frac{n^2}{\gamma+2n-1},\qquad \sigma^2=\frac{n^2(\gamma+n-1)^2}{(\gamma+2n-1)^2(\gamma+2n-2)}=\mu\frac{\mu+\gamma-1}{\gamma+2n-2}. \tag{3.11}$$

3.3. Propiedades

Para estudiar las propiedades de la distribución EBW distinguimos tres casos: (1) $\alpha\in\mathbb{R}^+$; (2) $\alpha\in\mathbb{R}^-\setminus\mathbb{Z}^-$; (3) $\alpha\in\mathbb{Z}^-$.

3.3.1. Caso $\alpha\in\mathbb{R}^+$

Teniendo en cuenta la expresión (3.5), puede establecerse el siguiente resultado.

Teorema 3.3.1. *La distribución $EBW(\alpha,\rho)$ con $\alpha,\rho>0$ es una distribución $UGW(\alpha,\alpha,\rho)$.*

Demostración. Si consideramos $\alpha=\beta>0$ y $\lambda=1$ en la ecuación (2.12) y utilizamos (3.1) es fácil ver que la ecuación (3.5) coincide con la fmp de una distribución $UGW(\alpha,\alpha,\rho)$ (ver (1.14)). □

Este teorema nos dice que la distribución EBW con $\alpha>0$ puede verse como un caso particular biparamétrico de la distribución UGW. Así pues, se obtiene de forma directa el siguiente corolario:

Corolario 3.3.1. *La distribución $EBW(\alpha,\rho)$ con $\alpha,\rho>0$ hereda las propiedades de la distribución UGW.*

A continuación enumeramos cuáles son estas propiedades:

1. La distribución $EBW(\alpha,\rho)$ es una doble mixtura de una distribución de Poisson. Concretamente:
   - $X|\lambda\sim\mathcal{P}(\lambda)$
   - $\lambda\sim Gamma(\alpha,v)$ con función de densidad

   $$f(\lambda|\alpha,v)=\frac{1}{\Gamma(\alpha)v^\alpha}\lambda^{\alpha-1}e^{-\lambda/v},\quad \lambda>0,\quad \alpha,v>0,$$

   Por tanto, $X|\alpha,\nu\sim BN(\alpha,v)$ con fmp

   $$f(x|\alpha,v)=\frac{1}{x!}\frac{\Gamma(x+\alpha)}{\Gamma(\alpha)}\left(\frac{1}{1+v}\right)^\alpha\left(\frac{v}{1+v}\right)^x,\quad x=0,1,\dots$$

   - $v|\alpha,\rho\sim Beta(\alpha,\rho)$ de tipo II con función de densidad

   $$f(v|\alpha,\rho)=\frac{\Gamma(\alpha+\rho)}{\Gamma(\alpha)\Gamma(\rho)}v^{\alpha-1}(1+v)^{-(\alpha+\rho)},\quad v>0,\quad \alpha,\rho>0$$

2. Al ser una mixtura de Poisson, la EBW con $\alpha>0$ es siempre sobredispersa.

3. Converge a la distribución de Poisson cuando ρ y α^2 tienden a infinito con el mismo orden de convergencia.

4. Como consecuencia de la mixtura, la varianza de X puede descomponerse como suma de tres componentes conocidas como aleatoriedad, riesgo y predisposición, respectivamente:

$$\sigma^2 = \frac{\alpha^2}{\rho-1} + \frac{\alpha^2(\alpha+1)}{(\rho-1)(\rho-2)} + \frac{\alpha^3(\alpha+\rho-1)}{(\rho-1)^2(\rho-2)}. \tag{3.12}$$

   Estas componentes ya son claramente distinguibles, puesto que al ser $\alpha = \beta$ en la distribución UGW, se elimina uno de los parámetros (no son intercambiables) y, por tanto, desaparece el problema de la identificación de las componentes de la varianza. En consecuencia, la distribución EBW con $\alpha > 0$ resuelve el inconveniente que presenta la UGW sin necesidad de disponer de información adicional.

   Para determinar el efecto que cada parámetro ejerce sobre las componentes de la varianza hemos calculado la proporción de varianza explicada por cada una de ellas dividiendo entre el valor de σ^2 dado en (3.7)

$$1 = \frac{(\rho-1)(\rho-2)}{(\alpha+\rho-1)^2} + \frac{(\alpha+1)(\rho-1)}{(\alpha+\rho-1)^2} + \frac{\alpha}{\alpha+\rho-1},$$

   y hemos representado estas componentes (en porcentaje) considerando primero el parámetro α fijo y ρ variable y, posteriormente, α variable y ρ fijo. Los gráficos se incluyen en la Figura 3.1.

   Se observa que cuando α es fijo (gráficos de la primera columna), cuanto mayor es ρ más importante es la predisposición. Por su parte, cuando ρ es fijo, cuanto mayor es α más importante es la aleatoriedad. Además, si α y ρ aumentan con el mismo orden de convergencia, la predisposición no supera el 50 % de la varianza total, mientras que las otras dos componentes tienden al 25 % cada una.

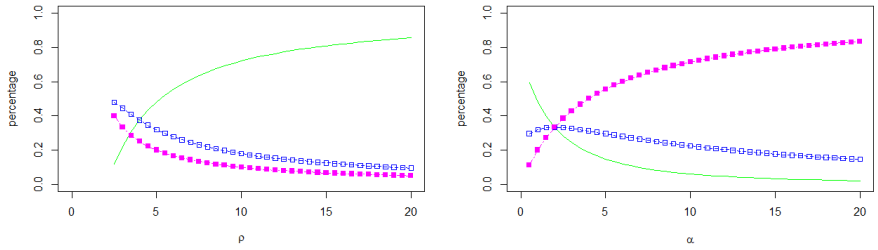
La Propiedad 1 permite expresar la distribución EBW con $\alpha > 0$ como una doble mixtura de una Poisson, pero también como una mixtura simple de una distribución binomial negativa y una distribución beta tipo II o como una mixtura simple de una distribución de Poisson con una distribución gamma-beta tipo II, tal y como se desarrolla en la siguiente proposición.

Proposición 3.3.1. *La distribución $EBW(\alpha,\rho)$ puede expresarse como una mixtura simple de una distribución de Poisson con una distribución gamma-beta tipo II. Específicamente:*

- $X|\lambda \sim \mathcal{P}(\lambda)$
- $\lambda \sim Gamma - BetaII(\alpha,\rho)$ *con función de densidad*

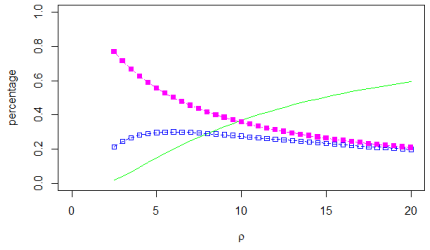
$$g(\lambda|\alpha,\rho) = \frac{\lambda^{\alpha-1}\Gamma(\alpha+\rho)^2}{\Gamma(\alpha)^2\Gamma(\rho)} U(\alpha+\rho,1,\lambda), \quad \lambda>0, \quad \alpha,\rho>0;$$

donde $U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt}t^{a-1}(1+t)^{b-a-1}dt$ con $Re(a), Re(z) > 0$ y $Re(\cdot)$ la función parte real, es la función hipergeométrica confluente de segunda clase, también conocida como función de Kummer, función de Tricomi o función de Gordon Abramowitz and Stegun (1972).

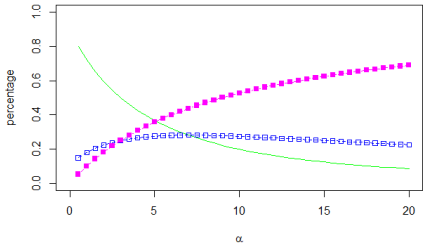


(a) $\alpha = 0.5$

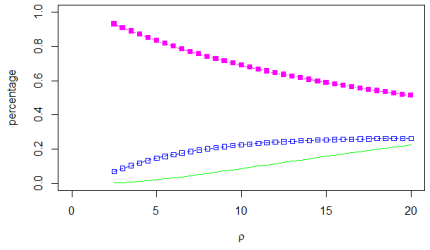
(b) $\rho = 2.5$



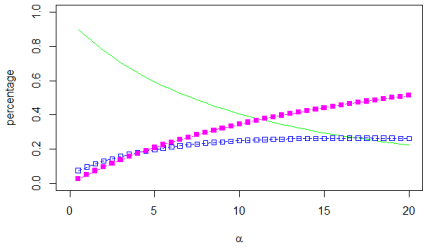
(c) $\alpha = 1$



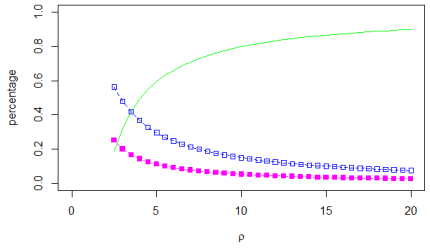
(d) $\rho = 5$



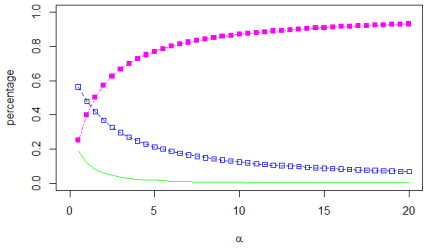
(e) $\alpha = 5$



(f) $\rho = 10$



(g) $\alpha = 20$



(h) $\rho = 20$

Figura 3.1: Componentes de la partición de la varianza (en %) de una $EBW(\alpha, \rho)$: aleatoriedad (linea verde), riesgo (linea azul) y predisposición (linea morada). ρ varía de 2.1 a 20 en los gráficos (a), (c), (e) y (g). α varía de 0.1 a 20 en los gráficos (b), (d), (f) y (h).

Demostración. Basta con obtener la función de densidad de la mixtura gamma-beta tipo II. Para ello, consideramos $\lambda|\alpha, v \sim Gamma(\alpha, v)$ con función de densidad

$$g(\lambda|\alpha, v) = \frac{1}{\Gamma(\alpha)v^{\alpha}}\lambda^{\alpha-1}e^{-\lambda/v}, \quad \lambda > 0, \quad \alpha, v > 0$$

y $v \sim BetaII(\rho, \alpha)$ con función de densidad

$$h(v|\rho, \alpha) = \frac{\Gamma(\alpha+\rho)}{\Gamma(\alpha)\Gamma(\rho)}v^{\alpha-1}(1+v)^{-(\alpha+\rho)}, \quad v > 0, \quad \alpha, \rho > 0.$$

Entonces, la función de densidad de la mixtura gamma-betaII, $\lambda|\alpha, \rho$, se obtiene como

$$\begin{aligned} g(\lambda|\alpha, \rho) &= \int_0^{\infty} g(\lambda|\alpha, v)h(v|\rho, \alpha)dv \\ &= \int_0^{\infty} \frac{1}{\Gamma(\alpha)v^{\alpha}}\lambda^{\alpha-1}e^{-\lambda/v}\frac{\Gamma(\alpha+\rho)}{\Gamma(\alpha)\Gamma(\rho)}v^{\alpha-1}(1+v)^{-(\alpha+\rho)}dv \\ &= \frac{\lambda^{\alpha-1}\Gamma(\alpha+\rho)}{\Gamma(\alpha)^2\Gamma(\rho)}\int_0^{\infty} e^{-\lambda/v}v^{-1}(1+v)^{-(\alpha+\rho)}dv \\ &= \frac{\lambda^{\alpha-1}\Gamma(\alpha+\rho)^2}{\Gamma(\alpha)^2\Gamma(\rho)}U(\alpha+\rho, 1, \lambda), \quad \lambda > 0, \quad \alpha, \rho > 0. \end{aligned}$$

□

Finalmente, en la Figura 3.2 se muestran algunos perfiles de la distribución EBW con $\alpha > 0$. Claramente, a medida que aumenta ρ aumenta la probabilidad del 0. Además, cuando α crece, el perfil cambia de $J-$traspuesta a acampanado, debido a que la moda se desplaza hacia la derecha.

3.3.2. Caso $\alpha \in \mathbb{R}^-\backslash\mathbb{Z}^-$

Cuando el parámetro α es negativo (no entero), la única condición sobre el parámetro γ es que sea positivo, de modo que la distribución EBW puede verse como un caso particular de una distribución CTP, ya que esta última surge cuando el polinomio $L(x)$ en (2.10) tiene raíces complejas conjugadas $\alpha = a + ib$ y $\beta = a - ib$, que pueden ser iguales si $b = 0$. Específicamente, se tiene el siguiente resultado.

Teorema 3.3.2. *Si $\alpha < 0$, la distribución $EBW(\alpha, \gamma)$ con $\gamma > 0$ es una distribución $CTP(\alpha, 0, \gamma)$.*

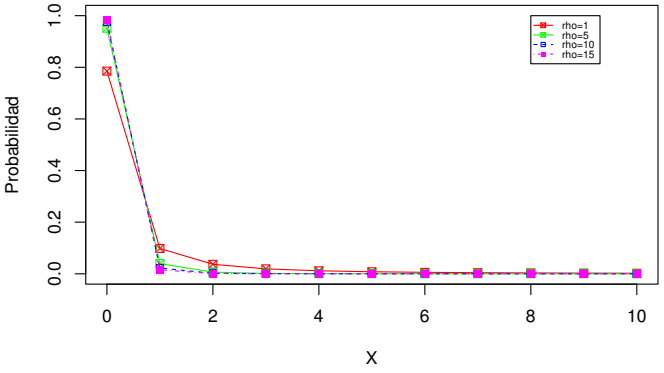
Demostración. De acuerdo con (2.20) la expresión de la fmp de la distribución $CTP(\alpha, 0, \gamma)$ está dada por

$$f(x) = \frac{\Gamma(\gamma-\alpha)^2}{\Gamma(\alpha)^2\Gamma(\gamma-2\alpha)} \cdot \frac{\Gamma(\alpha+x)^2}{\Gamma(\gamma+x)\Gamma(x+1)}, \quad x = 0, 1, \ldots$$

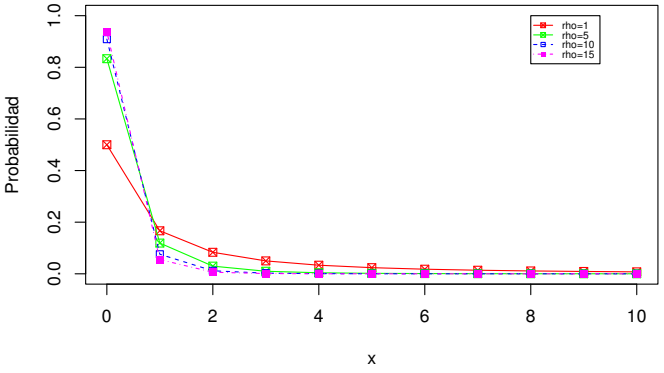
que claramente coincide con la expresión (3.2) de la fmp de una $EBW(\alpha, \gamma)$. □

Nota 3.3.1. *Este resultado se extiende de forma directa al caso $\alpha > 0$ puesto que el primer parámetro de la distribución CTP puede ser cualquier número real.*

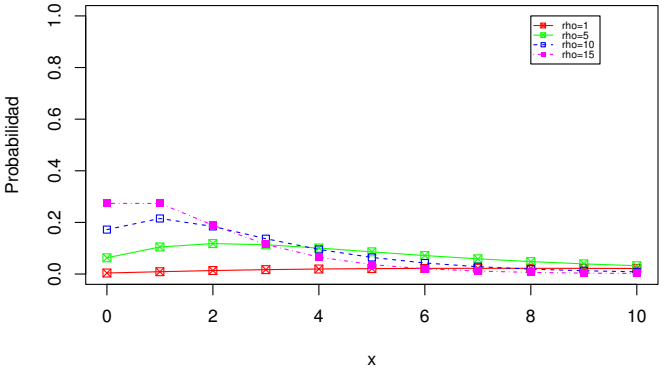
Corolario 3.3.2. *La distribución $EBW(\alpha, \gamma)$ con $\alpha < 0$ hereda las propiedades de la distribución CTP.*



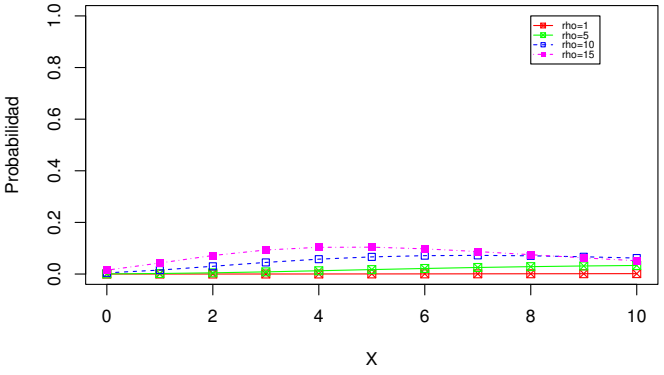
(a) $\alpha = 0.5$



(b) $\alpha = 1$



(c) $\alpha = 5$



(d) $\alpha = 10$

Figura 3.2: Perfiles de la distribución EBW con $\alpha < 0$

En consecuencia, las propiedades de la distribución EBW con $\alpha < 0$ son las siguientes:

1. Si $\frac{(\alpha-1)^2}{\gamma-2\alpha+1} \in \mathbb{Z}$, la distribución tiene dos modas consecutivas en los valores

$$\frac{(\alpha-1)^2}{\gamma-2\alpha+1} - 1 = \frac{\alpha^2-\gamma}{\gamma-2\alpha+1} \quad \text{y} \quad \frac{(\alpha-1)^2}{\gamma-2\alpha+1}.$$

   En caso contrario, la distribución tiene una única moda en 0 si $\alpha^2 < \gamma$ o en $\left[\frac{(\alpha-1)^2}{\gamma-2\alpha+1}\right]$, donde $[\cdot]$ representa la parte entera. En consecuencia, la fmp de la EBW con $\alpha < 0$ tiene forma acampanada o de J traspuesta.

2. La distribución EBW con $\alpha < 0$ puede ser infra-, equi- o sobredispersa. Concretamente, es

   - infradispersa cuando $\alpha \leq -1$ o cuando $-1 < \alpha < -0.5$ y $\gamma > \dfrac{3\alpha^2+4\alpha+1}{2\alpha+1}$.
   - equidispersa cuando $-1 < \alpha < -0.5$ y $\gamma = \dfrac{3\alpha^2+4\alpha+1}{2\alpha+1}$.
   - sobredispersa cuando $\alpha \geq -0.5$ o cuando $-1 < \alpha < -0.5$ y $\gamma < \dfrac{3\alpha^2+4\alpha+1}{2\alpha+1}$.

   En definitiva, una condición necesaria (pero no suficiente) para que sea infradispersa es que $\alpha < -0.5$.

3. Una condición suficiente para que esta distribución sea infinitamente divisible es que $\alpha > -0.5$ y $\gamma > \alpha^2/(1+2\alpha)$. Como consecuencia, una distribución EBW infradispersa nunca puede ser infinitamente divisible.

4. La distribución EBW con converge a:

   - la distribución de Poisson, $\mathcal{P}(\mu)$, cuando γ y $\alpha^2 \to \infty$ con el mismo orden de convergencia.
   - la distribución normal, $\mathcal{N}(\mu, \sigma)$, cuando γ y α tienen el mismo orden de convergencia.

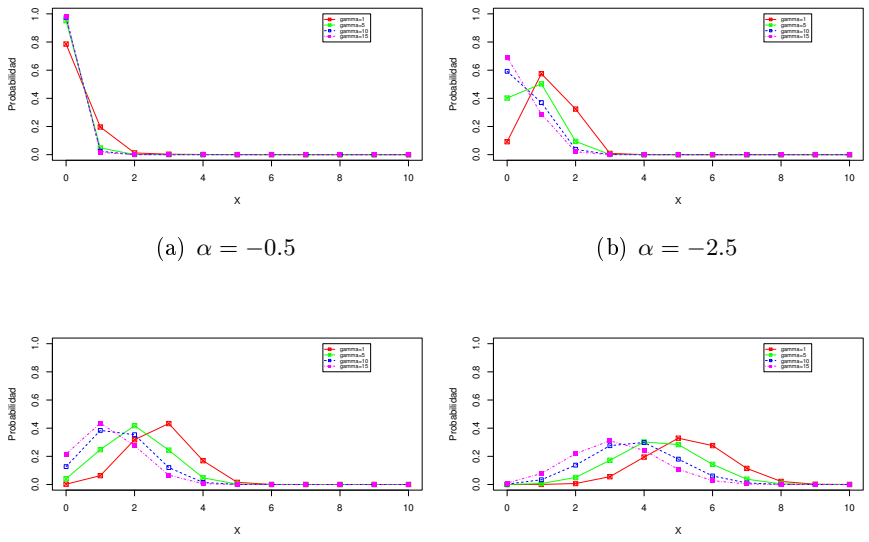
5. Al igual que la distribución CTP, la distribución EBW con $\alpha < 0$ no puede expresarse como una mixtura de una distribución de Poisson, por lo que no puede establecerse en este caso un resultado de descomposición de la varianza.

Por último, en la Figura 3.3 se incluyen algunos perfiles de la distribución EBW con α negativo no entero. Se observa cómo a medida que γ disminuye, se reduce la probabilidad del 0 y la moda se desplaza hacia la derecha; mientras que cuando α disminuye, el perfil pasa de $J-$traspuesto a acampanado, puesto que la moda aumenta.

3.3.3. Caso $\alpha \in \mathbb{Z}^-$

Teniendo en cuenta la expresión de la media dada en (3.11), puede establecerse el siguiente resultado.

Proposición 3.3.2. *La esperanza de la distribución $EBW(n, \gamma)$ con $n \in \mathbb{N}$ y $\gamma > 0$ alcanza un máximo en $\frac{n^2}{2n-1}$. Además, la media es menor que 1 si y sólo si $\gamma > (n-1)^2$.*



(a) $\alpha = -0.5$ (b) $\alpha = -2.5$

(c) $\alpha = -5.5$ (d) $\alpha = -10.5$

Figura 3.3: Perfiles de la distribución EBW con $\alpha < 0$

Demostración. En primer lugar, basta tener en cuenta que μ es una función decreciente de γ. Por tanto, como $\gamma > 0$:

$$\lim_{\gamma \to 0^+} \mu = \lim_{\gamma \to 0^+} \frac{n^2}{\gamma + 2n - 1} = \frac{n^2}{2n-1}.$$

Para demostrar la segunda afirmación, consideramos la expresión de $\mu = \frac{n^2}{\gamma+2n-1}$:

$$\mu < 1 \Leftrightarrow n^2 < \gamma + 2n - 1 \Leftrightarrow n^2 + 2n - 1 = (n-1)^2 < \gamma.$$

□

El IA correspondiente a esta distribución finita está dado por:

$$IA = \frac{\sigma^2}{\mu} = \frac{\mu \frac{\mu+\gamma-1}{\gamma+2n-2}}{\mu} = \frac{\mu + \gamma - 1}{\gamma + 2n - 2} \tag{3.13}$$

Esta distribución es siempre infradispersa, ya que $\alpha \leq -n \leq -1$ y, por tanto, $IA < 1$.

En la Figura 3.4 se observa cómo varía el IA en función de γ para distintos valores de n. El IA crece a medida que γ aumenta. Cabe resaltar, además, que la infradispersión es mayor cuanto mayor es el valor de n. De hecho, puede establecerse el siguiente resultado:

Proposición 3.3.3. *Si* $X \sim EBW(n, \gamma)$,

$$\lim_{\gamma \to \infty} IA = 1, \quad \lim_{\gamma \to 0^+} IA = \frac{1}{4}(n \textit{ fijo}), \quad \lim_{n \to \infty} IA = \frac{1}{4}(\gamma \textit{ fijo}).$$

Demostración. Teniendo en cuenta (3.13) y la expresión de la media dada en (3.11),

$$IA = \frac{n^2 + \gamma^2 + O(n) + O(\gamma)}{4n^2 + \gamma^2 + O(n) + O(\gamma)},$$

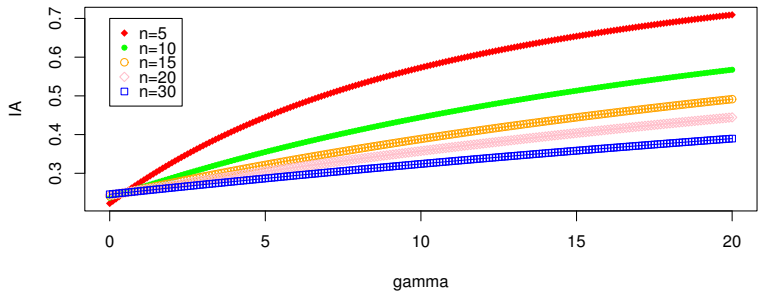


Figura 3.4: IA de la distribución EBW para distintos valores de n y γ

de donde se obtienen dos de los límites mencionados. En cuanto al segundo,

$$\lim_{\gamma \to 0^+} IA = \frac{n^2 - (-1)(2n-1)}{(2n-2)^2} = \frac{n-1)^2}{4(n-1)^2} = \frac{1}{4}$$

□

La Figura 3.5 muestra algunos perfiles de la distribución EBW finita para distintos valores de n y γ, todos ellos acampanados o con forma de J traspuesta. Como se puede observar, la moda disminuye a medida que aumenta el valor de γ.

3.4. Estimación

Tradicionalmente, para abordar el problema de la estimación de los parámetros de las distribuciones discretas que verifican una ecuación funcional, y en las que se pueden encontrar sistemas de ecuaciones en recurrencias para los momentos, el método más "natural" es el de los momentos, considerando momentos respecto al origen o bien momentos factoriales. Así pues, en primer lugar, estimamos los parámetros de la distribución EBW mediante el método de los momentos y, posteriormente, mediante el método de máxima verosimilitud (MV) considerando como valores iniciales las estimaciones anteriores. En el caso $\alpha > 0$ aplicamos también el algoritmo de maximización de la esperanza, ampliamente conocido como algoritmo EM.

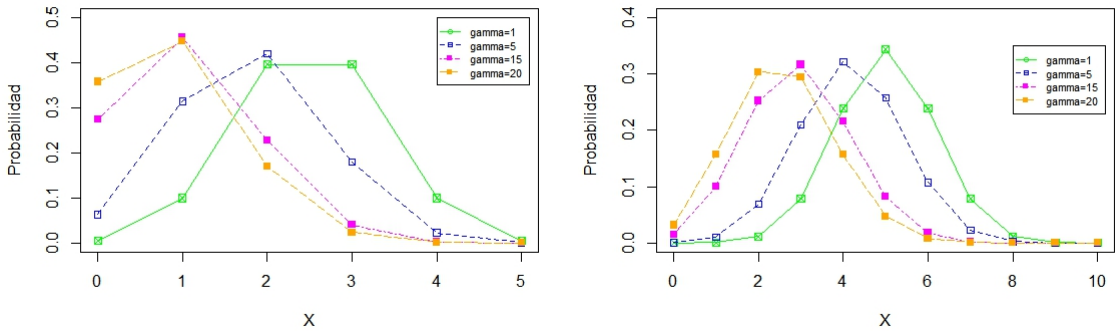
3.4.1. Mediante el método de los momentos

En primer lugar, para aplicar el método de los momentos resolvemos el sistema de ecuaciones dado en (3.3) y (3.4). Para ello, despejamos por ejemplo $\gamma - 2\alpha - 1 = \alpha^2/\mu$ de (3.3) y sustituimos en (3.4). Así, se obtiene

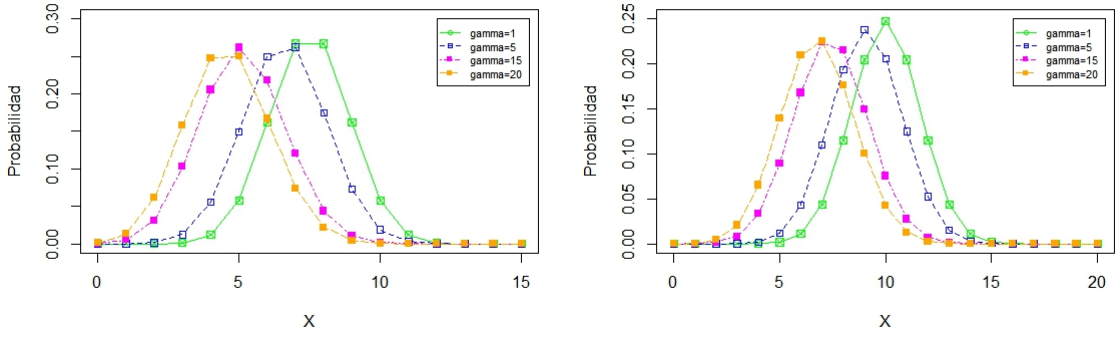
$$\sigma^2 = \mu \frac{(\alpha+\mu)^2}{\alpha^2 - \mu},$$

que equivale a la ecuación en α:

$$\alpha^2(\sigma^2 - \mu) - 2\mu^2\alpha - \mu(\mu^2 + \sigma^2) = 0$$



(a) $n = 5$ (b) $n = 10$



(c) $n = 15$ (d) $n = 20$

Figura 3.5: Perfiles de la distribución EBW de rango finito para distintos valores de n y γ

con dos posibles soluciones

$$\alpha_1 = \frac{\mu^2 + \sqrt{\mu^4 + \mu(\sigma^2 - \mu)(\mu^2 + \sigma^2)}}{\sigma^2 - \mu}, \tag{3.14}$$

$$\alpha_2 = \frac{\mu^2 - \sqrt{\mu^4 + \mu(\sigma^2 - \mu)(\mu^2 + \sigma^2)}}{\sigma^2 - \mu}. \tag{3.15}$$

A continuación, reemplazamos los valores de μ y σ^2 por sus homólogos muestrales, $\overline{x}$ y s^2, obteniendo así dos posibles estimaciones para α mediante el método de los momentos:

$$\widehat{\alpha}_1 = \frac{\overline{x}^2 + \sqrt{\overline{x}^4 + \overline{x}(s^2 - \overline{x})(\overline{x}^2 + s^2)}}{s^2 - \overline{x}},$$

$$\widehat{\alpha}_2 = \frac{\overline{x}^2 - \sqrt{\overline{x}^4 + \overline{x}(s^2 - \overline{x})(\overline{x}^2 + s^2)}}{s^2 - \overline{x}}.$$

Proposición 3.4.1. *Si los datos presentan sobredispersión, $\widehat{\alpha}_1 > 0$ y $\widehat{\alpha}_2 < 0$; mientras que si los datos son infradispersos, $\widehat{\alpha}_1, \widehat{\alpha}_2 < 0$.*

Demostración. El numerador de $\widehat{\alpha}_1$ es siempre positivo, de modo que su signo depende del signo del denominador. Por tanto, si los datos son sobredispersos, esto es, $s^2 - \overline{x} > 0$, entonces $\widehat{\alpha}_1 > 0$; mientras que si los datos son infradispersos, esto es, $s^2 - \overline{x} < 0$, entonces $\widehat{\alpha}_1 < 0$.

En cuando a $\widehat{\alpha}_2$, si los datos son sobredispersos ($s^2 > \overline{x}$), su signo coincide con el signo del numerador. Como $s^2 - \overline{x} > 0$, entonces $\overline{x}^4 + \overline{x}(s^2 - \overline{x})(\overline{x}^2 + s^2) > \overline{x}^4$. Tomando la raiz cuadrada, $\sqrt{\overline{x}^4 + \overline{x}(s^2 - \overline{x})(\overline{x}^2 + s^2)} > \overline{x}^2$, de donde el numerador de $\widehat{\alpha}_2$ es negativo.

Por el contrario, si los datos son infradispersos ($s^2 < \overline{x}$), el signo de $\widehat{\alpha}_2$ es opuesto al signo del numerador. Como $s^2 - \overline{x} < 0$, entonces $\overline{x}^4 + \overline{x}(s^2 - \overline{x})(\overline{x}^2 + s^2) < \overline{x}^4$. Tomando la raiz cuadrada, $\sqrt{\overline{x}^4 + \overline{x}(s^2 - \overline{x})(\overline{x}^2 + s^2)} < \overline{x}^2$, de donde el numerador de $\widehat{\alpha}_2$ es positivo. □

Una vez estimado el parámetro α, la estimación de γ se obtiene como

$$\widehat{\gamma} = \widehat{\alpha}^2/\overline{x} + 2\widehat{\alpha} + 1. \tag{3.16}$$

Proposición 3.4.2. *Existen dos posibles estimaciones para el parámetro γ si y sólo si:*

- $0 < \overline{x} < 1$, *o bien*
- $\overline{x} > 1$ *y* $\widehat{\alpha} < -\overline{x} - \sqrt{\overline{x}(\overline{x} - 1)}$ *o* $\widehat{\alpha} > -\overline{x} + \sqrt{\overline{x}(\overline{x} - 1)}$.

Demostración. La restricción $\widehat{\gamma} > 2\widehat{\alpha}$ se verifica directamente, de modo que la única condición que hay que imponer a $\widehat{\gamma}$ es que sea positivo, esto es,

$$\widehat{\alpha}^2/\overline{x} + 2\widehat{\alpha} + 1 > 0 \Leftrightarrow \widehat{\alpha}^2 + 2\widehat{\alpha}\overline{x} + \overline{x} > 0.$$

Se trata de un polinomio de segundo grado en $\widehat{\alpha}$, cuyas raices existen y son distintas si y solo si $\overline{x}(\overline{x} - 1) > 0 \Leftrightarrow \overline{x} > 1$ y son $-\overline{x} + \sqrt{\overline{x}(\overline{x} - 1)}$ y $-\overline{x} - \sqrt{\overline{x}(\overline{x} - 1)}$. Este polinomio puede verse además como una parábola en $\widehat{\alpha}$ con las ramas hacia arriba, de modo que $\widehat{\gamma} > 0$ si y sólo si $\widehat{\alpha} < -\overline{x} - \sqrt{\overline{x}(\overline{x} - 1)}$ o bien $\widehat{\alpha} > -\overline{x} + \sqrt{\overline{x}(\overline{x} - 1)}$. □

En el caso finito, la estimación de γ mediante el método de los momentos se obtiene reemplazando μ por su homólogo muestral $\overline{x}$ en (3.11), de modo que

$$\widehat{\gamma} = \frac{n^2}{\overline{x}} - 2n + 1.$$

3.4.2. Mediante el método de máxima verosimilitud

Distinguimos aquí entre las dos parametrizaciones consideradas para la distribución EBW con (α, ρ), ambos positivos, o con (α, γ), $\alpha < 0$ y $\gamma > 0$.

Dada una muestra aleatoria simple $x_1, x_2, \ldots, x_n$ de tamaño n, las estimaciones máximo-verosímiles de $\alpha > 0$ y $\rho > 0$ se obtienen maximizando la función de log-verosimilitud dada por:

$$\begin{aligned} \ln L_{x_1,\ldots,x_n}(\alpha, \rho) = cte &+ \sum_{i=1}^{n} \left[2 \ln \Gamma(\alpha + x_i) - \ln \Gamma(\rho + 2\alpha + x_i)\right] \\ &- n \left[2 \ln \Gamma(\alpha) - 2 \ln \Gamma(\rho + \alpha) + \ln \Gamma(\rho)\right], \end{aligned} \tag{3.17}$$

con $cte = -\sum_{i=1}^{n} \ln(x_i!)$.

En el caso, $\alpha < 0$ y $\gamma > 0$, la función de log-verosimilitud tiene la expresión:

$$\begin{aligned} \ln L_{x_1,\ldots,x_n}(\alpha, \gamma) = cte + &\sum_{i=1}^{n} \left[2 \ln \Gamma(\alpha + x_i) - \ln \Gamma(\gamma + x_i)\right] \\ &- n \left[2 \ln \Gamma(\alpha) - 2 \ln \Gamma(\gamma - \alpha) + \ln \Gamma(\gamma - 2\alpha)\right]. \end{aligned} \tag{3.18}$$

Ambas expresiones se maximizan utilizando métodos numéricos que dependen fuertemente de los valores iniciales. En el paquete `cpd` (Olmo-Jiménez et al., 2022) se realiza la estimación máximo-verosímil de los parámetros de la distribución EBW mediante la función `optim()` del paquete `MASS` de R Team (2023) y permite seleccionar como valores iniciales los que proporciona el método de los momentos.

En el caso finito, con $\alpha = -N \in \mathbb{Z}^-$, la función de log-verosimilitud correspondiente es:

$$\ln L_{x_1,\ldots,x_n}(\gamma) = cte + \sum_{i=1}^{n} \left[2 \ln \binom{N}{x_i} - \ln \Gamma(\gamma + x_i)\right] + n \left[2 \ln \Gamma(\gamma + N) - \ln \Gamma(\gamma + 2N)\right].$$

3.4.3. Mediante el algoritmo EM

El algoritmo de maximización de esperanzas (EM) desarrollado por Dempster et al. (1977) es un enfoque muy útil para el cálculo iterativo de estimaciones MV en problemas con datos faltantes. Sin embargo, este algoritmo no solo se puede aplicar a casos incompletos (datos faltantes, distribuciones truncadas, observaciones censuradas o agrupadas), sino también en modelos estadísticos donde no es tan evidente que los datos estén incompletos McLachlan and Krishnan (2008). Este es el caso de las distribuciones que se obtienen como mixturas, ya que se puede considerar que el proceso de mixtura produce datos faltantes.

Sea X una variable aleatoria cuya distribución depende del parámetro θ. Si consideramos que θ es una variable aleatoria, se puede obtener una nueva distribución denominada mixta o distribución mixtura. Específicamente, si $f(x|\theta)$ es la función de densidad de X y θ es una variable aleatoria con función de distribución $G(\theta|\varphi)$, la distribución incondicional de X está dada por

$$f(x|\varphi) = \int_{\theta} f(x|\theta) dG(\theta|\varphi)$$

La distribución de $X|\varphi$ es la distribución mixtura, mientras que la distribución de $\theta|\varphi$ es la distribución mixtante.

En la formulación de la mixtura, los valores no observados son las realizaciones del parámetro no observado θ_i para cada x_i, de modo que el algoritmo EM maximiza la log-verosimilitud $\ln \mathrm{L}(\varphi|x_1,\ldots,x_n) = \ln f(x_1,\ldots,x_n|\varphi)$, maximizando iterativamente

$$E\left[\ln \mathrm{L}(\varphi|\theta_1,\ldots,\theta_n)\right] = \ln g(\theta_1,\ldots,\theta_n|\varphi).$$

Por lo tanto, en el paso E de la iteración $(k+1)$-ésima, se calcula la esperanza

$$E\left[\ln L(\varphi|\theta_1,\ldots,\theta_n)\,|x_1,\ldots,x_n,\varphi^{(k)}\right] = E\left[\ln g(\theta_1,\ldots,\theta_n|\varphi)\,|x_1,\ldots,x_n,\varphi^{(k)}\right] \quad (3.19)$$

con respecto a la distribución condicionada $g\left(\theta|x_1,\ldots,x_n,\varphi^{(k)}\right)$. A continuación, en el paso M, se maximiza sobre φ. En resumen, en el paso E, es necesario calcular la esperanza condicional de algunas funciones de $\theta_1,\ldots,\theta_n$ y luego, en el paso M, maximizar la verosimilitud de la densidad mixtante.

McLachlan y Krishnan McLachlan and Krishnan (2008) establecen el algoritmo EM para mixturas con los siguientes pasos:

Paso E) A partir de las estimaciones $\varphi^{(k)}$ obtenidas en la iteración k-ésima, se calculan los pseudovalores $t_{ij} = E\left[h_j(\theta_i)\,|X_i,\varphi^{(k)}\right]$ for $i = 1,\ldots,n$ y $j = 1,\ldots,m$, donde $h_j(\cdot)$ son determinadas funciones.

Paso M) Utilizar los pseudovalores t_{ij} del paso E para maximizar la verosimilitud de la distribución mixtante y obtener las estimaciones actualizadas $\varphi^{(k+1)}$.

Los pasos E y M se repiten iterativamente hasta que se cumpla una condición de parada.

En el caso de las mixturas que pertenecen a la familia exponencial, las funciones $h_j(\theta)$ coinciden con los estadísticos suficientes necesarios para la estimación MV de la distribución mixtante. De este modo, el cálculo de los pseudovalores t_{ij} es más sencillo, especialmente si los estadísticos suficientes son de la forma θ^r y la distribución de $X|\varphi$ pertenece a la familia de distribuciones de series de potencias Sapatinas (1995).

Aplicación del algoritmo EM a la distribución EBW

Como vimos en la Proposición 3.3.1 la distribución EBW con $\alpha > 0$ puede obtenerse como una mixtura de una distribución binomial negativa y una distribución beta tipo II. Esto permite aplicar el algoritmo EM a esta distribución para obtener estimaciones MV de α y β. El principal inconveniente es que el cálculo de las funciones $h_j(\theta_i)$ no resulta sencillo, al no pertenecer la distribución EBW a la familia exponencial.

Dada una muestra aleatoria $\theta_1,\ldots,\theta_n$, la función de log-verosimilitud correspondiente a $\theta \sim BetaII(\alpha,\rho)$ es

$$\begin{aligned}\ln L(\alpha,\rho|\theta_1,\ldots,\theta_n) = {} & n\ln\Gamma(\alpha+\rho) - n\ln\Gamma(\alpha) - n\ln\Gamma(\rho) \\ & + (\alpha-1)\sum_{i=1}^{n}\ln\theta_i - (\alpha+\rho)\sum_{i=1}^{n}\ln(1+\theta_i).\end{aligned}$$

De este modo, la ecuación (3.19) tiene la expresión

$$\begin{aligned} E_{\alpha^{(k)},\rho^{(k)}}\left[\ln L\left(\alpha,\rho|\theta_1,\ldots,\theta_n\right)|x_1,\ldots,x_n\right] &= n\ln\Gamma(\alpha+\rho)-n\ln\Gamma(\alpha)-n\ln\Gamma(\rho) \\ &\quad+(\alpha-1)\sum_{i=1}^{n}E_{\alpha^{(k)},\rho^{(k)}}\left[\ln\theta_i|x_i\right] \\ &\quad-(\alpha+\rho)\sum_{i=1}^{n}E_{\alpha^{(k)},\rho^{(k)}}\left[\ln\left(1+\theta_i\right)|x_i\right] \end{aligned} \tag{3.20}$$

y, por tanto, $t_{i1}=E_{\alpha^{(k)},\rho^{(k)}}\left[\ln\theta_i|x_i\right]$ y $t_{i2}=E_{\alpha^{(k)},\rho^{(k)}}\left[\ln\left(1+\theta_i\right)|x_i\right]$.

En consecuencia, el algoritmo EM para la distribución EBW con $\alpha>0$ se reduce a los siguientes pasos:

Paso E) Calcular los pseudovalores $t_{ij}=E\left[h_j\left(\theta_i\right)|X_i,\alpha^{(k)},\rho^{(k)}\right]$ para $i=1,\ldots,n$ y $j=1,2$, donde $h_1\left(\theta_i\right)=\ln\theta_i$ y $h_2\left(\theta_i\right)=\ln\left(1+\theta_i\right)$.

Para ello, hemos de tener en cuenta los siguientes resultados.

Proposición 3.4.3. *Dada $X|\theta\sim NB(\alpha,\theta)$ con $\theta\sim BetaII(\alpha,\rho)$, entonces $\theta|X\sim BetaII(X+\alpha,\alpha+\rho)$.*

Demostración. Aplicando el teorema de Bayes,

$$\begin{aligned} \pi(\theta|x) &\propto \pi(x|\theta)\pi(\theta) \\ &\propto \left(\frac{1}{1+\theta}\right)^{\alpha}\left(\frac{\theta}{1+\theta}\right)^{x}\theta^{\alpha-1}(1+\theta)^{-(\alpha+\rho)}=\theta^{(x+\alpha)-1}(1+\theta)^{-[(x+\alpha)+(\alpha+\rho)]}. \end{aligned}$$

Esta expresión es el núcleo de una distribución $BetaII$ con parámetros $x+\alpha$ y $\alpha+\rho$ y, por tanto, $\theta|X\sim BetaII(X+\alpha,\alpha+\rho)$. □

Proposición 3.4.4. *Dada $X\sim BetaII(\alpha,\beta)$, entonces $E(\ln X)=\psi(\alpha)-\psi(\beta)$ y $E\left[\ln(1+X)\right]=\psi(\alpha+\beta)-\psi(\beta)$, siendo $\psi(\cdot)$ la función digamma definida como*

$$\psi(x)=\frac{\partial\ln\Gamma(x)}{\partial x}=\frac{1}{\Gamma(\alpha)}\frac{\partial\Gamma(\alpha)}{\partial\alpha}.$$

Demostración. La función $B(\alpha,\beta)$ se define como

$$B(\alpha,\beta)=\int_0^{\infty}x^{\alpha-1}(1+x)^{-(\alpha+\beta)}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Derivando los dos miembros de la igualdad con respecto a β se tiene que

$$\begin{aligned} -\int_0^{\infty}\ln(1+x)\cdot x^{\alpha-1}(1+x)^{-(\alpha+\beta)}dx &= \Gamma(\alpha)\left[\frac{\Gamma'(\beta)\Gamma(\alpha+\beta)-\Gamma(\beta)\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)^2}\right] \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\left[\psi(\beta)-\psi(\alpha+\beta)\right] \end{aligned} \tag{3.21}$$

La ecuación (3.21) también puede expresarse como

$$\underbrace{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\int_0^{\infty}\ln(1+x)\cdot x^{\alpha-1}(1+x)^{-(\alpha+\beta)}dx}_{E[\ln(1+X)]}=\psi(\alpha+\beta)-\psi(\beta). \tag{3.22}$$

Si derivamos ahora los dos miembros de la igualdad con respecto a α, tenemos

$$\int_0^\infty \left[\ln x \cdot x^{\alpha-1}(1+x)^{-(\alpha+\beta)} - \ln(1+x)x^{\alpha-1}(1+x)^{-(\alpha+\beta)}dx\right] = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\left[\psi(\alpha) - \psi(\alpha+\beta)\right]$$

que puede escribirse como

$$\underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^\infty \left[\ln x \cdot x^{\alpha-1}(1+x)^{-(\alpha+\beta)} - \ln(1+x)x^{\alpha-1}(1+x)^{-(\alpha+\beta)}dx\right]}_{E(\ln X)-E[\ln(1+X)]} = \psi(\alpha) - \psi(\alpha+\beta).$$

De aquí, teniendo en cuenta (3.22), se tiene que $E(\ln X) = \psi(\alpha) - \psi(\beta)$. □

En consecuencia, los pseudovalores son:

$$t_{i1} = \psi\left(x_i + \alpha^{(k)}\right) - \psi\left(\alpha^{(k)} + \rho^{(k)}\right) \tag{3.23}$$

$$t_{i2} = \psi\left(x_i + 2\alpha^{(k)} + \rho^{(k)}\right) - \psi\left(\alpha^{(k)} + \rho^{(k)}\right). \tag{3.24}$$

Paso M) Maximizar la función de verosimilitud de la distribución BetaII usando los pseudovalores t_{i1} y t_{i2} dados en (3.23) y (3.24) del paso E. Esto se puede realizar a través del algoritmo de maximización de esperanza-condicional (ECM) Meng and Rubin (1993), que es una iteración de Newton Raphson de un paso adelante, para resolver las ecuaciones de MV dadas por:

$$\begin{aligned}\psi(\alpha+\rho) - \psi(\alpha) + \bar{t}_1 - \bar{t}_2 &= 0\\ \psi(\alpha+\rho) - \psi(\rho) - \bar{t}_2 &= 0,\end{aligned}$$

donde $\bar{t}_j = \sum_{i=1}^n t_{ij}/n, j = 1, 2$. Por lo tanto, las estimaciones de α y ρ se actualizan de la siguiente forma:

$$\alpha^{(k+1)} = \alpha^{(k)} - \frac{\psi\left(\alpha^{(k)} + \rho^{(k)}\right) - \psi\left(\alpha^{(k)}\right) + \bar{t}_1 - \bar{t}_2}{\psi'\left(\alpha^{(k)} + \rho^{(k)}\right) - \psi'\left(\alpha^{(k)}\right)} \tag{3.25}$$

$$\rho^{(k+1)} = \rho^{(k)} - \frac{\psi\left(\alpha^{(k+1)} + \rho^{(k)}\right) - \psi\left(\alpha^{(k)}\right) - \bar{t}_2}{\psi'\left(\alpha^{(k+1)}+\rho^{(k)}\right) - \psi'\left(\alpha^{(k)}\right)}. \tag{3.26}$$

El algoritmo finaliza cuando se cumple un criterio de parada.

3.5. Similitudes con la UGW

Como ya se mencionó en el capítulo anterior, la estructura de la distribución UGW permite intercambiar los dos primeros parámetros. Además, estos aparecen en forma de producto en la fmp, en los momentos e incluso en la descomposición de la varianza, lo que provoca que, a menudo, las estimaciones máximo-verosímiles de estos parámetros sean prácticamente idénticas. De hecho, dada una distribución $UGW(a, k, \rho)$, existe una distribución $EBW(\alpha, \rho)$ con $\alpha = \sqrt{ak}$ y el mismo parámetro ρ, muy cercana a la primera. Para demostrar esta afirmación, hemos calculado la divergencia máxima de Kullback-Leibler, KL,

(Burnham and Anderson, 2002) entre las dos distribuciones de probabilidad para distintos valores de a, k y ρ, y la hemos representado gráficamente tal y como se muestra en la Figura 3.6. Hemos considerado la misma escala en todos los ejes con el fin de poder compararlos entre sí. En general, se observa que:

1. La divergencia aumenta cuando k se aleja de a.
2. No obstante, esta diferencia es menos relevante a medida que ρ aumenta.
3. En general, la divergencia es muy baja.

Para poner de manifiesto la similitud existente entre estas dos distribuciones, hemos simulado $M = 1000$ muestras de tamaño $N = 100$, $N = 300$ y $N = 500$ de una distribución UGW con distintos valores de sus parámetros y, para cada muestra, hemos estimado los parámetros de las distribuciones UGW y EBW mediante el método de máxima verosimilitud. Para ello, hemos implementado en R nuestras propias funciones para el cálculo de la fmp y de ajuste de la distribución UGW, utilizando la función `optim` del paquete `stats`. Concretamente, hemos usado el método L-BFGS-B que permite restricciones en el espacio paramétrico. En cuanto a la distribución EBW, hemos utilizado las funciones del paquete `cpd`. En ambos casos, se han considerado como valores iniciales las estimaciones obtenidas mediante el método de los momentos. Para cada grupo de 1000 muestras se ha calculado el porcentaje de ajustes obtenidos para el modelo EBW (es decir, en los que se ha alcanzado la convergencia del método de estimación), así como el porcentaje de estos que son *mejores* que los obtenidos para el modelo UGW usando tres métodos de selección:

- El criterio de información de Akaike o AIC (es preferible el modelo con menor AIC).
- El test de bondad de ajuste de la χ^2 para contrastar H_0 : Los datos proceden de un modelo EBW frente a H_1: Los datos no proceden de un modelo EBW, utilizando la función `chisq.test2` del paquete `cpd` de R.
- El test de bondad de ajuste de Kolmogorov-Smirnov para distribuciones discretas Arnold and Emerson (2011) utilizando la función `ks.test` del paquete `dgof` de R.

Los resultados se incluyen en la Tabla 3.1 a excepción de los del test de bondad de ajuste de Kolmogorov-Smirnov, ya que todos los p-valores son mayores que 0.05. Puede observarse que, en la mayoría de los casos, la distribución EBW, con un parámetro menos, modeliza los datos incluso mejor que la propia distribución UGW a partir de la cual se han generado. En consecuencia, la distribución EBW con $\alpha > 0$ se comporta de forma similar a la UGW pero con un grado de libertad más.

	Ajustes EBW obtenidos			$< AIC$			$p-$valor> 0.05		
	N			N			N		
$UGW(a,k,\rho)$	100	300	500	100	300	500	100	300	500
$(0.5,1,2.5)$	94.4	93.4	95.9	92.6	89.3	86.4	95.4	94.1	93.3
$(0.5,10,2.5)$	99.8	100	100	53.4	25.1	9.9	86.1	86.8	90.5
$(0.5,10,20)$	95.2	93.8	94.4	99.1	95.5	89	95.4	95.3	93.3
$(1.5,3,2.5)$	100	100	100	88.3	88.8	87.2	91.2	87.5	84.3
$(1.5,3,25)$	97.9	99	98.5	100	100	99.9	95.9	94.3	93.7
$(1.5,20,25)$	100	100	100	96.9	82.3	74.3	95.5	94.3	92.9
$(4,6,2.5)$	99.9	100	100	90	89.8	91.1	82.8	77	73.9
$(4,6,10)$	100	100	100	96.3	91.1	92.3	93.7	91.4	91.4
$(4,6,50)$	95.8	96.5	97.6	100	99.9	99.8	94	91	91.8

Tabla 3.1: Porcentaje de: (a) ajustes EBW obtenidos a partir de datos generados de una UGW; (b) ajustes EBW con menor AIC que el proporcionado por el ajuste UGW; (c) muestras que proceden de un modelo EBW al 5 % según el contraste de bondad de ajuste de la χ^2

.

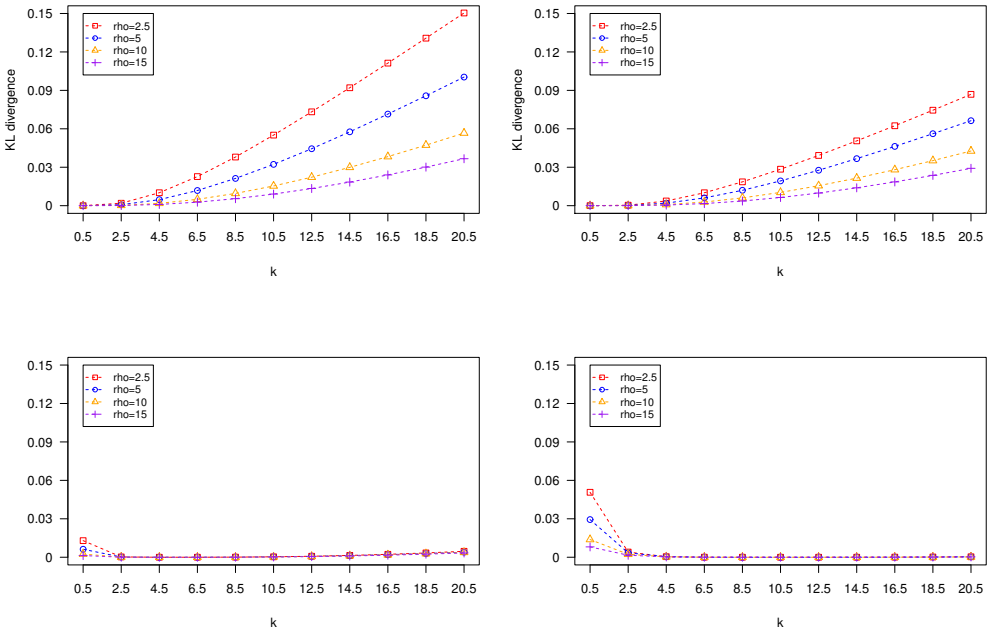


Figura 3.6: Máxima divergencia de Kullback-Leibler entre las distribuciones $UGW(a,k,\rho)$ y $EBW(\sqrt{ak},\rho)$ con $a = 0.5, 1, 5$ y 10 (de izquierda a derecha).

Capítulo 4

Comparación de CTP y EBW con otras distribuciones

4.1. Introducción

En este capítulo se comparan las distribuciones CTP y EBW con las distribuciones usuales para datos de conteo infra y sobredispersas, con objeto de analizar las similitudes y/o diferencias entre los modelos. En primer lugar, nos centramos en la CTP y, posteriormente, en la EBW. En ambos casos la comparación se realizará a través de los perfiles de las distribuciones, de la diferencia en la probabilidad del valor 0 y de la divergencia de Kullback-Leibler. Así pues, el capítulo se estructura en dos grandes secciones: comparación con la CTP y comparación con la EBW, dedicando aquí un breve espacio a la comparativa entre la EBW finita y su homóloga biparamétrica, la distribución BN.

4.2. Comparación de la distribución CTP

Tal y como hemos comentado en el Capítulo 2, la distribución CTP permite modelizar conjuntos de datos tanto infra como sobredispersos. En el caso de una CTP sobredispersa, comparamos con los modelos BN, GP, CBP, UGW, CMP y HP; mientras que si es infradispersa, comparamos con los modelos CMP y HP, ya que la distribución GP, si bien puede ser infradispersa, lo es con rango finito.

4.2.1. A través de la fmp

En primer lugar representamos gráficamente la fmp de cada modelo considerando distintos valores de la media, μ, y la varianza, σ^2. Así, para cada distribución, hemos obtenido los valores de los parámetros en términos de estas dos características, lo que impondrá una serie de restricciones en algunas de estas distribuciones, ya que no siempre será posible que exista esa pareja de momentos. En el caso de las distribuciones CTP y UGW que poseen tres parámetros, hemos considerado fijos γ y k, respectivamente. Así pues,

- Para la $CTP(a, b, \gamma)$ con $a \in \mathbb{R}$, $b > 0$ y $\gamma > \text{máx}(0, 2a)$ fijo:

$$a = \frac{\mu(1-\gamma-\mu)+\sigma^2(\gamma-2)}{2\sigma^2}, \quad b = \sqrt{\mu(\gamma-2a-1)-a^2}.$$

Recordemos que $\gamma > 2a + 2$ para que existan tanto la media como la varianza. Esto equivale a que el parámetro fijo γ sea mayor que $1 - \mu$. Veámoslo:

$$\gamma > 2a + 2 \Leftrightarrow \gamma > 2\frac{\mu(1-\gamma-\mu)+\sigma^2(\gamma-2)}{2\sigma^2} + 2 = \frac{\mu(1-\gamma-\mu)+\sigma^2\gamma}{\sigma^2}$$
$$\Leftrightarrow \mu(1-\gamma-\mu) < 0 \Leftrightarrow 1-\gamma-\mu < 0 \Leftrightarrow \gamma > 1-\mu.$$

Claramente, el parámetro $b > 0$ existe sii el discriminante $\mu(\gamma - 2a - 1) - a^2 \geq 0$. Este discriminante es positivo sii

$$\mu > \frac{(\gamma-2)^2}{4} \quad \text{y} \quad \sigma^2 > \frac{\mu(\mu+\gamma-1)\left[\gamma+2\left(\mu-1-\sqrt{\mu(\mu+\gamma-1)}\right)\right]}{(\gamma-2)^2-4\mu}.$$

En particular, si $\gamma > 2$, b también existe cuando $0 < \mu < \frac{(\gamma-2)^2}{4}$ y σ^2 está comprendido entre

$$\frac{\mu(\mu+\gamma-1)\left[\gamma+2\left(\mu-1-\sqrt{\mu(\mu+\gamma-1)}\right)\right]}{(\gamma-2)^2-4\mu}$$

y

$$\frac{\mu(\mu+\gamma-1)\left[\gamma+2\left(\mu-1+\sqrt{\mu(\mu+\gamma-1)}\right)\right]}{(\gamma-2)^2-4\mu},$$

o bien

$$\mu = \frac{(\gamma-2)^2}{4} \quad \text{y} \quad \sigma^2 > \frac{\mu(\mu+\gamma-1)}{2(2\mu+\gamma-2)}.$$

El discriminante es nulo cuando $0 < \mu < 1, \sigma^2 > \mu(1-\mu)$ y

$$\gamma = \frac{-\mu^3 - 3\mu\sigma^2 + 2\sigma^4 \pm 2\sigma^3\sqrt{\mu(-\mu+\mu^2+\sigma^2)} + \mu^2(1+3\sigma^2)}{(\mu-\sigma^2)^2}$$

o bien cuando $\mu > 1, \sigma^2 > \frac{\mu\left[1-\mu+\sqrt{\mu(1-\mu)}\right]}{2}$ y

$$\gamma = \frac{-\mu^3 - 3\mu\sigma^2 + 2\sigma^4 + 2\sigma^3\sqrt{\mu(-\mu+\mu^2+\sigma^2)} + \mu^2(1+3\sigma^2)}{(\mu-\sigma^2)^2}.$$

- Para la $CBP(b, \gamma)$ con $b, \gamma > 0$

$$\gamma = 2 + \frac{\mu(\mu+1)}{\sigma^2-\mu}, \quad b = \sqrt{\mu(\gamma-1)}$$

$\gamma > 2$, de modo que existe la varianza y, además, puede calcularse el valor de b.

- Para la $UGW(a, k, \rho)$ con $a, k, \rho > 0$:

$$a = \frac{\mu(k\mu+\mu^2+\sigma^2)}{k(\sigma^2-\mu)-\mu^2}, \quad \rho = \frac{k^2\mu-\mu^2-k(\mu-\mu^2-2\sigma^2)}{k(\sigma^2-\mu)-\mu^2}.$$

La condición $a > 0$ se verifica sii

$$k(\sigma^2-\mu)-\mu^2 > \Leftrightarrow k > \frac{\mu^2}{\sigma^2-\mu}.$$

Esta restricción también garantiza que $\rho > 2$ puesto que

$$\rho = \frac{k^2\mu - \mu^2 - k(\mu - \mu^2 - 2\sigma^2)}{k(\sigma^2 - \mu) - \mu^2} > 2$$
$$\Leftrightarrow k^2\,\mu - \mu^2 - k(\mu - \mu^2 - 2\sigma^2) > 2k(\sigma^2 - \mu) - 2\mu^2 \Leftrightarrow k^2\mu - \mu^2 + k\mu + k\mu^2 > 0,$$

de modo que la varianza es finita.

- Para la $BN(k, p)$ con $k > 0$ y $0 < p < 1$

$$p = \frac{\mu}{\sigma^2}, \quad k = \frac{\mu^2}{\sigma^2 - \mu}.$$

- Para la $GP(\theta, \lambda)$ con $\theta > 0$ y máx$(-1, -\theta/m) < \lambda < 1$,

$$\theta = \mu\sqrt{\mu/\sigma^2}, \quad \lambda = 1 - \sqrt{\mu/\sigma^2}. \tag{4.1}$$

- Para la $HP(\gamma, \lambda)$ con $\gamma, \lambda > 0$ no es posible expresar los parámetros en función de μ y σ^2 ya que no existen expresiones explícitas para estos momentos. Por tanto, hemos resuelto las ecuaciones de definición de estos momentos numéricamente utilizando la librería `BB` de R (Varadhan and Gilbert, 2009). En el caso sobredisperso, valores elevados del IA se corresponden también con valores elevados de los parámetros, tal y como puede comprobarse empíricamente en las Tablas 4.1 y 4.2.

  En el caso infradisperso, hay que tener en cuenta que $0 < \mu - \sigma^2 < 1$. Para demostrarlo, utilizamos la relación de recurrencia entre los momentos de la distribución $HP(\gamma, \lambda)$, de modo que

$$\sigma^2 = \lambda + [\lambda - (\gamma - 1)]\,\mu - \mu^2 = \lambda + (\lambda - \gamma)\mu + \mu - \mu^2,$$

  y entonces

$$\mu - \sigma^2 = -\lambda - (\lambda - \gamma)\mu + \mu^2. \tag{4.2}$$

  Ademas, Sáez-Castillo and Conde-Sánchez (2013) demostraron que si $\gamma < 1$, entonces $\mu + (\gamma - 1) < \lambda < \mu \Leftrightarrow \gamma - 1 < \lambda - \mu < 0 \Leftrightarrow \mu - 1 < \lambda - \gamma < \mu - \gamma$. En consecuencia, de acuerdo con (4.2)

$$\mu - \sigma^2 < -\lambda - (\mu - 1)\mu + \mu^2 = -\lambda + \mu < 1 - \gamma < 1.$$

- Para la distribución $CMP(\lambda, v)$ con $\lambda > 0$, $v \geq 0$, tampoco hay expresiones explícitas para μ y σ^2 en términos de los parámetros. Sellers et al. (2011) proponen utilizar las siguientes fórmulas aproximadas

$$\mu \approx \lambda^{1/v} - \frac{v - 1}{2v}, \quad \sigma^2 \approx \frac{1}{v}\lambda^{\frac{1}{v}}. \tag{4.3}$$

  Sin embargo, estas expresiones son menos precisas cuando $v > 1$ (en cuyo caso la distribución es infradispersa) y $\lambda \leq 10^v$. Por tanto, en lugar de usar dichas expresiones, hemos obtenido los parámetros de la distribución CMP resolviendo numéricamente las ecuaciones de definición de los momentos μ y σ^2, mediante la librería `BB` de R. En el caso sobredisperso, valores elevados del IA (esto es, sobredispersión severa) se corresponden con valores de la media y de la varianza también muy elevados. Para

γ	λ	μ	IA	γ	λ	μ	IA
1.10	0.10	0.09	1.00	10.00	30.00	21.00	1.43
1.10	2.00	1.92	1.03	10.00	32.00	23.00	1.39
1.10	4.00	3.90	1.02	10.00	34.00	25.00	1.36
1.10	6.00	5.90	1.02	10.00	36.00	27.00	1.33
1.10	8.00	7.90	1.01	10.00	38.00	29.00	1.31
1.10	10.00	9.90	1.01	10.00	40.00	31.00	1.29
1.10	12.00	11.90	1.01	15.00	0.10	0.01	1.01
1.10	14.00	13.90	1.01	15.00	2.00	0.15	1.13
1.10	16.00	15.90	1.01	15.00	4.00	0.34	1.28
1.10	18.00	17.90	1.01	15.00	6.00	0.60	1.45
1.10	20.00	19.90	1.01	15.00	8.00	0.93	1.65
1.10	22.00	21.90	1.00	15.00	10.00	1.38	1.87
1.10	24.00	23.90	1.00	15.00	12.00	1.98	2.09
1.10	26.00	25.90	1.00	15.00	14.00	2.77	2.28
1.10	28.00	0.00	1.00	15.00	16.00	3.79	2.42
1.10	30.00	0.00	1.00	15.00	18.00	5.07	2.48
1.10	32.00	0.00	1.00	15.00	20.00	6.58	2.46
1.10	34.00	0.00	1.00	15.00	22.00	8.29	2.37
1.10	36.00	0.00	1.00	15.00	24.00	10.13	2.24
1.10	38.00	0.00	1.00	15.00	26.00	12.05	2.10
1.10	40.00	0.00	1.00	15.00	28.00	14.02	1.98
5.00	0.10	0.02	1.01	15.00	30.00	16.01	1.87
5.00	2.00	0.53	1.28	15.00	32.00	18.00	1.78
5.00	4.00	1.38	1.52	15.00	34.00	20.00	1.70
5.00	6.00	2.63	1.65	15.00	36.00	22.00	1.64
5.00	8.00	4.24	1.65	15.00	38.00	24.00	1.58
5.00	10.00	6.08	1.57	15.00	40.00	26.00	1.54
5.00	12.00	8.02	1.47	20.00	0.10	0.01	1.00
5.00	14.00	10.01	1.39	20.00	2.00	0.11	1.10
5.00	16.00	12.00	1.33	20.00	4.00	0.24	1.21
5.00	18.00	14.00	1.29	20.00	6.00	0.41	1.34
5.00	20.00	16.00	1.25	20.00	8.00	0.61	1.49
5.00	22.00	18.00	1.22	20.00	10.00	0.87	1.66
5.00	24.00	20.00	1.20	20.00	12.00	1.19	1.85
5.00	26.00	22.00	1.18	20.00	14.00	1.61	2.06
5.00	28.00	24.00	1.17	20.00	16.00	2.15	2.27
5.00	30.00	26.00	1.15	20.00	18.00	2.85	2.47
5.00	32.00	28.00	1.14	20.00	20.00	3.73	2.64
5.00	34.00	30.00	1.13	20.00	22.00	4.82	2.74
5.00	36.00	32.00	1.13	20.00	24.00	6.13	2.78
5.00	38.00	0.00	1.00	20.00	26.00	7.65	2.74
5.00	40.00	0.00	1.00	20.00	28.00	9.35	2.65
10.00	0.10	0.01	1.01	20.00	30.00	11.17	2.51
10.00	2.00	0.24	1.18	20.00	32.00	13.08	2.37
10.00	4.00	0.57	1.39	20.00	34.00	15.03	2.23
10.00	6.00	1.06	1.63	20.00	36.00	17.01	2.10
10.00	8.00	1.74	1.86	20.00	38.00	19.01	1.99
10.00	10.00	2.69	2.03	20.00	40.00	21.00	1.90
10.00	12.00	3.93	2.12	25.00	0.10	0.00	1.00
10.00	14.00	5.45	2.11	25.00	2.00	0.09	1.08
10.00	16.00	7.20	2.03	25.00	4.00	0.19	1.17
10.00	18.00	9.08	1.91	25.00	6.00	0.31	1.27
10.00	20.00	11.03	1.79	25.00	8.00	0.45	1.39
10.00	22.00	13.01	1.68	25.00	10.00	0.62	1.52
10.00	24.00	15.00	1.60	25.00	12.00	0.83	1.67
10.00	26.00	17.00	1.53	25.00	14.00	1.08	1.83
10.00	28.00	19.00	1.47	25.00	16.00	1.40	2.02

Tabla 4.1: Valores del IA para $hP(\gamma, \lambda)$ con $\gamma > 1$ y $\lambda > 0$

γ	λ	μ	IA	γ	λ	μ	IA
25.00	18.00	1.80	2.22	40.00	6.00	0.17	1.16
25.00	20.00	2.29	2.42	40.00	8.00	0.25	1.23
25.00	22.00	2.92	2.62	40.00	10.00	0.33	1.30
25.00	24.00	3.69	2.80	40.00	12.00	0.42	1.38
25.00	26.00	4.65	2.94	40.00	14.00	0.52	1.47
25.00	28.00	5.80	3.02	40.00	16.00	0.63	1.56
25.00	30.00	7.15	3.04	40.00	18.00	0.77	1.67
25.00	32.00	8.69	2.99	40.00	20.00	0.92	1.79
25.00	34.00	10.39	2.88	40.00	22.00	1.10	1.92
25.00	36.00	12.20	2.75	40.00	24.00	1.31	2.06
25.00	38.00	14.10	2.59	40.00	26.00	1.55	2.22
25.00	40.00	16.05	2.45	40.00	28.00	1.84	2.39
30.00	0.10	0.00	1.00	40.00	30.00	2.18	2.57
30.00	2.00	0.07	1.07	40.00	32.00	2.59	2.76
30.00	4.00	0.15	1.14	40.00	34.00	3.08	2.96
30.00	6.00	0.25	1.22	40.00	36.00	3.67	3.15
30.00	8.00	0.35	1.32	40.00	38.00	4.37	3.33
30.00	10.00	0.48	1.42	40.00	40.00	5.20	3.48
30.00	12.00	0.63	1.54	45.00	0.10	0.00	1.00
30.00	14.00	0.80	1.67	45.00	2.00	0.05	1.04
30.00	16.00	1.01	1.82	45.00	4.00	0.10	1.09
30.00	18.00	1.26	1.98	45.00	6.00	0.15	1.14
30.00	20.00	1.57	2.16	45.00	8.00	0.21	1.20
30.00	22.00	1.95	2.35	45.00	10.00	0.28	1.26
30.00	24.00	2.41	2.55	45.00	12.00	0.36	1.33
30.00	26.00	2.98	2.75	45.00	14.00	0.44	1.40
30.00	28.00	3.68	2.94	45.00	16.00	0.53	1.48
30.00	30.00	4.53	3.10	45.00	18.00	0.64	1.57
30.00	32.00	5.55	3.21	45.00	20.00	0.76	1.67
30.00	34.00	6.76	3.27	45.00	22.00	0.89	1.78
30.00	36.00	8.15	3.27	45.00	24.00	1.05	1.90
30.00	38.00	9.71	3.20	45.00	26.00	1.22	2.02
30.00	40.00	11.41	3.09	45.00	28.00	1.43	2.17
35.00	0.10	0.00	1.00	45.00	30.00	1.67	2.32
35.00	2.00	0.06	1.06	45.00	32.00	1.95	2.49
35.00	4.00	0.13	1.12	45.00	34.00	2.28	2.66
35.00	6.00	0.20	1.19	45.00	36.00	2.66	2.85
35.00	8.00	0.29	1.27	45.00	38.00	3.12	3.04
35.00	10.00	0.39	1.35	45.00	40.00	3.67	3.23
35.00	12.00	0.50	1.45	50.00	0.10	0.00	1.00
35.00	14.00	0.63	1.55	50.00	2.00	0.04	1.04
35.00	16.00	0.78	1.67	50.00	4.00	0.09	1.08
35.00	18.00	0.96	1.80	50.00	6.00	0.14	1.13
35.00	20.00	1.17	1.95	50.00	8.00	0.19	1.18
35.00	22.00	1.42	2.11	50.00	10.00	0.25	1.23
35.00	24.00	1.71	2.28	50.00	12.00	0.31	1.29
35.00	26.00	2.07	2.47	50.00	14.00	0.38	1.35
35.00	28.00	2.51	2.66	50.00	16.00	0.46	1.42
35.00	30.00	3.03	2.86	50.00	18.00	0.54	1.50
35.00	32.00	3.67	3.05	50.00	20.00	0.64	1.58
35.00	34.00	4.44	3.22	50.00	22.00	0.75	1.67
35.00	36.00	5.36	3.36	50.00	24.00	0.87	1.77
35.00	38.00	6.44	3.45	50.00	26.00	1.00	1.88
35.00	40.00	7.70	3.49	50.00	28.00	1.16	1.99
40.00	0.10	0.00	1.00	50.00	30.00	1.34	2.12
40.00	2.00	0.05	1.05	50.00	32.00	1.54	2.26
40.00	4.00	0.11	1.10	50.00	34.00	1.77	2.41

Tabla 4.2: Valores del IA para $hP(\gamma, \lambda)$ con $\gamma > 1$ y $\lambda > 0$ (continuación)

comprobar esto, basta fijarse en la expresión aproximada de la media, de modo que, si $v < 1$, se tiene que

$$\lim_{v \to 0^+} \lambda^{1/v} - \frac{v-1}{2v} = \infty,$$

tanto si λ es menor como mayor que 1. Así, por ejemplo, si $v = 0.1$ la media es, aproximadamente, del orden de $\lambda^{10} + 4.5 > 4.5$ (si $\lambda = 0.9, \mu \approx 4.8487$; pero si $\lambda = 1.1, \mu \approx 7.0937$).

Por otra parte, si consideramos la expresión aproximada del IA dada en (1.23) para $v < 1$, se observa que es una función decreciente de v, de modo que si v decrece el IA crece. En consecuencia, valores elevados del IA se corresponden con valores del parámetro $v < 1$ muy pequeños. De hecho,

$$\lim_{\lambda \to \infty} \frac{2\lambda^{1/v}}{2v\lambda^{1/v} - v + 1} = \frac{1}{v}$$

A modo de ilustración, incluimos los resultados de las Tablas 4.3 y 4.4. Podemos ver que para obtener valores de IA cercanos a 10, el parámetro v toma valores menores o iguales que 0.1 y la media correspondiente está por encima de 9765629.5.

Las Figuras 4.1-4.6 contienen los perfiles de la fmp de las distribuciones mencionadas para distintos valores de μ y σ^2 en un escenario sobredisperso ($\sigma^2 > \mu$). No hemos representado las distribuciones CMP ni HP en los casos de fuerte sobredispersión, debido a los inconvenientes discutidos con anterioridad. En la Figura 4.6 se han considerado distintos valores de k para garantizar la existencia de la media y la varianza. En general, se observa que el perfil de la CTP se diferencia de forma notable, mientras que los restantes modelos presentan un comportamiento parecido. La diferencia más significativa se aprecia en la moda de la distribución CTP: mientras que las demás distribuciones mantienen la moda en 0 o en 1, la CTP presenta una moda más elevada. Así, en tanto las demás distribuciones adoptan con frecuencia un perfil de $J-$traspuesta, la CTP tiende a adoptar el perfil acampanado con más rapidez, ya que su moda se desplaza hacia la derecha (crece) a medida que aumenta la media y la sobredispersión.

Por otra parte, cabe destacar que - en todos los casos - el modelo CTP presenta una probabilidad del valor 0 inferior a la de los restantes modelos y que se parece mucho a la probabilidad del 0 de la distribución de Poisson con la misma media, tal y como se observa en la Tabla 4.5, pudiendo incluso ser inferior a ella para algunas combinaciones de los parámetros.

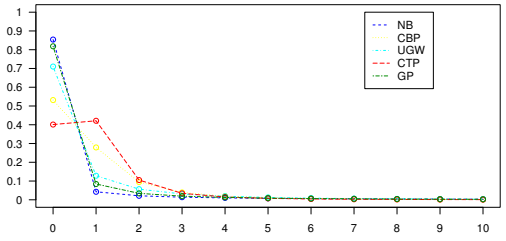
Del mismo modo, en la Figura 4.7 se incluyen perfiles infradispersos de las distribuciones HP, CMP y CTP teniendo en cuenta las condiciones para la obtención de los parámetros en función de la media y la varianza. Cada columna se corresponde con IA 0.4, 0.75 y 0.95, respectivamente. Se ha omitido la distribución HP en aquellos casos en los que no se cumple que $\mu - \sigma^2 < 1$, puesto que no puede darse la pareja de valores (μ, σ^2). Se observa que cuando hay fuerte infradispersión apenas hay diferencias entre los modelos, especialmente cuando la media es mayor. Estas diferencias se hacen más evidentes a medida que aumenta el IA, aunque los tres modelos, en general, son muy similares. La ventaja, por tanto, de la distribución CTP con respecto a la CMP y HP es la existencia de expresiones explícitas de la fmp y de los momentos en términos de sus parámetros, sin restricciones tan severas en cuanto al IA.

λ	v	μ	IA	λ	v	μ	IA
0.25	0.10	0.32	1.27	1.50	0.10	62.26	9.24
0.25	0.20	0.31	1.21	1.50	0.20	9.80	3.83
0.25	0.30	0.29	1.17	1.50	0.30	5.17	2.49
0.25	0.40	0.29	1.14	1.50	0.40	3.60	1.92
0.25	0.50	0.28	1.11	1.50	0.50	2.82	1.60
0.25	0.60	0.27	1.08	1.50	0.60	2.34	1.40
0.25	0.70	0.26	1.06	1.50	0.70	2.03	1.26
0.25	0.80	0.26	1.03	1.50	0.80	1.80	1.15
0.25	0.90	0.25	1.02	1.50	0.90	1.63	1.07
0.50	0.10	0.86	1.69	1.75	0.10	273.89	9.84
0.50	0.20	0.77	1.51	1.75	0.20	18.52	4.39
0.50	0.30	0.71	1.38	1.75	0.30	7.74	2.75
0.50	0.40	0.66	1.29	1.75	0.40	4.89	2.05
0.50	0.50	0.62	1.21	1.75	0.50	3.63	1.68
0.50	0.60	0.59	1.15	1.75	0.60	2.92	1.44
0.50	0.70	0.56	1.11	1.75	0.70	2.47	1.28
0.50	0.80	0.54	1.07	1.75	0.80	2.16	1.16
0.50	0.90	0.52	1.03	1.75	0.90	1.93	1.07
0.75	0.10	1.94	2.43	2.00	0.10	1028.50	9.96
0.75	0.20	1.52	1.90	2.00	0.20	34.04	4.69
0.75	0.30	1.29	1.62	2.00	0.30	11.32	2.94
0.75	0.40	1.14	1.44	2.00	0.40	6.48	2.16
0.75	0.50	1.03	1.32	2.00	0.50	4.55	1.74
0.75	0.60	0.95	1.23	2.00	0.60	3.55	1.48
0.75	0.70	0.88	1.15	2.00	0.70	2.93	1.30
0.75	0.80	0.83	1.09	2.00	0.80	2.52	1.17
0.75	0.90	0.79	1.04	2.00	0.90	2.22	1.08
1.00	0.10	4.62	3.85	2.25	0.10	3329.76	9.99
1.00	0.20	2.82	2.44	2.25	0.20	59.68	4.83
1.00	0.30	2.14	1.90	2.25	0.30	16.14	3.07
1.00	0.40	1.76	1.61	2.25	0.40	8.39	2.24
1.00	0.50	1.53	1.42	2.25	0.50	5.61	1.79
1.00	0.60	1.36	1.29	2.25	0.60	4.23	1.51
1.00	0.70	1.24	1.19	2.25	0.70	3.42	1.32
1.00	0.80	1.14	1.11	2.25	0.80	2.90	1.18
1.00	0.90	1.06	1.05	2.25	0.90	2.52	1.08
1.25	0.10	14.27	6.61	2.50	0.10	9541.24	10.00
1.25	0.20	5.21	3.11	2.50	0.20	99.67	4.90
1.25	0.30	3.37	2.20	2.50	0.30	22.40	3.15
1.25	0.40	2.57	1.77	2.50	0.40	10.67	2.30
1.25	0.50	2.12	1.52	2.50	0.50	6.78	1.83
1.25	0.60	1.83	1.35	2.50	0.60	4.97	1.53
1.25	0.70	1.62	1.23	2.50	0.70	3.94	1.34
1.25	0.80	1.46	1.13	2.50	0.80	3.28	1.19
1.25	0.90	1.35	1.06	2.50	0.90	2.83	1.08

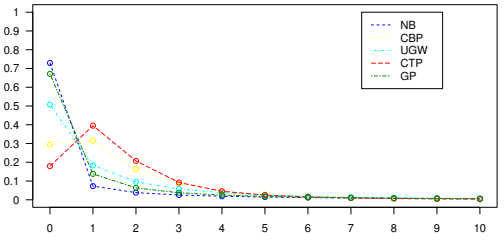
Tabla 4.3: Valores del IA para $CMP(\lambda, v)$ con $\lambda > 0$ y $v < 1$

λ	v	μ	IA	λ	v	μ	IA
2.75	0.10	24740.36	10.00	4.00	0.10	1048580.50	10.00
2.75	0.20	159.28	4.94	4.00	0.20	1026.00	4.99
2.75	0.30	30.32	3.20	4.00	0.30	102.76	3.19
2.75	0.40	13.32	2.35	4.00	0.40	32.76	2.44
2.75	0.50	8.09	1.86	4.00	0.50	16.51	1.94
2.75	0.60	5.75	1.55	4.00	0.60	10.42	1.61
2.75	0.70	4.47	1.35	4.00	0.70	7.47	1.38
2.75	0.80	3.68	1.20	4.00	0.80	5.79	1.22
2.75	0.90	3.14	1.09	4.00	0.90	4.72	1.10
3.00	0.10	59053.50	10.00	4.25	0.10	1922606.10	10.00
3.00	0.20	245.00	4.96	4.25	0.20	1388.58	4.99
3.00	0.30	40.12	3.23	4.25	0.30	125.51	3.30
3.00	0.40	16.36	2.38	4.25	0.40	37.99	2.45
3.00	0.50	9.52	1.88	4.25	0.50	18.57	1.94
3.00	0.60	6.59	1.57	4.25	0.60	11.49	1.62
3.00	0.70	5.03	1.36	4.25	0.70	8.12	1.39
3.00	0.80	4.08	1.20	4.25	0.80	6.23	1.22
3.00	0.90	3.45	1.09	4.25	0.90	5.05	1.10
3.25	0.10	131476.60	10.00	4.50	0.10	3405067.39	10.00
3.25	0.20	364.59	4.97	4.50	0.20	1847.28	4.99
3.25	0.30	52.02	3.26	4.50	0.30	151.61	3.31
3.25	0.40	19.81	2.40	4.50	0.40	43.71	2.46
3.25	0.50	11.08	1.90	4.50	0.50	20.76	1.95
3.25	0.60	7.48	1.58	4.50	0.60	12.61	1.62
3.25	0.70	5.61	1.37	4.50	0.70	8.79	1.39
3.25	0.80	4.50	1.21	4.50	0.80	6.68	1.22
3.25	0.90	3.76	1.09	4.50	0.90	5.38	1.10
3.50	0.10	275859.24	10.00	4.75	0.10	5847044.92	10.00
3.50	0.20	527.22	4.98	4.75	0.20	2420.07	5.00
3.50	0.30	66.27	3.27	4.75	0.30	181.32	3.31
3.50	0.40	23.68	2.42	4.75	0.40	49.93	2.46
3.50	0.50	12.76	1.92	4.75	0.50	23.07	1.96
3.50	0.60	8.41	1.59	4.75	0.60	13.76	1.62
3.50	0.70	6.21	1.37	4.75	0.70	9.48	1.39
3.50	0.80	4.92	1.21	4.75	0.80	7.14	1.23
3.50	0.90	4.08	1.09	4.75	0.90	5.71	1.10
3.75	0.10	549941.17	10.00	5.00	0.10	9765629.50	10.00
3.75	0.20	743.58	4.99	5.00	0.20	3127.00	5.00
3.75	0.30	83.10	3.29	5.00	0.30	214.91	3.32
3.75	0.40	27.99	2.43	5.00	0.40	56.66	2.47
3.75	0.50	14.57	1.93	5.00	0.50	25.51	1.96
3.75	0.60	9.40	1.60	5.00	0.60	14.96	1.63
3.75	0.70	6.83	1.38	5.00	0.70	10.19	1.40
3.75	0.80	5.35	1.22	5.00	0.80	7.61	1.23
3.75	0.90	4.40	1.09	5.00	0.90	6.04	1.10

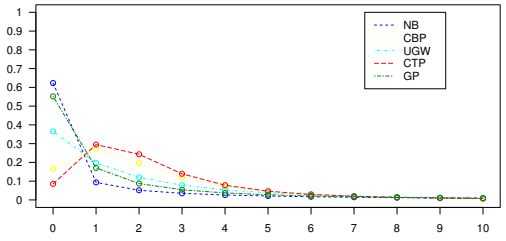
Tabla 4.4: Valores del IA para $CMP(\lambda, v)$ con $\lambda > 0$ y $v < 1$ (continuación)



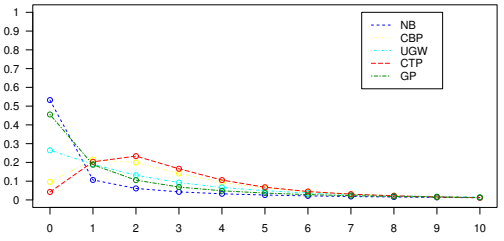
(a) $\mu = 1$



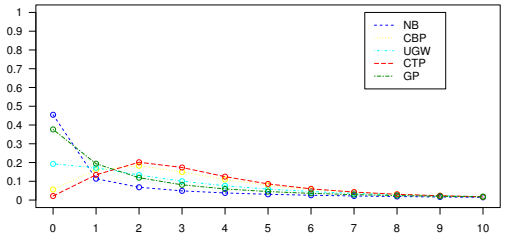
(b) $\mu = 2$



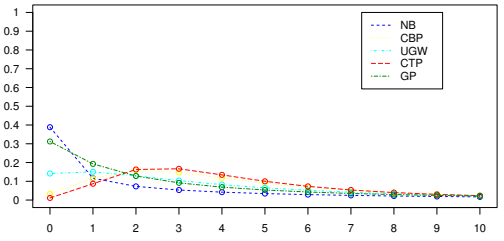
(c) $\mu = 3$



(d) $\mu = 4$

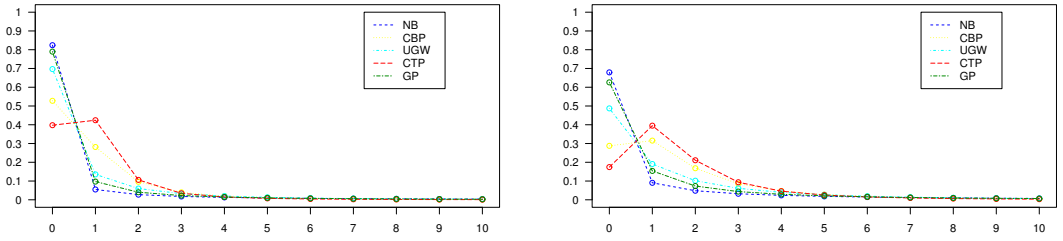


(e) $\mu = 5$

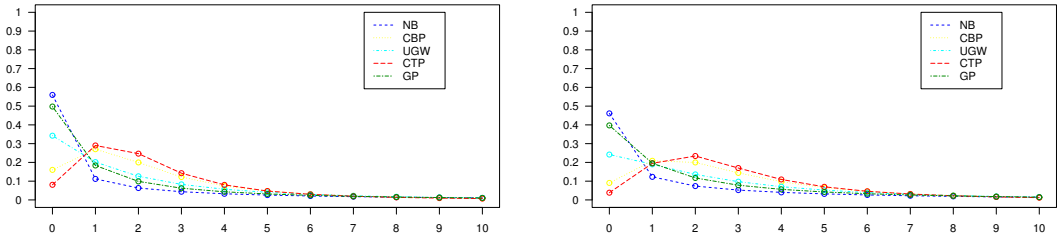


(f) $\mu = 6$

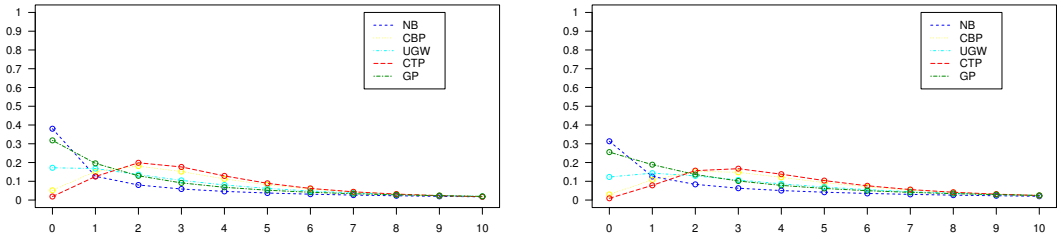
Figura 4.1: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y $UGW(\cdot, 5, \cdot)$ para distintos valores de μ e $IA = 20$



(a) $\mu = 1$ (b) $\mu = 2$

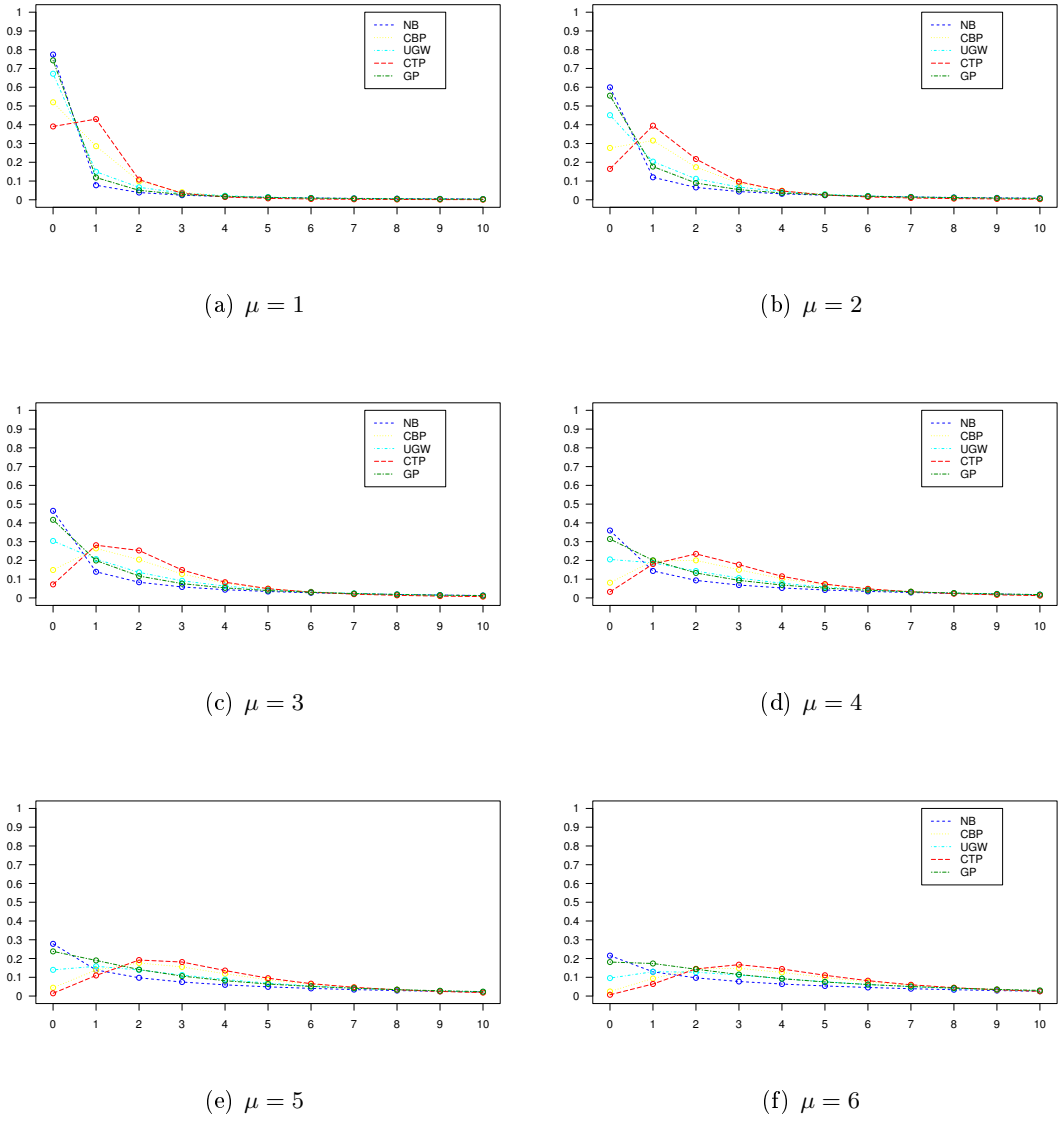


(c) $\mu = 3$ (d) $\mu = 4$



(e) $\mu = 5$ (f) $\mu = 6$

Figura 4.2: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y $UGW(\cdot, 5, \cdot)$ para distintos valores de μ e $IA = 15$

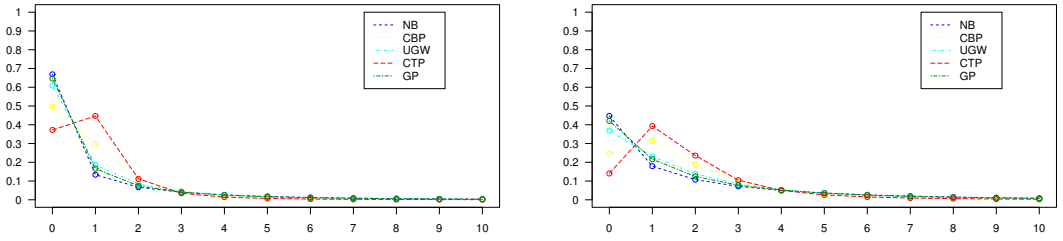


(a) $\mu = 1$ (b) $\mu = 2$

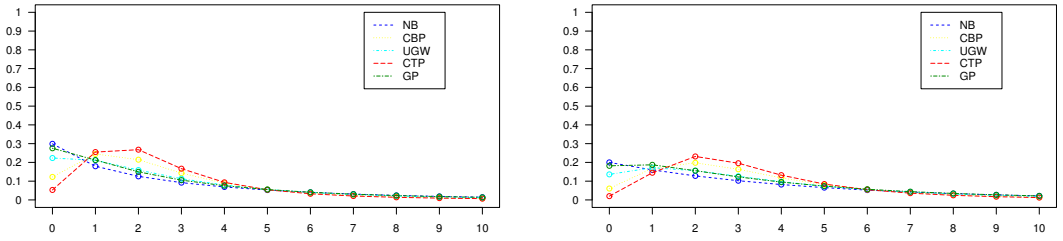
(c) $\mu = 3$ (d) $\mu = 4$

(e) $\mu = 5$ (f) $\mu = 6$

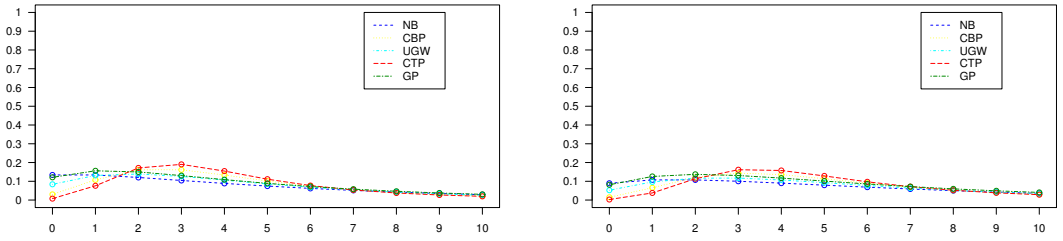
Figura 4.3: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y $UGW(\cdot, 5, \cdot)$ para distintos valores de μ e $IA = 10$



(a) $\mu = 1$ (b) $\mu = 2$

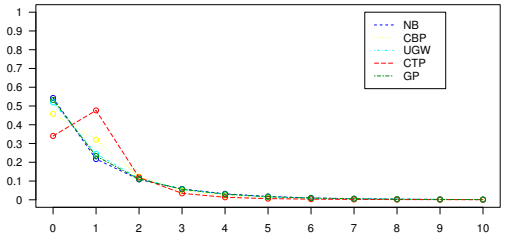


(c) $\mu = 3$ (d) $\mu = 4$

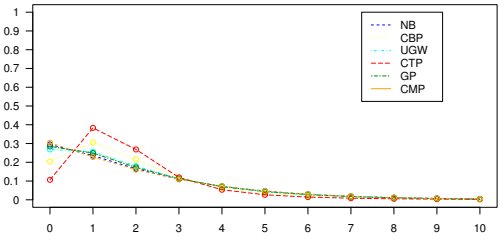


(e) $\mu = 5$ (f) $\mu = 6$

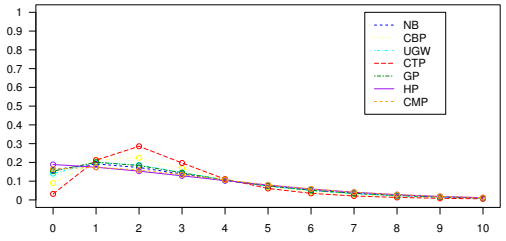
Figura 4.4: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y $UGW(\cdot, 5, \cdot)$ para distintos valores de μ e $IA = 5$



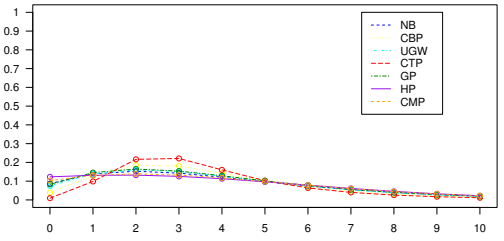
(a) $\mu = 1$



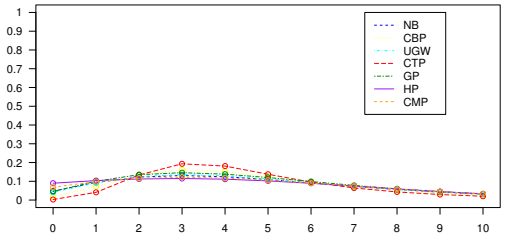
(b) $\mu = 2$



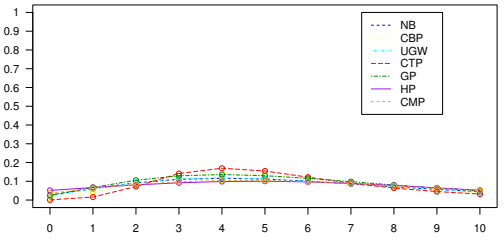
(c) $\mu = 3$



(d) $\mu = 4$

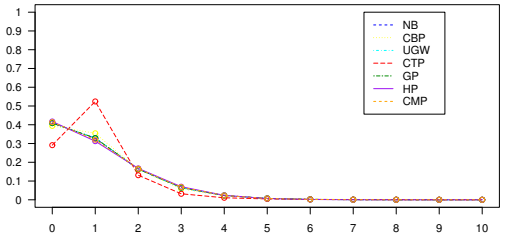


(e) $\mu = 5$

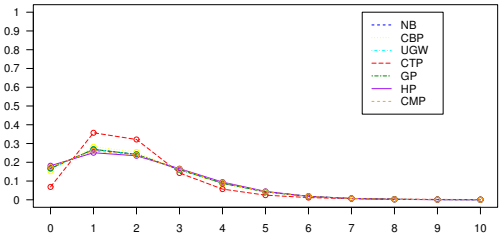


(f) $\mu = 6$

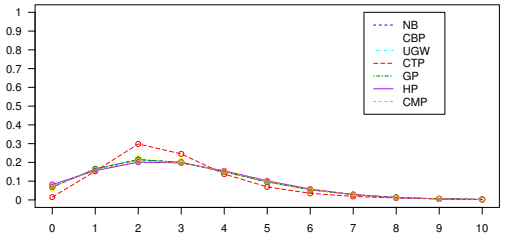
Figura 4.5: Perfiles de las distribuciones NB, CBP, $CTP(\cdot,\cdot,1)$, GP, HP, CMP y $UGW(\cdot,5,\cdot)$ para distintos valores de μ e $IA = 2.5$



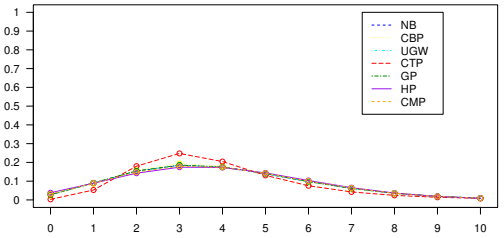
(a) $\mu = 1, k = 5$



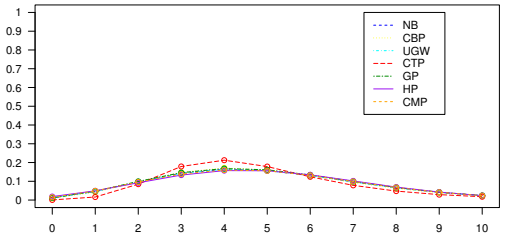
(b) $\mu = 2, k = 10$



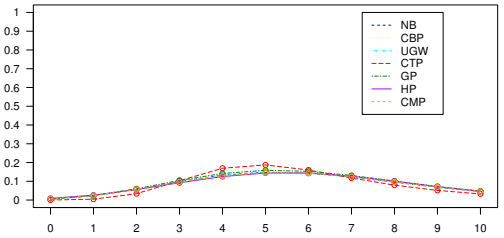
(c) $\mu = 3, k = 15$



(d) $\mu = 4, k = 20$

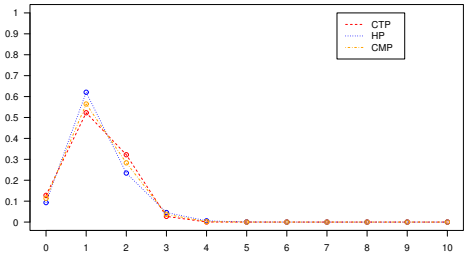


(e) $\mu = 5, k = 25$

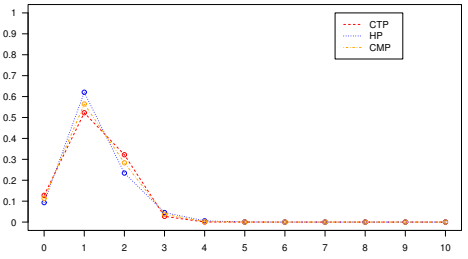


(f) $\mu = 6, k = 25$

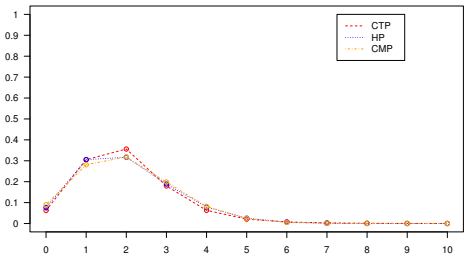
Figura 4.6: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP, HP, CMP y $UGW(\cdot, k, \cdot)$ para distintos valores de μ e $IA = 1.25$



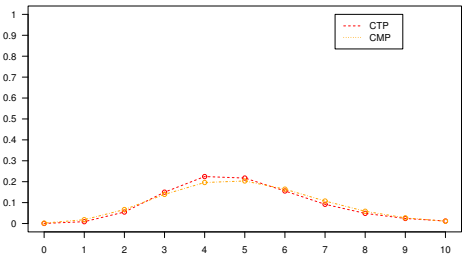
(a) $\mu = 1.25$



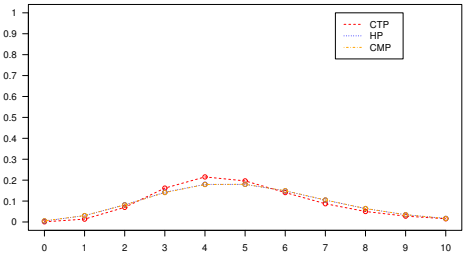
(b) $\mu = 2$



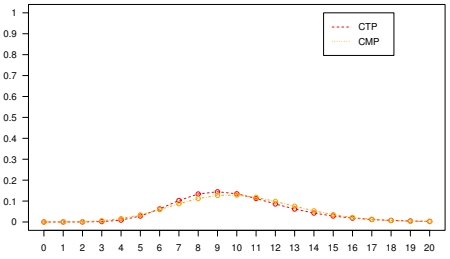
(c) $\mu = 2$



(d) $\mu = 5$



(e) $\mu = 5$



(f) $\mu = 10$

Figura 4.7: Perfiles de las distribuciones $CTP(\cdot, \cdot, 2.1)$, HP y CMP para distintos valores de μ y σ^2 en un escenario infradisperso

μ	σ^2	Poisson	CTP	μ	σ^2	Poisson	CTP
0.5	10	0.60653	0.72611	0.5	0.3	0.60653	0.52270
1.5	10	0.22313	0.34316	1.5	1	0.22313	0.12317
2.5	10	0.08208	0.12985	2.5	5	0.08208	0.08064
3.5	10	0.03020	0.03876	3.5	5	0.03020	0.01611
4.5	10	0.01111	0.00933	4.5	5	0.01111	0.00246
5.5	10	0.00409	0.00187	5.5	5	0.00409	0.00030
6.5	10	0.00150	0.00032	6.5	5	0.00150	0.00003
7.5	10	0.00055	0.00005	7.5	5	0.00055	0.00000
8.5	10	0.00020	0.00001	8.5	5	0.00020	0.00000
9.5	10	0.00007	0.00000	9.5	5	0.00007	0.00000

Tabla 4.5: $P(X = 0)$ en una Poisson y $CTP(\cdot, \cdot, 2.1)$ ambas con idéntica media μ y distintos valores de σ^2 para la CTP

4.2.2. A través de la divergencia de Kullback-Leibler

Otro criterio para estudiar la diferencia entre las distribuciones BN, UGW, GP, CBP, HP, CMP y CTP es mediante la divergencia de Kullback-Leibler (en adelante, KL), ya empleada para comparar la EBW y la UGW en el Capítulo 3.

La Figura 4.8 muestra los valores de la divergencia de KL entre la distribución CTP y las distribuciones NB, CBP, GP, UGW, HP y CMP y viceversa, en el caso sobredisperso, en términos de σ^2, y fijando diversos valores de μ y k. Además, para la CTP el valor de γ es 1.

En el caso infradisperso, se calcula la divergencia KL entre la distribución CTP y las distribuciones CMP y HP y viceversa, en función de σ^2, y fijando ciertos valores de μ. Los resultados se muestran en la Figura 4.9.

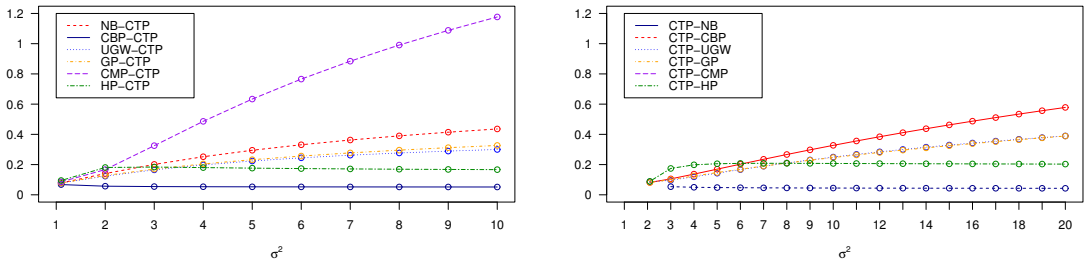
En general, se observa que las distancias con respecto a la CTP aumentan a medida que aumenta σ^2. Además:

- cuando hay sobredispersión, las distribuciones CTP y HP son las que más distan entre sí, seguidas de la CMP, NB, UGW y GP (las dos últimas prácticamente a la misma distancia).
- cuando hay infradispersión, tanto la HP como la CMP están muy cerca de la CTP, puesto que las distancias son menores que 0.1. Sin embargo, la distribución CTP está ligeramente más cerca de la CTP que la HP, si bien esta diferencia disminuye a medida que σ^2 aumenta.

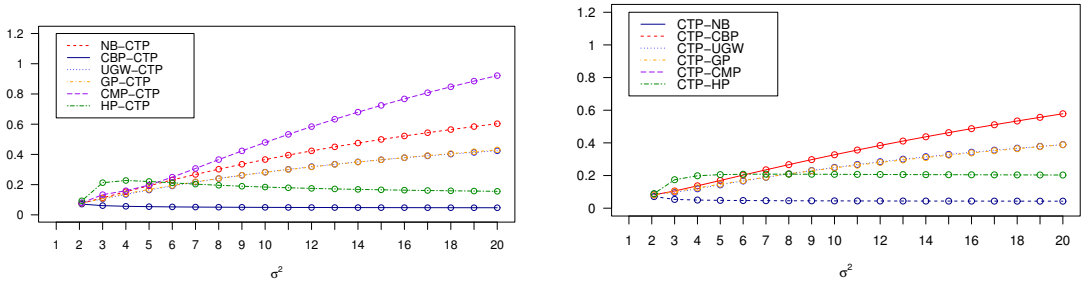
4.2.3. Estudio de simulación

Para finalizar esta primera sección de comparación con la distribución CTP hemos llevado a cabo un estudio de simulación. Concretamente hemos simulado $m = 1000$ muestras de tamaño $n = 100$, 300 y 500 para:

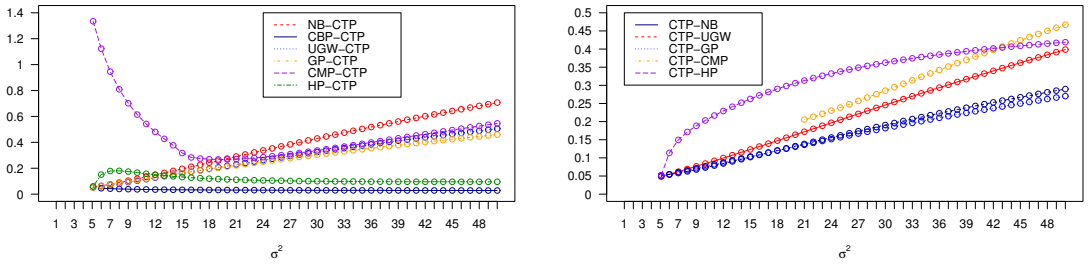
- una distribución CTP infradispersa con parámetros $a = -5, b = 4$ y $\gamma = 1$ ($\mu = 4.1$, $\sigma^2 = 1.8678$, $Moda = 4$)
- una distribución CTP sobredispersa con parámetros $a = -1, b = 3$ y $\gamma = 2$ ($\mu = 3.3333$, $\sigma^2 = 7.2222$, $Moda = 2$).



(a) $\mu = 1,\ k = 10.5$

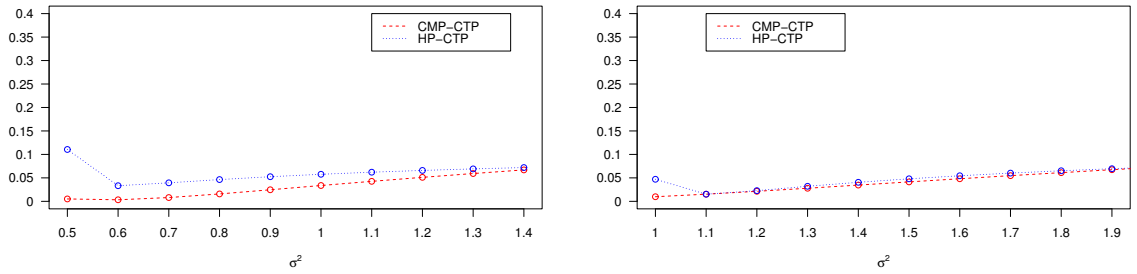


(b) $\mu = 2,\ k = 16.5$



(c) $\mu = 5,\ k = 32.5$

Figura 4.8: Divergencia KL entre la distribución CTP y las distribuciones BN, UGW, GP, CBP, HP y CMP, y viceversa (en un escenario sobredisperso)



(a) $\mu = 1.5$ (b) $\mu = 2$

Figura 4.9: Divergencia KL entre la distribución CTP y las distribuciones HP y CMP, y viceversa (en un escenario infradisperso)

En el escenario infradisperso, hemos intentado ajustar las distribuciones CMP, HP y CTP a los datos; mientras que en el escenario sobredisperso, hemos intentado ajustar las distribuciones $NB, CBP, UGW, GP, CMP, HP$ yCTP a los datos. Todas las estimaciones de los parámetros se han calculado mediante el método de máxima verosimilitud utilizando la función `fitdistr` de la librería `MASS` de R para la distribución NB, las funciones `com.log.density` y `dcom` de la librería `compoisson` para la distribución CMP y las funciones `fitcbp` y `fitctp` de la librería `cpd` para las distribuciones CBP y CTP. Para las restantes distribuciones, hemos implemetado nuestras propias funciones en R basándonos en la función `optim` de la librería `stats`.

Los resultados se resumen en las Tablas 4.6-4.10. En general, puede observarse que el tamaño muestral influye claramente en la identificación del modelo, lo que se hace más evidente cuando $n = 500$. Además:

- En el caso infradisperso (véase Tabla 4.6), el modelo CMP no es capaz de ajustar más del 20 % de las muestras; este porcentaje aumenta hasta casi el 90 % para el modelo hP, lo cual tiene sentido teniendo en cuenta que la restricción $0 < \mu - \sigma^2 < 1$ no se satisface. En definitiva, hay conjuntos de datos infradispersos que no pueden ser modelizados por los modelos usuales y sí por la CTP. Además, la CTP proporciona mejores ajustes que el resto de los modelos considerados a excepción de la CMP, que ajusta mejor el 28 % de las muestras generadas con $n = 500$ (Tabla 4.7). Sin embargo, como se muestra en la Tabla 4.9, las estimaciones del parámetro λ de la distribución CMP son bastante elevadas así como las desviaciones estándar. Con respecto a las estimaciones de a, b y γ en la CTP, los dos primeros están sesgados hacia la izquierda y el tercero hacia la derecha y con mayor dispersión.

- En el caso sobredisperso, prácticamente todas las muestras generadas pueden modelizarse mediante los modelos considerados (véase Tabla 4.6). No obstante, tal y como puede verse en la Tabla 4.8, pocos de estos ajustes son mejores que los obtenidos mediante la CTP según el criterio de información de Akaike (en adelante, AIC), específicamente, menos del 10 % para $n = 100$, menos del 1 % para $n = 500$ y ninguno para $n = 500$. Con respecto a las estimaciones de los parámetros (Tabla 4.10), las de a, b y γ están sesgadas a la derecha. Conviene señaar que las estimaciones de los parámetros de la distribución UGW, especialmente la de ρ, son muy elevadas, lo que sugiere la convergencia a la distribución BN. También las es-

	Infradispersión			Sobredispersión		
Modelo	$n = 100$	$n = 300$	$n = 500$	$n = 100$	$n = 300$	$n = 500$
CTP	100	100	100	99.8	100	100
NB	-	-	-	100	100	100
CBP	-	-	-	100	100	100
UGW	-	-	-	100	99.8	99.9
GP	-	-	-	100	100	100
CMP	72.5	76	78.8	99	99.9	99.9
HP	3.3	8.2	12.3	99.6	99.8	99.9

Tabla 4.6: Porcentaje de ajustes obtenidos para cada modelo utilizando los datos generados a partir de una distribución CTP

Modelo	$n = 100$	$n = 300$	$n = 500$
CMP	72.83	49.21	27.79
HP	0.00	0.00	0.00

Tabla 4.7: Porcentaje de ajustes con menor AIC que los obtenidos con la CTP para los datos infradispersos generados a partir de la CTP

timaciones medias de los parámetros de la hP son inusualmente altas, debido a las estimaciones muy altas obtenidas para las muestras 3, 51 y 50 de las 1000 muestras generadas para los tamaños muestral 100, 300 y 500, respectivamente. No obstante, los resultados no varían mucho si se excluyen estas muestras: $\widehat{\mu} = 3.3096(0.2448)$ y $\widehat{\sigma}^2 = 5.0033(1.6493)$ para $n = 100$, $\widehat{\mu} = 3.3228(0.1510)$ y $\widehat{\sigma}^2 = 5.2086(1.1906)$ para $n = 300$, y $\widehat{\mu} = 3.3265(0.1160)$ y $\widehat{\sigma}^2 = 5.2509(1.0192)$ para $n = 500$.

Hay que destacar que los modelos CMP y HP son capaces de reproducir la verdadera media de la CTP en ambos escenarios (el infradisperso y el sobredisperso), incluso con la misma precisión que el modelo CTP, pero no así la varianza.

4.3. Comparación de la distribución EBW

A continuación analizamos las diferencias existentes entre la distribución EBW, desarrollada en el Capítulo 3, y las distribuciones más usuales infra y sobredispersas, ya que este modelo permite ambos tipos de dispersión. Así pues, al igual que para la CTP, en el caso sobredisperso comparamos la EBW con la BN, UGW, CBP, CTP, GP, HP y CMP; mientras que en el caso infradisperso, nos limitamos a las distribuciones CTP, HP y CMP. Observaremos sus perfiles y calcularemos la divergencia de KL.

Modelo	$n = 100$	$n = 300$	$n = 500$
NB	9.20	0.20	0.00
CMP	4.24	0.00	0.00
HP	1.81	0.00	0.00

Tabla 4.8: Porcentaje de ajustes con menor AIC que los obtenidos con la CTP para los datos infradispersos generados a partir de la CTP

Modelo		$n = 100$	$n = 300$	$n = 500$
CTP	$\widehat{a}$	-5.9871(2.2284)	-5.3796(1.2104)	-5.1594(0.7838)
	$\widehat{b}$	3.4732(1.3184)	3.9404(0.5887)	3.9720(0.3399)
	$\widehat{\gamma}$	2.3830(3.8168)	1.5390(1.9126)	1.2016(1.1076)
CMP	$\widehat{\lambda}$	55.3060(38.6747)	41.4736(16.2561)	38.2364(10.4241)
	$\widehat{\nu}$	2.5969(0.3730)	2.4735(0.2279)	2.4362(0.1730)
HP	$\widehat{\gamma}$	0.1328(0.0273)	0.0451(0.0084)	0.0272(0.0053)
	$\widehat{\lambda}$	3.1623(0.1289)	3.1224(0.0793)	3.1246(0.0547)

Tabla 4.9: Medias y desviaciones estándar (en paréntesis) de las estimaciones máximo-verosímiles de los ajustes para los datos infradispersos generados a partir de la CTP

Modelo		$n = 100$	$n = 300$	$n = 500$
CTP	$\widehat{a}$	-0.7842(1.5811)	-0.9820(0.2413)	-0.9889(0.1543)
	$\widehat{b}$	3.2887(1.1622)	3.0705(0.5037)	3.0508(0.3515)
	$\widehat{\gamma}$	4.0683(11.6151)	2.2610(1.3874)	2.1583(0.8042)
NB	$\widehat{\theta}$	8.3126(14.2809)	5.6496(2.3654)	5.2818(1.4533)
	$\widehat{\mu}$	3.3384(0.2664)	3.3345(0.1599)	3.3357(0.1223)
UGW	$\widehat{a}$	16.9605(20.6932)	12.2113(4.4077)	11.5175(2.6403)
	$\widehat{k}$	16.9593(21.0021)	12.2113(4.4077)	11.5175(2.6403)
	$\widehat{\rho}$	233.0302(1535.8539)	52.6286(59.3604)	43.3452(23.6448)
GP	$\widehat{\theta}$	2.6033(0.2321)	2.5643(0.1339)	2.5526(0.1086)
	$\widehat{\lambda}$	0.2158(0.0876)	0.2292(0.0542)	0.2337(0.0440)
CMP	$\widehat{\lambda}$	1.8428(0.4813)	1.7337(0.2966)	1.6981(0.2398)
	$\widehat{\nu}$	0.5569(0.2037)	0.5226(0.1313)	0.5095(0.1097)
HP	$\widehat{\gamma}$	2116.2083(22791.8210)	419.5358(5123.7765)	15016.1431(342134.1560)
	$\widehat{\lambda}$	1667.5628(17904.3009)	331.6942(4008.3435)	11624.8836(264284.3991)

Tabla 4.10: Medias y desviaciones estándar (en paréntesis) de las estimaciones máximo-verosímiles de los ajustes para los datos sobredispersos generados a partir de la CTP

4.3.1. A través de la fmp

En primer lugar, representamos la fmp de estas distribuciones fijando los valores de la media y la varianza, y calculando los correspondientes valores de los parámetros tal y como se indica en la Subsección 4.2.1). En el caso de la EBW consideramos los valores dados en (3.14) y (3.16) tales que exista la varianza, esto es, $\gamma > 2\alpha + 2$.

Las Figuras 4.10-4.15 muestran los perfiles para las distribuciones mencionadas en un escenario sobredisperso. En todos ellos se aprecia que el comportamiento de la EBW sobredispersa es similar al de la UGW con la misma media y la misma varianza.

En cuanto a la probabilidad del 0 en la EBW, ésta es superior a la de la CTP y la EBW, pero no a la de los otros modelos sobredispersos.

Análogamente, en la Figura 4.16 se incluyen perfiles infradispersos de las distribuciones HP, CMP, CTP y EBW teniendo en cuenta las condiciones para la obtención de los parámetros en función de la media y la varianza. Cada fila de gráficos se corresponde con IA 0.4, 0.75 y 0.95, respectivamente. Para la CTP se ha fijado $\gamma = 2.1$ y se ha omitido la distribución HP en aquellos casos en los que no se cumple que $\mu - \sigma^2 < 1$, puesto que no puede darse la pareja de valores (μ, σ^2). Se observa que las distribuciones EBW, CMP y HP prácticamente se solapan, siendo la CTP la que difiere algo más, pero no de forma significativa.

4.3.2. A través de la divergencia de Kullback-Leibler

De nuevo calculamos la divergencia de KL en términos de σ^2, para distintos valores de μ, en este caso entre la distribución EBW y las distribuciones BN, GP, CBP, CTP, CMP y HP, y viceversa.

Los resultados se muestran en las Figuras 4.17, para el caso sobredisperso, y 4.18, para el caso infradisperso.

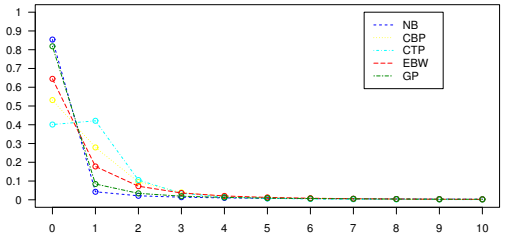
En general, se puede observar que:

- en una situación de sobredispersión la divergencia aumenta a medida que la variabilidad aumenta. Los modelos más alejados de la distribución EBW son el CBP y el HP; mientras que los más cercanos son el GP y el NB. No obstante, cabe resaltar que estas distancias son muy pequeñas, inferiores a 0.1, lo que muestra que la distribución EBW tiene una forma muy similar al resto de distribuciones.
- en una situación de infradispersión, sin embargo, la divergencia disminuye a medida que la varianza aumenta. Las distancias son algo mayores y la distribución HP - en general - es la más cercana a la EBW.

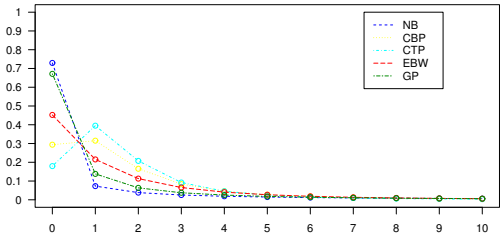
En definita, el modelo EBW tanto en su versión infradispersa como sobredispersa es muy similar a los restantes, con la ventaja de poseer expresiones explícitas de los momentos en términos de sus parámetros, proporcionar una explicación acerca del origen de la variabilidad de los datos bajo sobredispersión, gracias a la partición de la varianza y, en general, presentar menos problemas computacionales que la HP o CMP.

4.3.3. Estudio de simulación

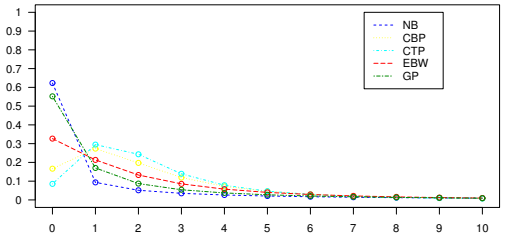
Para finalizar esta sección de comparación con la distribución EBW hemos llevado a cabo un estudio de simulación. Concretamente hemos simulado $m = 1000$ muestras de tamaño



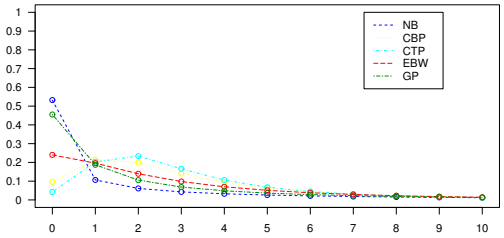
(a) $\mu = 1$



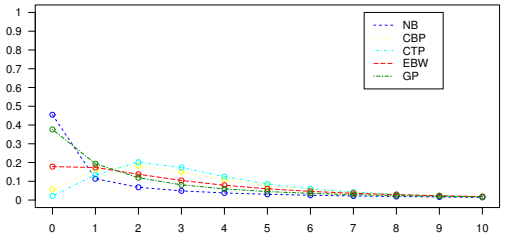
(b) $\mu = 2$



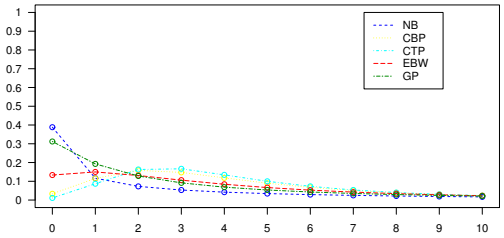
(c) $\mu = 3$



(d) $\mu = 4$

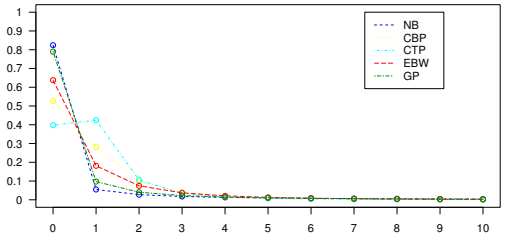


(e) $\mu = 5$

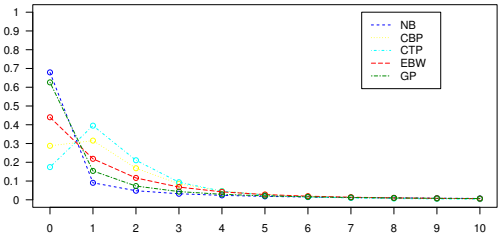


(f) $\mu = 6$

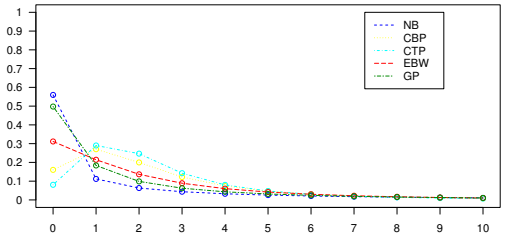
Figura 4.10: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y EBW para distintos valores de μ e $IA = 20$



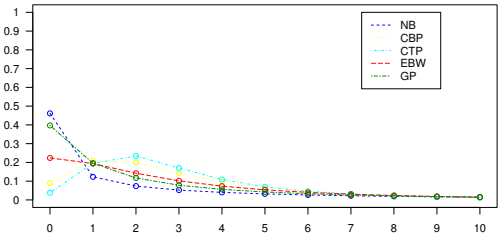
(a) $\mu = 1$



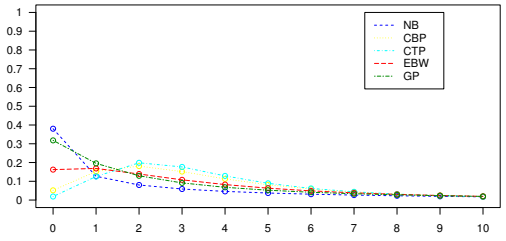
(b) $\mu = 2$



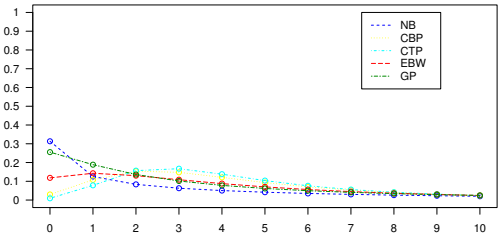
(c) $\mu = 3$



(d) $\mu = 4$

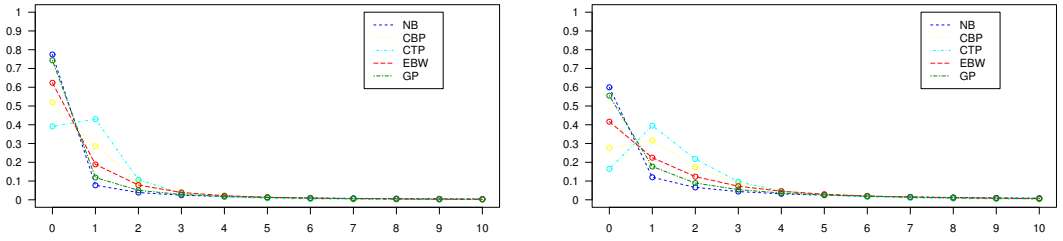


(e) $\mu = 5$

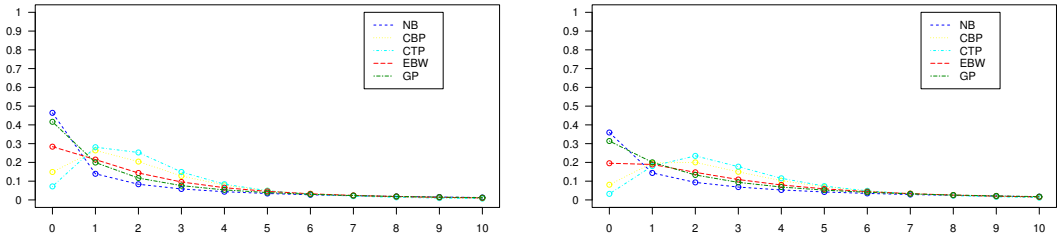


(f) $\mu = 6$

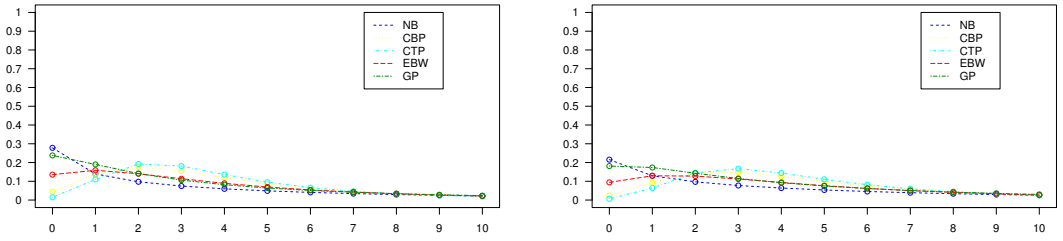
Figura 4.11: Perfiles de las distribuciones NB, CBP, $CTP(\cdot,\cdot,1)$, GP y EBW para distintos valores de μ e $IA = 15$



(a) $\mu = 1$ (b) $\mu = 2$

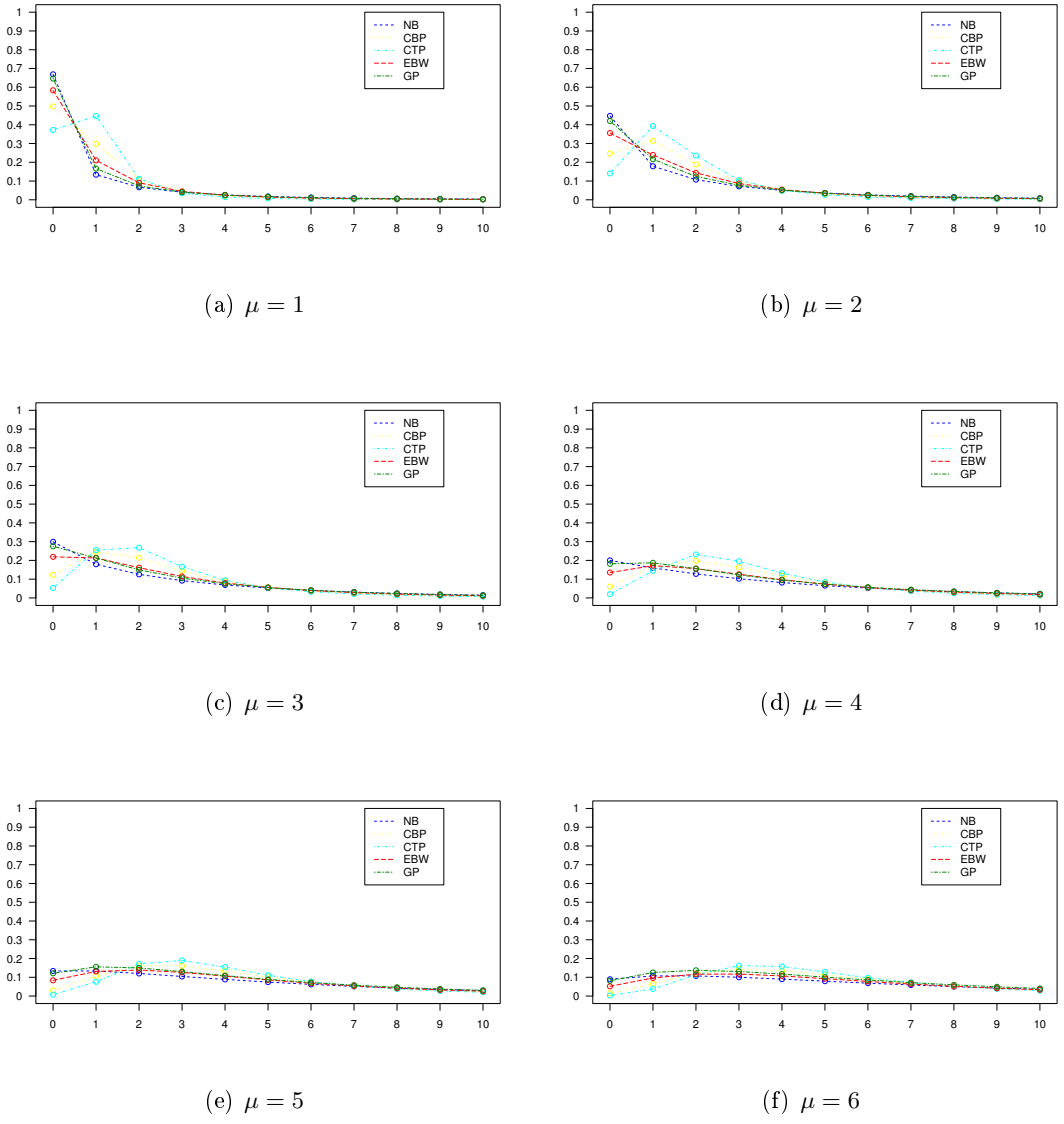


(c) $\mu = 3$ (d) $\mu = 4$



(e) $\mu = 5$ (f) $\mu = 6$

Figura 4.12: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y EBW para distintos valores de μ e $IA = 10$

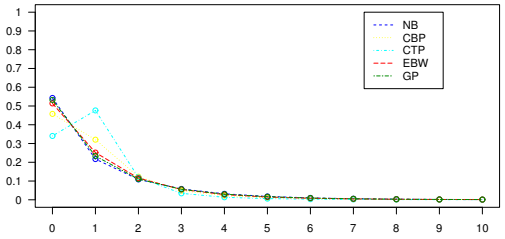


(a) $\mu = 1$ (b) $\mu = 2$

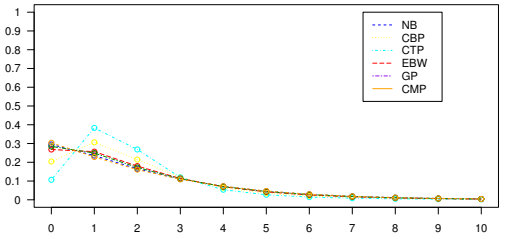
(c) $\mu = 3$ (d) $\mu = 4$

(e) $\mu = 5$ (f) $\mu = 6$

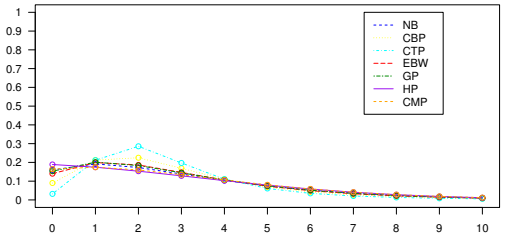
Figura 4.13: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y EBW para distintos valores de μ e $IA = 5$



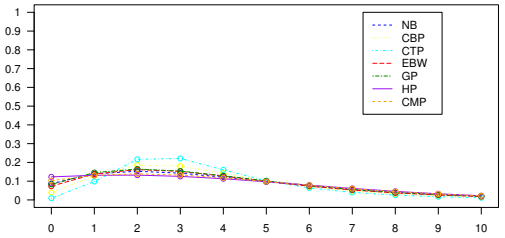
(a) $\mu = 1$



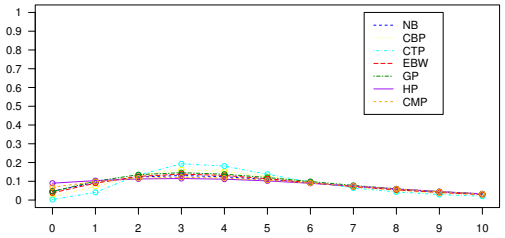
(b) $\mu = 2$



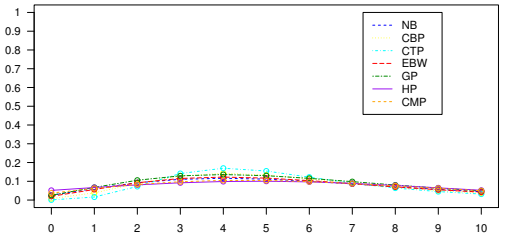
(c) $\mu = 3$



(d) $\mu = 4$

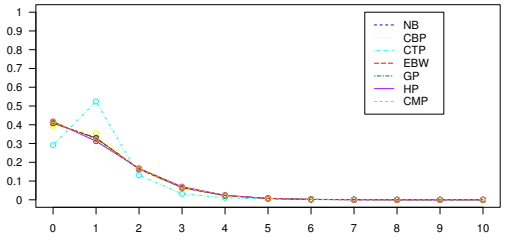


(e) $\mu = 5$

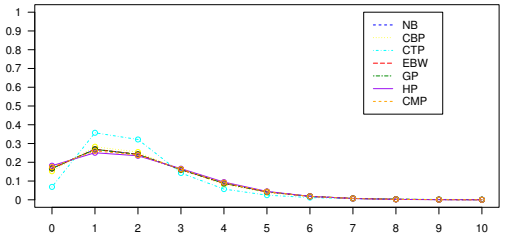


(f) $\mu = 6$

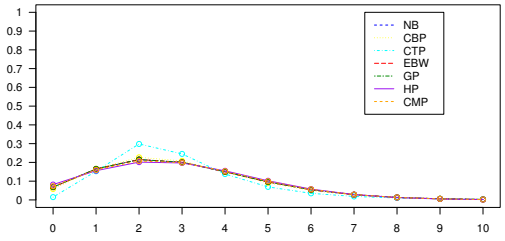
Figura 4.14: Perfiles de las distribuciones NB, CBP, $CTP(\cdot, \cdot, 1)$, GP y EBW para distintos valores de μ e $IA = 2.5$



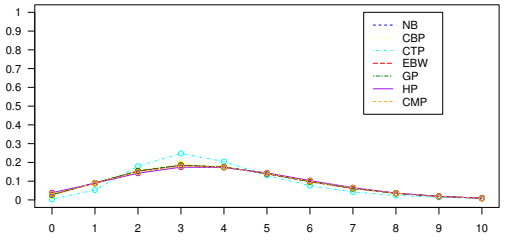
(a) $\mu = 1$



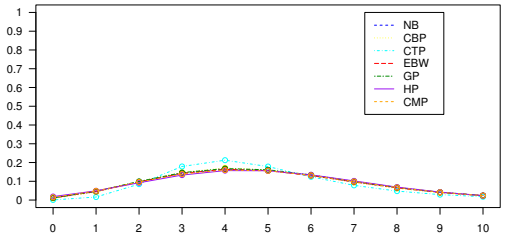
(b) $\mu = 2$



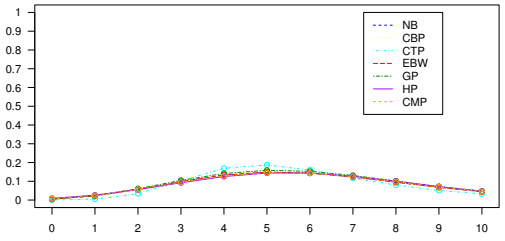
(c) $\mu = 3$



(d) $\mu = 4$

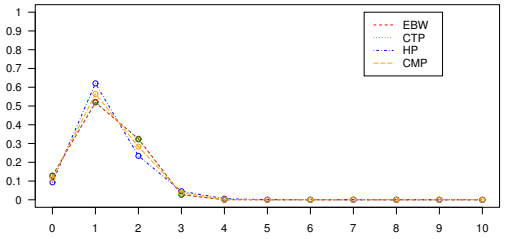


(e) $\mu = 5$

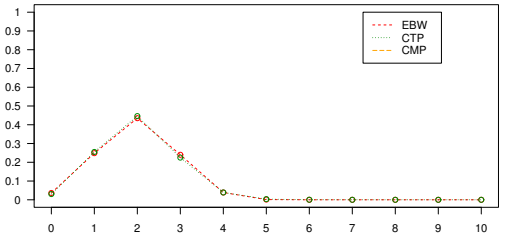


(f) $\mu = 6$

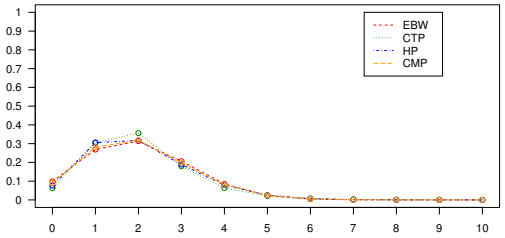
Figura 4.15: Perfiles de las distribuciones NB, CBP, $CTP(\cdot,\cdot,1)$, GP y EBW para distintos valores de μ e $IA = 1.25$



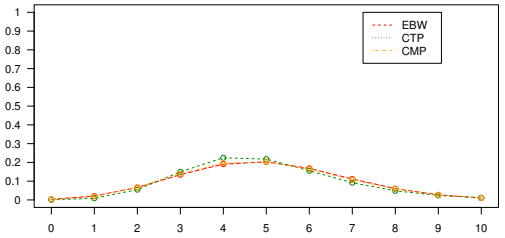
(a) $\mu = 1.25, \gamma = 2.1$



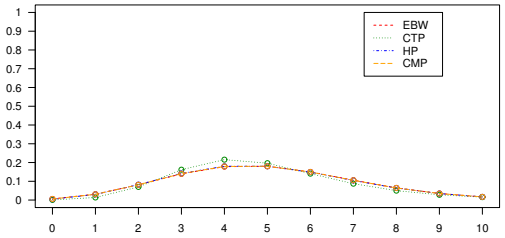
(b) $\mu = 2, \gamma = 2.1$



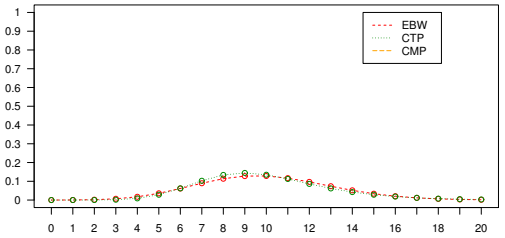
(c) $\mu = 5, \gamma = 2.1$



(d) $\mu = 2, \gamma = 2.1$

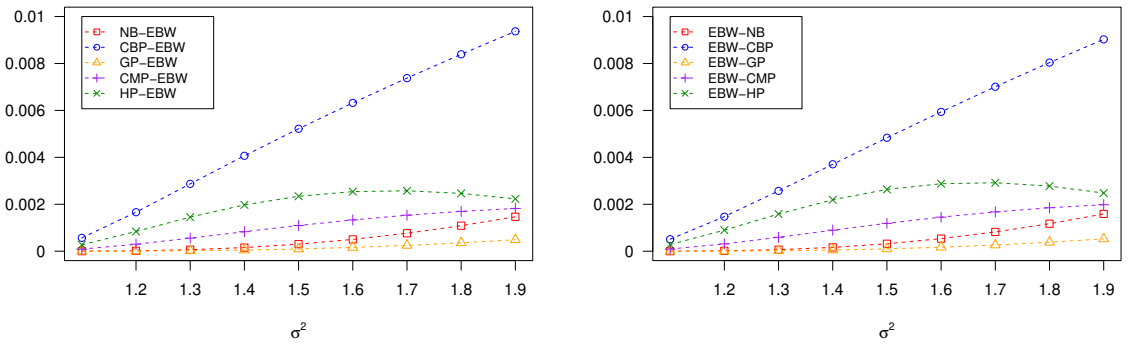


(e) $\mu = 5, \gamma = 2.1$

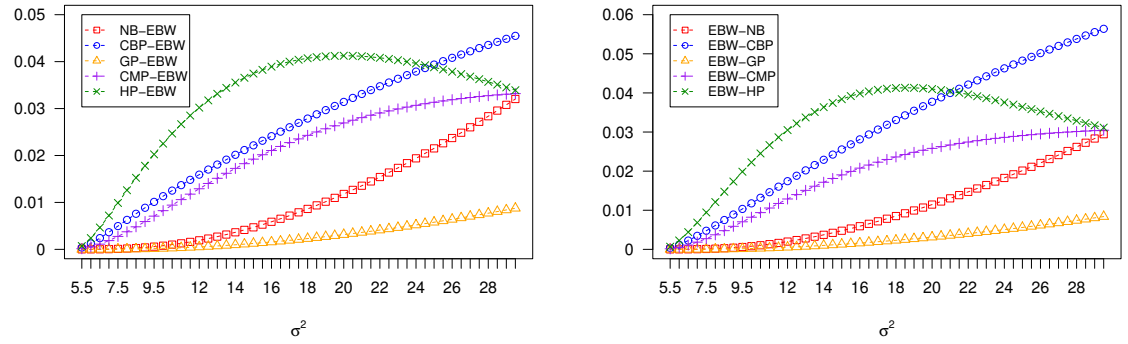


(f) $\mu = 10, \gamma = 2.1$

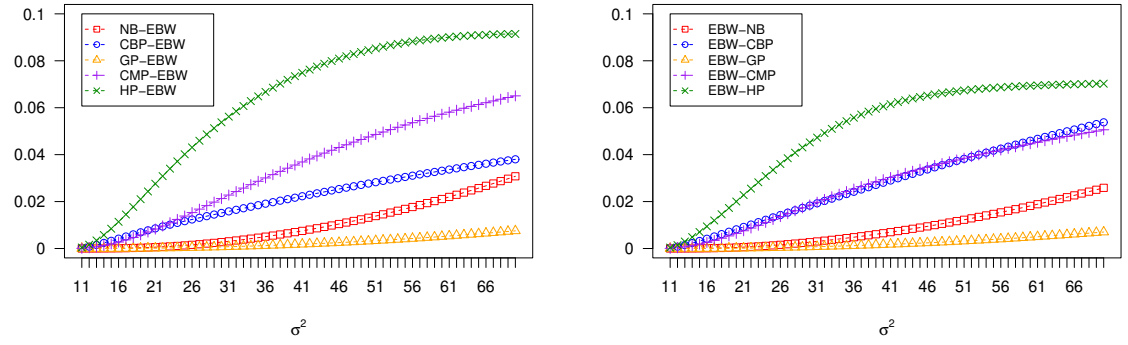
Figura 4.16: Perfiles de las distribuciones $CTP(\cdot, \cdot, 2.1)$, HP y CMP para distintos valores de μ y σ^2 en un escenario infradisperso



(a) $\mu = 1$

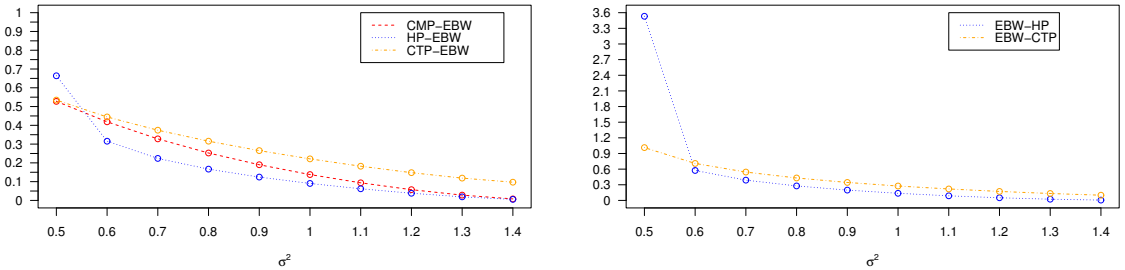


(b) $\mu = 5$

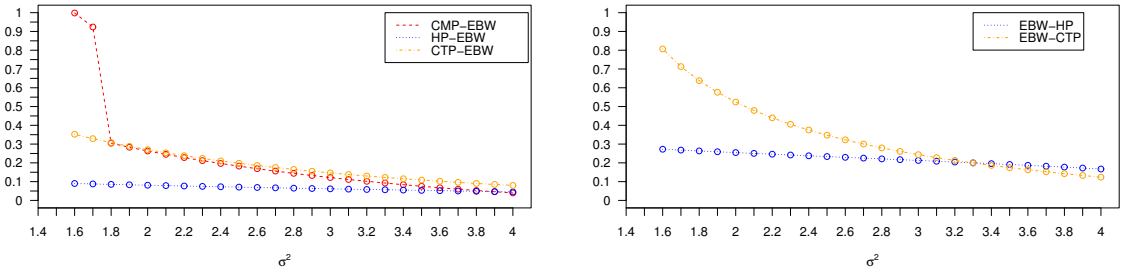


(c) $\mu = 10$

Figura 4.17: Divergencia de KL entre las distribuciones NB, CBP, GP, CMP, HP y EBW (y viceversa) en una situación de sobredispersión



(a) $\mu = 1.5$



(b) $\mu = 5$

Figura 4.18: Divergencia de KL entre las distribuciones CTP, CMP, HP y EBW (y viceversa) en una situación de infradispersión

$n = 100$, 300 y 500 para:

- una distribución EBW infradispersa con parámetros $\alpha = -4.5$ y $\gamma = 1$ ($\mu = 2.25$, $\sigma^2 = 0.6328$, $IA = 0.2813$, $Moda = 2$)
- una distribución EBW sobredispersa con parámetros $\alpha = -0.5$, $\gamma = 1.1$ ($\mu = 0.2272$, $\sigma^2 = 0.7438$, $IA = 3.2727$, $Moda = 0$)

En el escenario infradisperso, hemos intentado ajustar las distribuciones CMP, HP y EBW a los datos; mientras que en el escenario sobredisperso, hemos intentado ajustar las distribuciones $NB, CBP, CTP, GP, CMP, HP$ yEBW a los datos. Todas las estimaciones de los parámetros se han calculado mediante el método de máxima verosimilitud utilizando las funciones de R citadas en la Subsección 4.2.3, además de la función `fitebw cpd` para la distribución EBW.

Los resultados se resumen en las Tablas 4.11-4.15. De nuevo se observa que en general, el tamaño muestral influye claramente en la identificación del modelo, lo que se hace más evidente cuando $n = 500$. Además:

- En el caso infradisperso (véase Tabla 4.11), el modelo CMP no es capaz de ajustar casi el 20 % de las muestras con $n = 100$ y más del 25 % cuando $n = 500$. El modelo HP, por su parte, no es capaz de ajustar casi la mitad de las muestras con $n = 100$, si bien este porcentaje se incrementa considerablemente cuando aumenta el tamaño muestra. En cuando al modelo CTP, al tratarse de modelo más general que el EBW se ajusta prácticamente a todas las muestras. No obstante, tal y como puede verse en la Tabla 4.12, estos ajustes son peores que los obtenidos mediante la distribución EBW: la CMP ajusta mejor el 20.57 % de las muestras generadas con $n = 100$, pero sólo alrededor del 6 % para $n = 500$, la CTP no llega al 5 % y la HP no mejora ningún ajuste. La Tabla 4.14 muestras las estimaciones máximo-verosímiles medias para cada parámetro, junto con sus desviaciones estándar para los tamaños de muestra considerados. Se observa que las estimaciones del parámetro λ de la distribución CMP son bastante elevadas así como las desviaciones estándar. Con respecto a las estimaciones de α y γ en la EBW, el primero está sesgado a la derecha y el segundo a la izquierda, pero con menor precisión.
- En el caso sobredisperso, prácticamente todas las muestras generadas pueden modelizarse mediante los modelos considerados (véase Tabla 4.11). No obstante, tal y como puede verse en la Tabla 4.13, pocos de estos ajustes son mejores que los obtenidos mediante la CTP según el criterio de información de Akaike (en adelante, AIC), específicamente, menos del 10 % para $n = 100$, menos del 1 % para $n = 500$ y ninguno para $n = 500$. Con respecto a las estimaciones de los parámetros (Tabla 4.15), las de a, b y γ están sesgadas a la derecha. Conviene señaar que las estimaciones de los parámetros de la distribución UGW, especialmente la de ρ, son muy elevadas, lo que sugiere la convergencia a la distribución BN. También las estimaciones medias de los parámetros de la hP son inusualmente altas, debido a las estimaciones muy altas obtenidas para las muestras 3, 51 y 50 de las 1000 muestras generadas para los tamaños muestral 100, 300 y 500, respectivamente. No obstante, los resultados no varían mucho si se excluyen estas muestras: $\widehat{\mu} = 3.3096(0.2448)$ y $\widehat{\sigma}^2 = 5.0033(1.6493)$ para $n = 100$, $\widehat{\mu} = 3.3228(0.1510)$ y $\widehat{\sigma}^2 = 5.2086(1.1906)$ para $n = 300$, y $\widehat{\mu} = 3.3265(0.1160)$ y $\widehat{\sigma}^2 = 5.2509(1.0192)$ para $n = 500$.

Hay que destacar que los modelos CMP y HP son capaces de reproducir la verdadera media de la EBW en ambos escenarios (el infradisperso y el sobredisperso), incluso con la

	Infradispersión			Sobredispersión		
Modelo	$n = 100$	$n = 300$	$n = 500$	$n = 100$	$n = 300$	$n = 500$
EBW	97.1	99.8	100	97.7	100	99.6
NB	-	-	-	100	100	99.1
CBP	-	-	-	76.9	82.5	83.5
CTP	100	99.8	99.9	99.9	99.9	99.7
GP	-	-	-	82.8	96	98.5
CMP	80.2	75.9	74.5	97.5	99.2	99.4
HP	53.5	86.6	94.8	93.5	96.3	97.4

Tabla 4.11: Porcentaje de ajustes obtenidos para cada modelo utilizando los datos generados a partir de una distribución EBW

Modelo	$n = 100$	$n = 300$	$n = 500$
CTP	4.7	4.7	4.7
CMP	20.57	8.56	5.91
HP	0.00	0.00	0.00

Tabla 4.12: Porcentaje de ajustes con menor AIC que los obtenidos con la EBW para los datos infradispersos generados a partir de la EBW

misma precisión que el modelo EBW, pero no así la varianza.

4.3.4. Comparación con la distribución Binomial

Finalmente, comparamos la distribución EBW de rango finito con la distribución binomial.

La Figura 4.19 contiene distintos perfiles para ambas distribuciones con idéntica media. Se puede observar cómo la distribución EBW tiene un valor modal mucho más diferenciado que la distribución binomial con la misma media. Además, la probabilidad en cero es inferior en la distribución EBW que en la distribución binomial.

La Tabla 4.16 contiene los resultados de la simulación realizada para comparar ambas distribuciones. Concretamente se han generado 1000 muestras de tamaño $n = 100$, $n = 300$ y $n = 500$ de una $EBW(n = 10, \gamma = 1)$ (correspondiente a una media igual a 5) y se han intentado modelizar mediante una $B(10, p)$ ($\widehat{p} = \overline{x}/10$). Ambas distribuciones se ajustan a todas las muestras, salvo para $n = 300$, donde la función de ajuste de la EBW finita no converge para 2 de las 1000 muestras generadas. En el 100 % de las muestras,

Modelo	$n = 100$	$n = 300$	$n = 500$
NB	17	2.7	0.71
CBP	12.61	14.91	11.38
GP	33.94	10.62	3.96
CMP	13.03	3.63	1.91
HP	14.97	3.63	2.36

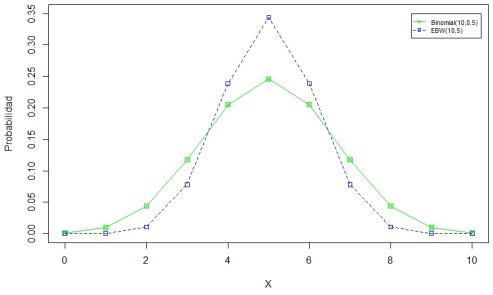
Tabla 4.13: Porcentaje de ajustes con menor AIC que los obtenidos con la EBW para los datos infradispersos generados a partir de la EBW

Modelo		$n = 100$	$n = 300$	$n = 500$
EBW	$\widehat{\alpha}$	-4.5228(0.3568)	-4.4896(0.2048)	-4.4995(0.1494)
	$\widehat{\gamma}$	1.1042(0.7562)	0.9951(0.3979)	1.0168(0.3044)
	$\widehat{\mu}$	2.2495(0.0794)	2.2503(0.0459)	2.2481(0.0348)
	$\widehat{\sigma}^2$	0.6344(0.0832)	0.6288(0.0482)	0.6324(0.0360)
CTP	$\widehat{a}$	-4.3059(0.5296)	-4.3923(0.2835)	-4.4091(0.2249)
	$\widehat{b}$	0.2814(0.5184)	0.2060(0.4020)	0.2137(0.3855)
	$\widehat{\gamma}$	2.3830(3.8168)	0.9159(0.4118)	0.9408(0.3199)
	$\widehat{\mu}$	2.2492(0.0786)	2.2498(0.0459)	2.2473(0.0347)
	$\widehat{\sigma}^2$	0.6324(0.0835)	0.6281(0.0479)	0.6311(0.0358)
CMP	$\widehat{\lambda}$	72.6403(75.2795)	59.9047(24.9410)	56.9149(18.0350)
	$\widehat{\nu}$	4.1274(0.6355)	4.1265(0.3546)	4.1054(0.2732)
	$\widehat{\mu}$	2.2489(0.0794)	2.2538(0.0467)	2.2524(0.0358)
	$\widehat{\sigma}^2$	0.6579(0.0930)	0.6507(0.0531)	0.6519(0.0401)
HP	$\widehat{\gamma}$	0.0471(0.0245)	0.0275(0.0144)	0.0250(0.0123)
	$\widehat{\lambda}$	1.2935(0.0797)	1.2815(0.0463)	1.2898(0.0349)
	$\widehat{\mu}$	2.2366(0.0788)	2.2480(0.0457)	2.2474(0.0347)
	$\widehat{\sigma}^2$	1.3150(0.0808)	1.2946(0.0475)	1.2898(0.0362)

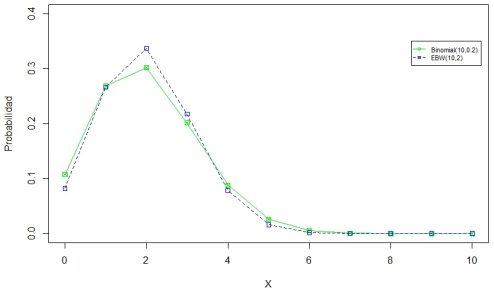
Tabla 4.14: Medias y desviaciones estándar (en paréntesis) de las estimaciones máximo-verosímiles de los ajustes para los datos infradispersos generados a partir de la EBW

Modelo		$n = 100$	$n = 300$	$n = 500$
EBW	$\widehat{\alpha}$	-0.5600(0.2374)	-0.4962(0.0823)	-0.4933(0.0451)
	$\widehat{\gamma}$	1.6328(1.5786)	1.1154(0.4679)	1.0849(0.1885)
CTP	$\widehat{a}$	-0.3685(4.2680)	-0.5172(2.1530)	-0.4738(1.1386)
	$\widehat{b}$	1.2839(4.4497)	0.4849(1.9306)	0.3301(0.7416)
	$\widehat{\gamma}$	188.7886(540.2925)	43.3876(257.3950)	11.0884(94.1402)
CBP	$\widehat{b}$	6.9041(10.5007)	3.3009(6.4207)	2.4019(4.5104)
	$\widehat{\gamma}$	737.8767(1404.9330)	256.0609(890.3768)	127.1294(580.2113)
NB	$\widehat{\theta}$	31.3350(44.2841)	16.3456(33.2056)	11.6175(45.9075)
	$\widehat{p}$	0.2354(0.0618)	0.2312(0.0391)	0.2300(0.0291)
GP	$\widehat{\theta}$	0.2141(0.0435)	0.2125(0.0268)	0.2127(0.0204)
	$\widehat{\lambda}$	0.0990(0.1214)	0.0760(0.0847)	0.0715(0.0691)
CMP	$\widehat{\lambda}$	0.2202(0.0517)	0.2074(0.0297)	0.2046(0.0233)
	$\widehat{\nu}$	3.0663(6.8995)	0.6842(1.8958)	0.4502(0.5259)
HP	$\widehat{\gamma}$	18859.3890(162055.7436)	70424.3041(659748.7613)	31404.2593(249439.4659)
	$\widehat{\lambda}$	3996.8241(31223.3327)	14018.8863(128959.6298)	6240.3331(48863.2801)

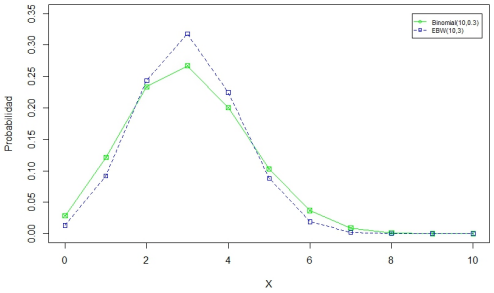
Tabla 4.15: Medias y desviaciones estándar (en paréntesis) de las estimaciones máximo-verosímiles de los ajustes para los datos sobredispersos generados a partir de la EBW



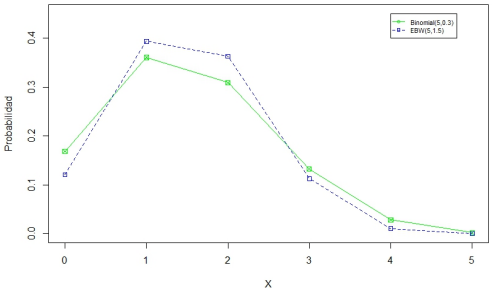
(a) $n = 10$ y $\mu = 5$



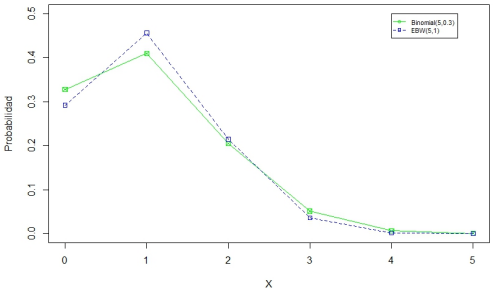
(b) $n = 10$ y $\mu = 2$



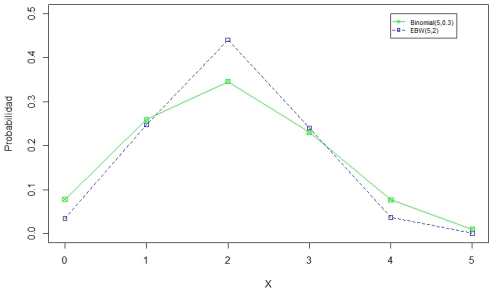
(c) $n = 10$ y $\mu = 3$



(d) $n = 5$ y $\mu = 1.5$



(e) $n = 5$ y $\mu = 1$



(f) $n = 5$ y $\mu = 2$

Figura 4.19: Perfil de la distribución EBW y Binomial con distintos n y μ

Modelo		$n = 100$	$n = 300$	$n = 500$
$EBW(10, \cdot)$	$\widehat{\gamma}$	1.0099(0.4530)	0.9886(0.2646)	1.0027(0.2058)
$Binomial(10, \cdot)$	$\widehat{p}$	0.4998(0.0116)	0.5003(0.0067)	0.4999(0.0052)

Tabla 4.16: Medias y desviaciones estándar (en paréntesis) de las estimaciones máximo-verosímiles de los ajustes para los datos generados a partir de $EBW(n = 10, \cdot)$ finita

independientemente del tamaño, el ajuste mediante la EBW finita presenta menor AIC que el correspondiente ajuste binomial.

Capítulo 5

Aplicación a datos reales

Una vez desarrollada la distribución EBW, estudiadas sus propiedades y habiendo sido comparada con otros modelos para datos de conteo (comparación que también se ha realizado para la distribución CTP), procedemos a aplicar ambos modelos en la modelización de datos reales tratados en la literatura. Algunos de ellos aparecen en artículos publicados por la doctoranda y sus directores.

5.1. Modelización del número de instalaciones en municipios andaluces tanto públicas como privadas

Una variable discreta que se encuentra en las bases de datos disponibles para los ciudadanos es el número de instalaciones existentes en cada municipio. Ejemplos de esta variable son el número de colegios, institutos, hospitales, bibliotecas y centros de salud, entre otros. Caba señalar que, en todos los casos, se trata de variables de conteo fuertemente sobredispersas (unos cuantos municipios presentan valores muy altos y la mayoría valores bajos, aunque no siempre con el valor modal en 0).

A priori hay una diferencia entre centros de titularidad privada o pública. Así, si nos referimos a los privados, se puede suponer que estas instalaciones están relacionadas con la rentabilidad, es decir, la empresa privada determina si es conveniente construir un determinado edificio o no. Esto conduce a una variable cuya forma es sobredispersa con la frecuencia de municipios sin ninguna instalación privada muy elevada. Un ejemplo de este tipo de variables es el número de entidades bancarias (sucursales) por municipio. En el caso de instalaciones de titularidad pública, la existencia o no de una de ellas en el municipio no se debe a razones de rentalibilidad de la instalación sino al estado de bienestar, cohesión territorial o interés electoral, como por ejemplo el número de colegios o de institutos por municipio. Así, un municipio pequeño puede tener instalaciones infrautilizadas. Estas situaciones provocan que el valor más habitual de la variable no sea el 0, sino el 1, además de que los valores altos son limitados. Este hecho introduce una diferencia en el tipo de distribución aconsejada en cada caso. Así, en el caso público la distribución más adecuada es la CTP, mientras que en el caso privado, una buena alternativa la constituye la distribución EBW. Datos de estas características han sido analizados en diferentes publicaciones de la doctoranda como, por ejemplo, Rodríguez Avi et al. (2020) o Cueva-López et al. (2022).

Para ilustrar lo mencionado anteriormente, se estudian las siguientes variables en los 773 municipios de Andalucía (España):

Educación pública	$\overline{x}$	s^2	IA	Q_1	Mediana	Q_3	Máximo
Preescolar	4.078	69.481	17.038	1	1	4	114
Primaria	2.96	46.56202	15.7304	1	1	3	94
Secundaria	1.659	11.6601	7.028	1	1	1	48
Bachillerato	0.9084	9.4095	10.3583	0	0	1	45
FPM	0.678	3.7601	5.5458	0	0	1	28
FPS	0.5908	5.7361	8.6935	0	0	0	41
Adultos	2.083	13.1854	6.33	1	2	2	60
Especial	1.521	15.0931	9.9231	0	1	2	54
Entidades públicas	$\overline{x}$	s^2	IA	Q_1	Mediana	Q_3	Máximo
Biblioteca	1.106	2.4118	2.1807	1	1	1	24
Salud	2.047	8.3262	4.0675	1	1	2	35
Educación privada	$\overline{x}$	s^2	IA	Q_1	Mediana	Q_3	Máximo
Primaria	0.68	15.29	22.36	0.00	0.00	0.00	61.00
Secundaria	0.61	12.60	20.60	0.00	0.00	0.00	54.00
Bachillerato	0.07	0.36	4.93	0.00	0.00	0.00	10.00
Turismo	$\overline{x}$	s^2	IA	Q_1	Mediana	Q_3	Máximo
Hotel 1 a 3	0.61	14.59	24.03	0.00	0.00	0.00	72.00
Hotel 4 a 5	0.40	11.65	28.90	0.00	0.00	0.00	71.00

Tabla 5.1: Resumen descriptivo del número de establecimientos públicos y privados

- Número de centros educativos públicos: preescolares (Preescolar), escuelas primarias (Primaria), instituto de educación secundaria obligatoria (Secundaria), instituto de educación secundaria no obligatoria (Bachillerato), formación profesional básica (FPM), formación profesional superior (FPS), escuelas de educación de adultos (Adultos) y escuelas de educación de régimen especial, es decir, escuelas de música y escuelas de lenguaje (Especial).

- Número de algunas otras instalaciones públicas: bibliotecas públicas (Biblioteca) y centros de salud (Salud).

- Centros educativos privados: colegios privados de primaria (Primaria), secundaria (Secundaria) y bachillerato (Bachillerato).

- Número de establecimientos turísticos: hoteles de 1 a 3 estrellas (Hotel 1 a 3) y hoteles de lujo, es decir, hoteles de 4 y 5 estrellas (Hotel 4 a 5).

Todos estos datos se refieren al año 2020 y han sido obtenidos a partir del Sistema de Información Multiterritorial de Andaluciía[1]. Para iniciar el análisis realizamos un estudio descriptivo de los datos en cuestión. La Tabla 5.1 muestra dicho resumen donde se explicitan los valores de la media, la varianza, la razón entre varianza y media (IA), los cuartiles primero y tercero (Q_1 y Q_3), la mediana y el máximo de los datos observados.

Para todas estas variables se ajustan los modelos de Poisson, BN, GP, UGW, CTP, EBW, CMP y HP, lo que permite seleccionar el mejor de entre ellos en cada caso.

[1] `https://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index2.htm`

Educación pública	*Poisson*	*BN*	*GP*	*UGW*	*CTP*	*EBW*	*CMP*	*HP*
Preescolar	6545.7	3402.0	3283.7	3176.2	**2768.4**	3174.2	3415.2	3415.4
Primaria	5558.1	3031.5	2916.1	2812.2	**2154.1**	2810.3	3033.9	3034.9
Secundaria	3313.0	2362.5	2309.4	2240.7	**1772.6**	2238.7	2387.6	2387.7
Bachillerato	2632.2	1683.1	1654.8	1631.2	**1618.1**	1629.2	1792.1	
FPM	2053.2	1471.3	1454.8	1439.5	**1430.6**	1437.5	1536.7	1536.8
FPS	1870.7	1202.9	1187.1	1179.6	1178.9	**1177.6**	1424.9	1425.7
Adultos	3001.7	2585.2	2549.3	2488.2	**2376.3**	2486.2	2633.3	2634.3
Especial	2984.6	2236.3	2197.9	2164.8	**2151.0**	2162.8	2296.8	2297.1
Entidades públicas	*Poisson*	*BN*	*GP*	*UGW*	*CTP*	*EBW*	*CMP*	*HP*
Bibliotecas	2162.5	1859.5	1850.3	1836.9	**1414**.4	1834.9	1906.9	1879.2
Salud	2962.8	2508.1	2473.3	2430.9	**1856.6**	2428.9	2594.5	2615.2
Educación privada	*Poisson*	*BN*	*GP*	*UGW*	*CTP*	*EBW*	*CMP*	*HP*
Primaria	1488.9	1035.9	1017.1	1014.5	1014.5	**1012.5**	1544.0	1544.1
Secundaria	1447.5	945.5	931.6	931.9	931.9	**929.9**	1452.4	1452.7
Bachillerato	1017.6	262.0	260.9	263.0	263.0	**261.0**	368.2	368.2
Turismo	*Poisson*	*BN*	*GP*	*UGW*	*CTP*	*EBW*	*CMP*	*HP*
Hotel 1 a 3	4715.9	907.4	**899.8**	904.3	904.3	902.3	1446.6	1446.7
Hotel 4 a 5	3039.6	511.6	**510.0**	515.5	515.5	513.5	1143.6	1143.7

Tabla 5.2: Valores de AIC para los ajustes de las variables de instalaciones públicas y privadas de los municipios andaluces.

5.1.1. Ajuste de las variables

A través de un programa implementado en R se estiman los parámetros de las distribuciones mencionadas con anterioridad para ver cuál o cuáles son las que proporcionan un mejor ajuste. La comparación se efectúa a través del índice de información de Akaike (AIC), que se define como

$$AIC = 2p - 2\ln(\mathfrak{L})$$

donde $\mathfrak{L}$ es el máximo valor de la función de verosimilitud para el modelo estimado y p es el número de parámetros estimados en el modelo estadístico correspondiente. Se basa en la entropía de información y se ofrece una estimación relativa de la información perdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos. No sirve para decidir si un ajuste es adecuado o no de manera absoluta, sino que sólo se utiliza para comparación de modelos, de modo que el preferido es el que tiene el valor mínimo en el AIC. Por lo tanto el AIC no solamente recompensa la bondad de ajuste, sino también incluye una penalización, que es una función creciente del número de parámetros estimados. Esta penalización desalienta el sobreajuste (aumentando el número de parámetros libres en el modelo mejora la bondad del ajuste, sin importar el número de parámetros libres en el proceso de generación de datos). Para más información véase, por ejemplo, Hu (2007) o Cavanaugh and Neath (2019).

La Tabla 5.2 muestra el AIC obtenido para cada ajuste, resaltando en negrita el mejor valor obtenido. Se puede ver que la distribución CTP es la distribución que mejor se ajusta a los datos correspondientes a las instalaciones públicas, salvo los centros de formación profesional superior. El resto de variables se ajusta mejor con la distribución EBW, excepto en el caso de los hoteles, en donde el mejor AIC se alcanza con la distribución GP.

Se puede observar cómo el valor del AIC en el caso de la EBW es dos unidades menor

		Esperadas			
X	*Observadas*	*CTP*	*UGW*	*GP*	*EBW*
0	93	93.067	227.302	223.505	227.302
1	437	435.846	186.187	177.737	186.187
2	67	74.282	114.591	113.000	115.590
3	31	26.648	64.975	67.517	64.975
4	17	13.205	36.080	39.544	36.080
5	6	7.775	20.093	23.042	20.093
6	6	5.042	11.338	133.436	11.338
7	2	3.519	6.511	7.860	6.511
8	3	2.582	3.811	4.616	3.811
9	2	1.968	274	2.724	2.275
10-11	3	2.787	2.130	2.570	2.232
12-19	5	4.956			
20-48	4	5.346			
		$\widehat{a} = -0.5558$ (0.044)	$\widehat{a} = 3.8591$ (0.6338)	$\widehat{\lambda} = 1.1082$ (0.0444)	$\widehat{a} = 10.4628$ (0.2838)
		$\widehat{b} = 0.4232$ (0.0922)	$\widehat{k} = 3.8591$ (0.6338)	$\widehat{\theta} = 0.3319$ (0.0204)	$\widehat{\rho} = 10.4628$ (1.4646)
		$\widehat{\gamma} = 0.1042$ (0.0292)	$\widehat{\rho} = 10.4628$ (1.4616)		
Estadístico χ^2		3.7308	530.687	486.95	489.950
$p-$valor		0.713	0.0000	0.0000	0.000

Tabla 5.3: Frecuencias observadas, esperadas, parámetros estimados (errores estándar en paréntesis) y contraste de bondad de ajuste χ^2 para el número de institutos de secundaria por municipio en Andalucía

que el correspondiente para la dsitribución UGW, lo que ratifica el hecho de que el modelo biparamétrico ajusta tan bien como el triparamétrico (valor de la verosimilitud) pero es menos penalizado en el AIC al tener un parámetro menos.

A continuación, se desarrolla un estudio más pormenorizado de algunas de estas variables.

5.1.2. Centros de educación pública

La primera variable analizada es el número de institutos públicos de educación secundaria por municipio de Andalucía. Cabe destacar que el valor modal de dicha variable es 1, siendo este mucho mayor que el resto de las frecuencias de los valores modales de la variable (aglutina más del 50 % de la población de municipios). En primer lugar, en la Tabla 5.2 se muestran las frecuencias esperadas correspondientes a los ajustes proporcionados por las distribuciones CTP, UGW, EBW y GP, ya que son los que tienen mejores AIC. Esta tabla también incluye las estimaciones máximo-verosímiles (MV) de los respectivos parámetros, sus errores estándar (entre paréntesis) y los resultados correspondientes al contraste de bondad de ajuste χ^2 de Pearson.

Se observa que no se rechaza la hipótesis nula (H_0: Los datos proceden de una distribución concreta) en el contraste de bondad de ajuste únicamente en el caso de la distribución CTP (con un $p-$valor de 0.713). La Figura 5.1 dibuja de manera exacta la forma de la distribución de la variable mostrando las frecuencias esperadas a partir de las distribuciones CTP y EBW ajustadas (esta última por tener mejor AIC entre las que no se ajustan a los

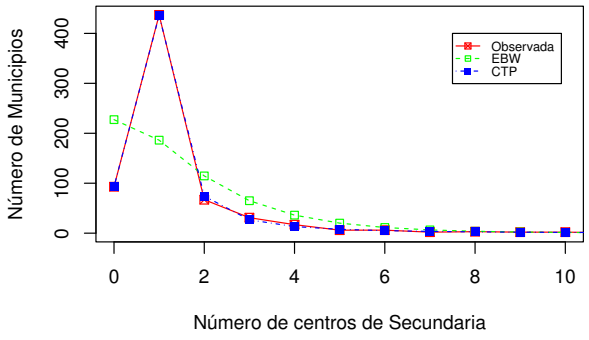


Figura 5.1: Frecuencias observadas y esperadas para el número de institutos con educación secundaria obligatoria en los municipios de Andalucía

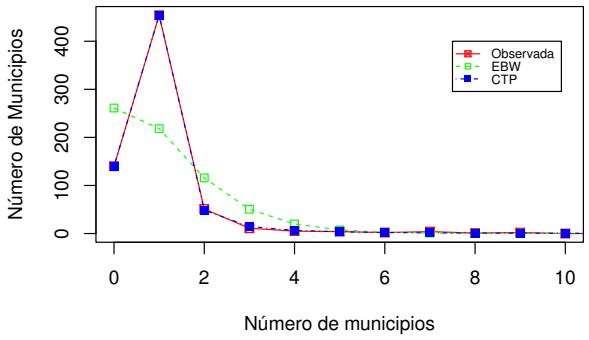


Figura 5.2: Frecuencias observadas y esperadas para el número de bibliotecas públicas en los municipios de Andalucía

datos). Es de destacar cómo la CTP es la única capaz de reproducir el valor modal en 1, y que es el causante de la sobredispersión, mientras que el resto de distribuciones ajustadas sobreestiman la frecuencia del 0.

5.1.3. Bibliotecas públicas

Centrándonos en el número de bibliotecas públicas en los municipios andaluces, se puede observar que las distribuciones ajustadas que tienen un mejor AIC son las mismas que en el caso anterior. Además, se vuelve a cumplir que el valor modal (de nuevo 1) tiene una frecuencia mucho mayor que el resto de los valores de la variable.

A partir de los datos del contraste de bondad de ajuste (Tabla **??**), no hay evidencia para rechazar la hipótesis nula (H_0: Los datos proceden de una distribución concreta) sólo en el caso de la distribución CTP (con un $p - valor$ de 0.5103). Por su parte, en la Figura 5.2 se observa que el modelo CTP es capaz de reproducir de forma fidedigna el perfil que presenta la variable.

		Esperadas			
X	*Observadas*	*CTP*	*UGW*	*GP*	*EBW*
0	140	139.993	261.079	255.57	261.079
1	454	454.142	218.202	220.931	218.202
2	51	48.548	115.707	118,917	115.707
3	11	14.619	50.457	51.490	50.458
4	5	6.538	19.904	23.348	19.904
5	4	3.569	7.438	6.963	2.702
6	3	2.197	2.702	2.330	2.702
7-25	9	7.392			2.100
		$\widehat{a} = -0.6459$ (0.0642)	$\widehat{a} = 7.406$ (2.05)	$\widehat{\lambda} = 0.9793$ (0.0382)	$\widehat{a} = 7.406$ (1,0689)
		$\widehat{b} = 0.3523$ (0.1121)	$\widehat{k} = 7.406$ (2.051)	$\widehat{\theta} = 0.1093$ (0.0170)	$\widehat{\rho} = 50.8151$ (14.5102)
		$\widehat{\gamma} = 0.1669$ (0.0512)	$\widehat{\rho} = 50.8151$ (14.5397)		
Estadístico χ^2		2.3114	390.791	411.025	390.791
$p-$valor		0.5103	0.0000	0.0000	0.0000

Tabla 5.4: Frecuencias observadas y esperadas, estimaciones MV de los parámetros (errores estándar entre paréntesis) y contraste de bondad de ajuste χ^2 para el número de bibliotecas públicas por municipio en Andalucía

	$\overline{x}$	s^2	IA	Q_1	Q_2	Q_3	Max
Huelgas	0.99	0.74	0.7467	0	1	1	4

Tabla 5.5: Resumen descriptivo del número de huelgas realizadas

5.2. Huelgas en la industria minera

Se modelizan los datos referentes al número de huelgas realizadas en un periodo de cuatro semanas en el sector minero de Reino Unido durante los años $1948-1959$ (Kendall, 1961). La Tabla 5.5 contiene un resumen descriptivo de estos datos, donde se observa que presentan infradispersión, con un cociente varianza-media igual a 0.7467. En consecuencia, se ajustan las distribuciones infradispersas CTP, CMP y HP. Para cada ajuste se calculan las correspondientes frecuencias esperadas, el AIC y se aplica el contraste de bondad de ajuste χ^2. Los resultados se incluyen en la Tabla 5.6.

Se concluye que la distribución CTP es la que proporciona un mejor ajuste para este conjunto de datos, dado que presenta un valor AIC inferior a las otras dos distribuciones, así como mayor p-valor asociado al contraste de bondad de ajuste. Además, la frecuencia esperada que proporciona para el primer valor modal de la variable, es decir, 1, es la más próxima a la frecuencia observada, situación que se repite para los restantes valores observados de la variable. Para visualizar gráficamente esta afirmación, se incluye la Figura 5.3 con las frecuencias observadas y esperadas.

5.3. Conatos de incendios en los municipios de Andalucía

A continuación consideradamos la variable "número de conatos de incendios por municipio en Andalucía". Según la Dirección General de Protección Civil y Emergencias del Ministerio Interior del Gobierno de España, un conato es un "fuego pequeño en sus orígenes,

		Esperadas			
X	*Observadas*	CTP	CMP	HP	EBW
0	46	47.0256	47.4920	46.2770	49.48638
1	76	72.6824	70.4330	73.3368	66,581
2	24	28.5577	30.6571	28.7903	32.1434
3	9	6.0448	6.5141	6.4501	7.0309
≥ 4	1	1.6895	0.9038	1.1459	0.7244
		$\widehat{a} = -1.5429$ (0.7662)	$\widehat{\lambda} = 1.4830$ (0.2521)	$\widehat{\gamma} = 0.3293$ (0.1115)	$\widehat{\alpha} = -7.0659$ (2.4353)
		$\widehat{b} = 1.8604$ (0.8021)	$\widehat{\nu} = 1.7686$ (0.2945)	$\widehat{\lambda} = 0.5219$ (0.1065)	$\widehat{\gamma} = 37.1084$ (29.8034)
		$\widehat{\gamma} = 3.7796$ (3.7612)			
AIC		**382.0883**	380.0228	379.0609	381.41
Estadístico χ^2		2.6273	2.8914	1.9220	4.2697
$p - valor$		0.1050	0.2356	0.3825	0.1183

Tabla 5.6: Frecuencias observadas, esperadas, estimaciones MV de los parámetros (erroes estándar entre paréntesis), AIC y contraste de bondad de ajuste χ^2 para el número de huelgas en la industria minera

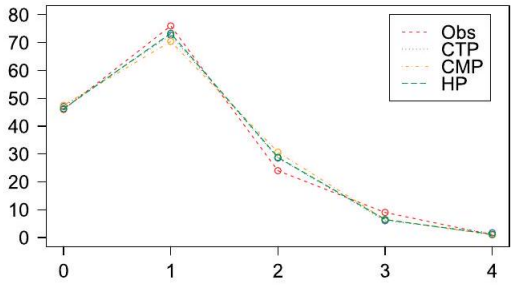


Figura 5.3: Frecuencias esperadas y observadas para el número de huelgas en Reino Unido

	$\overline{x}$	s^2	IA	Q_1	Q_2	Q_3	Max
Conatos de incendios	3.53	28.97	8.21	1	2	4.75	56

Tabla 5.7: Resumen descriptivo para el número de conatos de incendio en los municipios andaluces

	CTP	BN	GP	CBP
AIC	3253.45	3292.70	3263.29	3278.42
Estadístico χ^2	23.21	45.28	44.25	60.06
$p-$valor	0.06	0.00	0.00	0.00
	UGW	CMP	HP	EBW
AIC	3254.24	3302.99	3303.30	3252.24
Estadístico χ^2	24.06	54.15	60.25	24.06
$p-$valor	0.05	0.00	0.00	0.06

Tabla 5.8: Valores de AIC y el contraste de bondad de ajuste χ^2 para los conatos de incendios en los municipios andaluces.

fácilmente controlable pero que, si se le deja evolucionar, puede dar lugar a un incendio". Los datos, correspondientes al periodo $2001-2014$, han sido extraidos del Banco de datos de la Naturaleza del Ministerio de transición ecológica y reto demográfico, eliminando los municipios que no tuviesen superficie forestal. Para comenzar, se realiza una descripción de la variable, mostrando estadísticos descriptivos como la media aritmética, la varianza, el IA, los cuartiles y el máximo de los datos (Tabla 5.7). Claramente, los datos presentan una fuerte sobredispersión. Ajustamos los modelos sobredispersos CTP, BN, GP, UGW, CMP, HP y EBW. Los resultados obtenidos se recogen en la Tabla 5.8. A partir de ella, se puede determinar que no hay evidencia para rechazar la hipťesis nula en el contraste de bondad de ajuste (H_0: La variable se distribuye según una distribución concreta) para las distribuciones EBW, CTP y UGW. De estas, la que posee menor AIC es la distribución EBW.

La Tabla 5.9 contiene las frecuencias observadas y esperadas para cada modelo estimado, así como los correspondientes parámetros estimados por MV y sus errores estándar (entre paréntesis). A pesar de que la distribución que mostraba mejor valor AIC e incluso un $p-$valor más elevado era la distribución EBW, es la CTP la que presenta frecuencias esperadas más cercanas a las observadas, en general. Este comportamiento también se observa en la Figura 5.4.

Finalmente, para el modelo EBW sobredisperso estimado obtenemos la descomposición de la varianza en aleatoriedad, riesgo y predisposición. En porcentaje, estas componentes son 10.2 %, 33.6 % y 56.2 %, respectivamente. En consecuencia, la aleatoriedad no es muy importante, siendo la predisposición la más relevante, con más del 50 % de la variabilidad. Esto puede deberse a las características particulares de cada municipio que estaán relacionadas con el número de conatos habidos durante este periodo, así como de otras causas que se puedan desconocer.

5.4. Poema turco

Se considera la longitud de cada palabra (en número de sílabas) en el poema turco *Gidisat* escrito por Ercüment Behzat Lâv y disponible en Wimmer et al. (1994).

		Esperadas		
X	*Observadas*	*CTP*	*EBW*	*UGW*
0	150.00	154.49	159.31	159.31
1	161.00	144.67	139.42	139.42
2	94.00	104.90	101.69	101.69
3	73.00	72.71	71.87	71.87
4	39.00	50.66	51.04	51.04
5	34.00	35.97	36.79	36.79
6	41.00	26.11	27.00	27.00
7	20.00	19.37	20.17	20.17
8	12.00	14.65	15.33	15.33
9	11.00	11.29	11.83	11.83
10	10.00	8.83	9.26	9.26
11	4.00	7.02	7.34	7.34
12	6.00	5.65	5.89	5.89
13	4.00	8.38 4.60	8.68	8.68
14	2.00			
15-22	19.00	15.16	15.33	15.33
23-30	8.00	5.02	4.79	4.79
31-56	4.00	5.14	4.26	4.26
		$\widehat{a} = 1.8796$ (0.8138)	$\widehat{\alpha} = 2.749$ (0.162)	$\widehat{a} = 2.7495$ (1.5848)
		$\widehat{b} = 1.5803$ (0.5091)	$\widehat{\rho} = 3.139$ (0.317)	$\widehat{k} = 2.7495$ (1.5848)
		$\widehat{\gamma} = 6.440$ (0.5091)		$\widehat{\rho} = 3.1392$ (0.3172)

Tabla 5.9: Frecuencias observadas, esperadas, estimaciones MV de los parámetros (errores estándar entre paréntesis) y contraste de bondad de ajuste χ^2 para el número de conatos de incendios por municipio en Andalucía

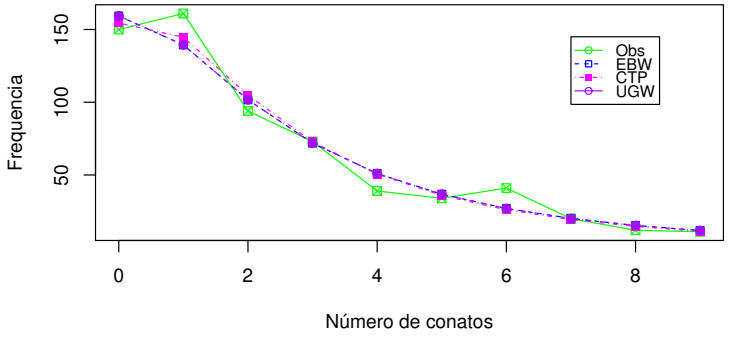


Figura 5.4: Frecuencias esperadas y observadas del número de conatos de incendios.

	$\overline{x}$	s^2	IA	Q_1	Q_2	Q_3	Max
X	1.58	1.17	0.74	1	2	2	5

Tabla 5.10: Estadísticos descriptivos para el número de sílabas -1

		Esperadas			
X	*Observadas*	EBW	CTP	CMP	HP
1	64	61.24	61.24	59.69	61.20
2	131	136.23	136.23	141.87	145.24
3	122	121.68	121.68	118.70	112.60
4	61	56.93	56.93	53.92	52.17
5	13	15.27	15.27	15.88	17.27
≥ 6	3	2.66	2.66	3.94	5.55
		$\widehat{\alpha} = -10.530$ (2.144)	$\widehat{a} = -10.530$ (2.158)	$\widehat{\lambda} = 2.377$ (0.276)	$\widehat{\gamma} = 0.485$ (0.099)
		$\widehat{\gamma} = 49.843$ (24.257)	$\widehat{b} = 0.001$ (14.254)	$\widehat{v} = 1.506$ (0.137)	$\widehat{\lambda} = 1.151$ (0.104)
			$\widehat{\gamma} = 49.843$ (24.416)		
	AIC	1158.3	1160.3	1160.7	1164.4
	Estadístico χ^2	1.000	1.000	2.914	6.014
	$p-$valor	0.801	0.606	0.405	0.111

Tabla 5.11: Frecuencias observadas y esperadas, estimación de los parámetros (errores estándar), AIC y el contraste de bondad de ajuste χ^2 para el número de sílabas por palabras del poema turco

Para que el rango de la variable X: "número de sílabas" empiece en 0, se considera $X-1$ como si los datos se generasen añadiendo 1 a la distribución inicial. Un resumen descriptivo de estos datos aparece en la Tabla 5.10. Como se observa, estos datos son infradispersos, ya que la razón entre la varianza y la media es inferior a 1, concretamente 0.74.

Por tanto, se estiman los modelos EBW, CTP, HP y CMP a partir de los datos. La Tabla 5.11 contiene las estimaciones MV de los parámetros estimados con sus errores estándar (entre paréntesis), los valores del AIC, las frecuencias observadas y esperadas de los valores de la variable y el estadístico, junto con el $p-$valor, del contraste de bondad de ajuste χ^2 de Pearson.

Se observa que los ajustes de las distribuciones CTP y EBW son prácticamente iguales, ya que el parámetro a de la CTP no es significativo. Por ello, se ha suprimido el modelo CTP de la Figura 5.5 que contiene las frecuencias observadas y esperadas de los valores de la variable. Finalmente, se puede determinar que la mejor distribución que permite modelizar estos datos es la EBW, ya que - aunque las frecuencias esperadas son iguales que las proporcionadas por la CTP, el $p-valor$ es mayor, al tener más grados de libertad (se estima un parámetro menos).

5.5. Número de granjas ecológicas en los municipios de Andalucía

Huete-Morales and Marmolejo-Martín (2020) propusieron la distribución UGW para la modelización de las granjas ecológicas en Andalucía. Utilizamos los datos proporcionados

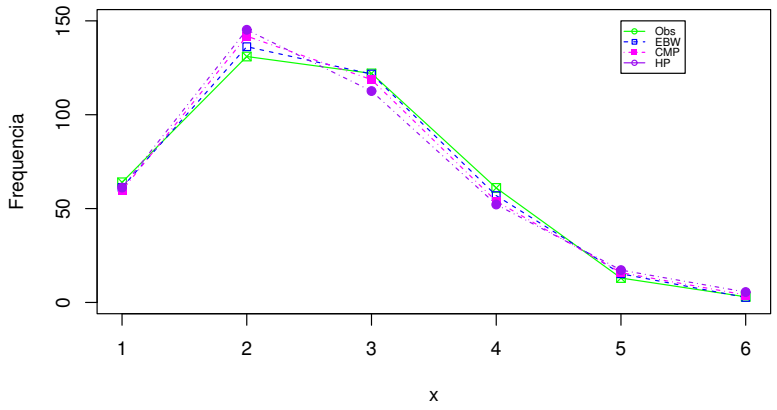


Figura 5.5: Frecuencias observadas y esperadas del número de sílabas por palabras del poema turco.

	$\overline{x}$	s^2	IA	Q_1	Q_2	Q_3	Max
Granjas	2.16	38.37	17.75	0.00	0.00	1.00	70.00

Tabla 5.12: Estadísticos descriptivos para el número de granjas ecológicas en los municipios andaluces

en dicho artículo para ajustar los modelos CTP y EBW. Un resumen descriptivo para los datos de la variable "número de granjas ecológicas en los municipios andaluces" aparece en la Tabla 5.12. Estos datos presentan uns sobredispersión servera, por lo que las distribuciones con las que vamos a modelizar estos datos son BN, GP, UGW, CTP, CMP y EBW.

En la Tabla 5.13 se incluyen los resultados obtenidos para estos ajustes, junto con los de la UGW proporcionados por los autores del artículo mencionado. Las estimaciones de los parámetros de la distribución UGW son $\widehat{a} = 0.3608$, $\widehat{k} = 4.3361$ y $\widehat{\rho} = 1.6262$. A pesar de que el $p-$valor asociado al contraste de bondad de ajuste de la UGW es mejor que en el resto de los ajustes, el mejor AIC se corresponde con el ajuste del modelo EBW.

A continuación, en la Tabla 5.14 se muestran las frecuencias observadas y esperadas correspondientes a los modelos ajustados con $p-$valor mayor que 0.05 (se excluye la CMP).

La distribución UGW reproduce de forma más precisa el valor modal en 0. Sin embargo,

	CTP	BN	GP	CBP
AIC	2522.6799	2542.5578	2515.1661	2541.6347
Estadístico χ^2	14.1742	45.5476	13.9177	39.3382
$p-$valor	0.2897	0.0001	0.4559	0.0002
	UGW	$UGW^{(*)}$	CMP	EBW
AIC	2522.6799	2544.4012	3042.8248	2520.6799
Estadístico χ^2	14.1743	9.2303	405.9935	14.1736
$p-$valor	0.2897	0.6831	0.0000	0.3617

Tabla 5.13: Valores del AIC y del contraste de bondad de ajuste χ^2 para el número de granjas ecológicas en los municipios andaluces. $(*)$ Ajuste en Huete-Morales and Marmolejo-Martín (2020)

		Esperadas			
X	*Observadas*	*EBW*	$UGW^{(*)}$	*GP*	*CTP*
0.00	456.00	452.00	457.25	459.33	451.99
1.00	122.00	125.96	113.13	110.86	125.96
2.00	49.00	57.38	56.09	52.79	57.37
3.00	32.00	32.30	33.60	31.61	32.29
4.00	19.00	20.50	22.21	21.26	20.50
5.00	17.00	14.08	15.65	15.35	14.08
6.00	13.00	10.21	11.53	11.63	10.21
7.00	9.00	7.72	8.78	9.11	7.72
8.00	8.00	6.02	6.88	7.32	6.02
9.00	8.00	4.81	5.50	6.01	4.81
10.00	3.00	3.93	4.48	5.01	3.93
11.00	3.00	3.26	3.71	4.23	3.26
12.00	2.00	2.75	3.11	3.62	2.75
13.00	3.00	2.35	2.63	3.12	2.36
14.00	1.00	2.02	2.26	2.71	2.02
15.00	2.00	1.76	1.95	2.37	1.76
16.00	3.00	1.54	1.70	2.09	1.54
		$\widehat{\alpha} = 0.9239$ (0.0621)	$\widehat{a} = 0.3608$	$\widehat{\lambda} = 0.5166$ (0.0289)	$\widehat{a} = 0.9239$ (0.0621)
		$\widehat{\rho} = 1.2151$ (0.1135)	$\widehat{k} = 4.3361$	$\widehat{\theta} = 0.7611$ (0.0222)	$\widehat{b} = 0.000002$ (0.4506)
			$\widehat{\rho} = 1.6262$		$\widehat{\gamma} = 3.0629$ (0.2287)

Tabla 5.14: Frecuencias observadas y esperadas, estimaciones MV de los parámetros y errores estándar (entre paréntesis) para el número de granjas ecológicas en los municipios de Andalucía. $(*)$ Ajuste en Huete-Morales and Marmolejo-Martín (2020)

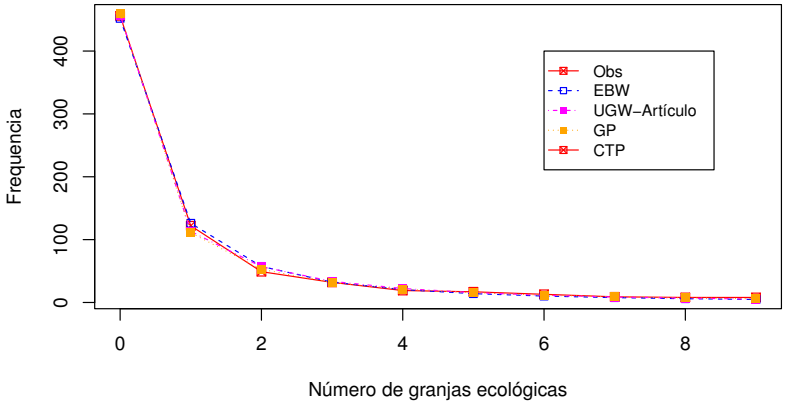


Figura 5.6: Frecuencias observadas y esperadas, estimación de los parámetros (errores estándar), AIC y contraste de bondad de ajuste χ^2 para el número de granjas ecológicas en los municipios de Andalucía

las demás distribuciones modelizan mejor el resto de valores modales. Cabe señalar que la distribución UGW ajusta mejor la cola de la distribución, mientras que la distribución EBW ajusta mejor los primeros valores de la variable. Esto también queda reflejado en la Figura 5.6.

Capítulo 6

Futuras vías de investigación

6.1. Introducción

La generación y aplicaciones de modelos de probabilidad para datos de conteo es un campo de investigación que lleva abierto desde los albores del Cálculo de Probabilidades y que se mantiene muy vivo aún en nuestros días, como lo pueba la cantidad de artículos dedicados a tal fin con el advenimiento de nuevos modelos. Cada modelo desarrollado y estudiado lleva implícito el estudio y comprensión de las propiedades inherentes a dicho modelo, de modo que, una vez modelizados unos datos reales mediante un modelo concreto, tales propiedades pueden ser utilizadas para una mejor interpretación y conocimiento del fenómeno estudiado. Por ejemplo, si se analiza el número de afectados por una enfermedad en un territorio, el hecho de que los datos puedan ajustarse por una distribución de Poisson o por una distribución binomial negativa, cambia drásticamente el modo en que las autoridades sanitarias han de enfrentarse a ella.

En este sentido, esta tesis se centra en el desarrollo de un modelo concreto, perteneciente a la familia de distribuciones gaussianas discretas, que cubre un caso fronterizo entre modelos ampliamente estudiados: los derivados del caso en el que el polinomio L tiene dos raíces reales y los derivados del caso en el que las raices de ese polinomio son complejas conjugadas. El caso especial es aquél en el que el polinomio L tiene una raíz doble. Esto se traduce en un modelo biparamétrico que hereda propiedades de modelos con tres parámetros, que permite modelizar datos tanto sobredispersos como infradispersos y que, en el caso de raíz doble positiva, propociona un modelo similar al de la distribución univariante generalizada de Waring, pero con la ventaja de ser una versión biparametria obtenida a través de una doble mixtura de una Poisson, y que elimina la indeterminación en el cálculo de la partición de la varianza no debida al "puro azar". Además, desde el punto de vista de su aplicabilidad, proporciona valores de verosimilitud parecidos a los ajustes con una UGW, pero con más grados de libertad, tanto en el cálculo del AIC como para la decisión de aceptación o rechazo en el contraste de bondad de ajuste. A su vez, el caso de una raíz doble negativa puede verse como una extensión de la distribución de Waring al caso infradisperso. Por último, también permite desarrollar un modelo infradisperso para datos de conteo finito, cuando el primer parámetro es un entero negativo.

6.2. Futuras vías de extensión

A partir de las distribuciones desarrolladas en los capítulos anteriores, se plantean diferentes posibilidades de extensión de los resultados obtenidos.

6.2.1. Generalización de los modelos obtenidos

Las distribuciones presentadas se han obtenido como solución de una ecuación en diferencias con coeficientes polinomiales. Tal como hemos visto, la función generadora de las distribuciones desarrolladas es una ${}_2F_1(\alpha, \beta, \gamma; 1)$. Por tanto, una primera expansión es el análisis de distribuciones discretas generadas cuendo λ es distinto de 1. Distribuciones generadas de tal manera pueden verse, por ejemplo, en Rodríguez-Avi et al. (2007) en el caso de rango infinito numerable y Rodríguez-Avi et al. (2007b) en el supuesto de datos de rango finito. La adición del nuevo parámetro proporciona una mayor flexibilidad a las distribuciones generadas, lo que permiten un mejor ajuste a los datos reales con los que se trabaja, pero adolecen de la falta de teoremas de sumación, lo que impide la expresión de los parámetros de manera exacta mediante ecuaciones y la necesidad de obtener las probabilidades de manera numérica y aproximada, con el orden de precisión que se desee, mediante métodos computacionales. Además, en el caso de la serie infinita, si $0 < \lambda < 1$ la función de Gauss es convergente siempre, por lo que se eliminan restricciones en los parámetros y además, la existencia de todos los momentos está garantizada. En el caso finito, λ puede ser cualquier número real positivo, y las probabilidades y momentos se pueden calcular de manera exacta.

Así, se está empezando a analizar el caso finito con la introducción del nuevo parámetro λ, por lo que la función generatriz viene dada en términos de la función de Gauss ${}_2F_1(-n, -n, \gamma; \lambda)$. Este nuevo parámetro proporciona una mayor versatilidad al modelo finito descrito, y permite un mayor rango de aplicabilidad, en competencia con modelos similares, como la distribución beta-binomial. La introducción de λ real está relacionado con un desplazamiento del peso de la probabilidad hacia los extremos de la distribución. En el caso de la CTP, también se puede estudiar el caso en que la función generatriz se exprese en términos de la función de Gauss con parámetros complejos ${}_2F_1(a + bi, a - bi, \gamma, \lambda)$, con $0 < \lambda < 1$. En ambos casos, es necesario un análisis completo de los modelos generados, comparación con otras distribuciones, desarrollo de técnicas de estimación y del software adecuado para su uso por parte de los investigadores.

6.2.2. Modelos inflados de ceros

Los modelos inflados de ceros se corresponden con modelos para datos de conteo en los cuales existe una gran cantidad de valores 0, que pueden ocurrir de manera estructural o aleatoria. Un ejemplo consiste en estudiar el número de hijos por familia. En este caso, el número de ceros puede tener un origen aleatorio - parejas que aún no han tenido hijos, por ejemplo, pero que pueden tenerlos - o no aleatorio: parejas que han decidido voluntariamente no tener hijos. En ese caso, la obtención de un valor 0 no se debe al azar. Este tipo de datos aparecen frecuentemente en análisis de encuestas, o en el número de visitas a los servicios de salud a menudo se incluyen muchos ceros que representan a los pacientes sin utilización durante un tiempo de seguimiento. Tales modelos proponen una mezcla de dos procesos: por un lado, se considera que un proceso binario genera los ceros estructurales; por otro lado, cuando un dato no es un cero estructural, el cero se genera como resultado de un proceso aleatorio modelado por una distribución de conteo, generalmente Poisson o binomial negativa (véase por ejemplo Sáez-Castillo and Conde-Sánchez (2013);

Baetschmann and Winkelmann (2013); Feng (2021); Hagen et al. (2023); Sengupta and Roy (2023)). En este tipo de datos, aun cuando el modelo subyacente resulte ser infradisperso, la acumulación de ceros, al disminuir la media, puede hacer que el modelo considerado de forma conjunta sea sobredisperso.

Dentro de este tipo de modelos, la vía de ampliación consiste en proponer como modelo aleatorio subyacente a las distribuciones desarrolladas, CTP y EBW. Para ello, hay que proponer procedimientos de determinación de la proporción de ceros estructurales, así como el estudio y análisis de las distribuciones obtenidas y los procedimientos de estimación.

Otra posible vía consiste en cambiar el 0 estructural por un valor bajo de la variable, por ejemplo, el 1. Esto puede verse en el caso de equipamientos públicos, ya que, bajo ciertas condiciones legales, debe existir 1 elemento al menos, de manera que el hecho de que exista no es aleatorio, sino determinístico.

6.2.3. Aplicación en modelos de regresión generalizados

Una aplicación importante de las distribuciones de conteo es la posibilidad de utilizarlas como distribuciones residuales para modelos de regresión lineales generalizados. En este caso, a la hora de explicar un fenómeno discreto en función de un conjunto de variables explicativas, se considera que cada celda (es decir, cada conjunto de observaciones con los mismos valores de las covariables utilizadas) presenta la misma distribución de conteo, con parámetros dependientes de las covariables. En este sentido el primer modelo que se empleó es la distribución de Poisson en donde la media, λ, viene expresada en función de las covariables

$$\lambda_x = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right),$$

lo que se traduce en una conversión directa del caso de la regresión lineal múltiple, en donde la hipótesis de normalidad de los errores implica que se deben exclusivamente al azar, al caso discreto infinito numerable.

Pero la imposición del modelo de Poisson incorpora la exigencia de la equidispersión en las celdas, lo que muchas veces es poco realista. Para flexibilizar esa exigencia, se han desarrollado otros modelos de regresión en los que la variable residual es otra distribuición de conteo, habitualmente sobredispersa. También se han desarrollado modelos de regresión basados en la distribución binomial negativa (véase, por ejemplo, Hilbe (2011)), en la distribución de Poisson generalizada (Cameron and Trivedi, 2013), en la distribución generalizada de Waring (Rodríguez-Avi et al., 2009), en la distribución CBP (Rodríguez-Avi and Olmo-Jiménez, 2017), en la distribución hyper-poisson generalizada (Sáez-Castillo and Conde-Sánchez, 2013; Khazraee et al., 2015) o modelos de regresión para variables infladas de ceros (Tanış et al., 2023; Sáez-Castillo and Conde-Sánchez, 2017), entre otros.

En este caso, la vía de investigación se abre a la hora de proponer modelos de regresión donde la distribución subyacente sea alguna de las dos desarrolladas, CTP y EBW. El estudio consiste en proponer los modelos, los métodos de estimación según los parámetros que sean explicados por las covariables (o la media y la varianza), procedimientos de validación y el desarrollo del software adecuado, vía librería implementada en R.

Bibliografía

Abdella, G. M., Kim, J., Al-Khalifa, K. N., and Hamouda, A. M. (2019). Penalized conway-maxwell-poisson regression for modelling dispersed discrete data: The case study of motor vehicle crash frequency. *Safety Science*, 120:157–163.

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Acu, A.-M. and Rasa, I. (2023). A discrete probability distribution and some applications. *Mediterranean Journal of Mathematics*, 20(1).

Adeoti, O. A., Malela-Majika, J.-C., Shongwe, S. C., and Aslam, M. (2022). A homogeneously weighted moving average control chart for conway–maxwell poisson distribution. *Journal of Applied Statistics*, 49(12):3090–3119.

Ahsan-ul Haq, M. and Zafar, J. (2023). A new one-parameter discrete probability distribution with its neutrosophic extension: mathematical properties and applications. *International Journal of Data Science and Analytics*.

Al-Babtain, A. A., Ahmed, A. H. N., and Afify, A. Z. (2020). A new discrete analog of the continuous lindley distribution, with reliability applications. *Entropy*, 22(6):603.

Andrews, G. E., Askey, R., and Roy, R. (2013). *Chapter 1. The Gamma and Beta functions*, pages 1–60. Cambridge University Press, New York.

Appell, P. and Kampe de Feriet, J. (1926). *Fonctions Hypergéometriques et Hyperspheriques: Polinômes D'Hermite*. Gauthier-Villars, Paris.

Ariza-López, F. and Rodríguez-Avi, J. (2015). Estimating the count of completeness errors in geographic data sets by means of a generalized waring regression model. *International Journal of Geographical Information Science*, 29(8):1394–1418.

Arnold, T. B. and Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3/2.

Baetschmann, G. and Winkelmann, R. (2013). Modeling zero-inflated count data when exposure varies: With an application to tumor counts. *Biometrical Journal*, 55(5):679–686.

Balakrishnan, N., Koutras, M. V., and Milienos, F. S. (2018). A weighted poisson distribution and its application to cure rate models. *Communications in Statistics - Theory and Methods*, 47(17):4297–4310.

Bardwell, G. E. and Crow, E. L. (1964). A two parameter family of hyper-poisson distributions. *J. Am. Stat. Assoc.*, 54:133–141.

Bedbur, S. and Kamps, U. (2023). Uniformly most powerful unbiased tests for the dispersion parameter of the conway–maxwell–poisson distribution. *Statistics & Probability Letters*, page 109801.

Berger, E., Larsen, J., Freudenberg, N., and Jones, H. E. (2022). Food insecurity associated with educational disruptions during the covid-19 pandemic for college students and the role of anxiety and depression. *Journal of American College Health*, 0(0):1–4. PMID: 35834743.

Bogdanov, Y. I., Bogdanova, N. A., Katamadze, K. G., Avosopyants, G. V., and Lukichev, V. F. (2020). Hyper-poisson photon statistics. *JETP Letters*, 111:543–548.

Bowman, K. O., Shenton, L. R., and Kastenbaum, M. A. (1991). Discrete Pearson Distributions. *Oak Ridge National Laboratory: Technical Report TM-11899*.

Boyer, C. B. (1950). Cardan and the pascal triangle. *The American Mathematical Monthly*, 57:387–390.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, 2nd edition.

Cameron, A. C. and Johansson, P. (1997). Count data regression using series expansions: with applications. *Journal of Applied Econometrics*, 12(3):203–223.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Castillo, J. D. and Pérez-Casany, M. (1998). Weighted poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, 50(3):567–585.

Castillo, J. D. and Pérez-Casany, M. (2005). Overdispersed and underdispersed poisson generalizations. *Journal of Statistical Planning and Inference*, 134:486–500.

Cavanaugh, J. E. and Neath, A. A. (2019). The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3):e1460.

Consul, P. C. (1989). *Generalized Poisson distributions: properties and applications*. M. Dekker.

Consul, P. C. and Famoye, F. (1989). Minimum variance unbiased estimation for the lagrange power series distributions. *Statistics*, 20(3):407–415.

Consul, P. C. and Jain, G. C. (1973a). A generalization of the poisson distribution. *Technometrics*, 15(4):791–799.

Consul, P. C. and Jain, G. C. (1973b). On some interesting properties of the generalized Poisson distribution. *Biometrische Zeitschrift*, 15(7):495–500.

Consul, P. C. and Shoukri, M. M. (1985). The generalized poisson distribution when the sample mean is larger than the sample variance. *Commun, Statist. - Simula. Computa*, 14(3):667–681.

Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates. *J. Ind. Eng.*, 12:132–136.

Cueva-López, V., Olmo-Jiménez, M. J., and Rodríguez-Avi, J. (2019). Em algorithm for an extension of the waring distribution. *Computational and Mathematical Methods*, 1(5):e1046.

Cueva-López, V., Olmo-Jiménez, M. J., and Rodríguez-Avi, J. (2021). An over and under-dispersed biparametric extension of the waring distribution. *Mathematics*, 9(2):170.

Cueva López, V., Rodriguez-Avi, J., Olmo-Jiménez, M. J., and Rodríguez-Reinoso, J. (2022). A modelling of the number of almazaras by municipality in andalusia. *Studies of Applied Economics*, 40(3).

Cueva-López, V., Rodriguez-Avi, J., Olmo-Jiménez, M. J., and Rodríguez-Reinoso, J. (2022). A modelling of the number of almazaras by municipality in andalusia. *Studies of Applied Economics*, 40(3).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Du, Z., Wang, C., Liu, C., Bai, Y., Pei, S., Adam, D. C., Wang, L., Wu, P., Lau, E. H. Y., and Cowling, B. J. (2022). Systematic review and meta-analyses of superspreading of SARS-CoV-2 infections. *Transboundary and Emerging Diseases*, 69(5).

Eggenberger, F. and Pólya, G. (1923). Über die statistik verketteter vorgänge. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, 3(4):279–289.

Erbayram, T. and Akdoğan, Y. (2023). A new discrete model generated from mixed poisson transmuted record type exponential distribution. *Ricerche di Matematica*.

Esmailian, M., Azimi, R., Gallardo, D. I., Nasiri, P., et al. (2023). New cure rate survival models generated by poisson distribution and different regression structures with applications to cancer data set. *Journal of Mathematics*, 2023.

Fajardo Caldera, M. A. (1985). *Generalizaciones de los Sistemas de Pearson Disretos*. PhD thesis, Universidad de Extremadura.

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of statistical distributions and applications*, 8(1):8.

Fermat, P. d., Fermat, S. d., and of Perga, A. (1679). *Varia opera mathematica*. apud Johannem Pech.

Fernández Coronado, N. A., García-García, J. I., Arredondo, E. H., and Araya Naveas, I. A. (2022). Epistemic configurations and holistic meaning of binomial distribution. *Mathematics*, 10(10).

García-García, J. I., Fernández Coronado, N. A., Arredondo, E. H., and Imilpán Rivera, I. A. (2022). The binomial distribution: Historical origin and evolution of its problem situations. *Mathematics*, 10(15).

Gning, L., Ndour, C., and Tchuenche, J. (2022). Modeling covid-19 daily cases in senegal using a generalized waring regression model. *Physica A: Statistical Mechanics and its Applications*, 597:127245.

Gómez-Déniz, E. (2010). Another generalization of the geometric distribution. *Test*, 19(2):399–415.

Gómez-Déniz, E., Pérez-Rodríguez, J. V., Reyes, J., and Gómez, H. W. (2020). A bimodal discrete shifted poisson distribution. a case study of tourists' length of stay. *Symmetry*, 12(3):442.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 83:255.

Guelfond, A. O. (1963). *Calcul des Différences Finies*. Dunod.

Gutiérrez-Jaimez, R. and Rodríguez-Avi, J. (1997). Family of Pearson Discrete Distributions Generated by the Univariate Hypergeometric Function ${}_3F_2\left(\alpha_1, \alpha_2, \alpha_3; \gamma_1, \gamma_2; 1\right)$. *Applied Stochastics Models and Data Analysis,*, 13:115–125.

Gutiérrez-Jáimez, R., Rodríguez-Avi, J., and Sáez-Castillo, A. J. (1999). Construction of Multivariate Discrete Distributions throught the ${}_2F_1$ Function of Matricial Argument. *IX International Symposium of Applied Stochastics Models and Data Analysis*.

Hagen, T., Reinfeld, N., and Saki, S. (2023). Modeling of parking violations using zero-inflated negative binomial regression: A case study for berlin. *Transportation Research Record*, page 03611981221148703.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Holmukhe, R. M., Jamdade, P. G., and Jamdade, S. G. (2022). Assessment of wind energy potential using poisson distribution model - a field study in india. *Journal of Statistics and Management Systems*, 25:971–981.

Hu, S. (2007). Akaike information criterion. *Center for Research in Scientific Computation*, 93:42.

Huete-Morales, M. D. and Marmolejo-Martín, J. A. (2020). The waring distribution as a low-frequency prediction model: A study of organic livestock farms in andalusia. *Mathematics*, 8(11):2025.

Irshad, M. R., Chesneau, C., Shibu, D. S., Monisha, M., and Maya, R. (2022). A novel generalization of zero-truncated binomial distribution by lagrangian approach with applications for the covid-19 pandemic. *Stats*, 5(4):1004–1028.

Irwin, J. O. (1963). The place of mathematics in medical and biological statistics. *J. Roy. Statist. Soc. A*, 126:1–44.

Irwin, J. O. (1968a). The generalized waring distribution applied to accident theory. *J. Roy. Statist. Soc. A*, 131(2):205–225.

Irwin, J. O. (1968b). The generalized waring distribution. part i. *J. Roy. Statist. Soc. A*, 138:18–31.

Irwin, J. O. (1968c). The generalized waring distribution. part ii. *J. Roy. Statist. Soc. A*, 138:204–227.

Irwin, J. O. (1968d). The generalized waring distribution. part iii. *J. Roy. Statist. Soc. A*, 138:374–378.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*. Wiley, 3rd edition.

Jordan, C. (1965). *Calculus on finite differences*. Chelsea Publishing Company.

Kemp, A. W. and Kemp, C. D. (1975). Models for Gaussian hypergeometric distributions. In Patil, G. P., Kotz, S., and Ord, J. K., editors, *A Modern Course on Statistical Distributions in Scientific Work*, volume 1 of *Nato Science Series C*, pages 31–40.

Kendall, M. G. (1961). Natural law in social sciences. *J. Royal Stat. Soc. Series A*, 124:1–19.

Khazraee, S. H., Sáez-Castillo, A. J., Geedipally, S. R., and Lord, D. (2015). Application of the hyper-poisson generalized linear model for analyzing motor vehicle crashes. *Risk analysis*, 35(5):919–930.

Kuznetsov, V. A., Grageda, A., and Farbod, D. (2022). Generalized hypergeometric distributions generated by birth-death process in bioinformatics. *bioRxiv*, pages 2022–02.

Li, Y., Rahman, T., Ma, T., Tang, L., and Tseng, G. C. (2023). A sparse negative binomial mixture model for clustering rna-seq count data. *Biostatistics*, 24(1):68–84.

Link, A., McGrath, J. S., Zaimagaoglu, M., and Franke, T. (2022). Active single cell encapsulation using saw overcoming the limitations of poisson distribution. *Lab on a Chip*, 22:193–200.

Liu, R., Heo, I., Liu, H., Shi, D., and Jiang, Z. (2023). Applying negative binomial distribution in diagnostic classification models for analyzing count data. *Applied Psychological Measurement*, 47(1):64–75.

Lizama, C. and Ponce, R. (2023). Time discretization and convergence to superdiffusion equations via poisson distribution. *Communications on Pure and Applied Analysis*, 22(2):572–596.

Lu, J., Shen, S., Yuan, F.-T., Shao, Z., Hou, J., and Zheng, X. (2022). The chocolate chip cookie model: Dust geometry of milky way–like disk galaxies. *The Astrophysical Journal*, 938(2):139.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions, 2E*. John Wiley & Sons, Inc.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

Mitov, K. and Nadarajah, S. (2023). Entropy of some discrete distributions. *Methodology and Computing in Applied Probability*, 25(1):2.

Montmort, P. (1714). *Essay d'analyse sur les jeux de hazard*. C. Jombert.

Olmo-Jiménez, M. J. (2002). Contribución al estudio de distribuciones generadas por funciones hipergeométricas con argumentos complejos. Master's thesis, Universidad de Jaén.

Olmo-Jiménez, M. J., Rodríguez-Avi, J., and Cueva-López, V. (2018). A review of the ctp distribution: a comparison with other over- and underdispersed count data models. *Journal of Statistical Computation and Simulation*, 88(14):2684–2706.

Olmo-Jiménez, M. J., Vílchez-López, S., and Rodríguez-Avi, J. (2022). cpd: An r package for complex pearson distributions. *Mathematics*, 10(21).

Olmo-Jiménez, M. J., Vílchez-López, S., and Rodríguez-Avi, J. (2022). cpd: An r package for complex pearson distributions. *Mathematics*, 10(21):4101.

Ord, J. K. (1972). *Families of Frequency Distributions*. Griffin, London.

Pagliara, F. and Mauriello, F. (2020). Modelling the impact of high speed rail on tourists with geographically weighted poisson regression. *Transportation Research Part A: Policy and Practice*, 132:780–790.

Panaretos, J. and Xekalaki, E. (1986). Extension of the application of conway-maxwell-poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis*, 4:313–318.

Pearson, K. (1895). Memoir on skew variation on homogeneous material. *Phylosophical Transactions of the Royal Society of London*, 186(Series A):343–414.

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile ; précédées des Règles générales du calcul des probabilités*. Paris, Bachelier.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 27:311–324.

Ridout, M. S. and Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, 4(1):77–89.

Rivas, L. and Galea, M. (2020). Influence analysis for the generalized waring regression model. *Journal of Applied Statistics*, 47(1):1–27.

Rodrigues, J., de Castro, M., Cancho, V. G., and Balakrishnan, N. (2009). Com–poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, 139(10):3605–3611.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A., and Olmo-Jiménez, M. (2006). Extended waring bivariate distribution. In *Distribution Models Theory*, pages 221–231. World Scientific.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A., and Olmo-Jiménez, M. (2007). A new generalization of the waring distribution. *Computational Statistics & Data Analysis*, 51(12):6138–6150.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A., Olmo-Jiménez, M., and Martínez-Rodríguez, A. (2009). A generalized waring regression model for count data. *Computational Statistics & Data Analysis*, 53(10):3717–3725.

Rodríguez-Avi, J., Conde-Sánchez, A., and Sáez-Castillo, A. J. (2001). Distribuciones Generadas por la Función Hipergeométrica ${}_{p+1}F_p(\alpha_1, \ldots, \alpha_{p+1}; \gamma_1, \ldots, \gamma_p; \lambda)$. *Investigación Operacional,*, 22(2):114–124.

Rodríguez-Avi, J., Conde-Sánchez, A., and Sáez-Castillo, A. J. (2003a). A new class of discrete distributions with complex parameters. *Statistical Papers*, 44(1):67–88.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., and Olmo-Jiménez, M. J. (2003b). Estimation of parameters in gaussian hypergeometric distributions. *Communications in Statistics - Theory and Methods*, 32(6):1101–1118.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., and Olmo-Jiménez, M. J. (2004). A triparametric discrete distribution with complex parameters. *Statistical Papers*, 45(1):81–95.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., and Olmo-Jiménez, M. J. (2007). *An application of the Extended Waring distribution to model count data variables*, pages 35 – 42. World Scientific Publishing, New Jersey.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., and Olmo-Jiménez, M. J. (2007a). Gaussian hypergeometric probability distributions for fitting discrete data. *Communications in Statistics - Theory and Methods*, 36(3):453–463.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., and Olmo-Jiménez, M. J. (2007b). A generalization of the beta-binomial distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(1):51–61.

Rodríguez Avi, J., Gutiérrez Jaimez, R., and Conde Sánchez, A. (1999). Discrete Distributions Generated by the Hypergeometric Function ${}_4F_3$. *Applied Stochastics Models, Proceedings IX International Symposium Lisbon (ed. H. Barcelar-Nicolau, F. Costa-Nicolay y J. Jansen)*, pages 200–205.

Rodríguez-Avi, J. and Olmo-Jiménez, M. J. (2016). Algunas consideraciones sobre la distribución ctp. In *Investigaciones en métodos cuantitativos para la economía y la empresa: homenaje al profesor Rafael Herrerías Pleguezuelo*, pages 629–638. Editorial Universidad de Granada.

Rodríguez-Avi, J. and Olmo-Jiménez, M. J. (2017). A regression model for overdispersed data without too many zeros. *Statistical Papers*, 58(3):749–773.

Rodríguez-Avi, J., Olmo-Jiménez, M. J., Conde-Sánchez, A., and Sáez-Castillo, A. J. (2008). The ${}_3f_2$ with complex parameters as generating function of discrete distribution. *Communications in Statistics - Theory and Methods*, 37(19):3009–3022.

Rodríguez Avi, J., Rodríguez-Reinoso, J., and Cueva-López, V. (2020). Nuevas distribuciones discretas para la modelización de datos sobredispersos: Aplicación a variables municipales. In *XXXIII Congreso Internacional de economía aplicada Asepelt 2019: economía azul*, pages 691–705. Universidade de Vigo.

Roy, S., Tripathi, R. C., and Balakrishnan, N. (2023). A conway–maxwell–poisson type generalization of hypergeometric distribution. *Mathematics*, 11(3):762.

Rutherford, E., Chadwick, J., and Ellis, C. D. (1930). *Radiations from Radioactive Substances*. Cambridge University Press.

Rutherford, E., H., G., and Bateman, H. (1910). Lxxvi. the probability variations in the distribution of α particles. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 20:698–707.

Sáez-Castillo, A. and Conde-Sánchez, A. (2013). A hyper-poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis*, 61:148–157.

Sáez-Castillo, A. and Conde-Sánchez, A. (2013). A hyper-poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis*, 61:148–157.

Sáez-Castillo, A. J. (1997). *Polinomios Zonales y Funciones Hipergeométricas de Argumento Matricial*. Memoria de Licenciatura, Universidad de Granada.

Sáez-Castillo, A. J. (2001). *Generación de Distribuciones Multivariantes Discretas Mediante Funciones Hipergeométricas de Argumento Matricial*. Tesis Doctoral, Memoria de Licenciatura.

Sáez-Castillo, A. J. and Conde-Sánchez, A. (2017). Detecting over-and under-dispersion in zero inflated data with the hyper-poisson regression model. *Statistical Papers*, 58(1):19–33.

Sáez-Castillo, A. J., Conde-Sánchez, A., and Río, F. M. D. (2022). *DGLMExtPois: Double Generalized Linear Models Extending Poisson Regression*. R package version 0.2.1.

Sáez-Castillo, A. J., Vílchez-López, S., Olmo-Jiménez, M. J., Rodríguez-Avi, J., Conde-Sánchez, A., and Martínez-Rodríguez, A. M. (2021). *GWRM: Generalized Waring Regression Model for Count Data*. GWRM R package version 2.1.0.4.

Santhiya, S., Thilagavathi, K., et al. (2023). Geometric properties of analytic functions defined by the (p, q) derivative operator involving the poisson distribution. *Journal of Mathematics*, 2023.

Santos, D. d. S., Cancho, V., and Rodrigues, J. (2019). Hypothesis testing for the dispersion parameter of the hyper-poisson regression model. *Journal of Statistical Computation and Simulation*, 89(5):763–775.

Sapatinas, T. (1995). Identifiability of mixtures of power-series distributions and related characterizations. *Annals of the Institute of Statistical Mathematics*, 47(3):447–459.

Satheesh Kumar, C. and Harisankar, S. (2020). On some aspects of a general class of yule distribution and its applications. *Communications in Statistics-Theory and Methods*, 49(12):2887–2897.

Satheesh Kumar, C. and Ramachandran, R. (2020). On zero-inflated hyper-poisson distribution and its applications. *Communications in Statistics-Simulation and Computation*, 51(3):868–881.

Sellers, K., Lotze, T., and Raim, A. (2019). *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*. R package version 0.7.0.

Sellers, K. F. (2023). *The Conway–Maxwell–Poisson Distribution*, volume 8. Cambridge University Press.

Sellers, K. F., Borle, S., and Shmueli, G. (2011). The COM-poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28(2):104–116.

Sengupta, D. and Roy, S. (2023). Modelling zero inflated and under-reported count data. *Journal of Statistical Computation and Simulation*, pages 1–20.

Ser, G. (2022). Extreme variability modelling of overdispersed germination count experiments. *Chilean journal of agricultural research*, 82:619–627.

Shah, S. Q. A., Khan, F. Z., and Ahmad, M. (2022). Mitigating tcp syn flooding based edos attack in cloud computing environment using binomial distribution in sdn. *Computer Communications*, 182:198–211.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the conway-maxwell-poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54:127–142.

Sibuya, M. (1979). Generalized hypergeometric, digamma and trigamma distributions. *Annals of the Institute of Statistical Mathematics*, 31(3):373–390.

Sibuya, M. and Shimizu, R. (1981). The generalized hypergeometric family of distributions. *Annals of the Institute of Statistical Mathematics*, 33(2):177–190.

Singh, B. P., Singh, G., Das, U. D., and Maurya, D. K. (2021). An under-dispersed discrete distribution and its application. *Journal of Statistics Applications and Probability Letters*, 8(3):205–213.

Singh, S., Chawla, M., Prasad, D., Anand, D., Alharbi, A., and Alosaimi, W. (2022). An improved binomial distribution-based trust management algorithm for remote patient monitoring in wbans. *Sustainability*, 14(4).

Slater, L. J. (1966). *Generalized Hypergeometric Functions*. London: Cambridge University Press.

Steutel, F. W., Kent, J. T., Bondesson, L., and Barndorff-Nielsen, O. (1979). Infinite divisibility in theory and practice. *Scandinavian Journal of Statistics*, 6:57–64.

Student (1907). On the error of counting with a haemacytometer. *Biometrika*, 5:351–360.

Sáez-Castillo, A. J., Conde-Sánchez, A., and Martínez Del Río, F. (2023). Dglmextpois: Advances in dealing with over and under-dispersion in a double glm framework. *R Journal*. (in press).

Tang, J., Gao, F., Liu, F., Han, C., and Lee, J. (2020). Spatial heterogeneity analysis of macro-level crashes using geographically weighted poisson quantile regression. *Accident Analysis & Prevention*, 148:105833.

Tanış, C., Koç, H., and Pekgör, A. (2023). A new zero–inflated discrete lindley regression model. *Communications in Statistics-Theory and Methods*, pages 1–22.

Team, R. C. (2023). *R: A Language and Environment for Statistical Computing*.

Tyas, S. W., Puspitasari, L. A., et al. (2023). Geographically weighted generalized poisson regression model with the best kernel function in the case of the number of postpartum maternal mortality in east java. *MethodsX*, 10:102002.

Vagelas, I. and Leontopoulos, S. (2022). Modeling the overdispersion of pasteuria penetrans endospores. *Parasitologia*, 2:206–227.

Varadhan, R. and Gilbert, P. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, 32(4):1–26.

Vasquez, J. K., Rodrigues, J., and Balakrishnan, N. (2022). A useful variance decomposition for destructive waring regression cure model with an application to hiv data. *Communications in Statistics-Theory and Methods*, 51(20):6978–6989.

Vilchez-Lopez, S., Olmo-Jimenez, M. J., and Rodriguez-Avi, J. (2022). *cpd: Complex Pearson Distributions*. R package version 0.3.0.

Vílchez-López, S., Sáez-Castillo, A. J., and Olmo-Jiménez, M. J. (2016). GWRM: An r package for identifying sources of variation in overdispersed count data. *PLOS ONE*, 11(12):e0167570.

von Bortkiewicz, L. (1898). *Das Gesetz der Kleinen Zahlen*. Teubner.

Willmot, G. (1986). Mixed compound poisson distributions. *ASTIN Bulletin*, 16.

Wimmer, G., Köhler, R., Grotjahn, R., and Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1(1):98–106.

Witowski, V. and Foraita, R. (2018). *HMMpa: Analysing Accelerometer Data Using Hidden Markov Models*. R package version 1.0.1.

Xekalaki, E. (1984). The bivariate generalized waring distribution and its application to accident theory. *J. Royal Stat. Soc. Series A*, 147:488–498.

Xelakaki, E. (1983a). Infinite divisibility, completeness and regression properties of the univariate generalized waring distribution. *Ann. Inst. Stat. Math.*, 35:279–289.

Xelakaki, E. (1983b). The univariate generalized waring distribution in relation to accident theory: proneness, spells or contagion? *Biometrics*, 39:887–895.

Yang, Z., Hardin, J. W., and Addy, C. L. (2009). A score test for overdispersion in poisson regression based on the generalized poisson-2 model. *Journal of Statistical Planning and Inference*, 139(4):1514–1521.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, USA: Springer.