**Revista Colombiana de Estadística**
Junio 2009, volumen 32, no. 1, pp. 59 a 78

# Estudio de homogeneidad de la dispersión en diseño a una vía de clasificación para datos de proporciones y conteos

## Study of Homogeneity of the Dispersion in one way Classification Models with Proportions and Counts Data

Mario Alfonso Morales[1,a], Luis Alberto López[2,b]

[1]Departamento de Matemáticas y Estadística, Universidad de Córdoba, Montería, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

**Resumen**

En el modelamiento de datos donde se evidencia la presencia de sobredispersión, usualmente se asume que para todos los efectos de tratamiento el parámetro de sobredispersión es común. Esta situación no siempre se satisface antes; por el contrario, puede ser más frecuente que la variabilidad exhibida por los datos sea mayor que la variación teórica del modelo; debiendo incorporarse en el modelamiento esta situación. En este artículo se llevan a cabo desarrollos teóricos con los cuales se evidencia si es aceptable o no la hipótesis de homogeneidad del parámetro de dispersión entre tratamientos, cuando estos se ensayan en condiciones de uniformidad del material experimental y la respuesta de interés sea conteos o de proporciones, las cuales se modelan a través de la distribución Binomial Negativa (BN) y Beta Binomial (BB), respectivamente. Se usó la Prueba de Razón de Verosimilitud como criterio para decidir acerca de la hipótesis nula de homogeneidad en el parámetro de dispersión. Para determinar la eficiencia de la prueba propuesta, mediante simulación, con procedimientos algorítmicos desarrollados en R, se evaluó la potencia de las pruebas frente al supuesto de homogeneidad del parámetro de dispersión. Bajo el supuesto que los modelos BB y BN son correctos, se propone el ajuste de un modelo lineal generalizado ponderado como una alternativa para el análisis de datos de conteo y proporción con sobredispersión.

***Palabras clave*:** sobredispersión, distribución beta-binomial, distribución binomial negativa, modelo lineal generalizado ponderado, razón de verosimilitud.

[a]Profesor asistente. E-mail: mmorales@sinu.unicordoba.edu.co
[b]Profesor asociado. E-mail: lalopezp@unal.edu.co

**Abstract**

When analyzing data in the presence of over dispersion, the usual practice is to assume a common dispersion parameter to all observations. However, there are situations where the assumption of homogeneity of the dispersion parameter does not hold. In this paper we present theoretical developments that allow contrasting the assumption of homogeneity of the dispersion parameter between treatments, in a completely randomized design, with the responses of proportions and counts, modeled through the distributions beta-binomial and negative binomial respectively. The hypothesis is contrasted through the proof of the likelihood ratio.

Under the assumption that the beta-binomial and the negative binomial models are correct, it is proposed an adjustment of a generalized linear weight ed model as an alternative for the data analysis of counts and proportions when over dispersion is present. It is also evaluated, through simulation, the performance of the proposed proofs in terms of its power.

***Key words*:** Overdispersion, Proportions, Beta-binomial distribution, Negative binomial distribution, Likelihood ratio, Generalized linear models.

# 1. Introducción

En el análisis de datos relacionados con proporciones y conteos es común que haya presencia de sobredispersión, situación que se presenta cuando la varianza exhibida por los datos es mucho más grande que la que predice el modelo. En el caso de datos resultantes de experimentos de una vía de clasificación, en la literatura comúnmente se aborda el problema asumiendo homogeneidad del parámetro de dispersión, como puede verse en Crowder (1978) y Brooks (1978). En la parametrización que hacen del modelo beta-binomial, asumen que $\sigma_i^2 = \sigma^2$ para $i = 1, \ldots, K$.

Si bien es cierto que existen aplicaciones donde el supuesto de homogeneidad de la dispersión se satisface, como ocurre con los datos de germinación de las semillas de *Orobanche Cernua* cultivadas en tres medios, analizados por Crowder (1978), existen situaciones donde, de acuerdo con la naturaleza del experimento, es posible que se justifique suponer heterogeneidad en el parámetro de dispersión. Por ejemplo, Crowder (1979), refiriéndose al análisis de un conjunto de datos que reportan la presencia o ausencia de tumores en ratones expuestos a un cancerígeno, comenta que el coeficiente de correlación entre respuestas binarias está asociado al efecto de la herencia sobre la respuesta, y demuestra que, para ese caso, es apropiado asumir un coeficiente de correlación distinto para cada nivel de exposición al cancerígeno, con lo cual se evidencia sobredispersión.

De acuerdo con Prentice (1986), en muchas aplicaciones es natural asumir que la probabilidad en una respuesta binaria tenga un valor común $p$ dentro de la unidad experimental y que pares de observaciones tengan una correlación común $\phi$; estos supuestos dan origen al modelo beta-binomial. La falta de independencia entre respuestas binarias explica la presencia de sobredispersión y es posible que, por su naturaleza, dentro de cada tratamiento se tenga una correlación distinta

entre pares de respuestas binarias, ocasionando heterogeneidad de la dispersión entre tratamientos.

Se tiene entonces la necesidad de contar con herramientas teóricas que permitan modelar datos en presencia de sobredispersión, diagnosticar si los tratamientos se pueden considerar homogéneos en cuanto a la dispersión o si, por el contrario, es necesario tener en cuenta que cada tratamiento "aporta" de una forma distinta a la sobredispersión, debiendo hacer correcciones en cada caso.

En este artículo se proponen métodos para juzgar acerca de la hipótesis de homogeneidad de la dispersión, es decir, se establece una metodología que permite contrastar el juego de hipótesis

$$\begin{aligned} &H_0\colon \phi_1 = \phi_2 = \cdots = \phi_K = \phi \\ &H_1\colon \phi_i \neq \phi_{i'} \quad \text{para} \quad i \neq i', \quad i, i' = 1, \ldots, K \end{aligned} \tag{1}$$

cuando se tienen datos provenientes de experimentos de una vía de clasificación, concretamente dentro de un diseño completamente al azar, cuando la variable respuesta es de tipo conteo o proporción y modeladas por medio de las distribuciones BN y BB, respectivamente.

El parámetro $\phi$ en (1) para el caso de datos de proporción, corresponde, de acuerdo con Prentice (1986), al coeficiente de correlación entre respuestas binarias. Para el caso de datos de conteo, el parámetro $\phi$ se define como el inverso del parámetro de forma de la distribución gamma.

## 2. Sobredispersión

En muchas situaciones prácticas, cuando se ajusta un modelo lineal generalizado a un conjunto de datos, se puede observar un desvío mucho más grande que el esperado si el modelo fuera "correcto". La explicación más común a este fenómeno es que se tiene una forma estructural incorrecta para el modelo, es decir, no se han incluido los predictores apropiados; no se ha hecho la transformación o combinación apropiada de efectos; o no se han incluido todos los factores de ajuste en el modelo. Otra explicación común para un desvío grande es la presencia de un número pequeño de *outliers*.

Habiendo excluido todas las posibilidades anteriores, es posible que la variación de los datos sea mucho más grande aún que la predicha por el modelo. A este fenómeno se le conoce como sobredispersión, de acuerdo con Hinde & Demétrio (1998).

Las principales consecuencias de no tener en cuenta la sobredispersión, según Hinde & Demétrio (1998), son: errores incorrectos (subestimados), y por tanto la significación de los parámetros de regresión puede juzgarse de manera incorrecta; los cambios en el desvío asociados a los términos del modelo serían grandes y pueden conducir a la selección de un modelo demasiado complejo, lo cual lleva a que la interpretación del modelo sea incorrecta y las predicciones demasiado imprecisas.

### 2.1. Modelos de sobredispersión

Para ajustar datos en presencia de sobredispersión, el modelamiento estadístico presenta varias alternativas que surgen de los supuestos que se asumen para explicar este fenómeno. Según Hinde & Demétrio (1998), los modelos se pueden clasificar en dos grandes grupos: (i) Modelos de media varianza: que asumen una forma más general de la función de varianza, incluidos parámetros adicionales. Estos posiblemente no correspondan a una distribución de probabilidad específica para la respuesta, pero se pueden ver como una extensión del modelo básico; cuando esto sucede, los parámetros pueden estimarse usando métodos de cuasiverosimilitud; (ii) Modelos en dos etapas: los cuales asumen que el parámetro asociado a la respuesta no es fijo, sino que tiene alguna distribución de probabilidad conocida. Estos modelos conducen a un modelo de probabilidad compuesto; en principio, todos los parámetros pueden estimarse usando máxima verosimilitud. Sin embargo, en general, la distribución resultante no toma una forma simple.

Dos ejemplos concretos de modelos de dos etapas son el beta-binomial y el binomial negativo, de gran utilidad en el modelamiento de datos de proporciones y conteos respectivamente.

## 3. Prueba de la hipótesis por medio de la razón de verosimilitud

Para contrastar la hipótesis (1), cuando se tienen datos resultantes de un experimento bajo un diseño sin ninguna restricción en la aleatorización, asumiendo que la variable respuesta sigue una distribución beta-binomial en el caso de proporciones o distribución binomial negativa en el caso de datos de conteos, se procedió a hacer uso de la prueba de la razón de verosimilitud. En el desarrollo de la prueba, es necesario definir dos modelos. Uno donde se asume que la hipótesis alterna $H_1$ es verdadera, el cual se llamó modelo alternativo; y otro que es un caso particular del primero, cuando $H_0$ es verdadera, conocido como modelo nulo. Para decidir entre los dos modelos, se siguen estos pasos: (i) se estiman los parámetros de ambos modelos por el método de máxima verosimilitud; (ii) se obtiene la razón de verosimilitud:

$$\lambda = \frac{L\left(\widehat{\theta}_0; y\right)}{L\left(\widehat{\theta}_1; y\right)}$$

donde $L\left(\widehat{\theta}_0; y\right)$ y $L\left(\widehat{\theta}_1; y\right)$ son las verosimilitudes maximizadas bajo los modelos nulo y alternativo, respectivamente; y (iii) se decide entre los dos modelos según $P\left(\chi^2_{k-1} > -2\ln\lambda\right)$ sean mayor o menor que el nivel de significación $\alpha$ fijado previamente.

La estimación por máxima verosimilitud implica la solución del sistema de ecuaciones $\boldsymbol{u}\left(\widehat{\theta}\right) = 0$, donde $\boldsymbol{u}$ representa el vector de primeras derivadas de la log-verosimilitud. La solución del sistema no lineal de ecuaciones se hace mediante

el algoritmo de *Newton-Raphson*, que consiste en obtener

$$\widehat{\boldsymbol{\theta}}^{m+1} = \widehat{\boldsymbol{\theta}}^{m} - \left[\boldsymbol{H}^{m}\right]^{-1}\left(\widehat{\boldsymbol{\theta}}^{m}\right)u\left(\widehat{\boldsymbol{\theta}}^{m}\right) \tag{2}$$

donde $\widehat{\boldsymbol{\theta}}^{m}$ es el vector de estimaciones en el paso $m$ y $\boldsymbol{H}$ es la matriz de segundas derivadas de la log-verosimilitud estimada en el paso $m$. En las secciones que siguen se establecen los supuestos y la notación, se definen los modelos nulos y alternativos y se obtienen las funciones de log-verosimilitud. En las secciones 3.1 y 3.2 se ilustra el procedimiento para el caso particular de este trabajo.

## 3.1. Datos de proporción

Cuando se tiene el caso de datos de proporción, para efectos de este trabajo se deben tener en consideración los siguientes supuestos: se tienen datos provenientes de un experimento con diseño completamente al azar con $K \geq 2$ tratamientos; se cuenta con $r_i$ repeticiones dentro del tratamiento $i$; la $j$-ésima observación del tratamiento $i$ es $y_{ij}$ con $i = 1, \ldots, K$ y $j = 1, \ldots, r_i$; $y_{ij}$ es el número de éxitos en $m_{ij}$ ensayos tipo Bernoulli, por tanto $y_{ij}$ puede tomar los valores $0, 1, \ldots, m_{ij}$; $y_{ij}$ es la realización de una variable aleatoria $Y_{ij}$ tal que $Y_{ij} \mid p_i \sim \text{Bin}(m_{ij}, p_i)$ con $p_i \sim \text{Beta}(a_i, b_i)$, lo cual implica que la distribución no condicional de $Y_{ij}$ es beta-binomial (véase Hinde & Demétrio 1998).

El modelo bajo el supuesto que la hipótesis alternativa es cierta debe considerar lo siguiente: la variable respuesta $Y_{ij}$ se asume con distribución beta-binomial de parámetros $m_{ij}$, $a_i$ y $b_i$ y, por tanto, de acuerdo con Prentice (1986), se tiene que la probabilidad de éxito en el tratamiento $i$ es $\pi_i = a_i/(a_i + b_i)$ y el coeficiente de correlación entre respuestas binarias en dicho tratamiento es $\phi_i = \gamma_i(1+\gamma_i)^{-1}$ con $\gamma_i = (a_i + b_i)^{-1}$ para $i = 1, \ldots, K$. También se tiene que

$$\begin{aligned} \mathrm{E}\big(Y_{ij}/m_{ij}\big) &= \pi_i \\ \mathrm{var}\big(Y_{ij}/m_{ij}\big) &= \big[1 + \phi_i(m_{ij} - 1)\big]\pi_i(1 - \pi_i)/m_{ij} \end{aligned}$$

de esta forma se está suponiendo un valor $\phi_i$ para cada tratamiento. Modelamos las probabilidades de éxito $\pi_i$ mediante $g(\pi_i) = \eta_i$ con $\eta_i = \alpha_i$, con este enlace queda caracterizado las medias de celda en la parte sistemática. En resumen, el modelo alternativo es:

$$\begin{aligned} \mathrm{E}\big(Y_{ij}/m_{ij}\big) &= \pi_i \\ \mathrm{var}\big(Y_{ij}/m_{ij}\big) &= \big[1 + \phi_i(m_{ij} - 1)\big]\pi_i(1 - \pi_i)/m_{ij} \mathrm{y} g(\pi_i) = \alpha_i \end{aligned} \tag{3}$$

y la distribución de probabilidad de $Y_{ij}$, siguiendo a Prentice (1986) es:

$$P(Y_{ij} = y_{ij}) = \binom{m_{ij}}{y_{ij}} \prod_{s=0}^{y_{ij}-1} (\pi_i + \gamma_i s) \prod_{s=0}^{m_{ij}-y_{ij}} (1 - \pi_i + \gamma_i s) \Big/ \prod_{s=0}^{m_{ij}-1} (1 + \gamma_i s) \tag{4}$$

para $y_{ij} = 0, 1, \ldots, m_{ij}$ y como es usual, se adopta la restricción que $\prod_{i=0}^{x} c_i = 1$ para cualquier $x < 0$. Este modelo tiene $2K$ parámetros, que corresponden a $K$

proporciones de éxitos y $K$ coeficientes de correlación entre respuestas binarias, uno dentro de cada tratamiento.

El modelo bajo el supuesto que la hipótesis nula es cierta resulta de la práctica usual al tomar $(a_i + b_i)^{-1} = \gamma$ para $i = 1, 2, \ldots, K$. Bajo ese supuesto se tiene que el coeficiente de correlación entre respuestas binarias es $\phi = \gamma(1+\gamma)^{-1}$ para todo $i = 1, \ldots, K$. Es decir, va a ser constante para todos los tratamientos. Los demás supuestos permanecen como en el modelo alternativo. Con esto se tiene que el modelo queda caracterizado por:

$$\begin{aligned} \mathrm{E}\big(Y_{ij}/m_{ij}\big) &= \pi_i \\ \mathrm{var}\big(Y_{ij}/m_{ij}\big) &= \big[1 + \phi(m_{ij} - 1)\big]\pi_i(1-\pi_i)/m_{ij} \text{ y } g(\pi_i) = \alpha_i \end{aligned} \tag{5}$$

y la distribución de probabilidad de la respuesta $Y_{ij}$ tiene la forma de la ecuación (4), remplazando $\gamma_i$ por $\gamma$. Este modelo tiene $K+1$ parámetros, que corresponden a $K$ proporciones de éxito y el coeficiente de correlación entre respuestas binarias dentro de cada tratamiento, que se supone el mismo para todos los tratamientos. Se observa que $\phi_1 = \phi_2 = \cdots = \phi_k = \phi$, es decir, se asume que la hipótesis $H_0$ de (1) es cierta.

De la ecuación (4), tomando logaritmo natural, se obtiene la contribución de la observación $y_{ij}$ a la log-verosimilitud, dada por

$$\begin{aligned} l(\pi_i, \gamma_i \mid y_{ij}, m_{ij}) = \sum_{s=0}^{y_{ij}-1} \log(\pi_i + \gamma_i s) + \\ \sum_{s=0}^{m_{ij}-y_{ij}-1} \log(1 - \pi_i + \gamma_i s) - \sum_{s=0}^{m_{ij}-1} \log(1 + \gamma_i s) \end{aligned} \tag{6}$$

para $i = 1, \ldots, K$, $j = 1, \ldots, r_i$, las sumas con índice superior negativo se toman iguales a cero, lo cual puede ocurrir cuando $y_{ij} = 0$ o cuando $y_{ij} = m_{ij}$. Para más detalle, véase Crowder (1978). La log-verosimilitud de toda la muestra es $l(\boldsymbol{\theta}_1 \mid \boldsymbol{y}) = \sum_{i=1}^{K} \sum_{j=1}^{r_i} l\big(\pi_i, \gamma_i \mid y_{ij}; m_{ij}\big)$ con $\boldsymbol{\theta}_1 = [\pi_1, \ldots, \pi_K, \gamma_1, \ldots, \gamma_K]^t$ el vector de parámetros. La contribución de la observación $y_{ij}$ a la log-verosimilitud, en el caso del modelo nulo, tiene la misma forma de la ecuación (6); basta con reemplazar $\gamma_i$ por $\gamma$. En ese caso, el vector de parámetros es $\boldsymbol{\theta} = [\pi_1, \ldots, \pi_K, \gamma]^t$. En Morales (2008) se encuentran los detalles de la obtención del vector $\boldsymbol{u}$ y la matriz $\boldsymbol{H}$ necesarios para implementar el algoritmo de Newton-Raphson, de acuerdo con la ecuación (2).

**Ejemplo 1.** Se sometieron 58 ratas hembras a dietas deficientes en hierro, divididas en cuatro grupos. A un grupo de control se le proporcionó semanalmente inyecciones de suplemento de hierro para mantener su ingestión de hierro en niveles normales, mientras que a otro grupo se le proporcionó solo inyecciones de un placebo. A los otros dos grupos se les proporcionó menos inyecciones de suplemento de hierro que a los controles. Las ratas fueron preñadas, sacrificadas y tres semanas después se registró el número total de fetos y el número de fetos muertos en cada camada. Los datos, que se muestran en la tabla 1, fueron analizados por

Agresti (2002) usando cuasiverosimilitud. De acuerdo con la notación adoptada en este trabajo se tiene: $K = 4$, $r_1 = 31$, $r_2 = 12$, $r_3 = 5$, $r_4 = 10$, $y_{11} = 1$, $m_{11} = 10$, $y_{43} = 2$ y $m_{43} = 9$.

Tabla 1: Número de fetos ($N$) y números de fetos muertos ($R$) de ratas sometidas a dietas bajas en hierro. Fuente Agresti (2002).

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grupo no tratado (bajo en hierro) | | | | | | | | | | | | |
| N | 10 | 11 | 12 | 4 | 10 | 11 | 9 | 11 | 10 | 10 | 12 | 10 |
| R | 1 | 4 | 9 | 4 | 10 | 9 | 9 | 11 | 10 | 7 | 12 | 9 |
| N | 8 | 11 | 6 | 9 | 14 | 12 | 11 | 13 | 14 | 10 | 12 | 13 |
| R | 8 | 9 | 4 | 7 | 14 | 7 | 9 | 8 | 5 | 10 | 10 | 8 |
| N | 10 | 14 | 13 | 4 | 8 | 13 | 12 | | | | | |
| R | 10 | 3 | 13 | 3 | 8 | 5 | 12 | | | | | |
| Grupo 2: inyecciones de hierro solo en el día 7 ó 10 | | | | | | | | | | | | |
| N | 10 | 3 | 13 | 12 | 14 | 9 | 13 | 16 | 11 | 4 | 1 | 12 |
| R | 1 | 1 | 1 | 0 | 4 | 2 | 2 | 1 | 0 | 0 | 0 | 0 |
| Grupo 3: inyecciones de hierro en el día 0 y 7 | | | | | | | | | | | | |
| N | 8 | 11 | 14 | 14 | 11 | | | | | | | |
| R | 0 | 1 | 0 | 1 | 0 | | | | | | | |
| Grupo 4: inyecciones de hierro semanalmente | | | | | | | | | | | | |
| N | 3 | 13 | 9 | 17 | 15 | 2 | 14 | 8 | 6 | 17 | | |
| R | 0 | 0 | 2 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | | |

Las varianzas observadas dentro de cada tratamiento son, respectivamente, 9.867, 1.454, 0.3 y 0.722, mientras que las varianzas esperadas, asumiendo la distribución binomial para la respuesta, son: 1.932, 0.898, 0.386 y 0.476. Nótese que las varianzas observadas de los tratamientos 1, 2 y 4 son mucho más grandes que las esperadas, presentándose sobredispersión (véase Sudhir & Saha 2007). Al ajustar un modelo lineal generalizado usual, usando la distribución binomial y la función logit$(\cdot)$ como enlace, se obtiene un desvío residual de 173.45 con 54 grados de libertad, lo cual confirma que los datos presentan sobredispersión. Dado que la variable respuesta es el número de fetos muertos del total de fetos en cada camada, no es necesario verificar el supuesto de independencia entre respuestas binarias (muerto o vivo) en cada camada; luego la distribución beta-binomial es un buen modelo para el ajuste de estos datos. Los valores iniciales para implementar el algoritmo de estimación se obtienen a partir del ajuste del modelo lineal generalizado binomial usual, del cual se obtiene $\widehat{\alpha}_1 = 1.1440$, $\widehat{\text{ee}}(\widehat{\alpha}_1) = 0.1292$, $\widehat{\alpha}_2 = -2.1785$, $\widehat{\text{ee}}(\widehat{\alpha}_2) = 0.3046$, $\widehat{\alpha}_3 = -3.3322$, $\widehat{\text{ee}}(\widehat{\alpha}_3) = 0.7196$ y $\widehat{\alpha}_4 = -2.9857$, $\widehat{\text{ee}}(\widehat{\alpha}_4) = 0.4584$; aplicando la función inversa del enlace a $\widehat{\alpha}_i$, se obtienen las probabilidades de éxito $\widehat{\pi}_1 = 0.7584$, $\widehat{\pi}_2 = 0.1017$, $\widehat{\pi}_3 = 0.0345$ y $\widehat{\pi}_4 = 0.0481$. El valor del estadístico de Pearson es $X^2 = 154.707$, $n = 58$, $p = 4$; en este caso, $\phi_0 = 0.2$, por lo que se toman los valores iniciales $\gamma_{01} = \gamma_{02} = \gamma_{03} = \gamma_{04} = 0.2(0.2 + 1)^{-1} = 0.167$. Después del ajuste del modelo alternativo (3) se obtiene $\widehat{\pi}_1 = 0.777$, $\widehat{\pi}_2 = 0.103$, $\widehat{\pi}_3 = 0.032$, $\widehat{\pi}_4 = 0.062$, $\widehat{\phi}_1 = 0.3397$, $\widehat{\phi}_2 = 0.0398$, $\widehat{\phi}_3 = 2.23 \times 10^{-9}$ y $\widehat{\phi}_4 = 0.0658$ y, a partir del modelo nulo (5), se obtiene $\widehat{\pi}_1 = 0.793$, $\widehat{\pi}_2 = 0.146$, $\widehat{\pi}_3 = 0.074$, $\widehat{\pi}_4 = 0.071$ y $\widehat{\phi} = 0.2412$. La log–verosimilitud para el modelo nulo, con 5 parámetros, es $-93.46$, mientras que para el modelo alternativo, con 8 parámetros,

es $-88.40$; por tanto, $\lambda^* = -2 \times (-93.46 - (-88.40)) = 10.11$, el valor $p$ de la prueba es $P\left(\chi^2_3 > 10.11\right) = 0.01766$, así que a un nivel de significación de 2 % se rechaza la hipótesis nula, luego los datos deben ajustarse con un parámetro de sobredispersión distinto para cada tratamiento, es decir, $\phi$ no es igual mismo para todos los tratamientos.

## 3.2. Datos de conteo

En el caso de datos de conteo, para respuestas provenientes de ensayos sin restricción en la aleatorización de un experimento con $K \geq 2$ tratamientos, con $r_i$ repeticiones dentro del tratamiento $i$, la $j$-ésima observación del tratamiento $i$ es $y_{ij}$, con $i = 1, \ldots, K$ y $j = 1, \ldots, r_i$; $y_{ij}$ es el número de ocurrencias de un evento en una unidad de tiempo o espacio, por tanto $y_{ij}$ puede tomar los valores $0, 1, 2, \ldots$; $y_{ij}$ es la realización de una variable aleatoria $Y_{ij}$ tal que $Y_{ij} \mid \theta_i \sim \text{Pois}(\theta_i)$, donde se supone que $\theta_i$ es aleatorio con distribución gamma, de tal forma que la distribución no condicional de $Y_{ij}$ es binomial negativa (véase Hinde & Demétrio 1998).

Para definir el modelo alternativo se asume que $\theta_i \sim \Gamma(\kappa_i, \lambda_i)$; por tanto la variable respuesta $Y_{ij}$ $i = 1, \ldots, K$, $j = 1, \ldots, r_i$ tiene distribución binomial negativa con media y varianza dadas por

$$\begin{aligned} \mathrm{E}\big(Y_{ij}\big) &= \frac{\kappa_i}{\lambda_i} = \mu_i \\ \mathrm{var}\big(Y_{ij}\big) &= \mu_i + \frac{\mu_i^2}{\kappa_i} = \mu_i + \phi_i \mu_i^2 = (1 + \phi_i \mu_i)\mu_i \end{aligned}$$

donde $\phi_i = 1/\kappa_i$. Al modelar la media de la variable respuesta mediante la función de enlace $g(\mu_i) = \eta_i$ con $\eta_i = \alpha_i$, es decir, la parte sistemática corresponde al modelo de medias de celda, se tiene que el modelo bajo la hipótesis alternativa está caracterizado por

$$\begin{aligned} \mathrm{E}\big(Y_{ij}\big) &= \frac{\kappa_i}{\lambda_i} = \mu_i \\ \mathrm{var}\big(Y_{ij}\big) &= \mu_i(1 + \phi_i \mu_i) \\ g\big(\mu_i\big) &= \alpha_i \end{aligned} \tag{7}$$

y la distribución de probabilidad de la variable respuesta $Y_{ij}$ es, según Hinde & Demétrio (1998),

$$P\big(Y_{ij} = y_{ij}\big) = \frac{\Gamma\big(\kappa_i + y_{ij}\big)}{\Gamma(\kappa_i) y_{ij}!} \frac{\mu_i^{y_{ij}} \kappa_i^{\kappa_i}}{(\mu_i + \kappa_i)^{\kappa_i + y_{ij}}} \tag{8}$$

donde $y_{ij} = 0, 1, 2 \ldots$. Tomando logaritmo natural, la contribución a la log-verosimilitud de la observación $y_{ij}$ es

$$\begin{aligned} l\big(\mu_i, \kappa_i \mid y_{ij}\big) = y_{ij} \ln \mu_i + \kappa_i \ln \kappa_i - \\ \big(\kappa_i + y_{ij}\big) \ln(\kappa_i + \mu_i) + \mathrm{dlg}(y_{ij}, \kappa_i) - \ln y_{ij}! \end{aligned} \tag{9}$$

donde $\mathrm{dlg}(y_{ij},\kappa_i) = \ln\Gamma\big(\kappa_i+y_{ij}\big)-\Gamma(\kappa_i)$. La log-verosimilitud para toda la muestra es $l(\boldsymbol{\mu},\boldsymbol{\kappa} \mid \boldsymbol{y}) = \sum_{i=1}^{K}\sum_{j=1}^{r_i} l\big(\mu_i,\kappa_i \mid y_{ij}\big)$.

En este caso, el modelo nulo es el usual descrito por Hinde & Demétrio (1998), en el cual se tienen los mismos supuestos del modelo alternativo, pero los $\theta_i$, $i = 1,\ldots,K$ son variables aleatorias con distribución $\Gamma(\kappa,\lambda_i)$, lo que equivale a tener igual parámetro de dispersión $\phi$ para todos los tratamientos. De esa forma se tiene

$$\begin{aligned} \mathrm{E}\big(Y_{ij}\big) &= \frac{\kappa}{\lambda_i} = \mu_i \\ \mathrm{var}\big(Y_{ij}\big) &= \mu_i(1+\phi\mu_i) \\ g\big(\mu_i\big) &= \alpha_i \end{aligned} \tag{10}$$

Nótese que el modelo alternativo se reduce al modelo nulo cuando $\phi_1 = \cdots = \phi_K = \phi = \frac{1}{\kappa}$. La distribución de probabilidad para la variable respuesta $Y_{ij}$ tiene la misma forma de la ecuación (8); basta remplazar $\kappa_i$ por $\kappa$. De manera similar, la contribución de la observación $y_{ij}$ a la log-verosimilitud tiene la misma forma de (9), remplazando $\kappa_i$ por $\kappa$.

**Ejemplo 2.** Se realizó un experimento con un diseño completamente al azar, en el cual se sometieron larvas de bagre blanco (*Sorubim cuspicaudus*) a tres densidades de siembra; 100, 200 y 300 larvas por litro (tratamientos). Una de las variables medidas fue el número de larvas muertas después del cuarto día de cultivo. Las muertes ocurren bien sea por canibalismo o por estrés debido a la manipulación. Los datos, tomados de García & Pérez (2008), se muestran en la tabla 2. Las medias observadas dentro de cada tratamiento son, respectivamente, 139.3, 230.5 y 321.6 mientras que las varianzas son: 9526.6, 2797.1 y 6190.3, las cuales son mucho más grandes que los valores de las medias. Asumiendo la distribución de Poisson para la respuesta, se espera que los valores de las medias sean similares a los de las varianzas. Estos resultados permiten concluir que los datos presentan sobredispersión (véase Sudhir & Saha 2007). Al ajustar un modelo lineal generalizado Poisson usual, con la función logaritmo natural como enlace, se obtiene un desvío de 457.58 con 15 grados de libertad, evidenciando la presencia de sobredispersión. Las estimaciones de los parámetros de regresión y sus errores estándar son: $\widehat{\alpha}_1 = 4.93687$, $\widehat{\mathrm{ee}}(\widehat{\alpha}_1) = 0.03459$, $\widehat{\alpha}_2 = 5.44025$, $\widehat{\mathrm{ee}}(\widehat{\alpha}_2) = 0.02689$ y $\widehat{\alpha}_3 = 5.77352$, $\widehat{\mathrm{ee}}(\widehat{\alpha}_3) = 0.02276$ y al aplicar la transformación inversa se tiene $\widehat{\mu}_1 = 139.33$, $\widehat{\mu}_2 = 230.50$ y $\widehat{\mu}_3 = 321.67$. Asumiendo que la distribución binomial negativa es adecuada para el mecanismo aleatorio que generó estos datos, procedemos a verificar el supuesto $\phi_1 = \phi_2 = \phi_3$.

Tabla 2: Número de larvas de bagre blanco muertas al cuarto día de cultivo.

| Densidades de siembra de larvas | Número de larvas muertas | | | | | |
|---|---|---|---|---|---|---|
| 100 | 325 | 105 | 124 | 155 | 63 | 64 |
| 200 | 257 | 233 | 310 | 223 | 210 | 150 |
| 300 | 335 | 376 | 426 | 319 | 201 | 273 |

Después de estimar por máxima verosimilitud los parámetros del modelo alternativo (7), se obtiene $\widehat{\mu}_1 = 139.394$, $\widehat{\mu}_2 = 230.579$, $\widehat{\mu}_3 = 321.548$, $\widehat{\phi}_1 = 0.3162$, $\widehat{\phi}_2 = 0.0415$ y $\widehat{\phi}_3 = 0.0510$, la log-verosimilitud es $-100.2$. Los valores estimados para el modelo nulo son: $\widehat{\mu}_1 = 139.3236, \widehat{\mu}_2 = 230.4857, \widehat{\mu}_3 = 321.6874, \widehat{\phi} = 0.1359$, y la log-verosimilitud es $-104$; por tanto, $\lambda^* = -2(-104 - (-100.2)) = 7.6$ con $6 - 4 = 2$ grados de libertad, el valor $p$ de la prueba es 0.02207 así que a un nivel de significación de 3 % se rechaza la hipótesis nula $H_0$, luego estadísticamente el parámetro $\phi$ no es el mismo para los tres tratamientos.

# 4. Modelo lineal generalizado ponderado: una alternativa para el análisis

Nelder & Wedderburn (1972) demuestran que la estimación por máxima verosimilitud de los parámetros de un modelo lineal generalizado es equivalente a un proceso de Mínimos Cuadrados Ponderados Iterativos (MCPI) con función de pesos

$$w_i = \frac{1}{\mathrm{V}(\mu_i)} \left( \frac{d\mu_i}{d\eta_i} \right)^2 \tag{11}$$

y una variable dependiente modificada

$$Z_i = \eta_i + (Y_i - \mu_i) \left( \frac{d\eta_i}{d\mu_i} \right) \tag{12}$$

Los pesos $w_i$, dados en la ecuación (11), corresponden al inverso de la varianza de $Z_i$ y, por tanto, son tales que $w_i \operatorname{var}(Z_i) = 1$. Sin embargo, si la variable respuesta tiene una función de varianza de la forma $\operatorname{var}(Y_i) = \delta_i \mathrm{V}(\mu_i)$, entonces se tiene que $w_i \operatorname{var}(Z_i) = \delta_i$; con este último factor se evalúa la dispersión extra. Lo anterior sugiere, por analogía con el modelo lineal clásico (véase Ravishanker & Dey 2001), que en el ajuste de los datos se usen los valores $1/\delta_i$ como ponderaciones, de tal forma que se cancele el efecto del factor $\delta_i$, lo que es equivalente a usar $w_i^* = w_i/\delta_i$, en lugar de los pesos $w_i$ en la ecuación (11).

Así que con datos de proporción con presencia de sobredispersión y bajo el supuesto que el modelo beta-binomial es correcto, el razonamiento anterior conduce a que el proceso MCPI se realice con pesos $w_{ij}^* = w_{ij}/\delta_{ij}$ con $\delta_{ij} = [1 + \phi_i(m_{ij} - 1)]$, donde $w_{ij}$ son los pesos que se calculan, para cada observación, mediante la ecuación (11), los cuales resultan del modelo binomial usual. Cuando se tienen datos de conteos con sobredispersión y bajo el supuesto que el modelo binomial negativo es adecuado, usando MCPI para la estimación, los pesos correspondientes son $w_{ij}^* = w_{ij}/\delta_{ij}$ con $\delta_{ij} = (1 + \phi_i \mu_i)$ y $w_{ij}$ los pesos que se calculan para cada observación mediante (11), que resultan del modelo Poisson usual.

Los parámetros $\phi_i$ y $\mu_i$ no se conocen, pero después del proceso de prueba de la hipótesis de interés (1) se tienen las estimaciones por máxima verosimilitud, que se usarían en su lugar como estimaciones iniciales. La mayoría de los paquetes para el cálculo estadístico permiten la inclusión de ponderaciones en el ajuste de

un modelo lineal generalizado. En particular, en `R` (Development Core Team 2007) solo hay que agregar la opción `weights=1/deltai` en la función `glm()`, así que una vez se han estimado los pesos que se deben usar, es muy sencillo llevar a cabo el análisis ponderado.

La matriz de varianzas y covarianzas del vector de parámetros estimados $\widehat{\boldsymbol{\beta}}$ está dada por la inversa de la matriz de información (véase Dobson 2002), $\boldsymbol{\mathfrak{I}}^{-1} = (\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}$, donde $\boldsymbol{W} = \text{diag}(w_{ij})$ para $i = 1, \ldots, K$, $j = 1, \ldots, r_i$ y $\boldsymbol{X}$ es la matriz de diseño, la cual, de acuerdo con la estructura de los datos en el modelo de medias de celda, tiene la forma $\boldsymbol{X} = \bigoplus_{i=1}^{K} \mathbf{1}_{r_i}$, así que la matriz de información $\boldsymbol{\mathfrak{I}}$ es diagonal con elementos en la diagonal dados por $\mathfrak{I}_i = \sum_{j=1}^{r_i} w_{ij}$; por tanto, si se ajusta un modelo lineal generalizado usual, $\widehat{\text{var}}\left(\widehat{\beta}_i\right) = \left(\sum_{j=1}^{r_i} w_{ij}\right)^{-1}$, mientras que si se ajusta un modelo lineal generalizado ponderado con pesos $1/\delta_{ij}$, se tiene $\widehat{\text{var}}\left(\widehat{\beta}_i^*\right) = \left(\sum_{j=1}^{r_i} w_{ij}^*\right)^{-1}$, donde los $\widehat{\beta}_i^*$ hacen referencia a la estimación del parámetro $\beta_i$ a partir del ajuste con ponderaciones y $w_{ij}^* = w_{ij}/\delta_{ij}$. Como en este caso se tiene que $\delta_{ij} \geq 1$, entonces $w_{ij} \geq w_{ij}^*$ y, por tanto, $\sum_{j=1}^{ri} w_{ij} \geq \sum_{j=1}^{r_i} w_{ij}^*$, así que

$$\widehat{\text{var}}\left(\widehat{\beta}_i\right) = \frac{1}{\sum_{j=1}^{r_i} w_{ij}} \leq \frac{1}{\sum_{j=1}^{r_i} w_{ij}^*} = \widehat{\text{var}}\left(\widehat{\beta}_i^*\right)$$

con lo cual se concluye que la varianza estimada de los estimadores obtenida con el modelo ponderado es mayor o igual que la que se obtiene al ajustar un modelo generalizado usual sin ponderaciones. Es claro que la igualdad se presenta cuando $\delta_{ij} = 1$, es decir, cuando $\phi_i = 0$. Se observa, además, que si $\delta_{ij} = \delta_{ij'} = \delta_i =$ para $j \neq j'$, es decir, si los pesos dentro del tratamiento $i$ son todos iguales, entonces

$$\widehat{\text{var}}\left(\widehat{\beta}_i^*\right) = \frac{\delta_i}{\sum_{j=1}^{r_i} w_{ij}} = \delta_i \widehat{\text{var}}\left(\widehat{\beta}_i\right)$$

es decir, la varianza estimada del estimador del parámetro $\beta_i$ queda multiplicada por el factor $\delta_i$ o, lo que es igual, el error estándar queda multiplicado por $\sqrt{\delta_i}$. La regresión ponderada resuelve el problema de la subestimación de los errores estándar de los estimadores de regresión, lo cual, según Hinde & Demétrio (1998), es una de las consecuencias más graves cuando no se tiene en cuenta la sobredispersión, porque se podría evaluar de manera incorrecta la significancia de los parámetros.

Las estimaciones de los parámetros de regresión también se ven afectadas por el uso de las ponderaciones $1/\delta_{ij}$. A partir de $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{z}$, teniendo en cuenta la forma de las matrices $\boldsymbol{X}$ y $\boldsymbol{W}$, se demuestra que

$$\widehat{\beta}_i^* = \frac{\sum_{j=1}^{r_i} \frac{w_{ij}}{\delta_{ij}} z_{ij}}{\sum_{j=1}^{r_i} \frac{w_{ij}}{\delta_{ij}}}$$

mientras que si no se usan las ponderaciones se tiene $\widehat{\beta}_i = \frac{\sum_{j=1}^{r_i} w_{ij} z_{ij}}{\sum_{j=1}^{r_i} w_{ij}}$. Nótese que $\widehat{\beta}_i^* \approx \widehat{\beta}_i$ cuando $\delta_{ij} \approx 1$ y si se verifica que $\delta_{ij} = \delta_{ij'} = \delta_i$, entonces las estimaciones del modelo sin ponderar coinciden con las del modelo que usa las ponderaciones,

es decir, $\widehat{\beta}_i^* = \widehat{\beta}_i$. En el caso de datos de proporción, la igualdad se verifica cuando $m_{ij} = m_{ij'} = m_i$, para todo $j \neq j'$, es decir, el número de ensayos a partir de los cuales se obtienen los $y_{ij}$ éxitos es el mismo dentro del tratamiento $i$.

**Ejemplo 3** (continuación del ejemplo 1)**.** En seguida se muestran los resultados del ajuste de un modelo lineal generalizado binomial con la función logit$(\cdot)$ como enlace, pero usando los valores $1/\widehat{\delta}_{ij}$ como ponderaciones con $\widehat{\delta}_{ij} = 1+\widehat{\phi}_i(m_{ij}-1)$ $\widehat{\alpha}_1 = 1.2261$, $\widehat{\text{ee}}(\widehat{\alpha}_1) = 0.2736$; $\widehat{\alpha}_2 = -2.1736$, $\widehat{\text{ee}}(\widehat{\alpha}_2) = 0.3626$; $\widehat{\alpha}_3 = -3.3322$, $\widehat{\text{ee}}(\widehat{\alpha}_3) = 0.7196$ y $\widehat{\alpha}_4 = -3.0011$, $\widehat{\text{ee}}(\widehat{\alpha}_4) = 0.6097$. Además, se obtuvo un desvío de 53.685 con 54 grados de libertad, evidenciando un buen ajuste de los datos al modelo donde no hay homogeneidad del parámetro de dispersión.

De acuerdo con los resultados teóricos mostrados en esta sección, se observa que la estimación $\widehat{\alpha}_3$ y su error estándar estimado no varían con respecto al ajuste no ponderado (véase el ejemplo 1). Eso es debido a que $\widehat{\phi}_3 \approx 0$, lo cual implica que $\delta_{3j} \approx 1$ para todo $j = 1, \ldots, r_3$. Las estimaciones $\widehat{\alpha}_1$, $\widehat{\alpha}_2$ y $\widehat{\alpha}_4$ varían poco, pero los errores estándar se modifican sustancialmente; por ejemplo, el del tratamiento 1 pasa de 0.1292 a 0.2736, aproximadamente el doble. Una aproximación de este valor se obtiene a partir de $\left[1+\widehat{\phi}_1 \times (11-1)\right]^{1/2} \times 0.1292 = \sqrt{1+0.3397\times 10} \times 0.1292 = 2.0969 \times 0.1292 = 0.2709$, donde se ha tomado 11 como un valor promedio del número de fetos en ese tratamiento. Si el número de fetos de cada rata dentro del tratamiento 1 fuera igual, es decir, si se tuviera que $m_{1j} = m_{1j'}$ para todo $j \neq j'$, se obtendría exactamente el valor 0.2736. Se observa que las varianzas estimadas de los estimadores son más grandes en el modelo ponderado, como se esperaba, de acuerdo con los resultados de esta sección. A partir del ajuste se obtiene un desvío residual $D = 53.68$ con 54 grados de libertad, $p = 0.486$; el valor del estadístico $X^2$ de Pearson es 48.84 con 54 grados de libertad y $p = 0.673$. Ambas estadísticas indican que el modelo lineal generalizado ponderado proporciona un buen ajuste a los datos (véase Dobson 2002). El análisis de residuales, que se muestra en el apéndice, indica que no hay evidencia de violación de los supuestos del modelo y tampoco hay presencia de datos influyentes ni atípicos.

**Ejemplo 4** (continuación del ejemplo 2)**.** En seguida se presentan los resultados de ajustar un modelo lineal generalizado tipo Poisson con la función log$(\cdot)$ como enlace y los valores $1/\widehat{\delta}_{ij}$ con $\widehat{\delta}_{ij} = 1+\widehat{\phi}_i\widehat{\mu}_i$ como ponderaciones. En este caso $\delta_{1j} = 45.081$, $\delta_{2j} = 10.560$ y $\delta_{3j} = 17.403$ para todo $j = 1, 2, \ldots, 6$. Se obtiene un desvío de 18.169 con 15 grados de libertad y las siguientes estimaciones de los parámetros de regresión $\widehat{\alpha}_1^* = 4.93687$, $\widehat{\text{ee}}(\widehat{\alpha}_1^*) = 0.23222$, $\widehat{\alpha}_2^* = 5.44025$, $\widehat{\text{ee}}(\widehat{\alpha}_2^*) = 0.08738$ y $\widehat{\alpha}_3^* = 5.77352$ , $\widehat{\text{ee}}(\widehat{\alpha}_3^*) = 0.09496$. Se observa que las estimaciones de los parámetros no varían con respecto a las obtenidas a partir del ajuste sin ponderaciones (véase ejemplo 2) pero $\widehat{\text{ee}}(\widehat{\alpha}_1^*) = \sqrt{45.081} \times \widehat{\text{ee}}(\widehat{\alpha}_1) = 0.23222$, de manera similar $\widehat{\text{ee}}(\widehat{\alpha}_2^*) = \sqrt{10.560} \times \widehat{\text{ee}}(\widehat{\alpha}_2) = 0.08738$ $\widehat{\text{ee}}(\widehat{\alpha}_3^*) = \sqrt{17.403} \times \widehat{\text{ee}}(\widehat{\alpha}_3) = 0.09496$.

A partir del ajuste se obtiene un desvío residual $D = 18.17$ con 15 grados de libertad y $p = 0.253$; el valor del estadístico $X^2$ es 18.86 con 15 grados de libertad y $p = 0.22$. Ambas estadísticas indican que el modelo lineal generalizado ponderado proporciona un buen ajuste a los datos, al confrontarlo con el ajuste del modelo maximal (sobreparametrizado), (véase Dobson 2002). El análisis de residuales,

que se muestra en el apéndice, indica que no hay evidencia de violaciones de los supuestos del modelo y tampoco hay presencia de datos influyentes o atípicos.

# 5. Análisis de la potencia de la prueba

Dado que el procedimiento de prueba propuesto para la hipótesis (1) involucra métodos numéricos iterativos, no es posible obtener una expresión explícita para la función de potencia de la prueba. Para obtener una aproximación gráfica, se llevó a cabo un proceso de simulación.

Se generaron datos por un modelo con una vía de clasificación, con $K = 2, 3, 4$ y 5 tratamientos, a partir de la distribución beta-binomial, en el caso de proporciones, y de la distribución binomial negativa, en el caso de conteos, mediante las funciones `rbetabin()`[1] y `rnbinom()`, respectivamente, ambas funciones del paquete estadístico R. Se generaron 5, 10, 15, 20 y 25 datos, dentro del tratamiento $i$, para cada valor de $K$. Se estableció, para $i = 1, \ldots, K$, $\pi_i = 0.5$ en el caso de proporciones y $\mu_i = 12$ para conteos. Los valores de los parámetros $\phi$'s se establecieron de la siguiente forma: en el caso de datos de proporción se tomó $\phi_1 = 0.19$ y, para $i = 2, \ldots, K$, los valores de $\phi_i$ se establecieron de tal forma que la diferencia $\Delta\phi = \phi_{k+1} - \phi_k$ tome los valores 0, 0.01, 0.03, 0.04, 0.06, 0.07, 0.09, 0.10, 0.11, 0.13, 0.14, 0.16, 0.17, 0.19, 0.20. Por ejemplo, en el caso de $K = 5$ los valores de $\phi_i$ fueron $\phi_1 = \cdots = \phi_5 = 0.19$ (hipótesis $H_0$ verdadera) y para un $\Delta\phi = 0.20$ se tomó $\phi_1 = 0.19$, $\phi_2 = 0.39$, $\phi_3 = 0.59$, $\phi_4 = 0.79$, $\phi_5 = 0.99$ (hipótesis $H_0$ falsa). En el caso de datos de conteo se tomó $\kappa_1 = 1.1$ ($\phi_1 = 1/1.1 = 0.9090$) y los demás valores de $\kappa_i$, para $i = 2, \ldots, K$, se establecieron de tal forma que la diferencia $\Delta\kappa = \kappa_{k+1} - \kappa_k$ tome los valores 0.00, 0.15, 0.30, 0.45, 0.60, 0.75, 0.90, 1.05, 1.20, 1.35, 1.50, 1.65, 1.80, 1.95, 2.10. Por ejemplo, en el caso de $K = 5$ los valores de $\kappa_i$ fueron $\kappa_1 = \cdots = \kappa_5 = 1.1$, equivalente a $\phi_1 = \cdots = \phi_K = 0.9090$ (hipótesis $H_0$ verdadera); para un $\Delta\kappa = 1.20$ se tomó $\kappa_1 = 1.1$, $\kappa_2 = 2.30$, $\kappa_3 = 3.50$, $\kappa_4 = 4.70$, $\kappa_5 = 5.90$, lo que equivale a $\phi_1 = 0.9090$, $\phi_2 = 0.4348$, $\phi_3 = 0.2857$, $\phi_4 = 0.2128$, $\phi_5 = 0.1695$ (hipótesis $H_0$ falsa).

Una vez generados los datos, se lleva a cabo el procedimiento de prueba como se explica en la sección 3, tomando $\alpha = 0.05$. El proceso se repite 1000 veces y se cuenta las veces ($R$) que se rechaza la hipótesis nula. El valor $R/1000$ proporciona una estimación de la potencia de la prueba.

## 5.1. Resultados de las simulaciones

En el caso de datos de proporción, para 5 repeticiones y 2 tratamientos, la prueba reporta una proporción de rechazo de 0.073 y un valor de 0.096 para tres tratamientos y 5 repeticiones, es decir, una proporción de rechazo moderadamente mayor que el valor esperado de 0.05. Se observa que para 15 repeticiones y 5 tratamientos el valor de la proporción de rechazos es 0.06; con 5 tratamientos y 25 repeticiones, es 0.055. De los resultados de las simulaciones es claro que cualquiera

[1] Esta función pertenece a la librería `VGAM`

que sea el número tratamientos, a medida que aumenta el número de repeticiones la proporción de rechazos se acerca al valor 0.05, mostrando el comportamiento asintótico de la prueba.

En la figura 1 se muestra la función de potencia obtenida mediante las simulaciones para cada combinación de número de repeticiones y número de tratamientos. La línea horizontal punteada corresponde al valor 0.05. A partir del gráfico se evidencia que a medida que aumenta el número de repeticiones mejora la capacidad de las pruebas para detectar alejamientos de la hipótesis nula (comportamiento asintótico). Por ejemplo, cuando se tiene $\Delta\phi = 0.14$, que corresponde, en el caso de 5 tratamientos, a la parametrización $\pi_i = 0.5$ para todo $i = 1, \ldots, 5$ y $\phi_1 = 0.19$, $\phi_2 = 0.33$, $\phi_3 = 0.47$, $\phi_4 = 0.61$ y $\phi_5 = 0.75$ se observa, para 5 repeticiones, que la proporción de rechazos es de 80 %, para 10 repeticiones es de 94 %, con 15 repeticiones es de 98 %, con 20 repeticiones la proporción de rechazos es de 99.7 % y, por último, con 25 repeticiones la prueba rechaza el 100 % de las veces.

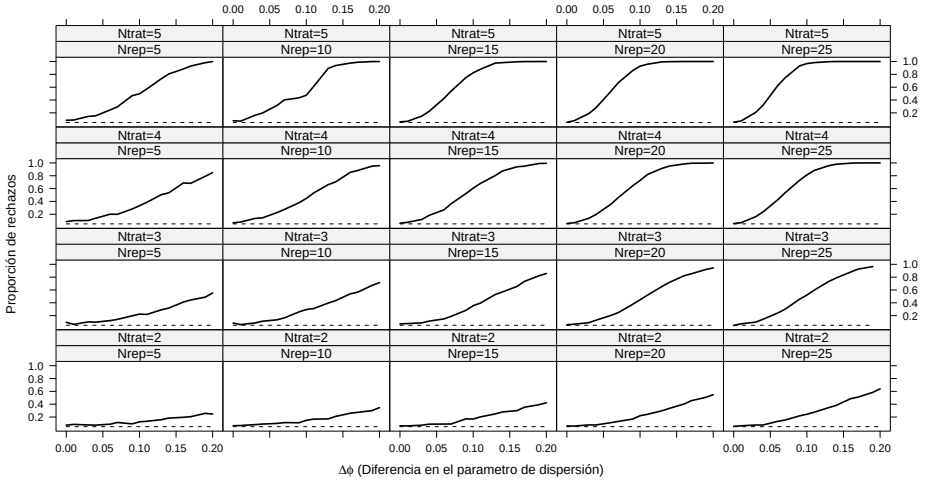


FIGURA 1: Función de potencia simulada: `Ntrat`, número de tratamientos. `Nrep`, número de repeticiones.

Con datos de conteo, la prueba reporta, para 5 repeticiones y 2 tratamientos, una proporción de rechazo de 0.077, y para tres tratamientos y 5 repeticiones, un valor de 0.079, es decir, una proporción de rechazo ligeramente mayor que el 5 % que se esperaba. Sin embargo, se observa que para 25 repeticiones y 5 tratamientos el valor de la proporción de rechazos es 0.045. De acá se induce que a medida que aumenta el número de repeticiones, la proporción de rechazos se acerca al valor 0.05, mostrando el comportamiento asintótico de la prueba.

A partir del gráfico de la función de potencia obtenida mediante las simulaciones, se evidencia un aumento de la potencia de la prueba a medida que aumenta el tamaño de la muestra. Por ejemplo, cuando se tiene $\Delta\kappa = 1.5$, que corresponde, en el caso de 4 tratamientos, a la parametrización $\mu_i = 12$ para todo $i = 1, \ldots, 4$

y $\kappa_1 = 1.1$, $\kappa_2 = 2.60$, $\kappa_3 = 4.10$, $\kappa_4 = 5.63$ o equivalentemente a $\phi_1 = 0.9090$, $\phi_2 = 0.3846$, $\phi_3 = 0.2439$, $\phi_4 = 0.1776$, se observó que para 5 repeticiones la proporción de rechazos es de 40.7 %; para 10, de 52.4 %; para 15, de 64.4 %; para 20, de 78.8 %; y para 25, de 87.6 %.

## 6. Conclusiones

Cuando se analizan datos de proporción o conteos en presencia de sobredispersión, en la literatura estadística lo usual es asumir un parámetro de dispersión $\phi$ común a todas las observaciones. Sin embargo, puede suceder que cada tratamiento tenga asociado su propio parámetro de sobredispersión. En este artículo se demuestra que hay situaciones en que es necesario verificar inicialmente la hipótesis de que el parámetro se mantiene constante a través de todos los tratamientos, es decir, hay homogeneidad de dicho parámetro. Una de las consecuencias de la sobredispersión es la subestimación de los errores estándar de los estimadores de regresión; una forma de evitar este problema es multiplicar los errores estándar por el factor $\delta$ de modo que la varianza de la variable respuesta sea $\delta\,\mathrm{V}(\mu)$, como lo sugieren Hinde & Demétrio (1998), pero si el parámetro $\delta$ no es igual para todas las observaciones, se corre el riesgo de sobreestimar unos errores estándar y subestimar otros.

Para el análisis de datos de proporción o conteos en presencia de sobredispersión, bajo el supuesto que los modelos beta binomial o binomial negativo son adecuados para el ajuste, es apropiado que se ajuste un modelo de regresión lineal generalizado ponderado con pesos $1/\delta_{ij}$, donde $\delta_{ij}$ es tal que $\mathrm{var}(Y_{ij}) = \delta_{ij}\,\mathrm{V}(\mu_i)$ con $\mathrm{V}(\mu_i)$ la función de varianza usual. Los valores $\delta_{ij}$ son desconocidos, pero en el caso de la distribución beta binomial o binomial negativa, después del procedimiento de prueba de la hipótesis propuesta en (1), se cuenta con la estimación de estos, la cual se usa en su lugar.

[Recibido: abril de 2008 — Aceptado: enero de 2009]

## Referencias

Agresti, A. (2002), *Categorical Data Analysis*, John Wiley & Sons, New Jersey, United States.

Atkinson, A. C. (1981), 'Two Graphical Displays for Outlying and Influential Observations in Regression', *Biometrika* **68**, 13–20.

Brooks, R. J. (1978), 'Approximate Likelihood-Ratio Test in the Analysis of Beta-Binomial Data', *Applied Statistics* **12**, 1589–1596.

Crowder, M. J. (1978), 'Beta-Binomial Anova for Proportions', *Applied Statistics* **27**(1), 34–37.

Crowder, M. J. (1979), 'Inference About the Intraclass Correlation Coefficient in the Beta-binomial ANOVA for Proportions', *Journal The Royal Statistical Society* **41**(2), 230–234.

Development Core Team, R. (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Dobson, A. J. (2002), *An Introduction to Generalized Linear Models*, Chapmann & Hall/CRC, New York, United States.

García, Y. B. & Pérez, J. A. (2008), Efecto de la densidad de siembra en la larvicultura de bagre blanco (*sorubim cuspicaudus*), Tesis de pregrado, Departamento de Acuicultura, Facultad de Medicina Veterinaria y Zootecnia, Universidad de Córdoba, Montería, Colombia.

Hinde, J. & Demétrio, C. (1998), *Overdispersion: Models and Estimation*, ABE, São Paulo, Brazil.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapmann & Hall/CRC, New York, United States.

Morales, M. A. (2008), Estudio de homogeneidad de la dispersión en un diseño completamente al azar con datos de proporciones y conteos, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.

Nelder, J. A. & Wedderburn, D. W. M. (1972), 'Generalized Linear Models', *Journal The Royal Statistical Society: Series A* **135**(3), 370–384.

Prentice, R. L. (1986), 'Binary Regresion Using an Extended Beta-Binomial Distribution with Discussion of Correlation Induced by Covariate Measurement Errors', *Journal of the American Statistical Association* **81**(394), 321–327.

Ravishanker, N. & Dey, D. K. (2001), *A First Course in Linear Model Theory*, Chapmann & Hall/CRC, New York, United States.

Sudhir, P. & Saha, K. (2007), 'The Generalized Linear Model and Extensions: a Review and some Biological and Environmental Applications', *Environmetrics* **18**, 421–443.

# Apéndice

## Análisis de residuales y técnicas de diagnóstico

En la figura 2 se muestran seis gráficos para el análisis de residuos y detección de puntos influyentes en el ajuste del modelo lineal generalizado binomial con ponderaciones (ejemplo 3). En el gráfico a, que muestra los desvíos residuales

(componentes de desvío) contra los valores ajustados se observa el dato $y_{11}$ como un posible atípico (*outlier*). En el gráfico b, que muestra los valores absolutos de los residuales contra los valores ajustados, no se observa tendencia alguna, por lo cual se concluye que la función de varianza asumida es adecuada, véase McCullagh & Nelder (1989). El gráfico del predictor lineal ajustado ($\widehat{\boldsymbol{\eta}}$) contra la variable dependiente ajustada ($\boldsymbol{z}$), (gráfico c), muestra una tendencia lineal de los puntos, con lo cual se concluye que la función de enlace logit$(\cdot)$ es apropiada (McCullagh & Nelder 1989). El gráfico de $h_{ii}$, que se muestra en el gráfico d, indica que la observación $y_{33}$ tiene un alto *leverage* y que podría ser una observación influyente, mientras que los gráficos e y f muestran que las observaciones $y_{16}$, $y_{25}$ y $y_{43}$ son las que más influencia ejercen sobre los parámetros de localización y escala. Sin embargo, todas las distancias de Cook en el gráfico medio normal (*half normal plot*) de la figura 3b se encuentran dentro de la banda. Por tanto, no hay evidencia de observaciones influyentes en el ajuste, véase Atkinson (1981). El gráfico de los residuales con bandas simuladas, que se muestra en la figura 3a, confirma la ausencia de datos atípicos de acuerdo con Atkinson (1981).

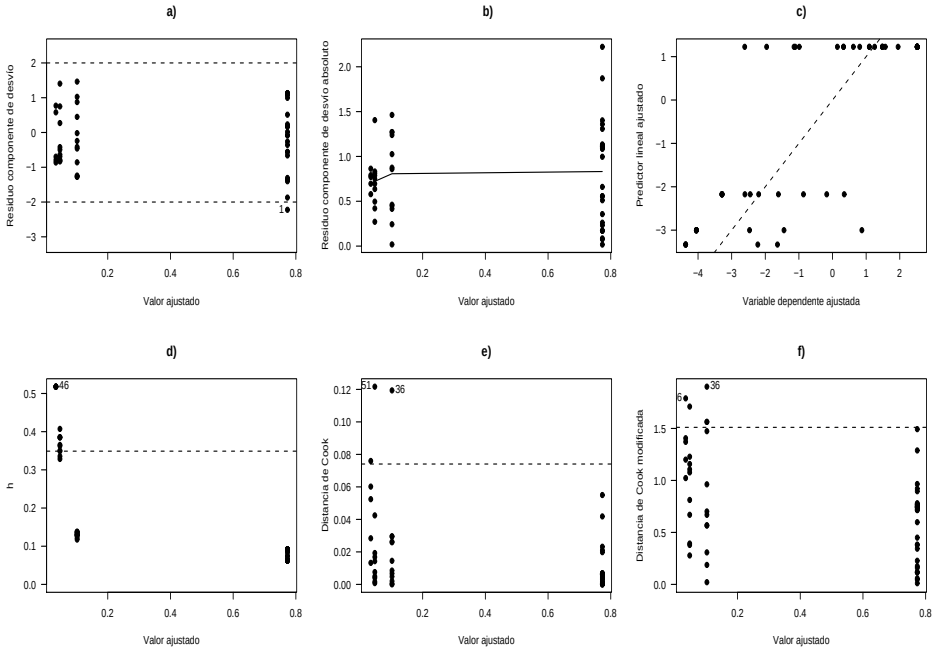


FIGURA 2: Gráficos para el análisis de residuales del modelo lineal generalizado binomial ponderado, ejemplo 3.

En la figura 4 se muestran seis gráficos para el análisis de residuos y detección de puntos influyentes en el modelo lineal generalizado Poisson. El gráfico a, que muestra los desvíos residuales (componentes de desvío), permite concluir que no hay evidencia de datos atípicos. En el gráfico b, que muestra los valores absolutos de los residuales contra los valores ajustados, no se observa tendencia alguna,

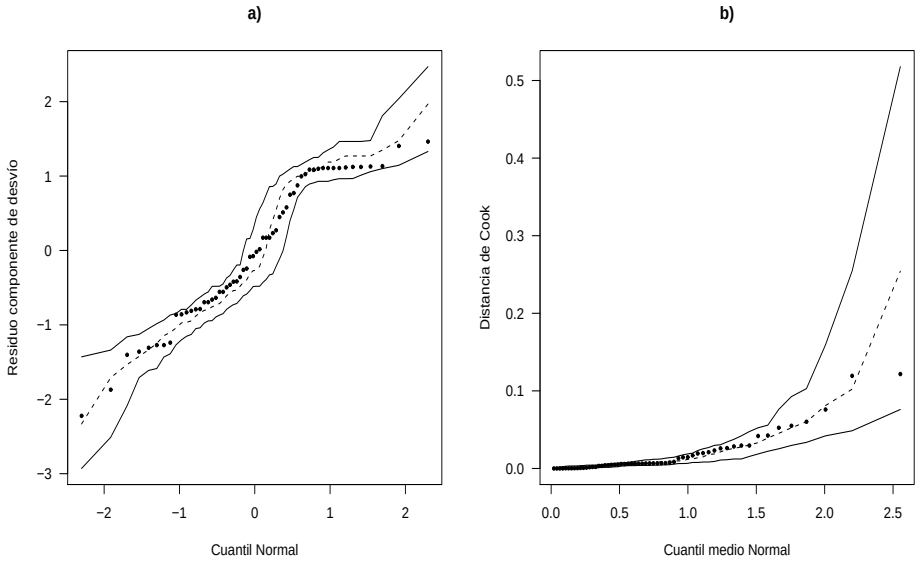


FIGURA 3: (a): gráfico de probabilidad normal con banda de confianza simulada para los residuos y (b): gráfico de probabilidad medio normal con banda de confianza al 95 % simulada para la distancia de Cook, (ejemplo 3).

por lo cual se concluye que la función de varianza asumida es adecuada, véase McCullagh & Nelder (1989); el gráfico del predictor lineal ajustado ($\widehat{\boldsymbol{\eta}}$) contra la variable dependiente ajustada ($\boldsymbol{z}$), gráfico c, muestra una tendencia lineal de los puntos, con lo cual se concluye que la función de enlace es apropiada (McCullagh & Nelder 1989). El gráfico d indica que no hay puntos con alto *leverage*, mientras que los gráficos e y f muestran que la observación $y_{11}$ es la que más influencia ejerce sobre los parámetros de localización y escala; sin embargo, todas las distancias de Cook en el gráfico medio normal (*half normal plot*) de la figura 5b se encuentran dentro de la banda simulada. Por tanto, no se tiene evidencia de observaciones influyentes en el ajuste (Atkinson 1981). El gráfico de los residuales con bandas simuladas, que se muestra en la figura 5a, confirma la ausencia de datos atípicos (Atkinson 1981).

## Código R

A continuación se presenta el código, en lenguaje R (Development Core Team 2007), que permite realizar los cálculos correspondientes a los ejemplos 2 y 4.

```
library(aod)
dia4<-c(325,105,124,155,63,64,
        257,233,310,223,210,150,
        335,376,426,319,201,273)
```

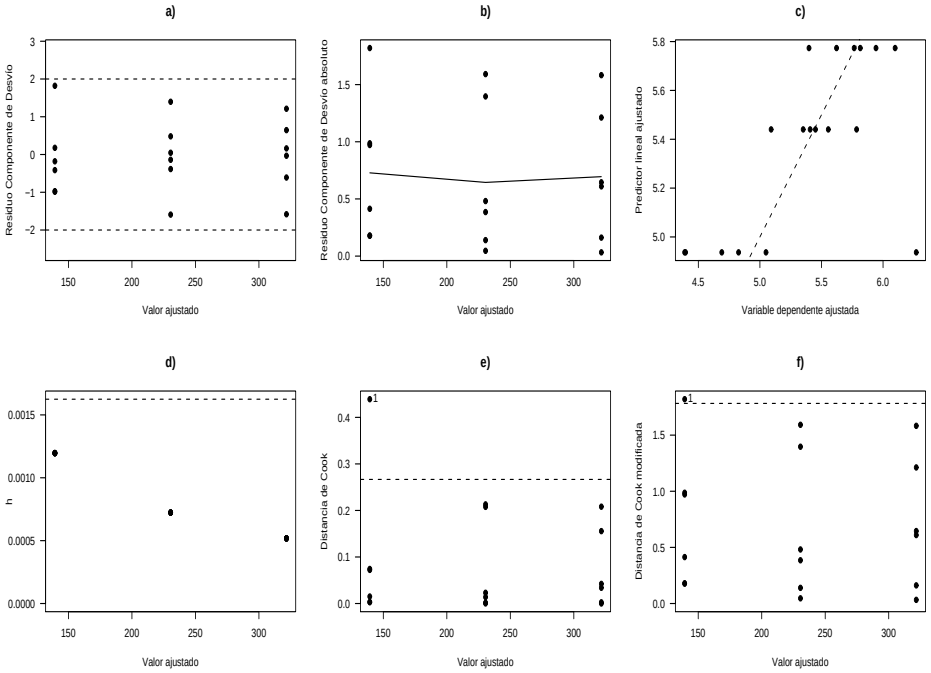


FIGURA 4: Gráficos para el análisis de residuales, del modelo lineal generalizado Poisson ponderado ejemplo 4.

```
Trat<-factor(rep(c("trat1","trat2","trat3"),c(6,6,6)))
mort<-data.frame(dia4=dia4,Trat=Trat)
# ajuste de un modelo lineal generalizado usual
mlgu4<-glm(dia4~-1+Trat,data=mort,
              family=poisson)
summary(mlgu4)
#ajuste de un modelo binomial negativo con un parámetro de
#dispersión por tratamiento
mod1<-negbin(dia4~-1+Trat,~Trat,data=mort)
summary(mod1)
# Mu's estimados
mu_est<-exp(mod1@fixed.param)
# deltas estimados
delta<-mod1@random.param
# K's estimados
k<-1/mod1@random.param
#phi's estimados
phi_est<-1+delta*mu_est
#ajuste de un modelo binomial negativo con un parámetro de
#dispersión común a todos los tratamientos
```

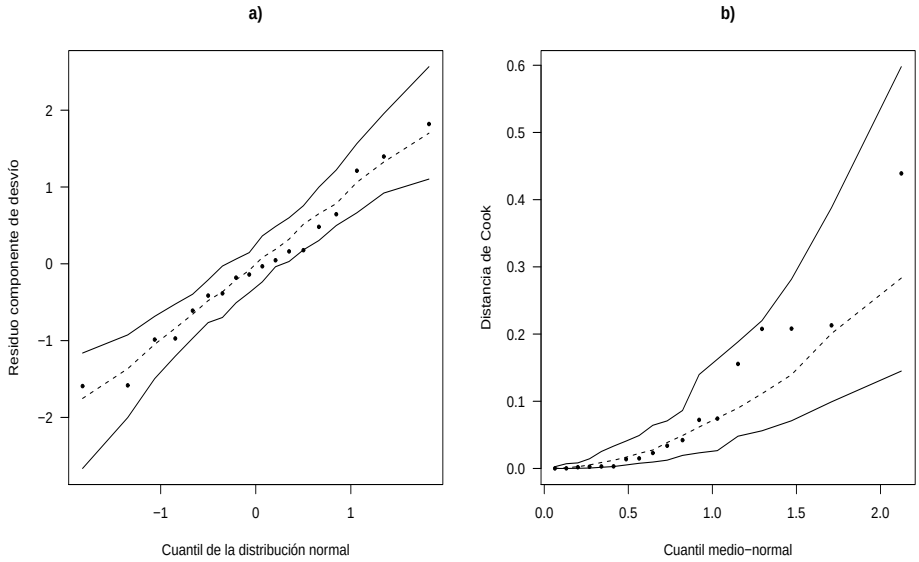


FIGURA 5: (a): gráfico de probabilidad normal con banda de confianza simulada para los residuos y (b): gráfico de probabilidad medio normal con banda de confianza al 95 % simulada para la distancia de Cook ejemplo 4.

```
mod0<-negbin(dia4~-1+Trat,~1,data=mort)
summary(mod0)
# prueba de la hipótesis
anova(mod0,mod1)
# Mu's estimados
mu<-exp(mod0@fixed.param)
# K estimado
k<-1/mod0@random.param
# Para ajustar un modelo lineal generalizado usual ponderado
repet<-as.numeric(by(mort$dia4, mort$Trat,length))
pesos <-rep(phi_est,repet)
modglm<-glm(dia4~-1+Trat,data=mort,family=poisson(link=log),weights=1/pesos)
summary(modglm)
```