MATE3523-MATE4523. Estadística No Paramétrica y Remuestreo. Proyecto I, 2017-I

Distribución del estadístico W^+ de Wilcoxon. Eficiencia de estimadores de centro de simetría.

El presente proyecto consta de dos partes. En la primera evaluamos la velocidad de convergencia de la aproximación normal a la distribución del estadístico W^+. En la segunda se comparan las eficiencias de los estadísticos t de Student, de signos, B, y de rangos signados, W^+, para evaluar la hipótesis nula $\theta = \theta_0$ contra la alternativa $\theta \neq \theta_0$ para el centro de simetría de la distribución de una muestra $Z_1, \ldots, Z_n$ al variar la distribución simétrica y el tamaño muestral.

Aproximación normal a la distribución de W^+

El archivo anexo, cnk.txt, contiene la función cnk(), que dado un valor de n retorna las cantidades $c_n(k)$ = número de subconjuntos de $\{1, 2, \ldots, n\}$ que suman k, para cada $k = 0, 1, \ldots, n(n+1)/2$ (véase el Prob. 2.4.1). Observe que 0 es uno de los valores de k. Estas cantidades permiten evaluar la distribución exacta de W^+, o de cualquier transformación lineal del mismo, usando el Corolario 2.4.12 del libro de texto. Por otro lado, vimos en clase (Clase 8) que

$$\frac{\sqrt{3n}(W^+ - \mathbf{E}W^+)}{\binom{n}{2}} \xrightarrow{\text{(d)}} \text{N}(0,1), \text{ cuando } n \to \infty \tag{1}$$

Para verificar la validez de la aproximación, sea F_v la f.d.a. verdadera del estadístico en (1) y sea F_a la f.d.a. Normal(0,1). Sea $\mathcal{R}$ el conjunto de valores posibles del estadístico en (1). Encuentre,

$$\max_{r \in \mathcal{R}} |F_v(r) - F_a(r)|.$$

Haga este cálculo para cada n = 10, 20, 50, 100, 200,... (incluya más valores de ser necesario) hasta llegar a que la máxima diferencia sea inferior a 0.005. Asimismo, para algunos de los valores de n haga un plot prob.-prob. (averiguar sobre este gráfico) para evaluar la calidad de la aproximación. Extraiga conclusiones.

Eficiencia de estadísticos para pruebas sobre el centro de simetría

Tenemos datos $Z_1, \ldots, Z_n$ i.i.d. de una distribución continua simétrica en θ. Para evaluar la hipótesis nula $\theta = \theta_0$ contra la alternativa $\theta \neq \theta_0$, podemos usar los siguientes estadísticos:

- t de student, dado por
$$t = \frac{\sqrt{n}(\overline{Z} - \theta_0)}{s}$$
siendo s la desviación estándar muestral.
- El estadístico de signos, $B = \sum_{i=1}^{n} \Psi(Z_i - \theta_0)$.
- El estadístico de rangos signados, $W^+ = \sum_{i=1}^{n} \Psi(Z_i - \theta_0) R_i^+$, donde los R_i^+ son los rangos absolutos de los $Z_i - \theta_0$.

Mediremos la eficiencia de cada estadístico mediante su potencia, es decir, la probabilidad de detectar que la H_0 no se cumple, cuando, efectivamente, la alternativa es cierta. Tomamos, sin pérdida de generalidad, $\theta_0 = \pi$ y para la alternativa, consideramos $\theta_a = \pi + 0.1, \pi + 0.25, \pi + 0.5$ y $\pi + 1$. Como tamaños de muestra tomamos $n = 20$, 30, 50 y 100. Consideramos las siguientes distribuciones simétricas para los datos (bajo H_0):

- N(θ,1),
- Unif($\theta - 1$,$\theta + 1$),
- Distribución doble exponencial de Laplace, con densidad
$$f(x) = \frac{1}{2} \exp(-|x - \theta|) \text{ para } x \in \mathbb{R},$$
- Datos con distribución $F(x - \theta)$, siendo F la distribución t de student con dos grados de libertad, y θ el centro que se quiere tener.
- Distribución de Cauchy con densidad
$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, \text{ para } x \in \mathbb{R}.$$

Para cada distribución simétrica, tamaño muestral, valor de la alternativa y estadístico:

- Generar muestra de la distribución simétrica, con centro en θ_a.

- Aplicar el estadístico para evaluar la hipótesis H_0 : el centro de simetría es $\theta = \pi$ contra la alternativa H_a : $\theta > \pi$ y decidir si el valor es significativo a nivel 5%.

- Repetir los pasos anteriores un total de 500 veces, almacenando la potencia como el porcentaje de veces que se rechaza H_0.

¿Para cuales distribuciones es más efectivo el t de Student? ¿Como es el desempeño de los estadísticos no paramétricos? ¿Cual de estos es preferible?

Su informe debe incluir (i) Una introducción a los problemas considerados, (ii) Los detalles de implementación de las simulaciones realizadas, (iii) Discusión de resultados y (iv) Conclusiones. Procure que su código en R sea vectorizado, en la medida de lo posible.

Valor del proyecto: 20 pts. Fecha de entrega de su informe, domingo 19-03-2017. Debe ser recibido antes de medianoche en la dirección electrónica ajquiroz@gmail.com.