Una prueba de ajuste χ^2 basada en Procesos Empíricos Transformados

Jorge Graneri *

abril de 1997

Resumen

En el presente trabajo se propone una prueba de ajuste χ^2 en la que el proceso empírico clásico es sustituído por un proceso empírico transformado. Se implementa la prueba en dos casos particulares y se compara su rendimiento con el de la prueba χ^2 clásica, obteniendose en ambos casos una mejora [1].

1 Introducción : Procesos Empíricos Transformados

A.Cabaña y E.M.Cabaña dan en [1] un método para llevar el proceso empírico de una muestra de variables i.i.d. en un proceso convergente en ley a un V-proceso de Wiener, siendo V una medida elegida apropiadamente. Además del interés que esto tiene en sí mismo, presenta una importante ventaja para la estadística, ya que permite trabajar con objetos que convergen a procesos de incrementos independientes. De esta forma puede pensarse en " hacer estadística a partir de Procesos Empíricos Transformados", tal como se hace a partir de los procesos empíricos clásicos. Por otra parte, la transformación propuesta en [1] no es única, lo cual nos permite elegir la más conveniente de acuerdo al problema planteado. El Proceso Empírico Transformado de una muestra $X_1, X_2, \ldots, X_n$ de variables i.i.d.(con función de distribución F) asociado a la función de distribución F_0, a la isometría T en $L_2(I\!R, dF_0)$, con rango ortonormal a la función constante 1 y a la *"función de pesos"* a con $\|a\|^2 = \int a^2(x)dF_0(x) = 1$, fue introducido en [1] y se define como :

*Centro de Matemática, Universidad de la República, Eduardo Acevedo 1139, 11200 Montevideo, Uruguay

[1]Este trabajo me fue propuesto por mi orientador: el Prof. Enrique Cabaña y es gracias a su generosidad y a su paciente guía que pude terminarlo.

$$w_n^{a,\mathcal{T}}(A) = \int \mathcal{T}(a\mathbb{I}_A)db_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{T}(a\mathbb{I}_A)(X_i) \tag{1}$$

donde $\mathbb{I}_A$ es la función indicatriz de A, $b_n(x) = \sqrt{n}(F_n(x) - F_0(x))$ es el proceso empírico y F_n es la función de distribución empírica de la muestra $X_1, X_2, \ldots, X_n$.
Si $F = F_0$, cuando n tiende a infinito, el proceso empírico b_n converge en ley a un F_0-puente browniano, mientras que el Proceso Empírico Transformado $w_n^{a,\mathcal{T}}(A)$ tiende en ley, bajo ciertas hipótesis, a un V-proceso de Wiener (ver [1]), esto es : proceso w^V gaussiano tal que :

$$E(w^V(A)) = 0 \ , \ \forall A \in \mathcal{B} \tag{2}$$

y

$$E(w^V(A)w^V(B)) = V(A \cap B) \ , \ \forall A, B \in \mathcal{B} \tag{3}$$

donde $\mathcal{B}$ es la $\sigma-$álgebra de Borel de $\mathbb{R}$ y V es una medida con densidad a^2 con respecto a F_0, o sea que para el conjunto A : $V(A) = \int_A a^2(x)dF_0(x)$.
Supongamos que queremos someter a prueba la hipótesis $\mathcal{H}_0 : F = F_0$ contra la sucesión de alternativas $\mathcal{H}_n : F = F^{(\delta/\sqrt{n})}$ contigua a $\mathcal{H}_0$, (ver [2]), con densidad $f^{(\delta/\sqrt{n})}$ con respecto a F_0 . Vamos a suponer que existe una función $k \in L^2(\mathbb{R}, dF_0)$ tal que :

$$\lim_{n\to+\infty} \left\| \frac{\sqrt{n}}{\delta}\left(\sqrt{f^{(\delta/\sqrt{n})}} - 1\right) - \frac{k}{2}\right\|_{L^2} = 0 \ , \ \delta \neq 0 \tag{4}$$

Esta última ecuación implica

$$\int k(x)dF_0(x) = 0 \tag{5}$$

y k es el límite en $L^1(\mathbb{R}, dF_0)$ de $\sqrt{n}/\delta(f^{(\delta/\sqrt{n})} - 1)$, cuando $n \to \infty$.
Bajo $\mathcal{H}_n$, $b_n(A)$ converge en distribución a un F_0-puente b^{F_0} más el sesgo determinístico $\delta \int_A k(x)dF_0(x)$ (ver [3]). Mientras que $w_n^{(a,\mathcal{T})}$ converge, bajo $\mathcal{H}_n$ en las condiciones mencionadas arriba, a un V-Wiener w^V más el sesgo determinístico $\delta \int k(x)\mathcal{T}(a\mathbb{I}_A)(x)dF_0(x)$. La propuesta del presente trabajo es, considerar una prueba de ajuste χ^2, en que los Procesos Empíricos Transformados reemplazan al proceso empírico clásico, y se eligen de modo de aumentar la potencia, para una sucesión dada de alternativas contíguas. Para diseñar la prueba debe introducirse una partición del espacio (en nuestro caso $\mathbb{R}$) de la forma $A_1, A_2, \ldots, A_m$. Siguiendo recomendaciones folcklóricas de la estadística

práctica tomaremos regiones equiprobables bajo la medida V (que será convenientemente elegida según las hipótesis de la prueba). Supongamos además, que las regiones son intervalos consecutivos de la recta. Consideremos el vector aleatorio

$$(w_n^{a,\mathcal{T}}(A_1), w_n^{a,\mathcal{T}}(A_2), \ldots, w_n^{a,\mathcal{T}}(A_m))$$

En virtud de lo anterior, bajo $\mathcal{H}_0$, este vector converge, cuando $n \to +\infty$ a una normal multivariada de media $(0, \ldots, 0)$ y varianza

$$\begin{pmatrix} \frac{1}{m} & 0 & \cdots & 0 \\ 0 & \frac{1}{m} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{m} \end{pmatrix}$$

Por lo tanto $m\sum_{i=1}^{m}\left(w_n^{a,\mathcal{T}}(A_i)\right)^2$ tiene, en el límite, distribución χ^2 con m grados de libertad. Mientras que, bajo la sucesión contigua de alternativas $\mathcal{H}_n : F = F^{(\delta/\sqrt{n})}$ el sesgo será :

$$\begin{aligned} m\delta^2\sum_{i=1}^{m}\left(\int k\mathcal{T}(a\mathbb{I}_{A_i})dF_0\right)^2 &= m\delta^2\sum_{i=1}^{m}\left(\int_{A_i} a\mathcal{T}^{-1}k dF_0\right)^2 \\ m\delta^2\sum_{i=1}^{m}\left(\int_{A_i}\frac{\mathcal{T}^{-1}k}{a}dV\right)^2 &= m\delta^2\sum_{i=1}^{m}\left(\int_{A_i} h dV\right)^2 \end{aligned} \tag{6}$$

donde $h = \frac{\mathcal{T}^{-1}k}{a}$. Esto sugiere una prueba, basada en la distribución límite del Proceso Empírico Transformado, en la que el sesgo sea máximo, para maximizar la potencia. Por Cauchy-Schwarz :

$$\sum_{i=1}^{m}\left(\int_{A_i} h dV\right)^2 \leq \sum_{i=1}^{m}\int_{A_i} h^2 dV \int_{A_i} dV = \frac{1}{m}\sum_{i=1}^{m}\int_{A_i} h^2 dV = \frac{\|k\|^2}{m} \tag{7}$$

ya que $\int h^2 dV = \int(\mathcal{T}^{-1}k)^2 dF_0 = \|k\|^2$. Por otra parte si tomamos $a = \frac{\mathcal{T}^{-1}(k)}{\|k\|}$, es decir $h = \|k\|$ tenemos que

$$\sum_{i=1}^{m}\left(\int_{A_i} h dV\right)^2 = \|k\|^2\sum_{i=1}^{m}\frac{1}{m^2} = \frac{\|k\|^2}{m} \tag{8}$$

y la conclusión es que, la función $a = \frac{\mathcal{T}^{-1}(k)}{\|k\|}$ es óptima, en el sentido en que maximiza el sesgo.

En la sección que sigue implementaremos la prueba χ^2 propuesta en dos casos particulares de alternativas contiguas : corrimientos en la media y cambios de dispersión, para la distribución normal.

2 La implementación de la prueba

En los dos ejemplos que siguen, utilizaremos la isometría $\mathcal{T}_{\mathrm{L}}$ definida en $L^2(\mathbb{R}, \Phi)$ por :

$$\mathcal{T}_{\mathrm{L}}\, g(x) = g(x) - \int_{-\infty}^{x} \frac{g(t)\varphi(t)}{1-\Phi(t)}\, dt \ , \ \forall g \in L^2(\mathbb{R}, \Phi) \tag{9}$$

Donde Φ es la distribución normal típica y φ la densidad correspondiente, la inversa de $\mathcal{T}_{\mathrm{L}}$ es

$$\left(\mathcal{T}_{\mathrm{L}}^{-1} h\right)(x) = h(x) + \frac{1}{1-\Phi(x)} \int_{-\infty}^{x} h(t)\varphi(t)\, dt \tag{10}$$

Para ver que $\mathcal{T}_{\mathrm{L}}$ es una isometría en $L^2(\mathbb{R}, \Phi)$, consultar [1]

Pasemos a los ejemplos mencionados

2.1 Caso 1. Corrimientos en la media

Sean $F_0(x) = \Phi(x)$ y $F_n(x) = \Phi\left(x - \delta/\sqrt{n}\right)$, donde Φ es la función de distribución normal típica. La función k será entonces :

$$k(x) = \lim_{n\to+\infty} \frac{\sqrt{n}}{\delta}\left(f^{(\delta/\sqrt{n})}(x) - 1\right) = \lim_{n\to+\infty} \frac{\sqrt{n}}{\delta}\left(\frac{\varphi(x-\delta/\sqrt{n})}{\varphi(x)} - 1\right) = x$$

Es posible demostrar de la función $k(x) = x$ cumple (4) . Elijamos, ahora la función a

$$a(x) = \mathcal{T}_{\mathrm{L}}^{-1} x = x + \frac{1}{1-\Phi(x)} \int_{-\infty}^{x} t\varphi(t)\, dt = x - \frac{\varphi(x)}{1-\Phi(x)} \tag{11}$$

Recordemos que queremos intervalos $A_1, A_2, \ldots, A_m$ equiprobables bajo la medida V, donde

$$V((-\infty, y]) = \int_{-\infty}^{y} a^2(x)\varphi(x)\, dx = \int_{-\infty}^{y} \left(x - \frac{\varphi(x)}{1-\Phi(x)}\right)^2 \varphi(x)\, dx =$$

$$\int_{-\infty}^{y} x^2\varphi(x)\, dx - \int_{-\infty}^{y} \frac{2x\varphi^2(x)}{1-\Phi(x)}\, dx + \int_{-\infty}^{y} \frac{\varphi^3(x)}{(1-\Phi(x))^2}\, dx =$$

$$= \Phi(y) - y\varphi(y) + \frac{\varphi^2(y)}{1-\Phi(y)} \tag{12}$$

Tomamos $m = 10$ y definimos los intervalos $A_j = (t_{j-1}, t_j]$, para $j = 1, \ldots, 10$ con la convención $t_0 = -\infty$ y $t_{10} = +\infty$, de modo que $V(A_j) =$

$\frac{1}{10}$, $j = 1, \ldots, 10$. Pasemos a evaluar el Proceso Empírico Transformado en las regiones en cuestión

$$\sqrt{n}\, w_n^{a,\mathcal{T}}(A_j) = \sqrt{n} \int \mathcal{T}(a\mathbb{I}_{A_j}) db_n = \sum_{i=1}^{n} \mathcal{T}(a\mathbb{I}_{A_j})(X_i) =$$

$$\sum_{i=1}^{n} \left(a(X_i)\mathbb{I}_{(t_{j-1},t_j]}(X_i) - \int_{-\infty}^{X_i} \frac{a(x)\mathbb{I}_{(t_{j-1},t_j]}(x)\varphi(x)}{1-\Phi(x)} dx \right) =$$

$$\sum_{i=1}^{n} \left(a(X_i)\mathbb{I}_{(t_{j-1},t_j]}(X_i) - \int_{X_i \wedge t_{j-1}}^{X_i \wedge t_j} \frac{a(x)\varphi(x)}{1-\Phi(x)} dx \right) =$$

$$\sum_{i=1}^{n} \left(a(X_i)\mathbb{I}_{(t_{j-1},t_j]}(X_i) - \left\{ \frac{-\varphi(x)}{1-\Phi(x)} \right\}_{x=X_i \wedge t_{j-1}}^{x=X_i \wedge t_j} \right) =$$

$$\sum_{i=1}^{n} \left(X_i\, \mathbb{I}_{(t_{j-1},t_j]}(X_i) + \frac{\varphi(t_j)}{1-\Phi(t_j)} \mathbb{I}_{\{X_i > t_j\}} - \frac{\varphi(t_{j-1})}{1-\Phi(t_{j-1})} \mathbb{I}_{\{X_i > t_{j-1}\}} \right) \tag{13}$$

La región crítica de nuestra prueba, para el nivel α será entonces

$$\left\{ \sum_{j=1}^{10} \left(w_n^{a,\mathcal{T}}(A_j) \right)^2 > \frac{1}{10} \chi_{10}^2(\alpha) \right\} \tag{14}$$

donde χ_{10}^2 es el percentil de orden $1-\alpha$ para la distribución ji-cuadrado con 10 grados de libertad.
Realizando una simulación, en la que se aplicaron la prueba ji-cuadrado clásica y la prueba aquí propuesta (ambas al nivel 0.05), a 5000 muestras de tamaño 100, para diferentes valores del apartamiento inicial δ, se obtuvieron los siguientes datos.

δ	χ^2 clásica	χ^2 con PETs
0	0,0466	0,0514
0,5	0,0560	0,0684
1,0	0,0812	0,1210
1,5	0,1222	0,1984
2.0	0,2236	0,3148
2.5	0,3568	0,4574
3,0	0,4942	0,6028
3,5	0,6418	0,7468
4,0	0,8040	0,8642
4,5	0,8976	0,9336
5,0	0,9528	0,9746

cuadro 1

Comparación de la proporción de rechazos para una y otra prueba con diferentes apartamientos iniciales, en el caso 1

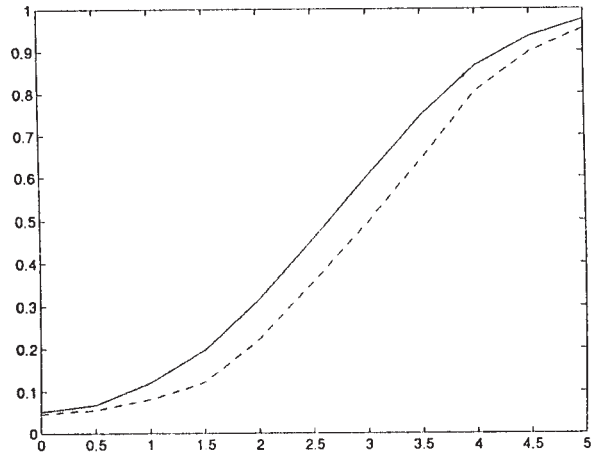

figura 1

`Caso 1.` Variación de la potencia en función del apartamiento inicial δ, la línea punteada corresponde a la prueba clásica y la otra a la prueba con el Proceso Empírico Transformado

2.2 Caso 2. Cambios de dispersión

Sean $F_0(x) = \Phi(x)$ y $F_n(x) = \Phi\left((1 - \frac{\delta}{\sqrt{n}})x\right)$, donde Φ es la función de distribución normal típica. La función k será entonces :

$$k(x) = \lim_{n\to+\infty} \frac{\sqrt{n}}{\delta}\left(f^{(\delta/\sqrt{n})}(x) - 1\right) =$$

$$= \lim_{n\to+\infty} \frac{\sqrt{n}}{\delta}\left(\frac{\varphi\left((1 - \frac{\delta}{\sqrt{n}})x\right)}{\varphi(x)} - 1\right) = x^2 - 1 \tag{15}$$

Es posible demostrar de la función $k(x) = x^2 - 1$ cumple (4).
Elijamos ahora, la función a, normalizando (ya que $\|k\|^2 = 2$) obtenemos

$$a(x) = \frac{1}{\sqrt{2}}\mathcal{T}_{\mathrm{L}}^{-1}(x^2 - 1) = \frac{1}{\sqrt{2}}\left(x^2 - 1 + \frac{1}{1 - \Phi(x)}\int_{-\infty}^{x}(t^2 - 1)\varphi(t)\,dt\right) =$$

$$= \frac{1}{\sqrt{2}}\left(x^2 - 1 - \frac{x\varphi(x)}{1 - \Phi(x)}\right) \tag{16}$$

Veamos ahora, cual es la medida V

$$V((-\infty, y]) = \int_{-\infty}^{y} a^2(x)\varphi(x)\,dx = \frac{1}{2}\int_{-\infty}^{y}\left(x^2 - 1 - \frac{x\varphi(x)}{1 - \Phi(x)}\right)^2\varphi(x)\,dx =$$

$$\frac{1}{2}\int_{-\infty}^{y}\left(x^4\varphi(x) + \varphi(x) + \frac{x^2\varphi^3(x)}{(1 - \Phi(x))^2} - 2x^2\varphi(x) - \frac{2x^3\varphi^2(x)}{1 - \Phi(x)} + \frac{2x\varphi^2(x)}{1 - \Phi(x)}\right)dx =$$

$$= 2\Phi(y) - y\varphi(y) - y^3\varphi(y) + \frac{y^2\varphi^2(y)}{1 - \Phi(y)} \tag{17}$$

Tomamos $m = 10$ y definimos los intervalos $A_j = (t_{j-1}, t_j]$ para $j = 1, \ldots, 10$ con la convención $t_0 = -\infty$ y $t_{10} = +\infty$, de modo que $V(A_j) = \frac{1}{10}$, $j = 1, \ldots, 10$.

Pasemos a evaluar el Proceso Empírico Transformado en las regiones en cuestión

$$w_n^{a,\mathcal{T}}(A_j) = \int \mathcal{T}(a\mathbb{I}_{A_j})db_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{T}(a\mathbb{I}_{A_j})(X_i) =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(a(X_i)\mathbb{I}_{(t_{j-1},t_j]}(X_i) - \int_{-\infty}^{X_i}\frac{a(x)\mathbb{I}_{(t_{j-1},t_j]}(x)\varphi(x)}{1-\Phi(x)}dx\right) =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(a(X_i)\mathbb{I}_{(t_{j-1},t_j]}(X_i) - \int_{X_i\wedge t_{j-1}}^{X_i\wedge t_j}\frac{a(x)\varphi(x)}{1-\Phi(x)}dx\right) =$$

$$\frac{1}{\sqrt{2n}}\sum_{i=1}^{n}\left((X_i^2-1)\,\mathbb{I}_{(t_{j-1},t_j]}(X_i) - \left\{\frac{-x\varphi(x)}{1-\Phi(x)}\right\}_{x=X_i\wedge t_{j-1}}^{x=X_i\wedge t_j}\right) =$$

$$\frac{1}{\sqrt{2n}}\sum_{i=1}^{n}\left((X_i^2-1)\,\mathbb{I}_{(t_{j-1},t_j]}(X_i) + \frac{t_j\varphi(t_j)}{1-\Phi(t_j)}\mathbb{I}_{\{X_i>t_j\}} - \frac{t_{j-1}\varphi(t_{j-1})}{1-\Phi(t_{j-1})}\mathbb{I}_{\{X_i>t_{j-1}\}}\right) \tag{18}$$

Realizando una simulación, en la que se aplicaron la prueba ji-cuadrado clásica y la prueba aquí propuesta (ambas al nivel 0.05), a 5000 muestras de tamaño 100, para diferentes valores del apartamiento inicial δ, se obtuvieron los siguientes datos.

δ	χ^2 clásica	χ^2 con PETs	δ	χ^2 clásica	χ^2 con PETs
0	0.0466	0.0454	1.6	0.2368	0.5768
0.2	0.0530	0.0610	1.8	0.3276	0.6742
0.4	0.0536	0.0946	2.0	0.4018	0.7716
0.6	0.0724	0.1400	2.2	0.5032	0.8482
0.8	0.0850	0.1970	2.4	0.5916	0.9092
1.0	0.1112	0.2698	2.6	0.6864	0.9508
1.2	0.1416	0.3588	2.8	0.7664	0.9748
1.4	0.1864	0.4680	3.0	0.8434	0.9910

cuadro 2

Comparación de la proporción de rechazos para una y otra prueba con diferentes apartamientos iniciales, en el caso 2

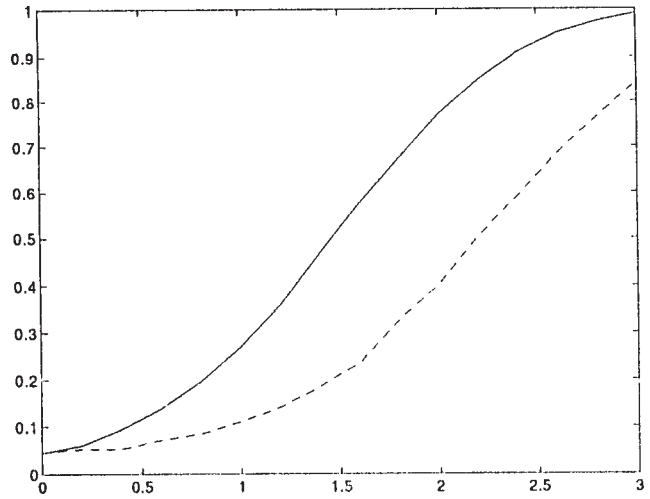

figura 2

Caso 2. Variación de la potencia en función del apartamiento inicial δ, la línea punteada corresponde a la prueba clásica y la otra a la prueba con el Proceso Empírico Transformado

2.3 Conclusiones

En ambos casos se ha obtenido un mejor rendimiento que el de la prueba χ^2 clásica. En el caso de cambios de dispersión la diferencia ha sido más notoria que en el caso de corrimientos en la media y eso es de alguna manera esperable, ya que en el primer caso, la prueba χ^2 clásica, tiene un buen rendimiento, mientras que en el segundo caso, no es tan bueno.

Referencias

[1] Cabaña A. y Cabaña E.M.: *Transformed Empirical Processes and Modified Kolmogorov Smirnov Tests for Multivariate Distributions,* Universitat de Barcelona, Mathematics Preprint Series, N.208 (1996) (Por aparecer en Ann. Statist.)

[2] Le Cam L., Yang G.L.: *Asymptotics in Statistics,* Springer-Verlag, New York (1990)

[3] Oosterhoff J. y Van Zwet W.R.: *A Note on Contiguity and the Hellinger Distance,* Contributions to Statistics, 157-166, Reidel, Dordrecht (1979)