Miscelánea Matemática **73** (2021) 21-29 SMM

DOI: https://doi.org/10.47234/mm.7302

Bondad de ajuste y las *look-alikes samples*

Leticia Gracia-Medrano Valdelamar
Instituto de Investigaciones en
Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
lety@sigma.iimas.unam.mx

1. Introducción

Con mucho gusto participo con esta contribución, aquí presento el trabajo en el tema de Bondad de Ajuste que realicé al lado de mi querido profesor, el Dr. Federico O'Reilly.

De manera muy coloquial las pruebas de bondad de ajuste son usadas para verificar que una muestra proviene de una cierta distribución. El tema de bondad de ajuste es amplio y conviene considerar por un lado si se trata probar una hipótesis simple o una compuesta; y por otro si se quiere probar un modelo de distribución para una variable continua o una discreta. Este trabajo se enfoca principalmente en una prueba de bondad de ajuste para distribuciones continuas y para el caso de hipótesis compuestas.

En el caso de distribuciones **continuas** el caso **simple**, la hipótesis nula correspondería a una distribución completamente especificada $H_0 : F = F_0$ y en el caso **compuesto** la hipótesis nula sería $H_0 : F \in F_0$, siendo F_0, una clase paramétrica con parámetro θ (de dimensión k).

Prácticamente todas las soluciones para el caso compuesto son adaptaciones de soluciones para el caso simple, en el que los parámetros son conocidos, sustituyéndolos por estimaciones hechas con la misma muestra, conocidos como métodos ***plug-in***.

Cualquier prueba de bondad de ajuste requiere de la disponibilidad de tablas o de algún método para evaluar (a través del uso de una α preespecificada y el correspondiente valor crítico, o del cálculo de la significancia) si existe o no evidencia para rechazar la hipótesis nula.

Palabras clave: Bondad de ajuste, *look alike samples*.

Existe literatura abundante relacionada con pruebas que utilizan la función de distribución empírica (pruebas EDF, presentadas más adelante), donde también presentan fórmulas correctivas para distintos valores de n intentando que los valores críticos asintóticos sean buenas aproximaciones [1, Cap. 4].

Existe otro enfoque diferente al de usar tablas basado en simulaciones. Varios artículos proponen el uso de algoritmos que simulan la distribución de la estadística de prueba, usando el llamado *bootstrap* paramétrico. Como ejemplos están [6] Henze y Klar que lo hacen para un caso continuo y en [5] Gurtler y Henze que se estudia uno discreto. En estos trabajos se hace uso de un límite basado en dos rubros, el primero corresponde a que el número de muestras *bootstrap* sea grande y el segundo a que el tamaño de muestra n, también sea grande y con ello logran que el valor crítico proveniente de la simulación sea parecido al asintótico.

En el caso compuesto, es importante señalar que la distribución de la estadística de bondad de ajuste varía, en general, al cambiar de familia nula, y en algunos casos, la distribución depende también de los parámetros. Este es el caso de la Gamma y de la Gaussiana-inversa, entre otras (véase [8] Lockhart y Stephens para la Gamma y [11] O'Reilly y Rueda para la Gaussiana-inversa).

En bondad de ajuste cuando los parámetros son considerados solo como ruido, el uso de la distribución condicionando con la estadística suficiente minimal de los parámetros (conocido como estimador Rao Blackwell), aparece como una opción muy apropiada. En 1970 Srinivasan [12] fue el primero en hacer uso del estimador Rao-Blackwell en lugar de la distribución *plug-in*, al estudiar la estadística de Kolmogorov-Smirnov. Un poco más adelante [9] Moore (1973), demostró que el proceso empírico derivado al usar la distribución *plug-in* o la Rao-Blackwell tienen el mismo límite para las distribuciones exponencial negativa, normal y uniforme. En [7] R. Lockhart, F. O'Reilly (2005) se demuestra que esto es cierto para la familia exponencial.

Si los procesos empíricos con la distribución Rao-Blackwell y la *plug-in* llevan a lo mismo, ¿qué ventaja habrá del uso de una sobre la otra?

En un contexto bastante general, la ventaja está en que el estimador Rao-Blackwell permite generar muestras con las mismas propiedades distribucionales de la muestra observada si la hipótesis nula es cierta. En la sección 3 se verá que estas muestras son condicionalmente independientes e idénticamente distribuidas y son referidas como *look alike samples*.

En la sección 4 se utiliza a la distribución Gaussiana-inversa para ilustrar el procedimiento y se presenta un breve estudio de simulación

donde se compara la potencia de este método con las potencias obtenidas en [4] Gracia-Medrano y O'Reilly (2004), se puede ver que hay un incremento en ellas al utilizar el procedimiento condicional.

Finalmente en la sección 5 se hacen algunos comentarios generales.

2. Estadísticas EDF para pruebas de bondad de ajuste

El problema de bondad de ajuste se puede plantear como sigue:

Sean $X_1, X_2, \ldots, X_n$ *vaiid* cuya función de distribución es F y se desea contrastar las siguientes hipótesis:

$$H_o : F \in \mathcal{F}_o$$
$$H_1 : F \notin \mathcal{F}_o \text{ o bién } F \in \mathcal{F}' - \mathcal{F}_o$$

donde $\mathcal{F}_o$ es una clase de funciones de probabilidad paramétricas, es decir $\mathcal{F}_o = \{F_\theta : \theta \in \Omega\}$. Por ejemplo $\mathcal{F}_o$ podría ser la familia exponencial y $\mathcal{F}'$ sería la familia de todas las distribuciones continuas.

Las estadísticas propuestas con mayor respaldo teórico utilizan distancias o disimilitudes entre F_0 y F_n; siendo esta última la función de distribución empírica, $F_n(x) = \frac{\#X_i's \leq x}{n}$.

Entre las estadísticas más conocidas están:

- la Kolmogorov-Smirnov denotada D_n,
$$D_n = \sup_x \sqrt{n}|F_n(x) - F_0(x)|.$$
- la Cramér-von Mises, denotada ${W_n}^2$,
$${W_n}^2 = n \int_R (F_n(x) - F_0(x))^2 dF_0(x),$$
- la Anderson-Darling, denotada ${A_n}^2$,
$${A_n}^2 = n \int_R (F_n(x) - F_0(x))^2 w(x) dF_0(x),$$
con
$$w(x) = \frac{1}{F_0(x)(1 - F_0(x))},$$

En la literatura estas estadísticas se conocen como EDF (*Empirical Distribution Function*) y resultan ser funcionales del proceso empírico:

$$\xi_n(x) = \sqrt{n}\{F_n(x) - F_0(x)\},$$

Véase [1] para más detalles de estas estadísticas.

Para el caso **simple** se encontró que la distribución asintótica de $\xi_n(x)$, está relacionada con un puente Browniano[1].

[1] movimiento Browniano $\{W(t), t \in [0,1] | W(1) = 0\}$

Para el caso **compuesto**, el método *plug-in* sugiere reemplazar a θ con $\hat{\theta}$, esto cambia a estudiar el proceso empírico con parámetros estimados:

$$\hat{\xi}_n(x) = \sqrt{n}\{F_n(x) - F(x, \hat{\theta})\},$$

Cuando se trabaja con parámetros estimados, bajo ciertas condiciones en el tipo de estimador utilizado (esencialmente que sea eficiente y asintóticamente normal, los llamados *BAN Best Asymptotically Normal*), el proceso empírico converge también a un proceso Gaussiano con media nula y una función de varianza-covarianza, que depende de la familia de distribuciones de que se trate (la establecida en la hipótesis nula) y también puede depender de los valores de los parámetros.

3. Generación de muestras semejantes o *look-alikes*

A continuación algo de notación y precisiones.

Denotando como $\tilde{F}_n(x)$ al estimador Rao-Blackwell de $F(x, \theta)$:

$$\tilde{F}_n(x) = P(X_i \leq x | T_n)$$

donde T_n es la estadística suficiente minimal.

En algunos casos T_n resulta ser **doble transitiva** esto es: que el conocer T_n y X_n es equivalente a conocer T_{n-1} y X_n. Por ejemplo si $T_n = \bar{X} = t$ y conozco $X_1, \ldots, X_n$ entonces $T_{n-1} = t - (X_n/n)$ y viceversa.

Aquí cabe hacer una observación, hay una redundancia al hablar de la distribución condicional de la muestra completa dada T_n.

Si la distribución tiene k parámetros, usualmente la dimensión de T_n es k. En realidad se identifica la distribución condicional de $n - k$ elementos de la muestra y los otros k elementos se derivan resolviendo un sistema de ecuaciones para que se cumpla que $T_n = t$ usando la doble transitividad.

En el siguiente teorema se verá que cuando $x_1, x_2, \ldots, x_n$ es una realización de $X_1, X_2, \ldots, X_n$, se puede generar una muestra $x_1^*, x_2^*, \ldots, x_n^*$ que resulta ser condicionalmente independiente dada $t_n = T(x_1, x_2, \ldots, x_n)$ y que además $t_n = T(x_1^*, x_2^*, \ldots, x_n^*)$. Por esto se le conoce como muestra semejante o *look-alike*.

Por facilidad de notación, los últimos $n - k$ elementos de la muestra x^* son los que se generarán primero en el teorema.

Teorema 3.1. *Bajo doble transitividad de la estadística suficiente T_n y cuando el número máximo de elementos de la muestra para el que no se tienen redundancias en la distribución condicional dada T_n es $n-k$, el procedimiento para obtener una muestra semejante x^* es el siguiente:*

- *Genero u_n, una realización de una variable aleatoria $U(0,1)$, defino $x_n^* = \tilde{F}^{-1}(u_n)$ (la inversa de la Rao-Blackwell) y recalculo t_{n-1} a partir de t_n y de x_n^* y la llamo t_{n-1}^*.*
- *Genero otra u_{n-1} independiente de u_n con $\tilde{F}_{n-1}^{-1}$ (la Rao-Blackwell condicionando con t_{n-1}^*) hago $x_{n-1}^* = F_{n-1}^{-1}(u_{n-1})$ y recalculo t_{n-2}^* a partir de t_{n-1}^* y x_{n-1}^* y continúo así hasta calcular x_{k+1}^* a partir de u_{k+1}*
- *Estos $n-k$ elementos de la muestra x^* forman una realización de la distribución condicional para cualesquiera $n-k$ elementos de la muestra original x dada t_n.*
- *Los k elementos restantes $x_1^*, x_2^*, \ldots, x_k^*$ se encuentran como la solución derivada de que $T_n(x_1^*, x_2^*, \ldots, x_n^*) = t_n$*

La prueba se obtiene usando el resultado de que para una T_n doblemente transitiva, la distribución condicional de X_r dadas $T_n, X_n, X_{n-1}, \ldots, X_{r+1}$ es la misma que la condicional de X_r dada T_r. Este resultado es el teorema 2.4 en [10] O'Reilly y Quesenberry (1973).

Este procedimiento es una aplicación secuencial en el contexto condicional del muy conocido resultado: Si G es una función acumulativa de distribución entonces $X^* = G^{-1}(U)$ se distribuye como G.

Ahora para hacer una prueba de bondad de ajuste para una muestra en particular, los pasos a seguir serían: primero se calcula sus t_n y A^2 de Anderson-Darling, llamándola A_o^2, enseguida se generan 1000 (o más) muestras semejantes siguiendo los pasos del teorema, se calculan las A^{*2} de cada una de estas 1000 muestras, se ordenan y finalmente se compara A_o^2 versus el porcentil $1-\alpha$ de las A^{*2}, si A_o^2 es mayor al porcentil se rechaza la hipótesis nula.

4. Ejemplo con la gaussiana inversa

La función de distribución es:

$$F(x;\mu,\lambda) = \Phi(R) + \Phi(L)\exp\left\{\frac{2\lambda}{\mu}\right\},$$

donde $R = -\left(\frac{\lambda}{x}\right)^{\frac{1}{2}} + \frac{(\lambda x)^{\frac{1}{2}}}{\mu}$, $L = -\left(\frac{\lambda}{x}\right)^{\frac{1}{2}} - \frac{(\lambda x)^{\frac{1}{2}}}{\mu}$ y Φ es función de distribución normal.

El estimador Rao-Blackwell es:

$$\tilde{F}_n(x) = \begin{cases} 0 & \text{si } x < l \\ 1 & \text{si } x > u \\ G_{n-2}(W) + \frac{n-2}{n}\left[1 + \frac{4(n-1)\hat{\lambda}}{n^2\hat{\mu}}\right]^{\frac{(n-3)}{2}} G_{n-2}(-W') & \text{en otro lado} \end{cases}$$

Donde $\hat{\mu}$ y $\hat{\lambda}$ son los estimadores máximo verosímiles, ambos funciones de la estadística suficiente minimal $T_n = (\sum_{i=1}^n X_i, \sum_{i=1}^n (\frac{1}{X_i}))$ que claramente cumple la doble transitividad.
G_{n-2} es una t de Student con $n-2$ grados de libertad.

Esta $\tilde{F}_n(x)$ esta bien definida para $n = 3$ $(k = 2)$

$$W = \frac{1}{C}\sqrt{n(n-2)}(\frac{x}{\hat{\mu}} - 1), \quad W' = \frac{1}{C}\sqrt{n(n-2)}(1 + \frac{n-2}{n}\frac{x}{\hat{\mu}}),$$

$$C = \sqrt{\frac{n}{\hat{\lambda}}(n - \frac{x}{\hat{\mu}})x - n(1 - \frac{x}{\hat{\mu}})^2},$$

l y u determinan el intervalo para las x donde C es no negativo.

Para el primer paso de la generación de las muestras semejantes, se calcula $t_n = (\sum_{i=1}^n x_i, \sum_{i=1}^n (\frac{1}{x_i})) = (t_1^n, t_2^n)$, para este ejemplo no hay una expresión explícita para $\tilde{F}^{-1}$ esta se invierte de manera numérica; y finalmente se utiliza un generador de variables uniformes, por ejemplo en lenguaje R sería $u_n <$ −runif(1).

Entonces se hace $x_n^* = \tilde{F}^{-1}(u_n)$ y $t_{n-1}^* = (t_1^n - x_n^*, t_2^n - (\frac{1}{x_n^*})) = (t_1^{(n-1)}, t_2^{(n-1)})$

Para el paso dos $u_{n-1} <$ −runif(1) (de manera automática los lenguajes cambian de semilla y esta uniforme resulta independiente de la anterior), $x_{n-1}^* = \tilde{F}^{-1}(u_{n-1})$, $t_{n-2}^* = (t_1^{(n-1)} - x_{n-1}^*, t_2^{(n-1)} - (\frac{1}{x_{n-1}^*}))$.

Y así sucesivamente hasta el paso $n-2$, las fórmulas son muy sencillas y al final se requiere que x_2^* y x_1^* satisfagan

$$\begin{aligned} t_n^* &= \left(\left[\sum_{i=3}^n x_i^* \right] + x_2^* + x_1^*, \left[\sum_{i=3}^n (\frac{1}{x_i^*}) \right] + \frac{1}{x_2^*} + \frac{1}{x_1^*} \right) \\ &= \left(t_1^{(3)} + x_2^* + x_1^*, t_2^{(3)} + \frac{1}{x_2^*} + \frac{1}{x_1^*} \right) \\ &= (t_1^n, t_2^n) \end{aligned}$$

Este sistema lleva a encontrar las raíces de una ecuación de segundo grado que resultan ser x_2^* y x_1^* que completan la muestra.

4.1 Simulación

En la primera parte se comprobó para el caso con $n = 20$ que las muestras semejantes tuviesen precisamente la misma distribución condicional dada T_n .

Se generaron 1000 muestras provenientes de una Gaussiana inversa con parámetros fijos y para cada una de las muestras se hizo lo siguiente:

1. se calculó t_n y su estadística A^2 Anderson-Darling, llamándola A_o^2.

2. usando esta t_n se generan 1000 *muestras semejantes* y para cada una de ellas se calcularon las estadísticas A^{*2} Anderson-Darling.
3. para determinar **la zona de rechazo** se ordenaron de menor a mayor las 1000 A^{*2} y se obtuvo el porcentil $1 - \alpha = 0.95$.
4. se comparó A_o^2 *versus* el porcentil calculado en el punto anterior; y se anotó si A_o^2 fue mayor a ese porcentil.

En esta comprobación se hicieron entonces un millón de simulaciones y se obtuvo que solo 42 de las 1000 muestras Gaussianas inversas calculadas el valor A_o^2 excedió el porcentil 0.95 de las A^{*2}, es decir en 42 casos fue rechazada la hipótesis de que provenía de una distribución Gaussiana inversa, coincidiendo con lo esperado.

La segunda parte fue un estudio de potencia. En la tabla 1 se presenta la potencia de la prueba usando tamaño de muestra $n = 20$. Se utilizaron cuatro distribuciones alternativas y en el último renglón aparece la distribución Gaussina Inversa. La A^2 semejante se compara frente la estadística presentada en [11] O'Reilly & Rueda (1992). Para las cuatro distribuciones alternativas la A^2 semejante tiene mejor potencia.

Tabla 1: Comparación de potencias.

Proporción de rechazos de 1000. $\alpha = .05$

Alternativa	A^2 semejante	A^2 O&R
Exponential escala = 1	.643	.62
Lognormal de $n(1/2, 1)$	.131	.10
Uniforme (0, 1)	.863	.85
Weibull escala = 1 forma= 2	.406	.38
I-G ($\mu = 1, \lambda = 8$)	.042	.05

Para A^2 «semejante» en cada simulación el valor crítico se encuentra con 1000 muestras «semejantes».
A^2 O&R usa el valor asintótico de O'Reilly & Rueda (1992),
2da columna tomada de Gracia-Medrano & O'Reilly (2004)

5. Comentarios

Los resultados de la sección 3 son válidos para distribuciones **continuas y discretas**, aunque para estas últimas hay una restricción respecto a la exactitud si se procede con nivel fijo α. Esto debido a lo discreto de la estadística T_n. No hay sin embargo inexactitud si se evalúa la significancia. En el artículo de [3] González-Barrios *et al* (2006) se presenta una prueba de bondad de ajuste usando la distribución condicional para el caso discreto.

En el caso de distribuciones continuas con parámetros de localización y/o escala existe este resultado de [2] Eaton (1983): Si T_n es equivariante suficiente y la estadística EDF es invariante a cambios de localización y/o escala entonces T_n y la estadística EDF son independientes. Se podría hacer la simulación condicional pero no es necesario pues es la misma que la incondicional. El proceso se simplifica mucho, solo se necesita generar muchas muestras y forzarlas a que den los valores del parámetro estimado, esto es $T_n^* = T_n$.

Finalmente un ejemplo de cómo la prueba EDF usando el estimador Rao Blackwell recupera un resultado muy conocido. Al probar la distribución $U(0,\theta)$ o la $U(\theta_1,\theta_2)$, el proceso empírico asociado es no nulo solo si el argumento x del proceso, está en $(0, X_{(n)})$, en el primer caso y para el caso con dos parámetros, solo si x está en $(X_{(1)}, X_{(n)})$, al condicionar con T_n estos procesos son proporcionales al proceso cuando se prueba el caso simple $U(0,1)$. Este resultado implica no hacer más cálculos pues se convierte en una prueba de bondad de ajuste con hipótesis nula simple, para la cual ya existen tablas para la estadísticas EDF más utilizadas, desde luego habiendo hecho antes para el primer caso la transformación a las nuevas $n-1$ observaciones $Y_{(i)} = X_{(i)}/X_{(n)}$ con $i = 1, \ldots, n-1$; y para el segundo caso a las $n-2$ transformadas, $Y_{(i)} = (X_{(i)} - X_{(1)})/(X_{(n)} - X_{(1)})$ con $i = 2, \ldots, n-1$.

Bibliografía

[1] R. D'Agostino y M. Stephens, *Goodness-of-Fit Techniques*, Marcel Dekker, 1986.

[2] M. Eaton, *Multivariate Statistics*, New York: John Wiley, 1983.

[3] J. González-Barrios, F. O'Reilly y R. Rueda, «Goodness of fit for discrete random variables using the conditional density», *Metrika*, vol. 64, núm. 1, 2006, 77–94.

[4] L. Gracia-Medrano y F. O'Reilly, «Transformations for testing the fit of the inverse-Gaussian distribution», *Commun. Statist. Theor. Meth.*, vol. 33, núm. 4, 2004, 919–924.

[5] N. Gurtler y N. Henze, «Recent and Classical goodness of fit tests for the Poisson distribution», *Journal of Statistical Planning and Inference*, núm. 90, 2000, 207–225.

[6] N. Henze y B. Klar, «Goodness of fit tests for the inverse Gaussian distribution based on the empirical Laplace transform», *Annals of the Institute of Statistical Mathematics*, vol. 54, núm. 2, 2002, 225–444.

[7] R. Lockhart y F. O'Reilly, «A note on Moore's conjecture», *Statistics & Probability Letters*, núm. 74, 2005, 212–220.

[8] R. Lockhart y M. Stephens, *Goodness-of-fit for the gamma distribution*, Technical Report, Department of Mathematics and Statistics, Simon Fraser University, 1985.

[9] D. Moore, «A note on Srinivasan's goodness-of-fit test», *Biometrika*, núm. 60, 1973, 209–211.

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

[11] F. O'Reilly y R. Rueda, «Goodness of fit for the inverse Gaussian distribution», *Canadian Journal Statistics*, vol. 20, núm. 4, 1992, 387–397.

[12] R. Srinivasan, «An approach to testing goodness of fit of incompletely specified distributions», *Biometrika*, vol. 57, núm. 3, 1970, 605–611.