Miscelánea Matemática **73** (2021) 39-47 SMM

DOI: https://doi.org/10.47234/mm.7304

Mi trabajo con el Dr. Federico O'Reilly

Silvia Ruiz-Velasco
Instituto de Investigaciones en
Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
silvia@sigma.iimas.unam.mx

1. Introducción

Con el Dr. Federico O'Reilly trabajeé en dos proyectos diferentes: uno de ellos fue el uso de distribuciones fiduciales para construir intervalos de confianza en distribuciones discretas, particularmente en distribuciones de la Familia de Series de Potencias este trabajo lo realizamos con el Dr. Edilberto Nájera y el otro tema fue el uso de muestras condicionalmente independientes o *look alike* para bondad de ajuste en modelos lineales generalizados, este trabajo lo realizamos con la Dra. Lizbeth Naranjo. Describiré brevemente ambas ideas.

2. Distribuciones fiduciales en la familia de series de potencias

Se considera un problema de inferencia en el que para el parámetro $\theta \in \Theta$ (Θ un intervalo) la estadística T es suficiente minimal y $G(t;\theta)$ su distribución. La distribución fiducial para θ (habiendo observado $T = t$) se define como $H(\theta;t) = 1 - G(t;\theta)$, distribución también conocida como confidencial o de significancia, Fisher(1930). La construcción de intervalos de confianza puede hacerse con H aún si esta posee una masa en un extremo.

La familia de series de potencias considerada es la definida en Johnson et al. (1992), en donde la binomial (Bernoulli), binomial negativa(geométrica) y Poisson son casos particulares.

Para esta familia, en donde la variable aleatoria toma valores en los enteros no negativos y el parámetro $\theta > 0$, la estadística suficiente minimal es la suma de n observaciones independientes idénticamente

Palabras clave: Estadística fiducial, modelos lineales generalizados, bondad de ajuste.

distribuidas. En estas distribuciones la estadística suficiente resulta ser miembro de la familia exponencial dada por

$$g(t,\eta) = a(\eta)b(t)\exp(t\eta),$$

con $\eta \in [\underline{\eta}, \overline{\eta}]$. Si la familia es regular el intervalo del espacio parametral es abierto. Si utilizamos esta distribución para generar un intervalo de confianza, de tamaño 2α tenemos que encontrar los valores de t tal que $\alpha = G(t_1,\eta)$ y $1-\alpha = G(t_2,\eta)$.

2.1 T discreta y η contínua

Consideremos primero el caso en donde η el parámetro canónico de la distribución exponencial es contínuo y la estadística suficiente T es discreta, si pensamos en el problema de la prueba de hipótesis $H_0 : \eta = \eta_0$ *vs* $H_1 : \eta > \eta_0$, entonces habiendo observado t parecería arbitrario elegir como nivel de significancia o valor-p la probabilidad del evento $[T > t]$ o la del evento $[T \geq t]$. Si T tiene una distribución contínua ambas definiciones coinciden y la definición de distribución fiducial es única,

Si la distribución de T es discreta podríamos definir dos distribuciones fiduciales, la que utiliza $G(t;\eta)$, es decir la función de distribución de T evaluada en t o la versión continua por la izquierda que llamaremos $G^-(t,\eta)$, las distribuciones fiduciales siendo estas cantidades restadas de uno. Las dos versiones de la distribución fiducial satisfacen todas las condiciones que estableció Fisher acerca de monotocidad y propiedades límite, Fisher(1930).

Dadas estás propiedades, lo que se propone en el artículo es definir una distribución fiducial mezclando las dos distribuciones fiduciales anteriormente definidas, es decir definimos la fiducial como

$$H_\gamma(\eta;t) = \gamma(1 - G(t,;\eta)) + (1-\gamma)(1 - G^-(t;\eta)) \quad \gamma \in [0,1].$$

Esta expresión induce una familia de distribuciones fiduciales, donde cada una tendrá como límites cero cuando $\eta \to \underline{\eta}$ y uno cuando $\eta \to \bar{\eta}$.

2.2 Familia de series de potencias

La definición de la familia de series de potencias tiene una densidad de la forma

$$f(x;\theta) = \frac{B_x\theta^x}{A(\theta)},$$

donde el parámetro $\theta > 0$ y la variable aleatoria X toma valores en el conjunto de enteros no negativos y

$$A(\theta) = \sum_x B_x\theta^x.$$

Esta familia de distribuciones es miembro de la familia exponencial natural con parámetro $\eta = log(\theta)$. Dada una muestra de tamaño n, la estadística suficiente minimal es $T = \sum_i X_i$.

2.2.1. Caso acotado

Un caso especial es cuando el espacio de posibles valores de la variable X esta acotado, es decir toma valores en el conjunto $\{0, 1, \ldots, K\}$ y en ese caso la estadística $V = nK - T$ también tiene una distribución que pertenece a la familia natural exponencial con parámetro θ^{-1}.

Lo mas importante de este resultado es que la inferencia que se obtiene sobre θ utilizando $T = t$, es la misma que la que se obtiene utilizando $V = nK - T$, es decir habiendo observado v.

Más aún se puede demostrar que si

$$g_T(t, \theta) = \frac{B_t \theta^t}{A(\theta)},$$

entonces

$$g_V(v; \theta^*) = \frac{B_v^* \theta^{*v}}{A^*(\theta^*)}$$

donde $\theta^* = \frac{1}{\theta}$,$B_v^* = B_{nK-v}$ y $(\theta^{nK})A^*(\theta^*) = A(\theta)$.

Teniendo todo la anterior en cuenta y considerando además que $G_T(t, \theta)$ es decreciente, excepto cuando $t = nK$ con límites 1 cuando $\theta \to 0$ y 0 cuando $\theta \to \bar{\theta}$, con $\bar{\theta}$ el límite superior de los posibles valores de θ (puede ser ∞). La fiducial propuesta, que además resulta ser única esta dada por

$$H(\theta; t) = \begin{cases} 1 - G_T(t; \theta) & \text{si } t = 0, \\ \frac{1}{2}(1 - G_T(t; \theta)) + \frac{1}{2}(1 - G^-{}_T(t; \theta)) & \text{si } t = 1, \ldots, nK - 1, \\ 1 - G^-{}_T(t; \theta) & \text{si } t = nK. \end{cases}$$

Es posible ver que si utilizamos la función de distribución para v y obtenemos la fiducial para θ^* y después la transformamos para θ obtendremos la misma fiducial obtenida anteriormente.

Por último, en cuanto a este tema, en el artículo se obtienen explícitamente las fiduciales para el caso Binomial (Bernoulli) y Uniforme discreta. En el caso Binomial, se hace un estudio para comprar la longitud esperada de los intervalos de confianza y cobertura para los intervalos de Clopper-Pearson, Fiducial y Jeffreys.

2.2.2. Caso no acotado

En el caso de que la estadística suficiente T pertence a la familia de potencias pero toma valores en los enteros no negativos, se tiene que

$$g_t(t,\theta) = \frac{B_t e^{t\theta}}{A(\theta)}$$

con $A(\theta) = \sum_{j=0}^{\infty} B_j e^{j\theta}$

Definimos a la distribución fiducial como el límite cuando $M \to \infty$ de las fiduciales corregidas que se obtienen al limitar el rango al conjunto de valores $\{0, 1, \ldots, M\}$. En el artículo se demuestra que la distribución fiducial propuesta queda dada por:

$$H(\theta;t) = \begin{cases} 1 - G_T(t;\theta) & \text{si } t = 0, \\ \frac{1}{2}(1 - G_T(t;\theta)) + \frac{1}{2}(1 - G^{-}{}_T(t;\theta)) & \text{si } t = 1, 2, \ldots, \end{cases}$$

En el artículo se trabaja la distribución Poisson, y se demuestra también que la misma fiducial se obtiene como el límite de la distribución fiducial de la estadística suficiente de la binomial. También se estudia el caso de la binomial negativa, se deja ver que en el caso de la Binomial Negativa existe la posibilidad de renombrar los éxitos y fracasos, sin embargo no estaríamos en la misma situación que en el caso Binomial, por lo que existe trabajo por hacer.

3. Muestras condicionalmente independientes para bondad de ajuste en modelos lineales generalizados

La idea es aplicar la construcción de muestras condicionalmente independientes en el contexto de Modelos Lineales Generalizados y obtener la distribución exacta para la devianza y la X^2 de Pearson, estadísticas de Bondad de Ajuste.

3.1 Modelos lineales generalizados

Los modelos lineales generalizados (MLG) propuestos por Nelder y Wedderburn (1972) están especificados por tres componentes:

- El *componente aleatorio*, observaciones independientes con distribución.

$$f(y_i;\theta_i) = \exp[(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i)] \tag{1}$$

- El *componente sistemático* o predictor lineal

$$\eta(\cdot) = \mathbf{X}\beta \tag{2}$$

- La *función liga*, función monótona y diferenciable que describe la relación entre la media de la i-ésima observación y su predictor lineal

$$\eta_i = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} \quad i = 1, \ldots, n \tag{3}$$

donde $\mu_i = E(y_i)$.

La función liga canónica se da cuando se cumple que $\theta = \eta$ y en este caso los parámetros desconocidos de la estructura lineal tienen estadísticas suficientes, dadas por:

$$t_n = \left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_{1i} y_i, \ldots, \sum_{i=1}^{n} x_{ki} y_i \right).$$

3.2 Bondad de ajuste

Uno de los criterios de bondad de ajuste más usados al ajustar un MLG es la devianza, dada por:

$$\sum_{i=1}^{n} 2w_i \{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i) \} \phi = D(y; \mu)\phi \tag{4}$$

donde $\tilde{\theta}$ representa el valor del parámetro que se obtiene en el modelo saturado.

Otra medida de discrepancia, que se usa es la X^2 de Pearson generalizada

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})} \tag{5}$$

la distribución asintótica de ambas estadísticas es una χ^2 con n-p grados de libertad. Las distribuciones condicionales de la devianza y la X^2 dadas las estadísticas suficientes, son asintóticamente normales (McCullagh, 1989).

3.3 Generación de muestras *look-alike*

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de la distribución $F(x; \theta)$, θ un vector de parámetros desconocidos, que asumiremos de dimensión k y T_n una estadística suficiente minimal para θ. Sea S una estadística, entonces la distribución condicional de la estadística $G(s|t_n)$, no dependerá del valor verdadero de θ. O'Reilly y Gracia-Medrano (2004), sugieren generar muestras tales que $T_n = t_n$, que ellos llaman *look-alike*, para estimar $G(s|t_n)$.

Lockhart, O'Reilly, y Stephens(1986) proponen el uso del estimador Rao Blackwell y el muestreador de Gibbs para generar estas muestras, en Naranjo et al.(2016) solo utilizamos la estimación Rao-Blackwell.

El concepto de la doble transitividad (ver O'Reilly y Quesenberry, 1973) puede usarse para generar muestras *look-alike*, que cuando se cumple nos dice que si t_m y x_m son conocidas, podemos encontrar t_{m-1} sin conocer explícitamente los valores de $x_1, x_2, \ldots, x_{m-1}$.

Hacemos una transformación de la muestra $x_1, x_2, \ldots, x_n$ y t_n a un nuevo conjunto de n variables, $X_{k+1}, X_{k+2}, \ldots, X_n$, $T_{n1}, T_{n2}, \ldots, T_{nk}$, y se trabaja con la densidad conjunta.

$$f(x_{k+1}, x_{k+2}, \ldots, x_n, t_{n1}, t_{n2}, \ldots, t_{nk}). \tag{6}$$

Una muestra *look-alike* $(x^*_{k+1}, x^*_{k+2}, \ldots, x^*_n)$ estará generada a partir de esta densidad conjunta en el orden de $x^*_n, x^*_{n-1}, \ldots, x^*_{k+1}$, y cuando estos valores sean conocidos, se encontrarán $x^*_1, x^*_2, \ldots, x^*_k$ resolviendo las k ecuaciones para $t_{n1}, t_{n2}, \ldots, t_{nk}$.

3.4 Estimación Rao-Blackwell de $F(x;\theta)$

La estimación de Rao-Blackwell de $F(x;\theta)$, basada en t_n, es $\tilde{F}_n(x|t_n) = P(X_j \leq x|t_n)$, donde x_j es un miembro de una muestra dada.

Sea $t^*_n = t_n$, se genera el primer elemento de la muestra look-alike x^*_n. El segundo elemento de la muestra look-alike, x^*_{n-1}, se genera a partir de la distribución $P(X_{n-1} \leq x|t^*_n, x^*_n)$. Usando la doble transitividad, se puede calcular t^*_{n-1}.

Como x^*_n es independiente de $x_1, x_2, \ldots, x_{n-1}$, y t^*_{n-1}, $P(X_{n-1} \leq x|t^*_n, x^*_n) = P(X_{n-1} \leq x|t^*_{n-1})$ es igual a $\tilde{F}_{n-1}(x|t^*_{n-1})$. Así x^*_{n-1} se genera a partir de $\tilde{F}_{n-1}(x|t^*_{n-1})$. Este proceso continúa hasta que se tenga x^*_{k+1}; entonces se calcula $x^*_1, x^*_2, \ldots, x^*_k$ a partir de las ecuaciones $t_{n1}, t_{n2}, \ldots, t_{nk}$ para completar la muestra.

3.5 Muestras *look alike* para modelos lineales generalizados

Dado que las estadísticas suficientes de los modelos lineales generalizados cumplen con la doble transitividad, para variables explicativas categóricas, vamos a suponer que tenemos una variable categórica con $k+1$ categorías, es posible utilizar el estimador Rao Blackwell para obtener una muestra *look alike*. Esto se hace reescribiendo la variable categórica como k variables indicadoras. Notando que la distribución

exponencial se puede reescribir como

$$
\begin{aligned}
& f_Y(y_i; \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) \\
&= \exp\left\{\left[y_i\theta_i - b(\theta_i)\right]/a(\phi) + c(y_i,\phi)\right\} \\
&= \exp\left\{\left[y_i\boldsymbol{x}_i'\boldsymbol{\beta} - b(\boldsymbol{x}_i'\boldsymbol{\beta})\right]/a(\phi) + c(y_i,\phi)\right\} \\
&= \begin{cases} \exp\left\{\left[(\beta_h+\beta_0)\,y_i - b(\beta_h+\beta_0)\right]/a(\phi) + c(y_i,\phi)\right\} & \text{si la categoría } h=1,\ldots,k \\ \exp\left\{\left[\beta_0 y_i - b(\beta_0)\right]/a(\phi) + c(y_i,\phi)\right\} & \text{si la categoría } h=k+1 \end{cases} \\
&= \exp\left\{\left[\beta_h^* y_i - b(\beta_h^*)\right]/a(\phi) + c(y_i,\phi)\right\},
\end{aligned}
$$

donde $\beta_h^* = \beta_h + \beta_0$ para $h = 1,\ldots,k$, y $\beta_h^* = \beta_0$ para $h = k+1$.

Si definimos

$$
z_{ih} = \begin{cases} x_{ih}\prod_{l\neq h}(1-x_{il}) & \text{if } h = 1,\ldots,k, \\ \prod_{l=1}^{k}(1-x_{il}) & \text{si } h = k+1. \end{cases}
$$

Estas variables permiten identificar la categoría de y_i, es decir, si y_i esta en la categoría h entonces $z_{ih} = 1$ y $z_{il} = 0$ para $l \neq h$; y si

$$
t_{nh}^z = \sum_{i=1}^{n} z_{ih} y_i,
$$

En este caso $t_{n1}^z, \ldots, t_{nk}^z, t_{n,k+1}^z$ corresponden a las estadísticas suficiente para las β_h^*. El super índice es para enfatizar que t_{nh}^z depende de las variables z_{ih}'s mientras que t_{nh} depende de las variables x_{ih}'s.

Es fácil ver que t_n^z es función de t_n ya que

$$
t_{nh}^z = \begin{cases} t_{nh} & \text{si } h = 1,\ldots,k, \\ t_{n,k+1} - \sum_{l=1}^{k} t_{nl} & \text{si } h = k+1, \end{cases}
$$

y $t_{n,k+1} = \sum_{l=1}^{k+1} t_{nl}^z$, por lo tanto $t_n^z = \left(t_{n1}^z, \ldots, t_{nk}^z, t_{n,k+1}^z\right)$ es también es una estadística suficiente (mínimal).

Más aún, las estadísticas t_{nh}^z's tienen una distribución que pertence a la familia exponencial de dispersión, con liga canónica $\theta_{T_{nh}^z} = \beta_h^*$, parámetro de dispersión $\phi_{T_{nh}^z} = \phi$, función cumulante $b_{T_{nh}^z}(\theta_{T_{nh}^z}) = \sum_{i=1}^{n} z_{ih} b(\theta)$, y funciones $a_{T_{nh}^z}(\phi_{T_{nh}^z}) = a(\phi)$ y $c_{T_{nh}^z}(t_{nh}^z, \phi_{T_{nh}^z})$.

La tabla 1 resume las propiedades de estas familia de dispersión exponencial.

3.6 Método Rao-Blackwell en modelos lineales generalizados

Para generar las muestras look-alike $y_1^*, \ldots, y_n^*$ es necesario calcular la distribución condicional $\widetilde{F}_j(y) = \widetilde{F}_j(y|t_j)$. La distribución está dada por $\widetilde{F}_j(y|t_j) = \int_{A_y} f_{Y_j}(w|T_j^z = t_j)\mathrm{d}w$, que corresponde a la integral de

NEF-QVF	$f_{Y_j}(w\|T_j^z = t_j)$
	$= \exp\left\{c_{Y_j}(w,\phi) + c_{T_{jh}^z - Y_j}\left(t_{jh} - w, \phi\right) - c_{T_{jh}^z}\left(t_{jh}, \phi\right)\right\}$
Normal	$w \sim \text{Normal}\left(t_{jh}\frac{1}{\sum_{i=1}^{j-1} z_{ih}+1}\ ,\ \sigma^2\frac{\sum_{i=1}^{j-1} z_{ih}}{\sum_{i=1}^{j-1} z_{ih}+1}\right)$
	σ^2 conocida
Gamma	$w/t_{jh} \sim \text{Beta}\left(\lambda\ ,\ \lambda\sum_{i=1}^{j-1} z_{ih}\right)$ λ conocida
Poisson	$w \sim \text{Binomial}\left(t_{jh}\ ,\ \frac{1}{\sum_{i=1}^{j-1} z_{ih}+1}\right)$
Binomial	$w \sim \text{Hipergeometrica}\left(\sum_{i=1}^{j-1} z_{ih}m_i + m_j\ ,\ t_{jh}\ ,\ m_j\right)$
	$m_1, \ldots, m_j$ fija
Binomial Negativa	$w \sim \text{Beta-Binomial}\left(\lambda\ ,\ \lambda\sum_{i=1}^{j-1} z_{ih}\ ,\ t_{jh}\right)$
	λ conocida

Cuadro 1. Distribuciones para usar el estimador Rao Blackwell en las muestras de los MLG.

Lebesgue evaluada in $A_y = \{w \in \mathcal{Y} : w \leq y\}$, donde $\mathcal{Y}$ es el soporte de la distribución de Y_j, la distribución condicional de Y_j dado T_j^z esta dada por

$$f_{Y_j}(w|T_j^z = t_j) = \sum_{h=1}^{k+1} \exp\left\{c_{Y_j}(w,\phi) + c_{T_{jh}^z - Y_j}\left(t_{jh} - w, \phi\right) - c_{T_{jh}^z}\left(t_{jh}, \phi\right)\right\} \times \mathbf{1}_{\{\text{categoría } h\}}(y_j). \tag{7}$$

La tabla 1 muestra las distribuciones para cinco distribuciones pertenecientes a la familia de dispersión exponencial.

Sin pérdida de generalidad, y con la intención de ser claros, suponemos que las categorías de respuesta de las primeras $k+1$ observaciones, $y_1, y_2, \ldots, y_k, y_{k+1}$, corresponden a las $h = 1, 2, \ldots, k, k+1$, respectivamente.

El primer paso es, y_n^* se genera de $\widetilde{F}_n^{-1}$, y t_{n-1}^* se calcula de t_n y y_n^*.

En el siguiente paso, y_{n-1}^* se genera de $\widetilde{F}_{n-1}^{-1}$, es decir de

$$f_{Y_{n-1}}(w|T_{n-1}^z = t_{n-1}^*) = \exp\left\{c_{Y_{n-1}}(w,\phi) + c_{T_{n-1,h}^z - Y_{n-1}}\left(t_{n-1,h}^* - w, \phi\right) - c_{T_{n-1,h}^z}\left(t_{n-1,h}^*, \phi\right)\right\}.$$

y t_{n-2}^* se calcula de t_{n-1}^* and y_{n-1}^*.

De esta manera se continua hasta que $y^*_{\nu+1}$ se genera de $\widetilde{F}^{-1}_{\nu+1}$,

$$f_{Y_{\nu+1}}(w|T^z_{\nu+1} = t^*_{\nu+1})$$
$$= \exp\left\{c_{Y_{\nu+1}}(w,\phi) + c_{T^z_{\nu+1,h}-Y_{\nu+1}}\left(t^*_{\nu+1,h} - w, \phi\right) - c_{T^z_{\nu+1,h}}\left(t^*_{\nu+1,h}, \phi\right)\right\}$$

y t^*_ν se calcula de $t^*_{\nu+1}$ y $y^*_{\nu+1}$.

Finalmente, los $\nu = k+1$ términos restantes $y^*_1, \ldots, y^*_\nu$ se obtiene de la solución única al resolver $T^z_n(y_1, \ldots, y_n) = T^z_n(y^*_1, \ldots, y^*_n)$.

Para su uso en bondad de ajuste lo que se hace es repetir este procedimiento M veces y en cada muestra calcular la estadística de Bondad de ajuste para la que se desea obtener su distribución exacta, utilizamos esta distribución empírica y localizamos el punto en el que se encuentra el valor de la estadística calculado con la muestra original, es decir el $p-valor$ estará dado por el número de valores de la estadística mayores al valor de la muestra original entre M.

En Naranjo et al. (2006) se presentan varios ejemplos con datos simulados para diferentes distribuciones y diferentes combinaciones de n y k y una aplicación con datos reales. Finalmente cuando las variables explicativas son contínuas se podría utilizar el muestreador de Gibbs para encontrar las muestras *look alike*, este también es un trabajo pendiente.

Bibliografía

[1] R. A. Fisher, «Inverse probability», en *Proceedings of the Cambridge Philosofical Society*, núm. 26, 1930, 528–535.

[2] N. L. Johnson, S. Kotz y W. K. Kemp, *Univariate Discrete Distributions*, Wiley, 1992.

[3] R. A. Lockhart, F. J. O'Reilly y M. A. Stephens, «Tests of fit based on normalized spacings», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 48, núm. 3, 1986, 344–352.

[4] P. McCullagh, «The conditional distribution of goodness-of-fit statistics for discrete data», *Journal of the American Statistical Association*, vol. 81, núm. 393, 1986, 104–107.

[5] E. Nájera, F. O'Reilly y S. Ruiz-Velasco, «Fiducial distribution in a power series family», *Communications in Statistics - Theory and Methods*, vol. 48, núm. 23, 2019, 5769–5808.

[6] L. Naranjo Albarrán, F. O'Reilly Togno y S. Ruiz-Velasco Acosta, *Muestras look-alike para modelos lineales generalizados y su uso en bondad de ajuste: la familia NEF-QVF*, Preimpresos 166, IIMAS, UNAM, 2016.

[7] J. Nelder y R. Wedderburn, «Generalized linear models», *Journal of the Royal Statistical Society. Series A*, vol. 135, núm. 3, 1972, 370–384.

[8] F. O'Reilly y L. Gracia-Medrano, «On the conditional distribution of goodness-of-fit tests», *Communications in Statistics. Theory and Methods*, núm. 35, 2006, 541–549.

[9] F. J. O'Reilly y C. Quesenberry, «The conditional probability integral transformation and applications to obtain composite chi-square goodness-of-fit tests», *The Annals of Statistics*, vol. 1, núm. 1, 1973, 74–83.