Nicolas Maldonado Baracaldo
201423809
n.maldonado10@uniandes.edu.co

Estadística No Paramétrica y Remuestreo (MATE3526)

Adolfo J. Quiroz

Proyecto 1

4 de octubre de 2022

En la teoría de estadística no paramétrica aparece una clase de estadísticos llamados U-*estadísticos*, de particular importancia pues permiten realizar tests de hipótesis acerca de parámetros de distribuciones desconocidas y en estos casos pueden llegar a ser no-paramétricos libres de distribución. En particular para muestras independientes $X_1, \ldots, X_m$ y $Y_1, \ldots, Y_n$ de dos poblaciones con distribuciones en alguna familia dada, a un parámetro γ se le llama *estimable de grado* (r, s) si r y s son los tamaños de muestra más pequeños para los cuales existe una función h tal que

$$\mathrm{E}[h(X_1, \ldots, X_r, Y_1, \ldots, Y_s)] = \gamma.$$

En dado caso, a la función h (la cual puede asumirse simétrica en los X_is y simétrica en los Y_js, pues incluso si no lo es se puede simetrizar) se le llama el *kernel* del U-estadístico y éste último se define como

$$U(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{\alpha \in A} \sum_{\beta \in B} h(X_{\alpha_1}, \ldots, X_{\alpha_r}, Y_{\beta_1}, \ldots, Y_{\beta_s}).$$

Más aún, el Teorema para U-Estadísticos de Dos Muestras nos garantiza que si $\mathrm{E}[h^2(X_1, \ldots, X_r, Y_1, \ldots, Y_s)]$ es finita, entonces $\sqrt{N}\,[U(X_1, \ldots, X_m, Y_1, \ldots, Y_n) - \gamma]$ tiene una distribución límite normal con media 0 y varianza $[r^2\zeta_{1,0}/\lambda] + [s^2\zeta_{0,1}/(1-\lambda)]$.[1, Capítulo 3]

A continuación se quiere evaluar, para muestras provenientes de distribuciones $F(x/\eta_1)$ y $F(x/\eta_2)$, el U-estadístico de diferencia de escala con kernel

$$h(x_1, x_2, y_1, y_2) = \Psi(|y_1 - y_2| - |x_1 - x_2|)$$

para H_0 la hipótesis $\eta_1 = \eta_2$ y H_a la hipótesis $\eta_1 < \eta_2$ usando el método de Monte Carlo, pues la estimación usando directamente U-estadísticos resulta computacionalmente demasiado costosa para grandes muestras.

(Todo lo que sigue se realizo usando `Mathematica`. El archivo `.nb` se adjunta al presente informe.)

1. Para empezar se definieron las funciones Ψ y h que se usarán para todos los cálculos. Se definieron los tamaños de las muestras, $m = 5000, n = 2000$, y se tomaron las muestras usando la función `RandomVariate` y proporcionando la distribución apropiada (`GammaDistribution`) con sus parámetros.

2. A continuación se definió el número de iteraciones a usar para Monte Carlo, $L = 2000$, y se procedió a usar el método para estimar $\zeta_{1,0}$ y $\zeta_{0,1}$. Se separó la muestra X en dos para esta parte, luego se tomaron subconjuntos de tamaño 3 y 4 (resp. 4 y 3) de cada mitad usando la función `RandomSample` y con ellos se calculó

   $$\operatorname{est}\zeta_{1,0} = h(x_1, x_2, y_1, y_2)h(x_1, x_3, y_3, y_4) - \gamma^2$$

   (resp. $\operatorname{est}\zeta_{0,1}$), luego se tomó el promedio de éstos como valor para $\zeta_{1,0}$ (resp. $\zeta_{0,1}$). Finalmente, usando los estimados anteriores, así como $\lambda = m/N$, se estimó $\tilde{\sigma}_U$.

3. Usando el mismo número de iteraciones que antes, se estimó mediante el método de Monte Carlo el U-estadístico con kernel h. Para esta parte se tomaron subconjuntos de X y de Y por separado, ambos de tamaño 2, de nuevo con la función `RandomSample` y se calculó $h(x_1, x_2, y_1, y_2)$, el cual luego se promedió para dar un estimado del U-estadístico. Por último se usó éste y el estimado anterior para calcular $\tilde{U} = \sqrt{N}(U - \gamma)/\tilde{\sigma}_U$, el cual según el teorema ya citado debería tener una distribución normal estándar.

4. En este punto se quería repetir todo el proceso 500 veces para evaluar la normalidad de $\tilde{U}$, por lo que resultó conveniente definir una nueva función `MC` que realizara todo esto dados los parámetros m, n, sX, sY, L, donde m, n, L tienen las mismas definiciones que en los procedimientos anteriores y sX, sY se refieren a los parámetros de escala a ingresar a `GammaDistribution` al momento de tomar las muestras X y Y, respectivamente (esto con el fin de que la misma función nos sirva para evaluar tanto la hipótesis nula como la alternativa). Teniendo esta función definida, se generaron los 500 valores usando la función `Table` con el llamado a la función MC con $m = 5000, n = 2000, sX = 2, sY = 2, L = 2000$. Se evaluó en este punto la normalidad mediante tres métodos,

   (i) Se generó un q-q plot usando la función `QuantilePlot` (figura 1).

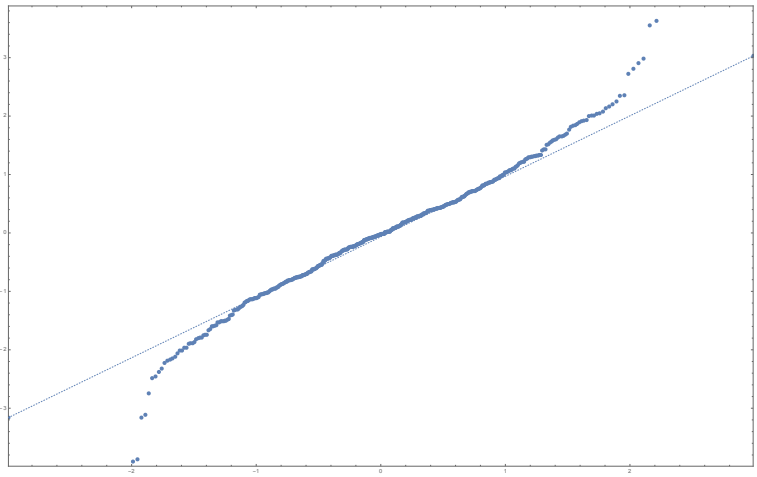

Figura 1: q-q plot obtenido para H_0.

   (ii) Se realizó el test de Cramér-von Mises usando la función `CramerVonMisesTest`, el cual arrojó el resultado "*The null hypothesis that the data is distributed according to the NormalDistribution[0, 1] is not rejected at the 5 percent level based on the Cramér-von Mises test.*".

   El test de Cramér-von Mises básicamente usa la suma de diferencias cuadradas entre las FDAs esperada y observada como estadístico de prueba.

   Se realizó también el test de Kolmogorov-Smirnov usando la función `KolmogorovSmirnovTest`, el cual arrojó el resultado "*The null hypothesis that the data is distributed according to the NormalDistribution[0, 1] is not rejected at the 5 percent level based on the Kolmogorov-Smirnov test.*".

   El test de Kolmogorov-Smirnov básicamente usa el supremo de los valores absolutos de las diferencias entre las FDAs esperada y observada como estadístico de prueba.

   (iii) Se contó el porcentaje de veces que $\tilde{U}$ sobrepasó el cuantil 95 % de la normal estándar. Para la corrida en que se realizó el conteo se obtuvo aquí que un $7{,}8\,\%$ de los datos excedían el cuantil.

5. Con base en las diversas evaluaciones de normalidad realizadas, es razonable concluir que bajo H_0 en efecto se tiene que $\tilde{U}$ calculado mediante Monte Carlo seguirá una distribución normal estándar, tal y como se quería.

6. A continuación se quería seguir básicamente el mismo proceso para la hipótesis alternativa. Se aprovechó acá que ya se había definido la función `MC`, y se generaron los 500 valores usando la función `Table` con el llamado a la función MC con $m = 5000, n = 2000, sX = 2, sY = 3, L = 2000$. Se evaluó en este punto la normalidad mediante los mismos tres métodos,

(i) Se generó un q-q plot usando la función `QuantilePlot` (figura 2).

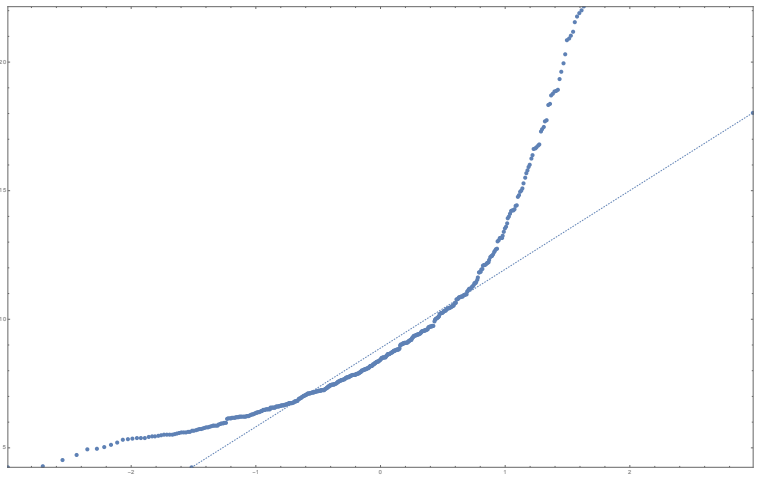

Figura 2: q-q plot obtenido para H_a.

(ii) Se realizó el test de Cramér-von Mises usando la función `CramerVonMisesTest`, el cual arrojó el resultado "*The null hypothesis that the data is distributed according to the NormalDistribution[0, 1] is rejected at the 5 percent level based on the Cramér-von Mises test.*".

Se realizó también el test de Kolmogorov-Smirnov usando la función `KolmogorovSmirnovTest`, el cual arrojó el resultado "*The null hypothesis that the data is distributed according to the NormalDistribution[0, 1] is rejected at the 5 percent level based on the Kolmogorov-Smirnov test.*".

(iii) Se contó el porcentaje de veces que $\tilde{U}$ sobrepasó el cuantil 95 % de la normal estándar. Para la corrida en que se realizó el conteo se obtuvo aquí que un $100\,\%$ de los datos excedían el cuantil.

7. Con base en las diversas evaluaciones de normalidad realizadas, es razonable concluir que bajo H_a no se tiene que $\tilde{U}$ calculado mediante Monte Carlo seguirá una distribución normal estándar, luego garantizamos que dadas muestras desconocidas, se podrá usar como test de hipótesis el estadístico así calculado, junto con algún tipo de criterio de normalidad.

Como ya se mencionó, el hecho de que hayamos podido usar el método de Monte Carlo para estimar $\tilde{U}$ en los dos casos, y que se tenga normalidad en su distribución únicamente cuando se cumple H_0, nos da entonces que este es un método sensato y razonable para evaluar las hipótesis H_0 y H_a relacionadas con el parámetro de escala de las dos muestras incluso en el caso en que no conozcamos la distribución de la que provienen. En adelante dadas dos muestras independientes se puede usar el método aquí presentado para estimar $\tilde{U}$ y luego un sencillo test de normalidad nos permitirá evaluar nuestras hipótesis.

Más allá de lo aquí realizado, esto también nos indica que incluso para muestras de tamaño muy grande los U-estadísticos no se vuelven inútiles. Si bien su cálculo directo ya no va a ser factible en estos casos, hemos visto que el uso del método de Monte Carlo no los hace menos poderosos. Claro está que aquí solo se evaluó para un U-estadístico particular, sin embargo los resultados son prometedores y fácilmente se podrían realizar evaluaciones similares para cualquier estadístico previo a su uso si igualmente se quiere estar seguro.

Referencias

[1] Ronald H. Randles and Douglas A. Wolfe. *Introduction to the Theory of Nonparametric Statistic*. Krieger Publishing Company, Malabar, Florida, reprint edition, 1991.