

Modelización del exceso de ceros en los datos de conteo

**Una nueva Perspectiva
sobre los enfoques de
modelización**



índice

Temas

0.Resumen del articulo

1.Introduccion

2.Ejemplo

3.Modelos Explicitos

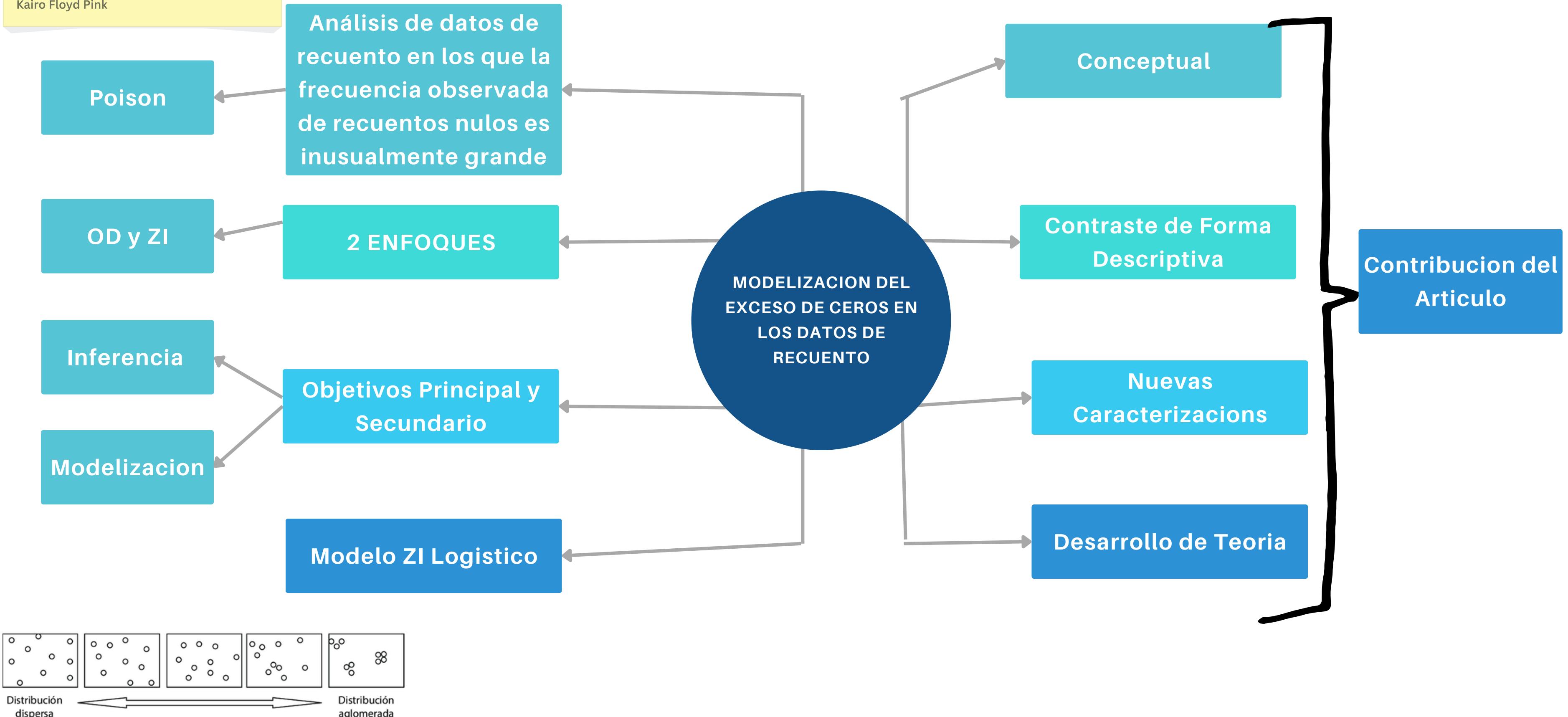
4.Modelos Implicitos

5.Paquetes R

La distribución de Poisson se caracteriza porque su esperanza y su varianza coinciden; cuando se ajusta un modelo de Poisson a un conjunto de datos puede ocurrir que ambos valores difieran de manera significativa. Se dice entonces que el modelo presenta sobredispersión

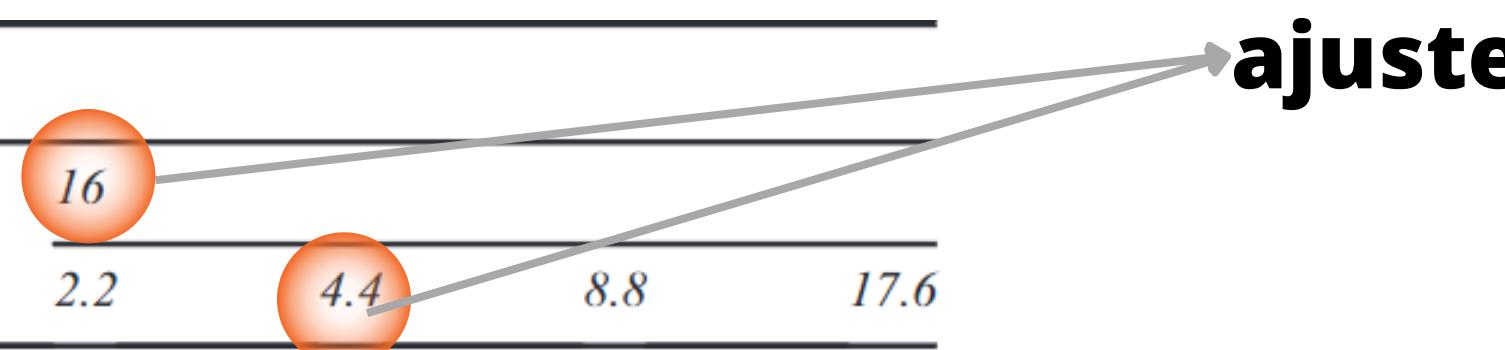
Kairo Floyd Pink

0. RESUMEN: MODELIZACION DEL EXCESO DE CEROS EN LOS DATOS DE CONTEO

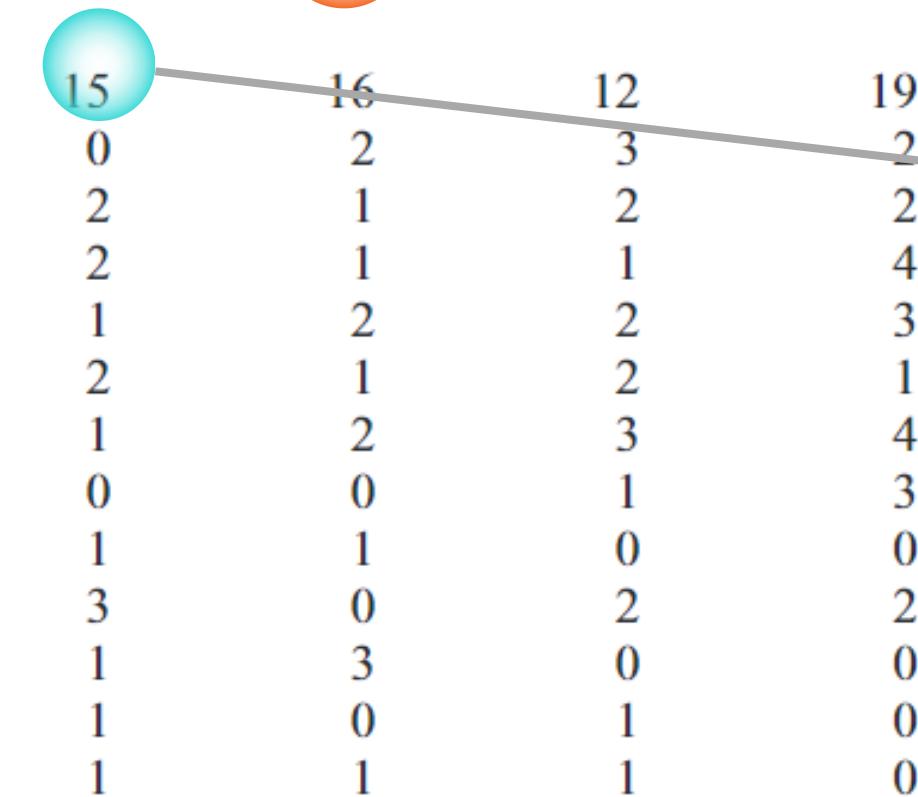


Distribuciones de frecuencia

	Photoperiod							
BAP (μM)	2.2	4.4	8.8	17.6	2.2	4.4	8.8	17.6
No. of roots	0	0	0	2	15	16	12	19
0	3	0	0	0	0	2	3	2
1	2	3	1	0	2	1	2	2
2	3	0	2	2	2	1	1	4
3	6	1	4	2	1	2	2	3
4	3	0	4	5	2	1	2	1
5	2	3	4	5	1	2	3	4
6	2	7	4	4	0	0	1	3
7	3	3	7	8	1	1	0	0
8	1	5	5	3	3	0	2	2
9	2	3	4	4	1	3	0	0
10	1	4	1	4	1	0	1	0
11	0	0	2	0	1	1	1	0
>12	13,17	13	14,14	14				



ajuste

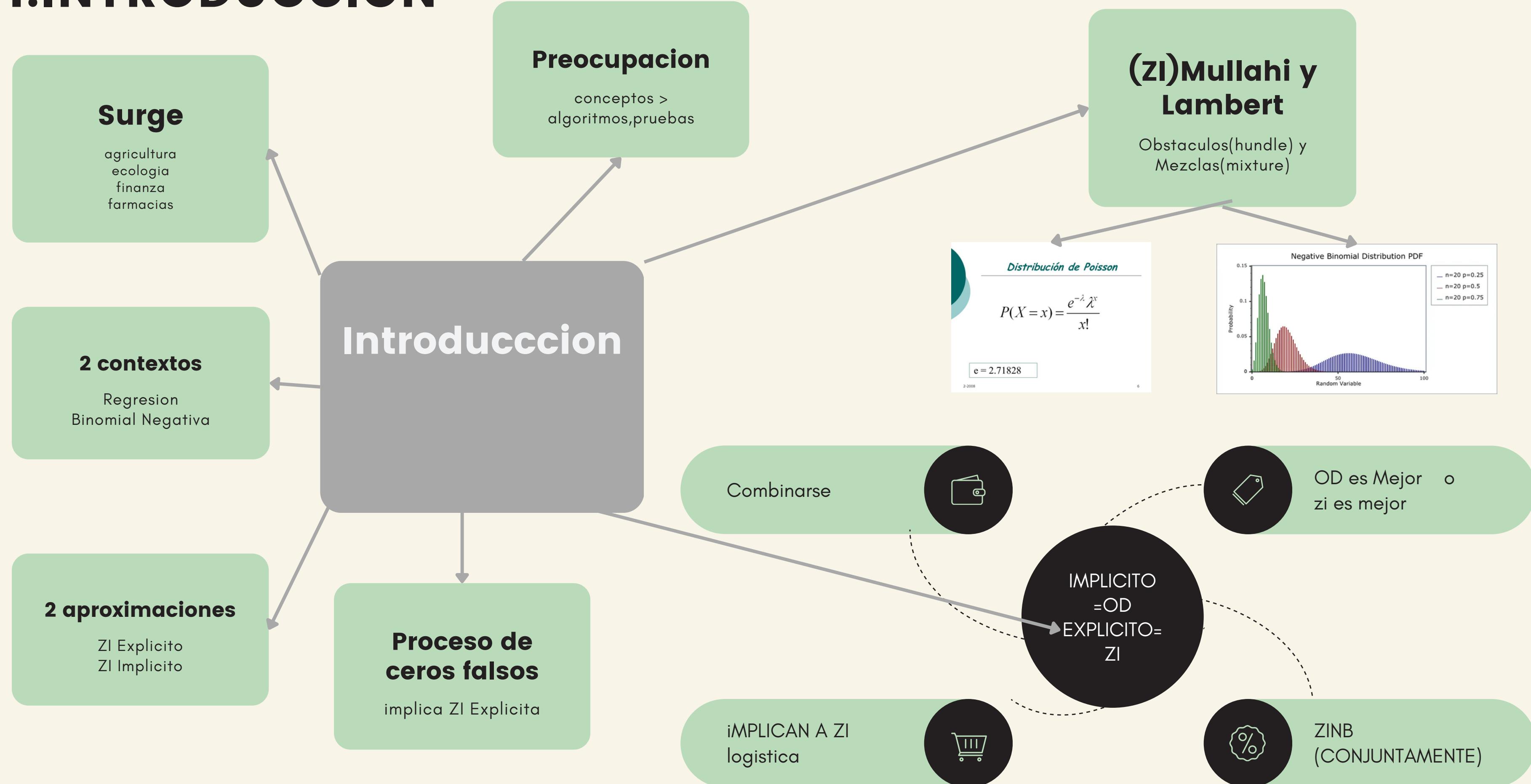


recuento

La tabla muestra el número de brotes que produjeron 0, 1, ..., 12 raíces, y los recuentos que superaron las 12 raíces se muestran individualmente en la última fila



1. INTRODUCCION



2.ejemplo

brotes micropagados
del cultivar de manzana
Trajan

La variable de respuesta es
el número de raíces

dos fotoperiodos
diferentes (8 y 16 h)

cuatro concentraciones
diferentes de la hormona
de crecimiento citoquinina
BAP 2.2 4.4 8.8 17.6

datos de la
manzana Trajan de
Ridout

diseño completamente
aleatorio con múltiples
réplicas en cada uno de
los ajustes

numero de ajustes = 8

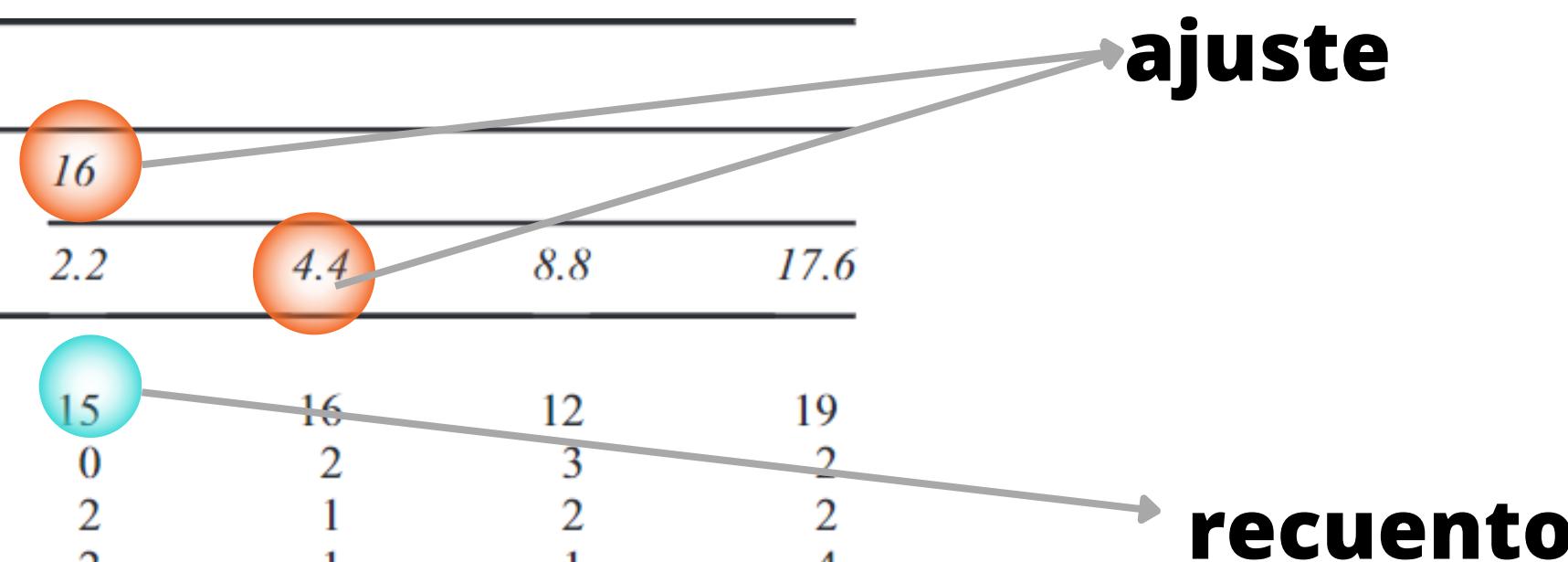
2 factores

numero de brotes = 270



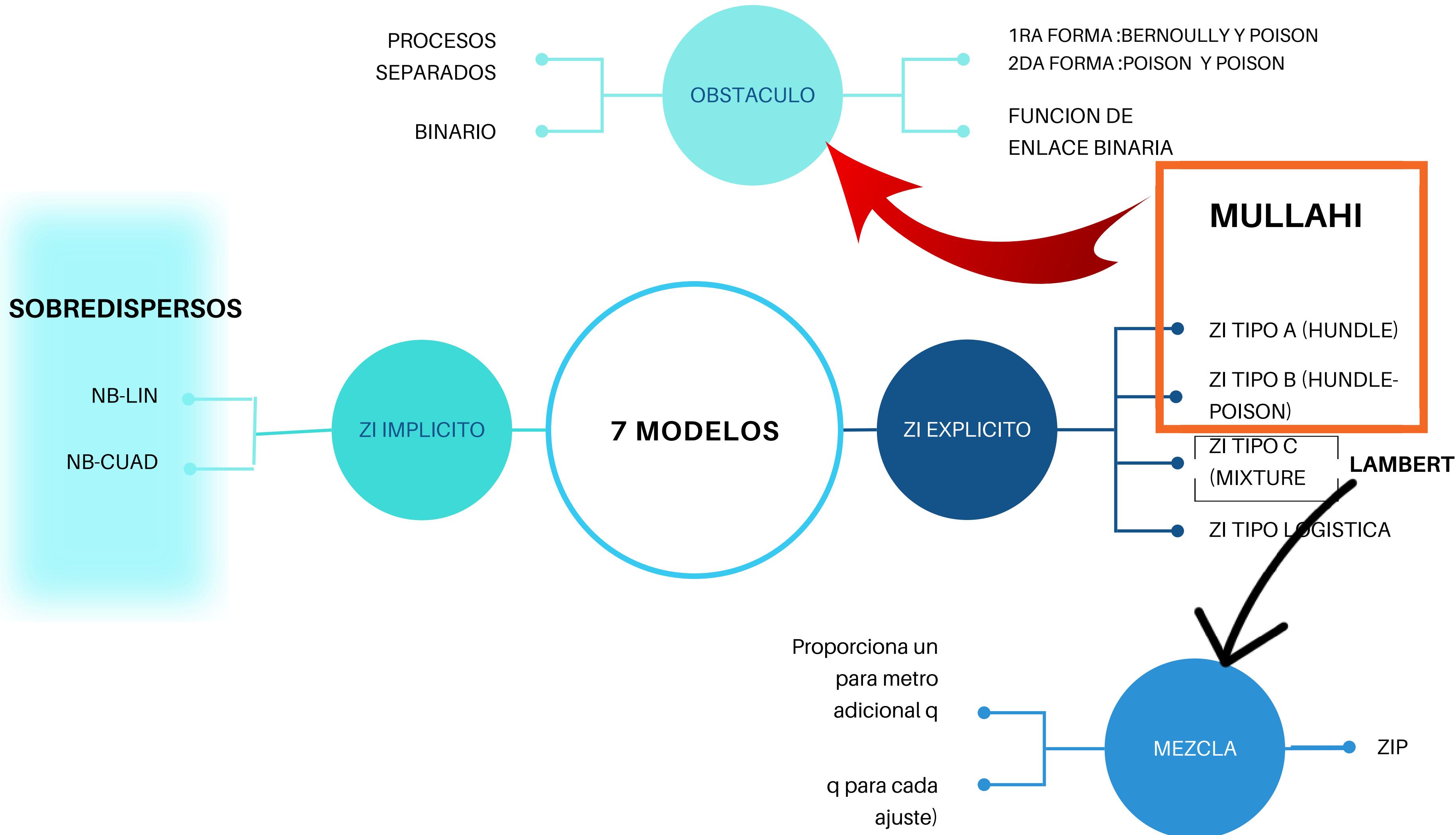
Distribuciones de frecuencia

		Photoperiod							
		8				2.2	4.4	8.8	17.6
BAP (μM)		2.2	4.4	8.8	17.6	2.2	4.4	8.8	17.6
No. of roots		0	0	0	2	15	16	12	19
0		0	0	0	2	0	2	3	2
1		3	0	0	0	2	1	2	2
2		2	3	1	0	2	1	2	2
3		3	0	2	2	2	1	1	4
4		6	1	4	2	1	2	2	3
5		3	0	4	5	2	1	2	1
6		2	3	4	5	1	2	3	4
7		2	7	4	4	0	0	1	3
8		3	3	7	8	1	1	0	0
9		1	5	5	3	3	0	2	2
10		2	3	4	4	1	3	0	0
11		1	4	1	4	1	0	1	0
12		0	0	2	0	1	1	1	0
>12		13,17	13	14,14	14				



La tabla muestra el número de brotes que produjeron 0, 1, ..., 12 raíces, y los recuentos que superaron las 12 raíces se muestran individualmente en la última fila





Distribuciones

$$\mu^+ = E_{\tilde{\pi}}[Y|Y>0] = E_{\pi}[Y|Y>0] = \lambda^+$$

$$\tilde{\pi}_0 = \gamma.$$

Función de vínculo	Fórmula	Uso
Identidad	μ	Datos continuos con errores normales (regresión y ANOVA)
Logarítmica	$\text{Log}(\mu)$	Conteos con errores de tipo Poisson
Logit	$\text{Log}(\frac{\mu}{n-\mu})$	Proporciones (datos entre 0 y 1) con errores binomiales
Recíproca	$\frac{1}{\mu}$	Datos continuos con errores gamma
Raíz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Funciones de potencia

$$\tilde{\pi}_0 = e^{-\alpha\lambda}$$

$$\tilde{\pi}_0 = (\pi_0)^{e^\gamma}$$

$$\gamma = \log(\alpha)$$

$$\log(\rho) = \gamma$$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$

$$\text{logit}(\tilde{\pi}_0) = \gamma + \text{logit}(\pi_0)$$

$$\tilde{\pi}_0 / (1 - \tilde{\pi}_0) = e^\gamma \pi_0 / (1 - \pi_0)$$

$$\tilde{\pi}_0 = e^\gamma \pi_0 / (1 + (e^\gamma - 1) \pi_0).$$

$$\pi_y(\mu, k) = \frac{\Gamma(k+y)}{y! \Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(\frac{\mu}{k+\mu}\right)^y$$

ZI TIPO A (HANDLE)

$$\text{logit}(\tilde{\pi}_0) = \gamma$$

- 0 En este caso, es ese valor de y el que hace $\tilde{\pi}_0 = p_0$, la proporción para el tipo A, el MLE $\hat{\lambda}^{(k)}$ se deriva de la media truncada $\bar{y}_+^{(k)}$

ZI TIPO B (HANDLE-POISON)

- 0 $\gamma = 0$ define el caso neutral, con y positivo y negativo
2 correspondientes a subinflación y sobreinflación

ZI TIPO C (MIXTURE) $\gamma \leq 0$

- 0 Cuando $\gamma = 0$, $\tilde{\pi}_0 = \pi_0$
3 $y > 0$, debemos escribir
 $\tilde{\pi}_0 = \max\{0, 1 - e^\gamma (1 - \pi_0)\}$.

ZI TIPO LOGISTICA

- 0 $\text{logit}(\tilde{\pi}_0) = \gamma + \text{logit}(\pi_0)$
4 $\hat{\mu}^{(k)} = \bar{y}^{(k)}$ aculos

0 POISON

- 0 Una variable aleatoria discreta X se dice que tiene una distribución de Poisson, con parámetro $\lambda > 0$, si tiene una función de masa de probabilidad dada por:^{9,60}

$$\tilde{\pi}_y(\lambda, \gamma) = \begin{cases} \tilde{\pi}_0 & \text{if } y = 0 \\ \frac{(1-\tilde{\pi}_0)}{(1-e^{-\lambda})} \pi_y(\lambda) & \text{if } y > 0 \end{cases}$$

$$E[Y|Y>0] = \lambda / (1 - e^{-\lambda})$$

$$E[Y] = (1 - \tilde{\pi}_0) / (1 - e^{-\lambda}) \lambda = \rho \lambda.$$

$$\rho = (1 - \tilde{\pi}_0) / (1 - e^{-\lambda}).$$

Una versión simplificada de esto, con un solo parámetro ZI adicional, toma los parámetros de Poisson como proporcionales con $\tilde{\pi}_0(\lambda, \alpha) = \pi_0(\alpha \lambda)$ con $\alpha < 1$ para inflación cero y donde $\alpha = 1$ se reduce a un modelo de Poisson común para conteos cero y distintos de cero. En esta versión de parámetros compartidos del modelo de obstáculos, ya

0 NB-LIN

- 0 $\log(\tilde{\pi}_0) = \phi^{-1} \log(1 + \phi) \log(\pi_0^P)$
6 $e^\gamma = \phi^{-1} \log(1 + \phi)$, el ZI inducido por NB-lin es exactamente como el tipo B para todos π_0^P y por lo tanto para todo μ

0 NB-CUAD

- 0 $\log(\tilde{\pi}_0) = -\phi^{-1} \log[1 - \phi \log(\pi_0^P)]$
7 para muy pequeños π_0^P (es decir, para π_0^P encima del 'codillo'), la función ZI $\pi_0(\pi_0^P, \phi)$ para NB-quad es aún más similar a la del tipo C.

Como otro modelo ZI explícito novedoso, presentamos el modelo 'ZI logístico', donde la alteración de la probabilidad cero se realiza a través de su razón de probabilidades. La distribución extendida resultante está en la familia exponencial, y para el modelo de interacción de datos de Trajan con los ocho $\lambda^{(k)}$ distintos y un solo parámetro ZI, las estadísticas suficientes son las medias de las celdas individuales, $\bar{y}^{(k)}$, y la proporción total de recuentos cero. En consecuencia, el modelo ajustado de máxima verosimilitud reproduce

Una variable aleatoria discreta X se dice que tiene una distribución de Poisson, con parámetro $\lambda > 0$, si tiene una función de masa de probabilidad dada por:^{9,60}

$$f(k; \lambda) = \Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

donde

- k es el número de ocasiones ($k = 0, 1, 2, \dots$)

A modo de comparación, también consideramos una parametrización diferente del modelo NB (que surge de una versión diferente de la mezcla de Poisson-gamma) que tiene una función de varianza lineal, $\mu(1 + \phi)$ y a la que nos referimos como NB-lin. Al igual que con todos los modelos de mezcla de Poisson, así como la sobredispersión, la probabilidad de cero se infla en comparación con el Poisson con ahora $\pi_0(\lambda, \phi) = 1 / \{(1 + \phi)^{1/\phi}\}^\lambda \geq e^{-\lambda}$, de nuevo con igualdad para $\phi = 0$. Este modelo no pertenece a la familia exponencial, e incluso cuando se ajusta el modelo de interacción de ocho parámetros completo para $\lambda^{(k)}$, no reproduce las medias de las celdas (Figura 1). Sin embargo, sí parece recuperar más estructura en las proporciones cero. El punto aquí es que diferentes modelos de recuento

aleatoria distribuida Gamma($1/\phi, 1/\phi$) con media 1 y varianza ϕ . La distribución NB resultante, a la que nos referimos como NB-quad, tiene una media $\mu = \lambda\phi$ y una función de varianza cuadrática $\mu + \phi\mu^2$ y para un valor fijo de ϕ está en la familia exponencial. La probabilidad de una observación cero es $\pi_0(\lambda, \phi) = (1 + \phi\lambda)^{-\phi^{-1}} \geq e^{-\lambda}$ con igualdad para $\phi = 0$, mostrando que, para $\phi > 0$, este modelo de hecho infla la probabilidad de una observación cero en comparación con la distribución de Poisson. Para el ajuste de datos de

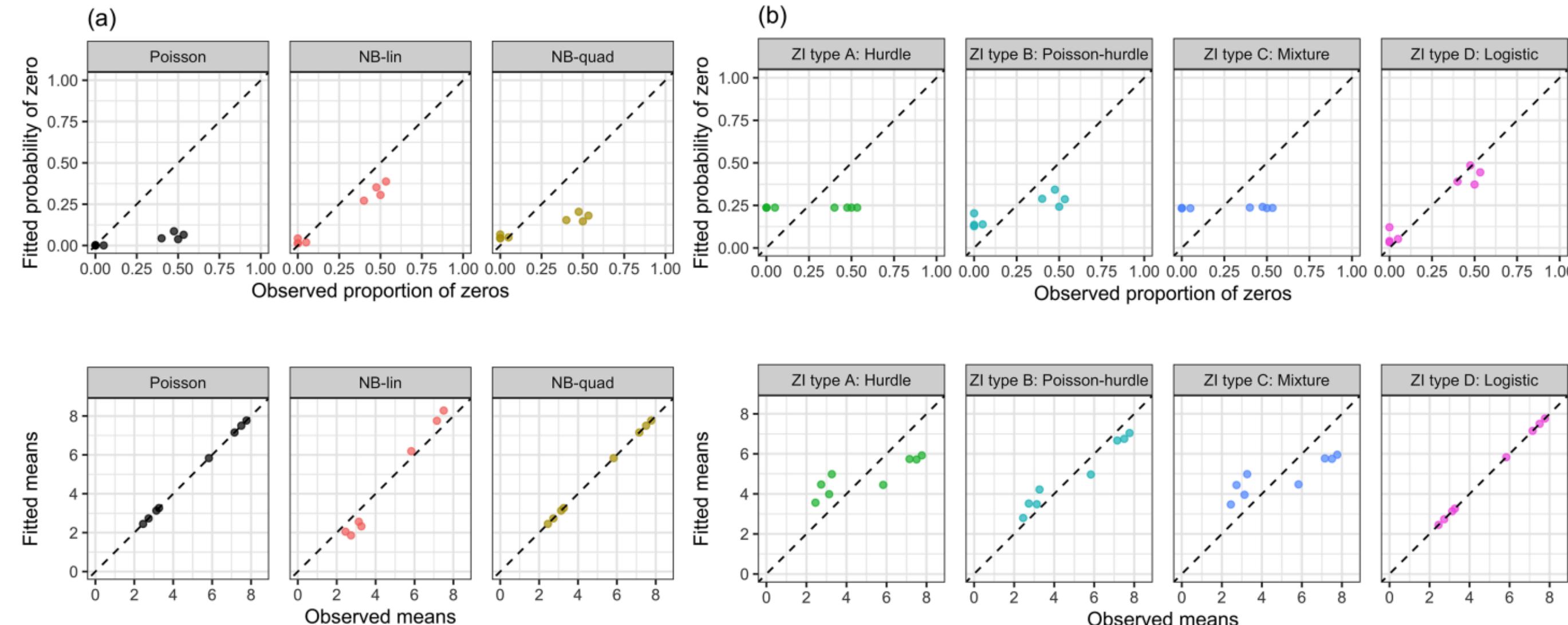
Proporción ajustada frente a la observada de ceros y medias

Se dice que un estimador es insesgado si la Media de la distribución del estimador es igual al parámetro

Kairo Floyd Pink

En la Figura 1(b), vemos que aquí los resultados son muy similares a los del modelo de obstáculos; esto se debe a que las medias de las celdas son generalmente grandes y, por tanto, las probabilidades cero del modelo de Poisson base, \exp^{-k} son pequeñas y contribuyen poco a las probabilidades cero ajustadas, que están dominadas por la constante estimada de inflación cero \hat{q} .

Kairo Floyd Pink



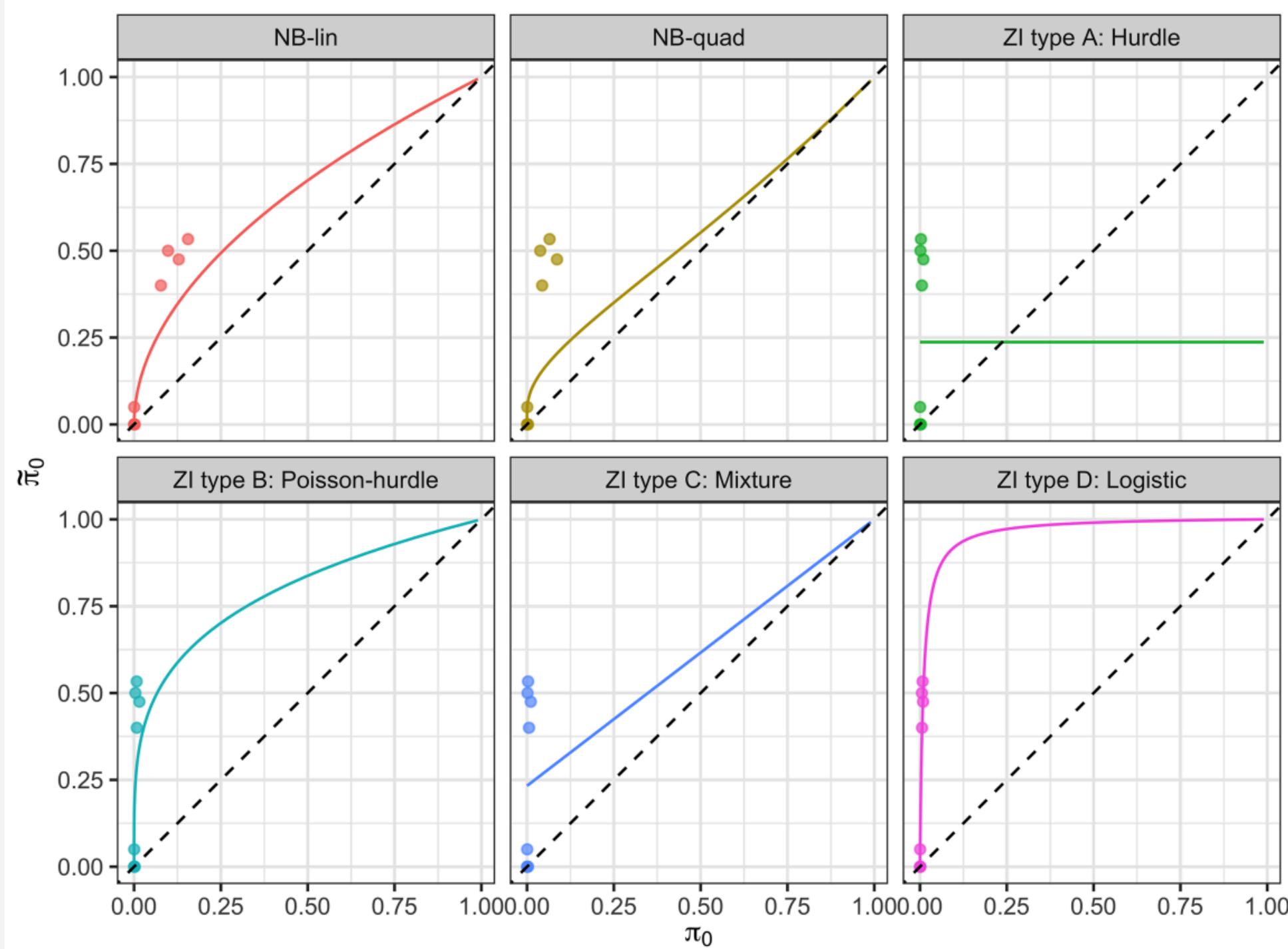
Como la distribución de Poisson truncada en cero pertenece a la familia exponencial, el modelo ajustado reproduce las medias de los datos truncados en cero; así, , $E[Y^{(k)}|Y^{(k)} > 0]$ se estima por la media truncada $y^{-{(k)}}$ para cada celda y, por lo tanto, los valores globales ajustados, $(1 - \pi_0^{\sim})y^{-{(k)}} = (1 - p_0)y^{-{(k)}}$ difieren de los valores de la muestra $y^{-{(k)}}$ como se ve en la figura 1(b).

Proporción ajustada frente a la observada de ceros y medias de cada combinación entre fotoperiodo y concentración hormonal en los datos de Trajano para (a) los modelos Poisson, NB-lin y NB-quad, y (b) los modelos ZI tipos A, B, C y D.

evidencia a priori de un exceso de ceros en comparación con un modelo de Poisson, especialmente para el fotoperíodo de 16 horas donde hay muchos ceros observados junto con algunos recuentos grandes de no ceros. Esto motiva la necesidad de explorar extensiones ZI del modelo básico de Poisson.

Kairo Floyd Pink

Probabilidad alterada de un cero $\tilde{\pi}_0$ frente a la probabilidad de un cero basada en la distribución de Poisson 0



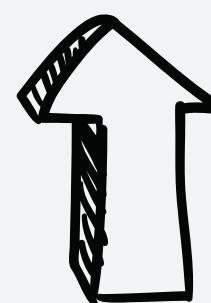
puntos corresponden a los pares $(\hat{\pi}_0^{(k)}, p_0^{(k)})$, $k \in \{1, \dots, 8\}$, es decir, las ocho combinaciones entre el

fotoperiodo y la concentración de hormonas. La [línea de identidad discontinua](#) es la base de Poisson suprimir el segundo parámetro. Aquí, nos referimos a una **distribución alterada** por el cero $\tilde{\pi}_y$ definida por una **función** $\tilde{\pi}_0(\pi_0, \gamma)$ con las consiguientes implicaciones para $\tilde{\pi}_y$, $y \neq 0$. Como desarrollamos a

3. Modelos Explicitos e Implicitos

Una tipología de la inflación cero y de la sobredispersión.

Type: Common names	Explicit ZI function
A: Basic hurdle, zero-altered, two-stage B: Poisson hurdle, zero-altered, two-stage C: ZIP, mixture, Lambert's mixture D: Logistic ZI (new—this paper)	$\text{logit}(\tilde{\pi}_0) = \gamma$ $\log(-\log(\tilde{\pi}_0)) = \gamma + \log(-\log(\pi_0))$ $\log(1 - \tilde{\pi}_0) = \gamma + \log(1 - \pi_0)$, $\gamma \leq 0$ $\text{logit}(\tilde{\pi}_0) = \gamma + \text{logit}(\pi_0)$
OD dist	<i>Implicit ZI function</i>
NB-lin Var(Y) = $\mu + \phi\mu$	$\log(\tilde{\pi}_0) = \phi^{-1} \log(1 + \phi) \log(\pi_0^P)$
NB-quad Var(Y) = $\mu + \phi\mu^2$	$\log(\tilde{\pi}_0) = -\phi^{-1} \log[1 - \phi \log(\pi_0^P)]$



$$g(\tilde{\pi}_0) = \gamma + g(\pi_0)$$

3.1.4 Inflado cero tipo D

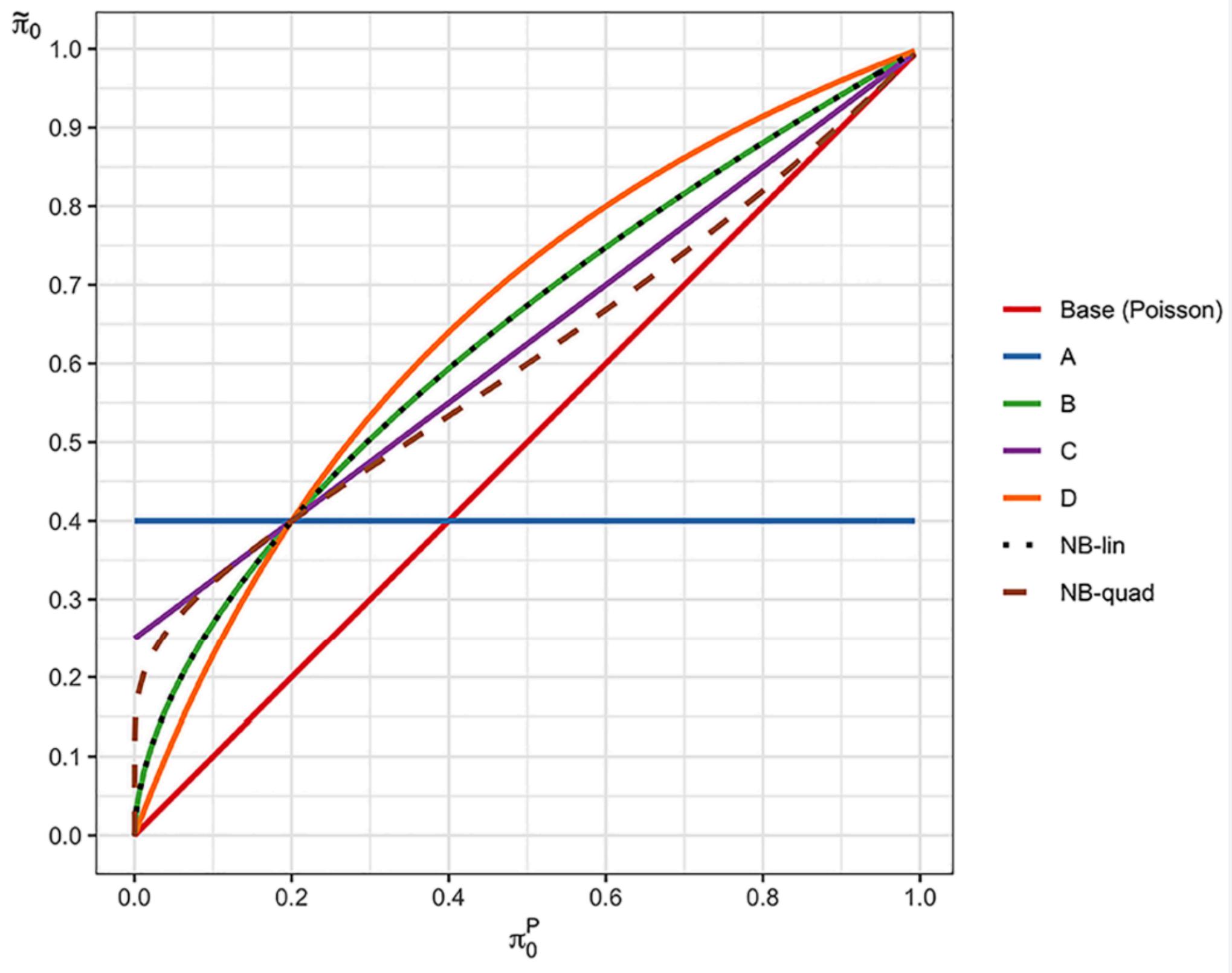
El tipo D, que es nuevo, se expresa más simplemente como

$$\text{logit}(\tilde{\pi}_0) = \gamma + \text{logit}(\pi_0)$$

para γ sin restricciones. De manera equivalente, tenemos $\tilde{\pi}_0 / (1 - \tilde{\pi}_0) = e^\gamma \pi_0 / (1 - \pi_0)$. La alteración ahora puede verse como una alteración multiplicativa de la razón de probabilidades de una cuenta cero. En forma cerrada, se puede escribir como $\tilde{\pi}_0 = e^\gamma \pi_0 / (1 + (e^\gamma - 1) \pi_0)$.

En la Figura 3, usamos esto para proporcionar un contraste diferente de las funciones. Allí se ve que los tipos B y C pueden ser notablemente similares para valores medianos a grandes de π_0 y, por lo tanto, podrían ser difíciles de distinguir en un diseño de datos que no incluyera valores pequeños de π_0 . Además, el tipo D se distingue del tipo B de dos maneras. En primer lugar, tiene una simetría (alrededor de la diagonal $\tilde{\pi}_0 = 1 - \pi_0$) que faltan los demás; pero esto puede ser de hecho desventajoso en algunos casos. Por ejemplo, puede tener una inclinación mayor que B para π_0 muy pequeño ; pero entonces, necesariamente, será mucho más plano para pequeños $1 - \pi_0$.

En segundo lugar, observamos que hay muchas más opciones en el diseño incluso de las funciones de un solo parámetro. $\tilde{\pi}_0 = \tilde{\pi}_0(\pi_0, \gamma)$ de lo que la literatura podría sugerir. Estos claramente se extienden más allá de los cuatro tipos que hemos usado para ilustración. Por lo general, involucran funciones de enlace clásicas $g(p)$, como log-log y logit



Relaciones teóricas cero-infladas (ZI). Trazamos poison(y=0) como Poisson frente a prob-ZI para los tipos ZI y sobredispersos . Los parámetros de cada distribución se ajustan de forma que cada línea (excepto la de Poisson) pase por el punto (0.2,0.4) para un rango de medias. Esto da lugar a diferentes ZI parámetros para cada distribución:

$$\gamma_A = -0.405, \gamma_B = -0.563, \gamma_C = -0.288 \text{ and } \gamma_D = 0.981,$$

, todos ellos proporcionados para que coincidan con la escala de la Tabla 2. Para los modelos binomiales negativos (NB) NB-lin=1.82 y NB-quad=1.13, de nuevo con referencia a la parametrización

Conclusión de modelos explícitos

Concluimos esta subsección haciendo dos observaciones. En primer lugar, el uso del cuadrado de la unidad para los gráficos de $\tilde{\pi}_0(\pi_0, \gamma)$ pone de relieve una característica peculiar del diseño de los datos de Trajano. No hay muestras correspondientes a medias pequeñas λ y, por tanto, grandes π_0 . Es esto lo que hace difícil distinguir cuál es el "mejor" modelo, a pesar de su otra característica peculiar -la replicación- que da acceso a datos condicionalmente iid y, por tanto, a las proporciones p_0 de ceros observados. Un ejemplo aún más extremo sería el de los datos totalmente iid. En este caso, veríamos que los cuatro tipos de ZI serían, de hecho, re-parametrizaciones de los demás y, por lo tanto, serían imposibles de distinguir. Las cuatro funciones se cruzarían en $\tilde{\pi}_0 = p_0$.

4. INFERENCIA

la estimación de λ

$$\log L = \sum_i \log (\tilde{\pi}_{y_i}) = \sum_i \ell_{y_i}(\mu(\lambda, \gamma), \gamma)$$

ecuaciones de puntuación siendo diferenciales de

con respecto a los parámetros

γ, λ_i o funciones de éstos como π^+ , u v/o los coeficientes β subyacentes a los λ términos
La distribución completa $\tilde{\pi}_y(\lambda, \gamma)$ puede escribirse en varias formas

$$\begin{aligned}\tilde{\pi}_y &= (\tilde{\pi}_0)^{\mathbb{I}\{y=0\}} (\rho \pi_y)^{1-\mathbb{I}\{y=0\}} = (\tilde{\pi}_0)^{\mathbb{I}\{y=0\}} \left(\frac{1-\tilde{\pi}_0}{1-\pi_0} \pi_y \right)^{1-\mathbb{I}\{y=0\}} = ((1-\tilde{\pi}_0)^{1-\mathbb{I}\{y=0\}} \tilde{\pi}_0^{\mathbb{I}\{y=0\}}) (\pi_y^+)^{1-\mathbb{I}\{y=0\}} \\ &= \left(\frac{\tilde{\pi}_0}{\rho \pi_0} \right)^{\mathbb{I}\{y=0\}} \rho \pi_y = e^{\gamma \mathbb{I}\{y=0\}} \left(\frac{\pi_y}{\pi_0} \right) e^{-\gamma \tilde{\pi}_0},\end{aligned}$$

e^γ denota la razón de probabilidades para $(\tilde{\pi}_0, \pi_0)$

$\pi_y^+ = \pi_y / (1 - \pi_0)$ es la forma truncada en cero de π_y

arroja luz sobre A

IMPLICAR

λ se enfoca únicamente en $\ell_y^+(\lambda)$

distribución truncada en cero $\pi_y^+(\lambda)$

El valor esperado DE

distribución truncada en cero es $(1 - \pi_0)^{-1} \lambda$

MLE es \bar{y}^+

para todos los pmfs en la familia exponencial, como Poisson y NB-quad

$$\hat{\lambda} = (1 - \hat{\pi}_0) \bar{y}^+$$

$$\begin{aligned}\text{iid} \quad \hat{\mu} &= \hat{\rho} \hat{\lambda} = \frac{1 - p_0}{1 - \hat{\pi}_0} (1 - \hat{\pi}_0) \bar{y}^+ = \bar{y}. \\ E_{\tilde{\pi}}[Y] &= \mu = \rho \lambda;\end{aligned}$$

observaciones iid
 $\hat{\pi}_0$ es tal que $\hat{\pi}_0(\hat{\lambda}) = p_0$

iid

caso del tipo A
 $\tilde{\pi}_0$ es independiente de $\hat{\pi}_0$

caso del tipo A

$\tilde{\pi}_0$ es independiente de $\hat{\pi}_0$

TRAJANO

$$\hat{\mu}^{(k)} = \hat{\rho}^{(k)} \lambda^{(k)} = (1 - p_0) \bar{y}^{+(k)} \neq (1 - p_0^{(k)}) \bar{y}^{+(k)} = \bar{y}^{(k)}$$

para Poisson, NB-quad y ZI tipo D con base de Poisson ambos conducen a los mismos estimadores de los valores esperados de las celdas. Por lo tanto, ZI tipo D puede ser una adición útil a los tipos ZI.

Además, el log-verosimilitud en la Ecuación (4) incluye una descomposición en la suma de dos términos, siendo $\ell_y^0(\gamma)$ y $\ell_y^+(\lambda)$. Y de hecho, si $\pi(y)$ está en la familia de parámetros m entonces $\tilde{\pi}_y$ está en la familia de parámetros $(m + 1)$. De ello se deduce inmediatamente que estadísticos suficientes para una muestra iid y_1, \dots, y_n de $\tilde{\pi}_y$ son $(\sum y_i, \sum \mathbb{I}\{y_i = 0\})$ siendo equivalente a (\bar{y}, p_0) y teniendo valores esperados ya conocidos aquí para ser $(\mu, \tilde{\pi}_0)$, recordando que $\mu = \rho \lambda$. Este resultado también está disponible en general al derivar $A(\lambda, \gamma)$. Y, siguiendo la observación anterior sobre la equivalencia, en el caso iid, de todos los tipos ZI de dos parámetros, \bar{y} y p_0 son estimadores MLE insesgados de μ y $\tilde{\pi}_0$, como ya se mostró arriba, a través de los argumentos de tipo A más engorrosos. Esta es la razón por la que ZI tipo D (con base Poisson) comparte esta propiedad con los modelos habituales de Poisson y NB-quad. Y además, podemos afirmar de inmediato que ZI tipo D con NB-quad como base también compartirá esta propiedad.

En regresión general, los λ i están dictados por coeficientes, siendo estos y el verdadero objetivo de la inferencia.

El primer término depende de $\mathbb{I}\{y_i = 0\}$ y proporciona la única información sobre y a través de $\tilde{\pi}_0$

el primer término generalmente proporciona alguna información sobre λ

El cálculo habitual conduce a funciones de puntuación cuya solución es el MLE para (λ, γ) (y para cualquier otro parámetro, ϕ)

$\tilde{\pi}_0$ es también una función de λ a través de π_0

el primer término generalmente proporciona alguna información sobre λ

$$\eta = \log(\lambda)$$

$$\log(\pi_y) = y\eta(\lambda) - A(\lambda) + c(y)$$

$$c(y) \quad c(0) = 0,$$

Conclusiones

En este documento, solo hemos examinado los fundamentos básicos de la regresión en presencia de ceros en exceso en los datos de conteo univariados. Nuestra primera contribución a esto es una perspectiva novedosa sobre el modelado de exceso de ceros en la regresión de datos de conteo. Esto es principalmente en la presentación de un enfoque al que nos hemos referido como ZI explícito. El problema del modelado, tal como se presenta aquí, es simplemente la elección de la función de enlace y la consiguiente estimación del único parámetro. Se presentan cuatro de estas opciones simples; son posibles otras y extensiones a variaciones de dos parámetros. Los dos enfoques analizados (ZI explícito y ZI implícito mediante el uso de distribuciones como la NB) pueden, por supuesto, combinarse y todos los parámetros pueden modelarse mediante covariables.

Se podría argumentar que la verdadera dificultad es la vergüenza de elegir. La estimación de los parámetros ya no es un desafío (para recuentos univariados), dados los algoritmos informáticos modernos; ni es la identificación de los mejores, dada una métrica como el AIC, una métrica que puede no ser natural para algunos usuarios. El verdadero desafío es, de hecho, la identificación del modelo más útil dados los caprichos de ese término y la ubicuidad de posibles valores atípicos dentro de todos los conjuntos de datos.

segunda contribución es hacer más explícito el paralelismo entre los enfoques OD y ZI para modelar el exceso de ceros. Pero no hemos resuelto la elección entre los enfoques explícito (ZI) e implícito (OD), cuando la base es la Poisson. Por el contrario, vemos que, desde el punto de vista del exceso de ceros, un modelo ZI -tipo B- se comporta exactamente como el NB-lin; y el NB-quad no se diferencia del ZI (dominante) tipo C. La diferencia entre estos dos últimos sólo es evidente para μ muy grandes, y sólo los datos cuidadosamente diseñados los diferenciarán. Sí difieren en lo que respecta a la relación media-varianza, pero sólo en detalle, ya que ambos presentan una relación cuadrática; y esto también sólo será aparente para μ muy grandes. Además, para los datos iid, los detalles técnicos de los procedimientos de inferencia habituales para NB-cuad no ayudan a diferenciar los modelos tan distintos. Además, para el caso más típico de la regresión, estos detalles sólo pueden volverse más difíciles. Lo que está claro es que cualquier intento de discriminar entre los modelos ZI explícitos e implícitos requerirá un conjunto de datos rico y variado y diagnósticos que se centren conjuntamente en la probabilidad cero ajustada y en el comportamiento de la cola superior. Esta discusión sugiere que la floreciente literatura sobre otras generalizaciones de la DO de Poisson -y sobre la inflación del cero- puede ser en sí misma prematura. Como contribución final, la Tabla 2 ofrece una alternativa a la confusa nomenclatura que se ha desarrollado.

5. PAQUETES R

Una lista de los paquetes de R que usamos para demostrar el modelado de regresión inflado con ceros

Nombre	Acercarse	Probabilidades admitidas	Tipos de covariables admitidos
zic	bayesiano	Poisson (y log-normal de Poisson)	Restringido para ser el mismo en componentes regulares y sin inflar. También incluye selección de variables.
pscl	frecuentista	Poisson, binomial negativa y geométrica (distribución de conteo), y binomial, Poisson, binomial negativa y geométrica (distribuciones infladas a cero; solo obstáculo)	Restringido para ser el mismo en componentes regulares y sin inflar
VGAM	frecuentista	Poisson, binomial negativo y geométrico (cero-inflado y obstáculo)	Los componentes de inflación cero no varían según la covariable, pero las fórmulas permiten spline y otras relaciones de tipo GAM
juego	frecuentista	Al menos 12 tipos diferentes (inflados a cero y ajustados)	Tiene la capacidad de modelar cada parámetro en una distribución por separado utilizando, por ejemplo, el marco GAM